

| | |
|----|------------------|
| 申报 | 系列：教师系列 科研为主型 |
| | 专业：畜牧 |
| | 职称：副教授 (科研型) |

业绩成果材料

(申报人的业绩成果材料包括论文、科研项目、获奖以及其他成果等)

单 位 (二级单位) 动物科学学院

姓 名 闫希亮

材料核对人：

单位盖章：

核对时间：

华南农业大学制

目 录

一、教学研究业绩

1. 教学研究项目：关于公布 2024 年华南农业大学研究生教育创新计划项目的立项通知（合同）及有关佐证材料.....1
4. 教学比赛证书.....7

二、科研项目

1. 主持：关于国家自然科学基金青年基金项目的立项通知（合同）及有关佐证材料.....8
2. 主持：关于国家自然科学基金面上项目的立项通知（合同）及有关佐证材料.....12
3. 主持：关于特定高校学科建设专项（人才引进类）项目（课题）任务书及有关佐证材料.....16
4. 主持：关于广州市科技计划项目的立项通知（合同）及有关佐证材料.....32
5. 主持：关于猪禽种业全国重点实验室开放课题合同书及有关佐证材料.....43
6. 主参：关于国家自然科学基金面上项目的立项通知（合同）及有关佐证材料.....50

三、论文、著作等

1. 检索证明.....68
2. 以第一作者发表本专业论文情况.....86
- 2.1. Construction of a web-based nanomaterial database by big data

| | |
|---|-----|
| curation and modeling friendly nanostructure annotations..... | 86 |
| 2.2. Converting Nanotoxicity Data to Information Using Artificial Intelligence and Simulation..... | 96 |
| 2.3. Prediction of Nano–Bio Interactions through Convolutional Neural Network Analysis of Nanostructure Images..... | 159 |
| 2.4. The Glutamatergic System Regulates Feather Pecking Behaviors in Laying Hens Through the Gut–Brain Axis..... | 168 |
| 3. 以通讯作者发表本专业论文情况..... | 186 |
| 3.1. Linking electron ionization mass spectra of organic chemicals to toxicity endpoints through machine learning and experimentation..... | 186 |
| 3.2. Implementing comprehensive machine learning models of multispecies toxicity assessment to improve regulation of organic compounds..... | 196 |
| 3.3. Unraveling the ecotoxicity of micro(nano)plastics loaded with environmental pollutants using ensemble machine learning..... | 207 |
| 3.4. Advanced Mass-Spectra-Based Machine Learning for Predicting the Toxicity of Traditional Chinese Medicines..... | 218 |
| 3.5. ILTox: A Curated Toxicity Database for Machine Learning and Design of Environmentally Friendly Ionic Liquids..... | 228 |
| 3.6. De novo Design of Biocompatible Nanomaterials Using Quasi-SMILES and Recurrent Neural Networks..... | 234 |

| | |
|---|-----|
| 3.7. Effect of Different Strategies for Modifying Graphene on the Adsorption and Gas Sensing of Trimethylamine: Insights from DFT Study..... | 244 |
| 3.8. MnN ₄ embedded zeolite-templated carbon for methylamine and trimethylamine sensing: Insights from DFT study..... | 254 |
| 3.9. DFT perspective of gas sensing properties of metal oxide nanocages toward trimethylamine: Effects of humidity, temperature and electric field..... | 262 |
| 3.10. Reaching the full potential of machine learning in mitigating environmental impacts of functional materials..... | 272 |
| 3.11. Comprehensive Interrogation on Acetylcholinesterase Inhibition by Ionic Liquids Using Machine Learning and Molecular Modeling..... | 291 |
| 3.12. Predicting cytotoxicity of binary pollutants towards a human cell panel in environmental water by experimentation and deep learning methods..... | 303 |
| 3.13. Developmental Toxicity of Fenbuconazole in Zebrafish: Effects on Mitochondrial Respiration and Locomotor Behavior... | 312 |
| 3.14. Unraveling the joint toxicity of transition-metal dichalcogenides and per- and polyfluoroalkyl substances in aqueous mediums by experimentation, machine learning and molecular dynamics..... | 323 |

四、科研成果

| | |
|--|-----|
| 1. 科技奖励证书..... | 336 |
| 2. 知识产权..... | 337 |
| 2.1. 专利授权证书：一种基于谱图分析的有机物生物毒性预测方法及系统..... | 337 |
| 2.2. 专利授权证书：一种基于实验和计算的二维纳米复合物毒性评价方法..... | 338 |
| 2.3. 专利授权证书：一种离子液体对乙酰胆碱酯酶的毒性预测方法及系统..... | 339 |

五、其他业绩

| | |
|--------------------------------|-----|
| 2. 个人荣誉..... | 340 |
| 2.1. 广州大学 2022 年“年度学术新锐” | 340 |
| 2.2. 第十次全国毒理学大会优秀论文..... | 341 |

华南农业大学文件

华南农研〔2024〕10号

关于公布 2024 年华南农业大学研究生教育 创新计划立项项目的通知

各学院、部处、各单位：

为深入实施研究生教育创新计划，进一步提高人才培养质量，结合高水平大学建设的有关工作，学校开展了 2024 年华南农业大学研究生教育创新计划项目的申报工作。

经个人申报、单位推荐、形式审查、专家评审和校内公示等程序，确定广东长隆集团有限公司等 10 个联合培养研究生示范基地，《现代汽车新技术》1 个一般性全英文课程建设项目，《农业机器人》等 12 个课程思政建设项目，《食品加工新技术研究与新产品研发专题》等 6 个示范课程理论课建设项目，《工程伦理学》等 3 个研究生在线开放课程建设项目，《智慧农业：关键技术与应用》等 6 个高水平研究生教材建设项目，《基于专题驱

动的《智慧农业理论与实践》教学方法探索》等 16 个专业学位研究生实践教学资源建设与培养模式改革研究项目，以及《面向智慧农业的《数字图像处理》课程案例库建设》等 10 个专业学位研究生课程案例库建设项目，共 64 个项目为 2024 年华南农业大学研究生教育创新计划立项项目（详见附件），现予公布。

各项目负责人应按照学校高水平大学建设专项资金的有关管理办法，合理使用项目经费，按照项目既定研究周期，严格执行研究计划，扎实推进研究工作，确保按时完成研究任务，实现预期目标。

附件：2024 年华南农业大学研究生教育创新计划项目立项
名单

华南农业大学
2024 年 4 月 25 日

（联系人：潘 科，电 话：85280189）

公开方式：主动公开

华南农业大学党政办公室

2024 年 4 月 26 日印发

附件

2024 年华南农业大学研究生教育创新计划 项目立项名单

| 项目类别 | 序号 | 合作单位 | 学科领域 | 校内 负责人 | 依托单位 |
|------------------|----|---------------------|-------------------|-----------|--------------|
| (一) 联合培养研究生示范基地 | 1 | 广东长隆集团有限公司 | 兽医、畜牧等 | 仝文宝 | 兽医学院 |
| | 2 | 广东省农业科学院作物研究所 | 作物学 | 冯发强 | 农学院 |
| | 3 | 中国水稻研究所 | 作物学 | 王少奎 | 农学院 |
| | 4 | 广州华农大智慧农业科技有限公司 | 计算机科学与技术 | 王春桃 | 数学与信息学院、软件学院 |
| | 5 | 东莞植物园 | 园艺 | 赵杰堂 | 园艺学院 |
| | 6 | 佛山市铁人环保科技有限公司 | 农业资源与环境、资源利用与植物保护 | 陈火君 | 资源环境学院 |
| | 7 | 广州市海珠湿地科研宣传教育中心 | 风景园林 | 李 晖 | 林学与风景园林学院 |
| | 8 | 广东东图规划科技有限公司 | 公共管理学 | 史传林 | 公共管理学院 |
| | 9 | 南方海洋科学与工程广东省实验室（湛江） | 计算机技术 | 黄 琼 | 数学与信息学院、软件学院 |
| | 10 | 广州市微生物研究所集团股份有限公司 | 食品科学与工程 | 方 祥 | 食品学院 |
| 项目类别 | 序号 | 课程名称 | 课程类型 | 负责人 | 所在单位 |
| (二) 一般性全英文课程建设项目 | 1 | 现代汽车新技术 | 专业选修课 | 肖博一 | 工程学院 |
| (三) 课程思政建设项目 | 1 | 农业机器人 | 专业选修课 | 王红军 | 工程学院 |
| | 2 | 农产品安全生产技术与应用 | 专业学位课 | 刘 婕 | 植物保护学院 |
| | 3 | 高级作物栽培分子生理（全英） | 专业选修课 | 张 慧 | 农学院 |
| | 4 | 农业生态与可持续耕作制度 | 专业选修课 | 王小龙 | 农学院 |
| | 5 | 生物组学大数据分析 | 专业选修课 | 张群洁 | 农学院 |

| | | | | | |
|-------------------|----|-------------------|-------------|-----|-----------------------|
| | 6 | 兽医临床实践 | 专业选修课 | 苏荣胜 | 兽医学院 |
| | 7 | 现代知识产权与保护 | 专业选修课 | 刘 涛 | 食品学院 |
| | 8 | 风景资源与旅游规划 | 专业选修课 | 林敏慧 | 林学与风景园林学院 |
| | 9 | 管理研究方法论 | 专业学位课 | 陈 灿 | 经济管理学院 |
| | 10 | 金融科技 | 专业选修课 | 莫易娴 | 经济管理学院 |
| | 11 | 领导科学专题 | 专业学位课 | 唐 斌 | 公共管理学院 |
| | 12 | 研究生心理素养与幸福人生 | 公共选修课 | 林 媛 | 党委学生工作部 (党委研究生工作部) |
| (四) 示范课程理论课建设项目 | 1 | 食品加工新技术研究与新产品研发专题 | 专业选修课 | 杜 冰 | 食品学院 |
| | 2 | 高级水产动物营养与饲料学 | 专业学位课、专业选修课 | 甘 炼 | 海洋学院 |
| | 3 | 现代管理学 | 专业学位课 | 郭 萍 | 经济管理学院 |
| | 4 | 公共预算与财政管理 | 专业学位课 | 武玉坤 | 公共管理学院 |
| | 5 | 农学概论 | 专业选修课 | 谢 萍 | 人文与法学学院 |
| | 6 | 生物组学大数据分析 | 专业选修课 | 张群洁 | 农学院 |
| (五) 研究生在线开放课程建设项目 | 1 | 工程伦理学 | 公共学位课 | 李高扬 | 水利与土木工程学院 |
| | 2 | 有限元与 ANSYS | 专业选修课 | 胡圣荣 | 水利与土木工程学院 |
| | 3 | 森林灾害防控技术及应用 | 专业选修课 | 单体江 | 林学与风景园林学院 |
| 项目类别 | 序号 | 教材名称 | 学科领域 | 负责人 | 所在单位 |
| (六) 高水平研究生教材建设项目 | 1 | 智慧农业：关键技术与应用 | 计算机科学与技术 | 黄 栋 | 数学与信息学院、软件学院 |
| | 2 | 深度学习基础与应用实践 | 计算机科学与技术 | 彭红星 | 数学与信息学院、软件学院 |
| | 3 | 水果产后处理技术与装备 | 农业工程 | 段洁利 | 工程学院 |
| | 4 | 岭南建筑与聚落防灾 | 建筑学 | 周彝馨 | 水利与土木工程学院 |
| | 5 | 植物线虫方法学实验教程 | 植物保护 | 文艳华 | 植物保护学院 |
| | 6 | 公共管理心理学 | 公共管理 | 贾海薇 | 公共管理学院 |

| 项目类别 | 序号 | 项目名称 | 学科领域 | 负责人 | 所在单位 |
|--------------------------------|----|---|--------------|-----|--------------------|
| (七) 专业学位研究生实践教学资源建设与培养模式改革研究项目 | 1 | 基于专题驱动的《智慧农业理论与实践》教学方法探索 | 农艺与种业 | 张 雷 | 农学院 |
| | 2 | 乡村振兴战略下林业专业硕士实践能力培养模式改革创新探索 | 林业 | 李青粉 | 林学与风景园林学院 |
| | 3 | 智慧养殖研究生专项建设与实践 | 畜牧 | 闫希亮 | 动物科学学院 |
| | 4 | 低碳农业背景下《生态环境材料学》交叉课程体系的构建 | 环境科学 | 朱雁平 | 资源环境学院 |
| | 5 | “三位一体”协同育人创新型乡村振兴人才培养模式探究 | 水产、渔业发展 | 赵会宏 | 海洋学院 |
| | 6 | 咸淡水水域环境的养护与治理教学改革 | 水产、渔业发展 | 王 俊 | 海洋学院 |
| | 7 | 产教融合食品工程专业学位研究生培养的导向性改革与实践 | 食品工程 | 黎 攀 | 食品学院 |
| | 8 | 科技赋能农业专业硕士班级培养模式 | 农业工程 | 岳学军 | 电子工程学院 (人工智能学院) |
| | 9 | 专业硕士学位论文质量提升策略探究 | 农林经济管理 | 贺梅英 | 经济管理学院 |
| | 10 | 基于国际视野的“双一流”涉农高校金融专业硕士培养模式改革研究 | 金融学 | 董 莹 | 经济管理学院 |
| | 11 | 参与式案例教学在公共管理硕士培养的应用及创新 | 公共管理 | 吴 彦 | 公共管理学院 |
| | 12 | 基于 SWOT 分析的专业学位研究生教育高质量发展策略研究——以华南农业大学 MPA 教育为例 | 教育管理 | 宋星洲 | 公共管理学院 |
| | 13 | 新时代法硕课程思政融入研究——以华南农业大学法硕课程教学为例 | 法学 | 林 友 | 人文与法学学院 |
| | 14 | AI 时代 MTI 翻译硕士的智能化实践教学教学改革研究 | 英语笔译 | 陈喜华 | 外国语学院 |
| | 15 | 双创驱动下艺术硕士专业工作室人才培养模式研究与实践 | 艺术学 | 盘湘龙 | 艺术学院 |
| | 16 | 农业数字化背景下高校研究生知识产权素养培育的创新模式研究 | 图情信息 | 刘 洋 | 图书馆 |
| (八) 专业学位研究生课程案例库建设项目 | 1 | 面向智慧农业的《数字图像处理》课程案例库建设 | 计算机应用技术 | 崔金荣 | 数学与信息学院、软件学院 |
| | 2 | 生物特征识别案例库 | 人工智能、新一代电子技术 | 代 芬 | 电子工程学院 (人工智能学院) |
| | 3 | 工商管理专业学位研究生 (MBA) 课程案例库 | 工商管理 | 杨学儒 | 经济管理学院 |
| | 4 | 投资银行学案例库 | 金融学 | 董 莹 | 经济管理学院 |

| | | | | |
|----|---------------------------------|---------|-----|---------|
| 5 | 数智时代财务管理教学案例库 | 会计学 | 周小春 | 经济管理学院 |
| 6 | 《农业政策学》案例库 | 农林经济及管理 | 彭东慧 | 经济管理学院 |
| 7 | 《管理研究方法论》课程农业管理案例库 | 农业管理 | 陈 灿 | 经济管理学院 |
| 8 | 公共部门网络舆情治理创新实践案例库 | 公共管理 | 赵国洪 | 公共管理学院 |
| 9 | 模拟法庭实践教学案例库 | 法学 | 刘万洪 | 人文与法学学院 |
| 10 | 关照人工智能技术发展的《中英语言对比与翻译》课程教学案例库建设 | 英语笔译 | 李 舸 | 外国语学院 |

荣誉证书

闫希亮 老师：

在动物科学学院 2023 年青年教师教学能力比赛中
荣获“一等奖”

特发此证，以资鼓励

动物科学学院

2023 年 12 月 22 日

国家自然科学基金资助项目批准通知

(包干制项目)

闫希亮 先生/女士:

根据《国家自然科学基金条例》、相关项目管理办法规定和专家评审意见,国家自然科学基金委员会(以下简称自然科学基金委)决定资助您申请的项目。项目批准号: 22106025, 项目名称: 基于原子尺度深度学习的纳塑料及其复合污染物构效关系和毒性预测研究, 资助经费: 30.00万元, 项目起止年月: 2022年01月至 2024年12月, 有关项目的评审意见及修改意见附后。

请您尽快登录科学基金网络信息系统(<https://isisn.nsfc.gov.cn>), **认真阅读《国家自然科学基金资助项目计划书填报说明》并按要求填写《国家自然科学基金资助项目计划书》(以下简称计划书)**。对于有修改意见的项目,请您按修改意见及时调整计划书相关内容;如您对修改意见有异议,须在电子版计划书报送截止日期前向相关科学处提出。

请您将电子版计划书通过科学基金网络信息系统(<https://isisn.nsfc.gov.cn>)提交,由依托单位审核后提交至自然科学基金委。自然科学基金委审核未通过者,将退回的电子版计划书修改后再行提交;审核通过者,打印纸质版计划书(一式两份,双面打印)并在项目负责人承诺栏签字,由依托单位在承诺栏加盖依托单位公章,且将申请书纸质签字盖章页订在其中一份计划书之后,一并报送至自然科学基金委项目材料接收工作组。纸质版计划书应当保证与审核通过的电子版计划书内容一致。**自然科学基金委将对申请书纸质签字盖章页进行审核,对存在问题的,允许依托单位进行一次修改或补齐。**

向自然科学基金委提交电子版计划书、报送纸质版计划书并补交申请书纸质签字盖章页截止时间节点如下:

1. **2021年10月22日16点:** 提交电子版计划书的截止时间(视为计划书正式提交时间);
2. **2021年10月29日16点:** 提交修改后电子版计划书的截止时间;
3. **2021年11月5日16点:** 报送纸质版计划书(其中一份包含申请书纸质签字盖章页)的截止时间

4. 2021年11月25日16点：报送修改后的申请书纸质签字盖章页的截止时间。

请按照以上规定及时提交电子版计划书，并报送纸质版计划书和申请书纸质签字盖章页，未说明理由且逾期不报计划书或申请书纸质签字盖章页者，视为自动放弃接受资助；未按要求修改或逾期提交申请书纸质签字盖章页者，将视情况给予暂缓拨付经费等处理。

附件：项目评审意见及修改意见表

国家自然科学基金委员会

2021年10月12日

附件：项目评审意见及修改意见表

| | | | | | |
|--|----------------------------------|-------|------|---------------------|-------|
| 项目批准号 | 22106025 | 项目负责人 | 闫希亮 | 申请代码1 | B0601 |
| 项目名称 | 基于原子尺度深度学习的纳塑料及其复合污染物构效关系和毒性预测研究 | | | | |
| 资助类别 | 青年科学基金项目 | | 亚类说明 | | |
| 附注说明 | | | | | |
| 依托单位 | 广州大学 | | | | |
| 直接费用 | 30.00 万元 | | 起止年月 | 2022年01月 至 2024年12月 | |
| 通讯评审意见： <1>具体评价意见： 一、该申请项目所关注的科学问题是否源于多学科领域交叉的共性问题，具有明确的学科交叉特征？请详细阐述判断理由并评价预期成果的科学价值。 申请项目以纳塑料及其复合污染物为研究对象，纳塑料及其复合污染物对人体健康存在极大风险，对其毒性数据、构型关系以及毒性预测模型的获得需要纳米材料、生物毒理、数学建模和人工智能等多领域的交叉融合。该申请项目通过组合化学合成吸附污染物的纳塑料库，毒理学方法检测获得其细胞毒性数据，数学建模方法模拟吸附行为构建其三维结构，并在此基础上运用人工智能手段探究纳塑料及其复合污染物的结构与其毒性效应的构效关系及预测模型，项目具有鲜明的学科交叉特性。项目的实施会为未来环境纳米材料精准构效关系分析提供理论支持。 二、请针对学科交叉特点评述申请项目研究方案或技术路线的创新性和可行性。 申请项目拟采取的研究方案具有较强的创新性，涉及到了材料、化学、生物、数学和计算科学的学科交叉，技术路线和研究方案具体详实，合理可行，而且已具有前期很好的研究基础，研究方案能够支撑研究目标的实现。 三、请评述申请人的多学科背景、研究专长和创新潜力。 申请者在纳米材料与生物体系相互作用的分子模拟和大数据分析方面有很好的研究基础和较强的创新潜力，能够确保项目的顺利完成。 四、其他建议 无 <2>具体评价意见： 一、该申请项目所关注的科学问题是否源于多学科领域交叉的共性问题，具有明确的学科交叉特征？请详细阐述判断理由并评价预期成果的科学价值。 申请人针对纳塑料的环境风险评估，结合化学合成、高通量筛选、机器学习、分子动力学模拟及复杂网络的多学科多种技术手段，探究了纳塑料结构与其毒性效应之间的定量关系。项目具有明确的学科交叉特征。 同时，环境中纳塑料的人体健康效应也是目前污染物研究的热点问题，预期获得的对纳塑料的表征、结构及理化数据库和毒性预测模型都具有较为重要的科学价值。 二、请针对学科交叉特点评述申请项目研究方案或技术路线的创新性和可行性。 1、项目研究方案可行，目标明确，研究内容贴合目标。申请人在机器学习方面已具备较为扎实的研究基础，在算法方面有良好的技术储备，具备较为完备的软硬件条件；同时，在化学和生物学方面也有很好的基础。 2. 项目研究方案和技术路线体现了化学、生物学及计算机科学的交叉融合，为纳塑料等环境纳米材料的毒性评价提供了一条可行的方案，具有较好的创新性。 三、请评述申请人的多学科背景、研究专长和创新潜力。 申请人的主要学科背景为化学，但后期研究工作中具有较显著的生物学及计算机科学交叉的研究特征；前期研究表明申请人具有较强的创新潜力。 | | | | | |

四、其他建议

1、 申请人的研究内容包括纳塑料的合成，表征、细胞活性测试，数据库构建以及深度学习预测模型构建等部分。研究内容对于青年基金三年期的项目来说体量很大，尽管申请人在前期已经有了很好的研究基础，但仍需要对项目的完成进度进行精确把控，确保项目完成。

2. 纳塑料颗粒数据库的构建是整个研究的基础，申请人设定的合成物质包括108种纳塑料及其复合污染物，后期对其进行表征及多项细胞效应的测试，其数据作为训练集数据。可以考虑加入申请人前期已经建立并发表的数据库数据共同建模。

<3>具体评价意见：

一、该申请项目所关注的科学问题是否源于多学科领域交叉的共性问题，具有明确的学科交叉特征？请详细阐述判断理由并评价预期成果的科学价值。

该项目针对纳塑料及其复合污染物的毒性预测开展研究，选题具有明显的需求导向，解决该问题需要化学、生物学、数学及计算机科学的多学科交叉，具有多学科交叉属性，项目的实施可为纳塑料的风险评价提供工具，也为其他纳米材料的毒性评价提供方法学的借鉴，具有很高的科学价值。

二、请针对学科交叉特点评述申请项目研究方案或技术路线的创新性和可行性。

该项目的研究方案和技术路线可行，具有一定的创新性。

三、请评述申请人的多学科背景、研究专长和创新潜力。

申请人具有化学、毒理学及计算机科学方面的知识背景，在纳米材料毒性预测方面取得创新性成果，具有较大的创新潜力，建议给予优先资助。

四、其他建议

修改意见：

化学科学部

2021年10月12日

国家自然科学基金资助项目批准通知

（预算制项目）

闫希亮 先生/女士：

根据《国家自然科学基金条例》、相关项目管理办法规定和专家评审意见，国家自然科学基金委员会（以下简称自然科学基金委）决定资助您申请的项目。项目批准号：22476056，项目名称：多任务深度学习设计高效去除水中典型PFAS的低毒MOF材料，直接费用：50.00万元，项目起止年月：2025年01月至2028年12月，有关项目的评审意见及修改意见附后。

请您尽快登录科学基金网络信息系统（<https://grants.nsfc.gov.cn>），**认真阅读《国家自然科学基金资助项目计划书填报说明》并按要求填写《国家自然科学基金资助项目计划书》（以下简称计划书）**。对于有修改意见的项目，请您按修改意见及时调整计划书相关内容；如您对修改意见有异议，须在电子版计划书报送截止日期前向相关科学处提出。

请您将电子版计划书通过科学基金网络信息系统（<https://grants.nsfc.gov.cn>）提交，由依托单位审核后提交至自然科学基金委。自然科学基金委审核未通过者，将退回的电子版计划书修改后再行提交；审核通过者，打印纸质版计划书（一式两份，双面打印）并在项目负责人承诺栏签字，由依托单位科研、财务管理等部门审核、签章并在承诺栏加盖依托单位公章，且将申请书纸质签字盖章页订在其中一份计划书之后，一并报送至自然科学基金委项目材料接收工作组。纸质版计划书应当保证与审核通过的电子版计划书内容一致。**自然科学基金委将对申请书纸质签字盖章页进行审核，对存在问题的，允许依托单位进行一次修改或补齐。**

向自然科学基金委提交电子版计划书、报送纸质版计划书并补交申请书纸质签字盖章页截止时间节点如下：

1. **2024年9月9日16点**：提交电子版计划书的截止时间；
2. **2024年9月16日16点**：提交修改后电子版计划书的截止时间；
3. **2024年9月23日**：报送纸质版计划书（一式两份，其中一份包含申请书纸质签字盖章页）的截止时间。
4. **2024年10月8日**：报送修改后的申请书纸质签字盖章页的截止时间。

请按照以上规定及时提交电子版计划书，并报送纸质版计划书和申请书纸质签字盖章页，逾期不报计划书或申请书纸质签字盖章页且未说明理由的，视为自动放弃接受资助；未按要求修改或逾期提交申请书纸质签字盖章页者，将视情况给予暂缓拨付经费等处理。

附件：项目评审意见及修改意见表

国家自然科学基金委员会

2024年8月23日

附件：项目评审意见及修改意见表

| | | | | | |
|--|-------------------------------|-------|------|---------------------|-------|
| 项目批准号 | 22476056 | 项目负责人 | 闫希亮 | 申请代码1 | B0601 |
| 项目名称 | 多任务深度学习设计高效去除水中典型PFAS的低毒MOF材料 | | | | |
| 资助类别 | 面上项目 | | 亚类说明 | | |
| 附注说明 | | | | | |
| 依托单位 | 华南农业大学 | | | | |
| 直接费用 | 50.00 万元 | | 起止年月 | 2025年01月 至 2028年12月 | |
| 通讯评审意见： <1>具体评价意见： 一、请评述该申请项目是否面向经济社会发展需要或国家需求背后的基础科学问题。请详细阐述判断理由。 环境中广泛存在的PFAS以及相关暴露所带来的负面健康影响备受关注，其未来的环境影响甚至可能被低估。研发其高效水处理功能材料，是治理新污染物PFAS的重要抓手。MOF被确定为一种可有效去除 PFAS 的材料类别。该项目运用分子模拟、机器学习和实验验证等多种手段，解析 MOF 材料结构组成与其吸附性能和毒性效应之间的定量关系，以实现水中典型 PFAS 的高效去除，研究内容隶属于面向经济社会发展需要背后的基础科学问题。 二、请评述申请项目所提出的科学问题的创新性与预期成果的科学价值。 该项目旨在结合理论计算和实验验证，设计能够高效去除水中典型PFAS的低毒MOF材料，以期作为生物相容性水处理功能材料设计提供预测工具和理论支持。提出拟解决的关键科学问题包括：如何构建高质量的 MOF 材料吸附性能和毒性效应数据集和如何提高 MOF 材料吸附性能的同时，降低其毒性效应。研究 MOF 结构、PFAS 性质和水基质对 PFAS@MOF 吸附的作用过程是决定项目实施的关键问题。预期利用多任务深度学习，解析MOF结构与其吸附性能和毒性之间的定量关系，进一步通过虚拟筛选和实验验证，设计出同时具有高吸附性能和低毒性的新型MOF材料，选题具有创新性和科学价值。 三、请评述该申请项目的研究基础与可行性；如有可能，请对完善研究方案提出建议。 申请人前期从事数据驱动的生物相容性功能材料设计方面研究，熟悉NAMD、GROMACS分子模拟软件，同时在利用机器学习、深度学习方法构建模型、预测纳米-生物效应方面研究经验丰富，取得的科研成果突出，为该申请项目实施奠定了良好基础。 申请人在申报项目书中未给出毒性参数的选择性原则；避免使用吸附剂处理中的二次污染问题，应探讨吸附剂再生条件。 四、其他建议 <2>具体评价意见： 一、请评述该申请项目是否面向经济社会发展需要或国家需求背后的基础科学问题。请详细阐述判断理由。 MOF材料是从水中去除PFAS的理想吸附剂。本项目围绕MOF材料的传统设计方法周期长成本高且存在二次污染风险的问题，采用理论计算和实验验证手段，设计用于高效去除水中典型PFAS的低毒MOF材料。总体上，本研究立意新颖，具有鲜明的需求和目标导向性，研究方案合理，研究手段先进，建议给予优先资助。 二、请评述申请项目所提出的科学问题的创新性与预期成果的科学价值。 本项目采用高通量分子模拟、大数据挖掘、多任务深度学习等多种计算手段，解析MOF结构与其吸附性能和毒性之间的内在关系，设计出同时具有高吸附性能和低毒性的新型MOF材料，具有较强的创新性。本项目的实施，有助于克服水处理过程中“一边治理一边污染”的困境，服务于饮用水安全保障。 三、请评述该申请项目的研究基础与可行性；如有可能，请对完善研究方案提出建议。 | | | | | |

项目申请人在生物相容性功能材料的设计、机器学习建模、纳米生物效应预测等方面具有良好的研究基础，可以保证工作的顺利开展，有望取得创新成果。

四、其他建议

<3>具体评价意见：

一、请评述该申请项目是否面向经济社会发展需要或国家需求背后的基础科学问题。请详细阐述判断理由。

PFAS是典型的新污染物，该项目针对水体中PFAS高效低毒的MOF去除材料，对PFAS的污染防治具有重要意义，有利于当前我国新污染物的风险防控。目前在PFAS去除材料分子设计方面，一个难点就是如何确保设计的MOF材料既能高效又要低毒，该项目应对了当前PFAS去除材料设计的迫切需要，具有重要的社会意义。

二、请评述申请项目所提出的科学问题的创新性与预期成果的科学价值。

本项目的关键科学问题是如何设计高效低毒的PFAS去除MOF材料。申请者创新性采用多任务机器学习并结合实验方法开展相关研究，项目具有极大的创新性，并且预期能够取得的成果在该领域具有重要的学术引领作用。

三、请评述该申请项目的研究基础与可行性；如有可能，请对完善研究方案提出建议。

申请人长期从事机器学习及纳米材料研究，在分子动力学模拟及机器学习方面具有很深的学术水平，特别在理论计算方面在Nature Communication等杂志发表多篇高水平论文，具备了较好的前期研究基础。论文研究方案设计合理，研究论证充分，能够保障项目的顺利实施。

四、其他建议

修改意见：

化学科学部

2024年8月23日

特定高校学科建设专项（人才引进类）

项目（课题）任务书

项目名称： 国猪免疫代谢特征形成的机制解析

课题名称： 抗菌肽智能设计及其在畜禽养殖中的应用

项目起止时间： 2024年01月01日至2025年12月31日

管理单位（甲方）： 华南农业大学

依托学院（乙方）： 动物科学学院

课题负责人（丙方）： 闫希亮 联系电话： 13884988366

课题联系人： 闫希亮 联系电话： 13884988366

华南农业大学
二〇二二年制

一、研究计划

(一) 主要研究内容及创新点

1. 主要研究内容

(1) 基于机器学习的抗菌肽定量构效关系研究

收集抗菌肽的公开序列数据和实验验证的活性数据，包括其抗菌谱、最低抑菌浓度（MIC）、细胞毒性等。对抗菌肽进行特征提取，考虑氨基酸组成、理化性质（如疏水性、电荷分布、分子量等）、序列模式（如二级结构、序列保守性）、三维结构信息等。采用多种机器学习算法（如支持向量机、随机森林、深度学习等）构建抗菌肽的定量构效关系（QSAR）模型，预测其抗菌活性与毒性。优化模型参数，通过交叉验证评估模型的泛化能力和预测精度。通过特征重要性分析，识别影响抗菌肽活性和安全性的关键因素，如氨基酸的特定排列模式、电荷分布或疏水性区域；探讨这些关键特征与抗菌机制的内在联系，为抗菌肽的设计提供理论依据。

(2) 新型抗菌肽虚拟筛选和设计

利用生成对抗网络（GAN）或变分自动编码器（VAE）生成多样化的抗菌肽序列库，涵盖不同长度、理化性质和序列特征的多种候选肽。结合进化算法或规则约束，优化生成的序列，以符合设计目标（如广谱抗菌性、高稳定性、低毒性等）。通过深度学习模型预测候选抗菌肽的抗菌活性、特异性和安全性。模型输入特征包括氨基酸组成、疏水性、电荷分布、二级结构等。利用多目标优化算法对候选肽进行筛选和优先级排序，筛选出高潜力的候选抗菌肽。使用分子动力学模拟评估候选抗菌肽在细菌膜上的作用机制，包括膜穿透性、孔形成能力及稳定性。

(3) 抗菌肽在畜禽养殖中的应用研究

研究抗菌肽在畜禽肠道中的作用机制，包括对肠道菌群结构的调控、病原菌抑制及有益菌促进效果。探讨抗菌肽在改善饲料利用率、促进畜禽生长性能和提高健康水平中的潜在作用。系统评估抗菌肽对畜禽机体的急性毒性、慢性毒性及免疫系统影响，确保其使用的安全性。研究抗菌肽对畜禽生殖能力、生长发育和代谢的长期影响。开发适用于饲料添加的抗菌肽制剂，包括微胶囊化、缓释技术及热稳定性改良。研究抗菌肽在不同饲料基质中的适配性及稳定性，确保其在加工、储存及运输过程中的功能保持。

2. 创新点

(1) 跨学科融合的智能设计方法和优化策略

将人工智能技术与抗菌肽设计深度结合，显著提升抗菌肽设计的效率和准确性。开发基于深度学习的预测模型，能够在海量序列中快速筛选出高效抗菌肽。结合分子动力学模拟和实验验证，提出了全新的抗菌肽序列优化方法，有效提高抗菌肽的稳定性和生物活性。

（2）应用导向的研发思路

以畜禽养殖的实际需求为目标，设计出广谱、低毒且易于工业化生产的抗菌肽产品。提出了一套针对饲料添加剂开发的抗菌肽评价体系，为大规模推广应用提供了技术支撑。

（3）全面的安全性及耐药性研究

系统分析抗菌肽对畜禽机体及环境的潜在影响，填补了抗菌肽长期使用安全性评估的研究空白。针对抗菌肽诱发耐药性的问题，提出了基于序列多样化的抗性规避策略。为替代抗生素的绿色养殖模式提供了新思路，推动了畜禽养殖行业的可持续发展。

（二）拟开展的研究在国际国内同领域所处的地位

本研究在国际上属于智能设计与畜禽养殖应用结合的前沿探索，在国内则是首次尝试全链条研究的系统化工作。无论从研究的创新性、应用的针对性还是产业化的潜力来看，都处于国内领先、国际先进的水平，为解决畜禽养殖领域的抗生素替代问题提供了新的技术方向和理论基础。

（1）国际领域

抗菌肽的智能设计目前在国际上是一个热点研究领域，特别是在机器学习和人工智能技术与生物学深度融合的背景下，已有诸多团队尝试利用深度学习、分子模拟等技术进行抗菌肽的虚拟筛选和优化设计。但是，目前国际研究多集中于人类医疗领域，如抗生素替代药物的开发和癌症治疗，针对畜禽养殖领域的应用研究较少且多处于初级阶段。本研究在特定应用场景的针对性和实际落地性上具有国际前沿性。

（2）国内领域

国内针对抗菌肽的研究多集中于实验筛选和初步性能评价，尚未广泛应用人工智能进行抗菌肽的智能化设计。本研究创新性地将深度学习、分子模拟与抗菌肽的设计和优化结合，填补了国内智能设计领域的空白，是国内在该方向上的重大技术突破。国内畜禽养殖业规模庞大，抗生素使用问题亟需解决，然而抗菌肽研究在养殖领域的应用开发尚

处于起步阶段，针对性研究较少。

（三）开展的研究对提升我国相关领域科技创新能力和发展战略性新兴产业等的主要作用

本研究将有助于提升我国在抗菌肽智能设计和应用领域的科技创新能力，为解决畜禽养殖中的抗生素替代问题提供绿色技术方案，并推动生物制品、生物农业等战略性新兴产业的快速发展。同时，这项研究在支撑国家食品安全、绿色转型和国际竞争力提升方面具有重要的现实意义和战略价值。

（1）提升我国生物科技创新能力

本研究将人工智能技术应用于抗菌肽设计与优化，推动了生物信息学与机器学习在抗菌肽领域的深度融合，填补了我国在抗菌肽智能化设计方面的技术空白。建立抗菌肽分子筛选与性能预测模型，显著提升了高效抗菌肽的设计速度和准确性，为未来生物医药、农业和环保领域的分子设计提供参考。

（2）推动畜禽养殖领域的绿色转型

研究开发绿色、安全、高效的抗菌肽饲料添加剂，作为抗生素的可行替代方案，助力我国减少抗生素使用，解决抗生素残留和细菌耐药性等问题。推动“无抗养殖”模式的规模化应用，有助于提升我国畜禽产品的国际市场竞争力，符合全球食品安全和环保趋势。

（3）服务国家“双碳”与可持续发展战略

抗菌肽的应用能够减少传统抗生素的环境污染和抗药性细菌的扩散，符合绿色环保和可持续发展的要求。通过改善养殖健康水平和降低疾病传播风险，为实现资源节约型、环境友好型养殖业提供技术保障。

（四）科研组织管理、国内外合作设想

（1）科研组织管理

由项目负责人和核心团队成员组成，负责整体研究计划制定、资源调配与项目进度把控。根据研究内容设立多学科研究子组，如数据处理组、算法开发组、生物实验组、产业转化组，明确任务分工，确保工作高效开展。建立共享数据库和智能设计平台，促进各组间数据流通和资源共享。制定明确的研究时间节点和成果目标，对每阶段的任务完成情况进行严格评估。根据评估结果动态调整研究重点，优化技术路线，确保研究方向符合应用需求。

（2）国内合作设想

与国内顶尖高校和科研院所合作开展基础研究，如抗菌肽的作用机制、结构与功能关系研究。结合各单位优势，搭建联合实验平台，共享资源与成果。与大型饲料企业、养殖企业合作，推动抗菌肽在饲料和畜禽养殖中的应用研究。通过产学研结合，开发抗菌肽制剂，推动产品化和规模化应用。联合生物信息学、分子生物学、农业科学等领域的专家，协同攻关，解决抗菌肽设计中的应用中的技术难题。

（五）个人能力提升、人才培养和团队建设

深入学习抗菌肽设计相关领域的前沿理论和技术，包括分子生物学、生物信息学、机器学习、分子模拟等，夯实多学科交叉基础。掌握最新的抗菌肽作用机制研究进展，提升对抗菌肽结构与功能关系的理解。学习并熟练掌握智能设计工具（如深度学习模型、分子模拟软件），提升在抗菌肽智能设计中的技术应用能力。加强实验技能，特别是抗菌活性测定、分子修饰与毒性评价等实验技术的应用水平。

培养既掌握抗菌肽基础研究技能，又具备智能设计、分子模拟与应用开发能力的复合型人才。通过课程培训、课题研究和技术交流，提升团队成员的跨学科协作能力。定期举办学术研讨会与技术沙龙，鼓励团队成员从多角度提出研究问题，激发创新灵感。加强科研伦理与规范教育，提升团队的学术素养和科学严谨性。

二、预期考核目标（参照人才引进合同指标填报）

1.教学任务：

(1)承担本科生及研究生的教学任务，每年至少完整讲授1门本科生课程，其中按照培养计划开设并计入学分的课程教学学时数不少于32。

(2)指导本科生创新创业实践或指导实践教学、毕业设计、毕业论文。

2.科研任务

(1) 以第一作者或通讯作者身份发表2篇SCI类论文

3.学科建设任务

协助课题组做好“智慧养殖”相关专业建设。

4.学科专业人才培养任务

按学校、学院规定培养或协助培养研究生。

三、经费预算

| 直接费用 | 经费额 (万元) | 用途说明 |
|--|----------|-------------------------|
| (1) 设备费 | 3.00 | 机器学习和分子模拟服务器购买 |
| (2) 材料费 | 10.00 | 抗菌肽合成材料购买 |
| (3) 测试化验加工外协费 | 17.00 | 抗菌测试以及超算机时租用 |
| (4) 实验室维修改造费 | 0.00 | 无 |
| (5) 出版/文献/信息传播/知识产权事务费/信息资料费(含科技查新费、复印打印、资料费等) | 2.00 | 文章出版版面费、专利申请等知识产权事务费 |
| (6) 数据采集费 | 0.00 | 无 |
| (7) 交通费 | 2.00 | 参加国内外相关学术会议注册费、交通费、住宿费等 |
| (8) 差旅费 | 2.00 | 参加国内外相关学术会议注册费、交通费、住宿费等 |
| (9) 劳务费 | 4.00 | 用于项目成员中没有工资性收入的研究生劳务费发放 |
| (10) 其他 | 0.00 | 无 |
| 合计 | 40.00 | - |
| 其他需说明的情况: 无 | | |

四、签约各方

管理单位（甲方）： 华南农业大学（盖章）

科研部门负责人（签章）：

胡传双

2024年12月3日

依托学院： 动物科学学院（盖章）

学院负责人：

（签字）

项目负责人：

（签字）

2024年12月04日

课题负责人（丙方）：

本人承诺由特定高校学科建设专项（Specific university discipline construction project）经费资助产出的相关科研成果，发表论文等成果将标注项目资助编号“2023B10564001”。

闫希亮（签字）

2024年11月30日

编号: HN20231001

华南农业大学引进人才 聘用合同

聘用单位: 华南农业大学

受聘人员: 闫希亮



华南农业大学人力资源处

填写说明

1. 本聘用合同书根据《中华人民共和国劳动合同法》《事业单位人事管理条例》《广东省用人单位用工管理制度》等制定，作为聘用单位与受聘人员签订聘用合同的文本。除本合同所列内容外，经聘用单位与受聘人员协商一致可增加有关条款。

本合同中空出的栏目，由双方协商确定的，须填写清楚。不需填写的栏目可写“无”或划上斜杠“/”。

2. 填写聘用合同书应当字迹清晰、工整，涂改处必须加盖校对章，否则无效。

3. 本聘用合同书应当由聘用单位和受聘人员双方当事人亲自签订。单位法定代表人因故确需代签的，应当经本书面委托，否则代签无效。

4. 本聘用合同书内的年、月、日均用阿拉伯数字填写。工资报酬等金额一律使用大写。

5. 本合同的未尽事宜，经双方协商一致，可在“双方约定的其他事项”中列明，或另行签订补充协议，作为本合同附件。补充协议效力等同于聘用合同，与聘用合同一并履行。

华南农业大学引进人才聘用合同

甲方：华南农业大学

乙方： 闫希亮

为保证甲方人才引进计划顺利实施，实现甲乙双方责权利的统一，保障甲乙双方的合法权益，根据《中华人民共和国教师法》、《华南农业大学人才引进和管理办法》的文件精神，经双方协商，订立本合同。

第一条 聘用岗位及聘期

甲方按 突出人才 层次引进乙方，聘用乙方为 动物科学 学院 智慧养殖 学科的 首聘副教授。

乙方在甲方的服务期至少为十年，聘期自 2023 年 10 月 1 日至 2033 年 9 月 30 日。其中，首个聘期为五年，聘期自 2023 年 10 月 1 日至 2028 年 9 月 30 日。首个聘期期满后，续聘工作按学校相关管理办法执行。

首聘专业技术职务 首聘副教授 聘期自 2023 年 10 月 1 日至 2026 年 9 月 30 日。（注：若作为高级职称调入者，则不需要填写首聘专业技术职务聘期）

首聘副教授（副研究员）实行“3+2”，即：如果3年内达到学校副教授（副研究员）职称评审条件，但是没有晋升的，首聘副教授（副研究员）聘期由3年调整为5年，即再延长2年，聘期内待遇仍按照首聘副教授（副研究员）兑现。如果3年内未达到副教授（副研究员）职称评审条件者，首聘副教授（副研究员）聘期结束，不再延长，也不再享受首聘副教授（副研究员）待遇。

第二条 乙方的岗位工作任务（各学院、各学科可根据实际情况修改）

乙方在首个聘期内应完成的工作任务（以甲方为第一完成单位，乙方为第一完成人）：

教育教学、科学研究业绩突出，个人入选珠江学者、广东特支计划等省级领军人才或国家级人才计划（项目）；或完成以下具体任务：

1. 教学任务（包括承担核心课程的讲授任务、必修课或选修课讲授任务等）：

根据《华南农业大学新进教师“双证”上讲台暂行实施办法》华南农办〔2020〕58号规定，原则上在报到入职1年内获得“双证”，持证上岗后需完成以下任务：

（1）承担本科生及研究生的教学任务，每年至少完整讲授1门本科生课程，其中按照培养计划开设并计入学分的课程教学学时数不少于32。

（2）指导本科生创新创业实践或指导实践教学、毕业设计、毕业论文。

2. 科研任务（包括科研项目、经费、论文、论著、奖励、专利等）：

（1）至少主持获得1项国家自然科学基金，主持到位科研经费60万元以上。

（2）以第一作者或通讯作者至少发表1篇T1类论文，或1篇T2类论文和1篇顶级期刊（需由学校人才引进考核工作小组同行专家进行论证后确定）论文，或3篇T2类论文，或6篇A类论文。

3. 学科建设任务：

协助课题组做好“智慧养殖”相关专业建设。

4. 学科专业人才培养任务（包括培养博士、硕士研究生，指导博士后研究人员、高级访问学者和青年教师等）：

按学校、学院规定培养或协助培养研究生。

5. 其他任务：

完成学院安排的其它工作任务。

第三条 权利和义务

（一）甲方权利

1. 甲方根据学校的有关规定以及第二条所规定的岗位工作任务，对乙方进行管理。

2. 甲方依照国家法律、法规及学校的有关规定，对乙方进行考核和奖惩。

（二）甲方义务

1. 依法保障和维护乙方应享有的各项权利。

2. 首个聘期内直聘为学术型和专业型硕士研究生导师

3. 为乙方提供必要的工作和生活条件

（1）科研配套经费：为乙方提供陆拾 万元（¥600, 000）人民币科研启动费。科研启动费根据乙方科研需要拨付，原则上期中考核前拨付的金额不超过合同约定金额的一半。中期考核不合格者剩余部分不予拨付。

（2）生活条件：①为乙方提供校内过渡租住房一套，过渡租住房的房租标准执行学校的现行相关政策与管理规定。②同时配套安家费伍拾万元（¥500, 000）人民币（税前），正式报到到岗后分2次给付。第一次拨付安家费的一半，有正式购房合同且中期考核合格后再拨付另一半；若中期考核不合格，则待期满考核合格后再拨付另一半。

（三）乙方权利

1. 乙方享受甲方按国家规定提供的工资、保险、福利等待遇，按照甲方的有关规定，在聘期内享受相关的校内津贴。

2. 享受甲方为其提供的工作和生活条件。

(四) 乙方义务

1. 认真遵守国家法律、法规和甲方的各项规章制度。

2. 乙方全职在甲方工作。

3. 乙方全面履行岗位职责，完成岗位工作任务；在第一个聘期内，不得与其他用人单位建立有偿兼职劳动关系；接受甲方的监督、考核及管理。

4. 乙方在甲方工作期间，所取得的工作成果，均按甲方知识产权管理规定执行。

5. 乙方若申请调出或辞职或被辞聘，须接受学校对给付的科研启动费进行财务审计。

第四条 考核

(一) 乙方应按甲方现行的年度考核办法参加甲方正常的年度考核，首个聘期内年度考核全部材料及结果同时报人事部门备案。

(二) 乙方在首个聘期，接受甲方的工作满三年中期考核和聘期期满考核。中期考核和聘期期满考核根据聘用合同岗位工作任务开展。

第五条 合同的变更与解除

(一) 乙方在服务期内如不能履行本合同所规定的岗位职责，中期或期满考核不合格或有违法违纪行为，甲方有权予以解聘，解除本合同，同时，乙方承担如下责任：乙方须将剩余的科研启动费和给付的全部安家费如数退还甲方，除此之外，如对甲方造成其他损失，甲方还有权对乙方追究损失赔偿。

(二) 乙方在服务期内提出辞聘的，需提前三个月向甲方提出书面申请，经甲方考核同意后乙方方可辞聘，但乙方承担相应

的违约责任：安家费按 10 年期剩余的未服务年限 \times 10 年平均数的标准交回甲方，配套的科研启动费余额由学校收回，同时，乙方应一次性缴纳服务期未满违约金，违约金标准为：服务期未满五年，缴纳人民币 10 万元；服务期超过五年（含五年），缴纳人民币 5 万元。

（三）因甲方不能履约而导致乙方无法开展工作的，乙方可提前三个月提出辞聘申请，甲方无法改进的，乙方在 3 个月期满后可单方解除合同，但乙方在解除合同的同时要退还甲方在引进时给付的按 10 年期剩余的未服务年限 \times 10 年平均数标准的安家费和剩余的科研启动费。

（四）凡发生上述第五条（一）、（二）、（三）款中任何一种情况的，甲方同时解聘由甲方安排乙方校内工作的配偶或停止对乙方配偶发放剩余的安置费。

（五）聘用期间如发生双方无法预见、无法防范而致使合同无法正常履行的事由，需要变更或解除合同的，聘用双方应按照国家有关规定妥善处理。

第六条 附则

（一）乙方同意，在其处于联系障碍状态（包括但不限于乙方因重病住院、发生意外事故、丧失人身自由等情形）时，委托紧急状态联系人（姓名）刘国红、身份证号码371402198906066125、通讯地址广东省广州市天河区五山街道华南农业大学嵩山区 20 栋 401、联系电话15053173015作为乙方的受委托人，该受委托人享有全权代理乙方处理本合同项下所涉一切问题的权限，包括但不限于与甲方进行协商、和解、代为收付有关款项及代为收发有关文书等。

（二）本合同一式三份，甲方、乙方及乙方所在学院各持一份。本合同于双方签字盖章并乙方到岗之日起生效。

（三）除发生不可抗力因素致使合同无法履行外，双方应严格履行合同中的各项条款，如发生争议，双方应协商处理，对合

同有关条款的变更，应征得对方同意。不愿协商、协商不成或者达成和解协议后不履行的，双方均有权向劳动争议仲裁委员会申请仲裁；对仲裁裁决不服的，可以向人民法院提起诉讼。

(四) 本合同如有未尽事项，应由双方协商，做出补充规定。补充规定与本合同具有同等效力。补充规定：

- 1、无
- 2、无
- 3、无

(此处以下无内容)

甲方：(盖章)

法定代表人：(或委托代理人)

2023 年 10 月 1 日

乙方：闫希亮

身份证号码(或护照号)：

370181199103036515

2023 年 10 月 1 日

所在学院(部)负责人：(单位行政公章)

2023 年 10 月 1 日

项目编号： 202201010541

基础与应用基础研究项目 合同书

项目名称： 基于端到端深度学习的纳塑料毒性预测研究

承担单位： 广州大学

项目负责人： 闫希亮

计划类别： 基础研究计划

专题名称： 基础与应用基础研究项目

支持方向： 一般项目（博士青年科技人员类）

组织单位： 广州大学

起止时间： 2022年04月01日 至 2024年03月31日

主管处室： 基础研究处

广州市科学技术局
(二〇二二年制)

填写说明

一、本合同书的项目编号由市科学技术局（以下简称市科技局）统一确定。

二、本合同书由申报书在后台自动转换生成，如有错漏之处需修正，请联系市科技局项目责任处室退回承担单位修正。

三、项目经费分为直接费用和间接费用。

基础与应用基础研究专题项目试点实施“包干制”，经费支出不设科目比例限制，由项目研究团队自主调剂使用，按照市科研项目经费“包干制”管理有关规定执行，同时应符合以下要求：

（1）经费支出应实际用于项目研究支出，使用范围限于设备费、材料费、测试化验加工费、燃料动力费、差旅/会议/国际合作与交流费、出版/文献/信息传播/知识产权事务费、劳务费、专家咨询费、依托单位管理费用、绩效支出以及其他合理支出。

（2）经费支出应按照市级财政科研项目资金开支范围和标准使用；

（3）间接费用是指项目承担单位在组织实施项目过程中发生的无法直接列支的相关费用，主要用于补偿项目承担单位为了项目研究提供的现有仪器设备及房屋，水、电、气、暖消耗，有关提高科研管理、服务能力等费用，以及绩效支出等。间接费用按照不超过项目直接费用扣除设备购置费后的一定比例核定，具体比例按《广州市财政局 广州市科学技术局 广州市审计局关于市级财政科研项目资金绩效提升和管理监督办法》规定确定。

（4）不得列支基建费；

（5）项目验收时应提交经费决算表

四、本合同书仅适用于广州市基础与应用基础研究项目

一、基本信息

| | | | | | | | | |
|-------|------|-------------|-------------|-------------|------|--------------------|------|------------------------|
| 项目负责人 | 姓名 | 闫希亮 | 证件类型 | 身份证 | 证件号码 | 370181199103036515 | 性别 | 男 |
| | 出生年月 | 1991年03月03日 | 民族 | 汉族 | 国籍 | 中国 | 学历 | 博士研究生 |
| | 学位 | 博士 | 学位授予国家(或地区) | 中国 | 职务 | 无 | 职称 | 讲师(高校) |
| | 所学专业 | 分析化学 | 手机号码 | 13884988366 | 办公电话 | 13884988366 | 电子邮箱 | yanxiliang1991@163.com |

| | | | | |
|--------|------|---------------------|--------------------|--------------------|
| 项目承担单位 | 单位名称 | 广州大学 | 统一社会信用代码或组织机构代码 | 124401007348911139 |
| | 注册时间 | 2000-07-07 | 单位类型 | 高等院校 |
| | 注册地址 | 广东省广州市番禺区外环西路230号 | | |
| | 办公地址 | 广东省广州市番禺区外环西路230号 | | |
| | 联系人 | 姓名 | 刘亚兰 | |
| | | 手机号码 | 15602407931 | |
| | | 电子邮箱 | yalanl@gzhu.edu.cn | |
| | 开户银行 | 工行广州大学城中环支行 | | |
| | 开户户名 | 广州大学 | | |
| | 银行帐号 | 3602114819100000192 | | |
| 研究平台 | | | | |

| | | | | |
|--------|---|--------------------------|------|-------------------------|
| 项目基本信息 | 项目名称 | 基于端到端深度学习的纳塑料毒性预测研究 | | |
| | 所属学科 | 化学-环境化学-理论环境化学-环境污染模式与预测 | | |
| | 申请金额 | 5.00万元 | 研究期限 | 2022年04月01日-2024年03月31日 |
| 项目摘要 | 纳塑料及其复合污染物存在极大地对人体健康风险，迫切需要对其毒性进行快速准确评估。然而，受限于纳塑料毒性数据缺乏，以及纳米结构信息提取片面化（如忽略真实粒径分布、表面配体/吸附污染物数目和位置等特征）和纳米描述符构建的难题，我们难以建立有效的预测模型。本项目旨在基于纳塑料毒性大数据、分子模拟和深度学习，从原子尺度精准探究纳塑料及其复合污染物完整三维结构和毒性效应之间的定量关系。拟通过组合化学合成吸附不同环境污染物的纳塑料颗粒库，并通过高通量筛选对纳塑料细胞效应进行测试，为深度学习提供高质量的纳塑料毒性大数据；同时利用分子动力学，对纳塑料自组装和表面吸附行为进行模拟，从原子尺度构建纳塑料及其复合污染物准确三维结构；通过图神经网络直接提取三维结构特征，并结合毒性大数据，建立无需描述符的端到端预测模型。项目的实施可为纳塑料风险评估提供预测工具和理论支持，并促进人工智能在毒理学研究中的应用和落地。 | | | |

二、项目预期成果

| | | | | | | | | |
|---------|-------------------------|----|--------|----|---------|----|------|----|
| 论文及专著情况 | 国家统计源刊物以上刊物发表论文（篇） | | 5 | | 科技报告（篇） | | 0 | |
| | 其中，被SCI/EI/ISTP收录论文数（篇） | | 3 | | 培养人才（人） | | 2 | |
| | 专著（册） | | 0 | | 引进人才（人） | | 0 | |
| 专利情况(项) | 发明专利 | | 实用新型专利 | | 外观设计专利 | | 国外专利 | |
| | 申请 | 授权 | 申请 | 授权 | 申请 | 授权 | 申请 | 授权 |
| | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 其他 | 无 | | | | | | | |

三、项目经费预算

本项目总投入： ¥（ 5.00 ）万元，其中，市财政补助经费：¥（5.00）万元，自筹经费：¥（ 0 ）万元。

(单位：万元)

| 1. 经费下达计划 | | | |
|-----------|------|--------|------|
| 资金来源 | 小计 | 市科技局经费 | 自筹资金 |
| 2023年 | 5.00 | 5.00 | / |
| 合计 | 5.00 | 5.00 | / |

注：本项目市科技局经费试点实施“包干制”，经费支出不设科目比例限制，由项目研究团队自主调剂使用，按照市科研项目经费“包干制”管理有关规定执行。

四、合同条款

第一条 甲、乙、丙方根据《中华人民共和国科学技术进步法》《广东省自主创新促进条例》《广州市科技创新条例》及《中华人民共和国民法典》等国家有关法规和规定，经协商一致，特订立本合同，作为甲、乙、丙方在合同执行中共同遵守的依据。

第二条 甲方、乙、丙三方应当严格履行《广州市科技计划项目管理办法》《广州市级财政科研项目资金绩效提升和管理监督办法》《广州市科技创新发展专项资金管理办法》《广州市科技计划项目经费“包干制”改革试点工作方案》《广州市科技计划项目全过程管理简政放权改革工作方案》的规定要求。

第三条 甲方应：

1. 根据财政经费预算安排，及时进行拨付项目经费。
2. 赋予乙方和丙方广州市科技业务管理阳光政务平台（以下简称阳光政务平台）的使用权限，保障丙方进行项目全过程管理的使用需求。
3. 根据甲方需要，在不影响乙方工作的前提下，定期或不定期对乙方项目的实施情况和经费使用情况进行检查或抽查。
4. 审核丙方提交的年度工作报告，制定下一年度的资金切块方案。
5. 对丙方进行周期绩效考核和检查评估，重新评估丙方资格。

第四条 乙方应：

1. 作为项目具体组织实施的责任主体，为本单位提供的与本项目有关的全部材料真实、合法、有效性负责，同意甲方向行业协会等第三方机构直接调取乙方与本项目相关的数据、信息、材料，包括但不限于工商登记信息、审计报告等；
2. 按照《合同书》规定的内容组织实施项目，接受并配合甲方、丙方以及各级财政、审计部门，或上述部门委托的机构进行评估、稽查、审计、检查和绩效评价，并按要求提供项目任务与预算执行情况和有关财务资料；
3. 按照市财政科技经费管理“包干制”相关要求对项目经费单独设账，专款专用；
4. 保证自筹资金按时到位和其它配套条件的落实；
5. 在项目研究开发过程中优先考虑使用“广东省科技资源共享服务平台”的仪器设备，项目购置的设备仪器若符合入网条件应及时办理入网手续对社会共用共享，提高设备仪器的使用率。按照《中华人民共和国采购法》要求，对符合政府采购范围的设备仪器，执行政府采购；
6. 项目合同执行期内需进行变更的，按照《广州市科技计划项目管理办法》《广州市级财政科研项目资金绩效提升和管理监督办法》《广州市科技创新发展专项资金管理办法》《广州市科技计划项目全过程管理简政放权改革工作方案》相关程序办理；
7. 项目合同执行期满后3个月内向丙方提出验收申请，并出具在广州注册会计市协会备案的验收专项审计报告，提前完成合同规定任务的可提前申请验收；
8. 按照相关规定，在项目验收时提交科技报告，办理《验收证书》和科技成果登记手续；

9. 在项目实施期间和项目结题验收后3年内，配合甲方开展对财政资金年度绩效跟踪，按照甲方要求提供相关信息和数据，完成年度报告填报任务；

第五条 丙方应：

1. 明确项目管理依据的管理办法或管理规程，承担项目全过程管理职责；
2. 自主安排立项评审和结题验收工作，充分利用阳光政务平台，推进项目全过程管理的网络化电子化，主动配合推行合同电子签章；
3. 严格落实信息公开制度，公示遴选和结题验收结果，并及时处理异议；
4. 及时报送相关材料，按广州市科学技术局要求，每年按时提交拟立项项目清单，报送年度工作总结；
5. 按广州市科学技术局要求配合开展绩效评价和监督检查工作；
6. 主动追回终止项目未使用和不合规支出的市财政科技经费；
7. 按照本单位相关项目管理办法组织项目验收工作，并按相关规定做好存档工作；
8. 协助甲方对项目的实施过程进行跟踪、检查和提供相关信息，并对所提供信息的客观真实性负责；
9. 负责监管乙方严格遵守本合同规定的任务；

第六条 甲方同意给予乙方人民币（5.00万）的资助，立项后一次性拨付。

第七条 合同终止：

1. 项目因故无法继续进行的，按照相关规定实施合同终止。
2. 发现存在以下情况之一的，立即启动终止程序：
 - ①因不可抗拒因素导致项目无法继续进行、没有必要继续进行或无法完成合同预期目标任务的；
 - ②不接受项目监督检查、检查不合格限期整改后仍未通过的或拒不配合项目验收工作的；
 - ③无正当理由项目合同执行期满后3个月以后仍未提交验收申请的；
 - ④项目承担单位已迁出本市，或已停止经营活动，或已注销的；
 - ⑤发现在项目申报、实施过程中有违法、欺骗等事实的；
 - ⑥存在其他导致项目不能正常实施的原因。
3. 合同终止由乙方提出申请，丙方审定，也可由丙方强制实施。具体由丙方按照《广州市科技创新发展专项项目全过程管理简政放权改革试点工作方案》的规定要求进行办理。
4. 合同终止后，乙方或承接乙方法定义务的责任人应停止使用该项目财政经费；上缴尚未使用和使用不符合规定的财政经费。

第八条 对合同正常执行期及项目整改期之外的经费开支，不属于财政项目经费列支范围。

第九条 在履行本合同的过程中，乙方发现可能导致项目失败或部分失败的情形时，应及时通知甲方，并采取适当措施减少损失，没有及时通知并采取适当措施，致使损失扩大的，应当就扩大的损失承担责任。

第十条 在履行本合同的过程中，如遇到市财政计划改变等不可抗力情况，甲方对所核拨经费的数量和时间可进行相应变更。

第十一条 成果转化：

1. 本项目技术成果及知识产权的归属、转让和实施技术成果所产生的经济利益的分享，除另有约定外，按国家和省、市有关规定执行；正式发表的论文、论著应标注“广州市科技计划项目资助”字样及项目编号；项目所取得的技术成果和知识产权应优先广州产业化或推广转让。

第十二条 属技术保密的项目，经协商订立如下技术保密条款：

1. 本合同书保密内容范围为：本合同及其补充协议和附件、乙方因履行本合同所接触或知晓的甲方工作秘密（包括但不限于甲方的任何技术性资料、以及甲方为完成本合同提供的任何其他信息资料并且在提供时未说明是公开信息的）、在合同履行过程中，乙方接触到的，或履行合同产生的任何国家、商业、工作信息（包括但不限于计算机系统数据信息、审计工作资料、技术文档及相关敏感资料等）。

2. 本合同书保密期限为：\

3. 乙方及乙方人员（包括但不限于项目组人员、乙方雇员、代理人、顾问等工作人员，下同）采取有效的保密措施以避免泄露给任何第三方；乙方增强对项目组人员的保密教育，每年至少开展一次保密自查，并与可能知悉保密内容的人员签订技术保密保护协议，确保项目组人员遵守保密协议，乙方应保密义务不得低于本合同书的约定；甲乙双方应建立技术保密制度。

4. 乙方在合同履行的过程中，对接触到的相关信息，乙方及项目组人员承担保密责任；乙方应将项目组人员的身份证复印件、劳动合同、学历职称证明、项目经验等资料提供给甲方，更换项目负责人时需事先征得甲方书面同意并提交上述资料。

5. 在本合同有效存续期间及合同终止后，未经甲方事先的书面同意，不得以任何方式记录、复制、拍摄、摘抄、收藏、公布、发表、公开、披露、散播本合同项下保密信息的任何部分，或对其加以任何形式的利用或使用；乙方及乙方人员未经甲方书面同意不得私自下载、拷贝计算机内项目相关数据信息，不得擅自携带记载项目内容的载体（例如移动硬盘、U盘等）和打印资料外出，严禁将工作系统的程序、口令等泄露给他人。

6. 属技术保密的项目必须经相关负责技术保密部门审查、批准后，方可发表或用于境外合作与交流。

7. 乙方应当制定泄密应急预案，一旦发现本单位持有的国家科学技术秘密可能泄露或者已经泄露，应当在24小时内向甲方报告，同时启动应急预案，并协助有关部门查处泄密事件。

8. 乙方应严格遵守国家、省市规定的其他技术保密相关法律、法规和政策。在项目实施过程中，乙方或项目合作单位及其相关人员违反科学技术保密管理规定，给国家安全和利益造成损害的，应当依法追究单位和相关人员的法律责任。

第十三条 廉洁责任

甲方、丙方、评审机构及其工作人员不得索取、收受利益相关方财物或其他不正当利益，严格遵守中央八项规定精神及其实施细则。

乙方应严格遵守国家、省、市关于科技专项经费使用的有关法律、法规，相关政策以及廉洁建设的各项规定，积极开展人员廉洁从业教育，防范科技项目组成员在科研活动中出现“法律、行政法规、部门规章或规范性文件规定的其他相关违规行为”。

第十四条 科研诚信和科技伦理要求

乙方应建立健全促进科研诚信和科技伦理的规章制度，落实以下职责：

1. 建立健全本单位学术论文发表诚信承诺制度、科研过程可追溯制度、科研成果检查和报告制度等成果管理制度。对本项目形成的科研成果的署名、研究数据真实性、实验可重复性等进行诚信审核和学术把关。防范科技项目组成员在项目申报、研发过程中出现提供虚假信息或材料，抄袭、剽窃他人科研成果，捏造、变造或篡改科研数据等违反科研诚信和科技伦理要求的情形。

2. 加强对科技项目参加人员的科研诚信和科技伦理教育，督促科技项目组成员恪守科学道德准则，遵守科研活动规范。对在科研诚信和科技伦理方面存在问题情节较严重的，应及时调整出项目团队并及时以书面形式报告甲方；

3. 加强对项目合作单位的科研诚信管理，正确履行管理、指导、监督职责，全面落实科研诚信和科技伦理要求；

4. 乙方或项目合作单位及其相关人员被记入科研严重失信行为数据库或相关社会领域信用“黑名单”，乙方应及时以书面形式报告甲方；

5. 乙方应严格遵守国家、省市规定的其他科研诚信管理和科技伦理相关法律、法规和政策。

6. 其他：在项目实施过程中，对乙方或项目合作单位及其相关人员有严重违背科研诚信和科技伦理要求的行为，甲方和相关部门可依照相关法律、法规规定对乙方采取责令改正、终止或撤销项目并追回财政性资金、记入科研诚信严重失信行为数据库等处理处罚措施。

第十五条 争议解决

因本合同书所产生的争议，各方应友好协商解决；协商不成的，各方同意由本合同签订地人民法院管辖。

第十六条 书面通知与送达

甲方在本合同履行过程中向乙方或丙方发出或者提供的所有书面通知、文件、文书、资料等，均以本合同所列明的乙方或丙方地址送达。乙方或丙方如果迁址，应当书面通知甲方；未履行书面通知义务的，甲方按原地址邮寄相关材料即视为已履行送达义务。

第十七条 鼓励开展科普工作

鼓励项目承担单位和人员结合科研任务对适合进行科学普及的项目内容加强科普工作。

本合同一式四份，各份具有同等效力。甲方和丙方各存一份，乙方存二份。本合同签订各方均负有相应的法律责任，不受机构、人事变动而影响。

说明：本《合同书》中，凡是三方约定无需填写的条款，在该条款的空白处划（\）。

附件：项目承担单位（乙方）及项目负责人承诺书

承诺书

本单位/本人作为广州市科技计划项目承担单位/项目负责人，将严格遵守广州市科技计划管理相关规定，严格履行自身责任，加强对项目组人员及合作单位的管理，在此郑重承诺：

（一）确保与本项目有关的全部材料真实、合法、有效，未侵犯其他方知识产权等权利；

（二）严格遵守《广州市科技计划项目管理办法》《广州市级财政科研项目资金绩效提升和管理监督办法》《广州市科技创新发展专项资金管理办法》《广州市科技计划项目全过程管理简政放权改革工作方案》等相关规定，实施项目和经费管理；

（三）严格遵守国家、省、市关于科研诚信和科技伦理的有关法律、法规，相关政策以及各项规定，加强项目实施过程中的科研诚信及科技伦理管理，恪守科研道德准则。

（四）_____

如有违反，本单位/本人愿意接受相关部门做出的各项处理决定，包括但不限于终止项目，停拨或核减经费，追回项目经费，取消一定期限广州市科技计划项目申报资格，记入科研诚信严重失信行为数据库，将不良行为向社会公开以及主要责任人接受相应党纪政纪处理等。

项目承担单位签章：

日期：2022.6.7

项目负责人签章：

日期：

闫希亮
2022.6.7

合同书各方签章

广州市科学技术局（甲方）：广州市科学技术局

项目经办人：李磊

联系电话：020-83124052

责任处室负责人：莫雪华



项目承担单位（乙方）：广州大学

二级部门：广州大学大湾区环境研究院（含珠江三角洲水质安全与保护协同创新中心）

项目负责人：闫希亮

项目经费汇入账号

帐户名：广州大学

帐号：3602114819100000192

开户银行：工行广州大学城中环支行

财务负责人：徐凌军

财务负责人联系电话：020-9366182



组织单位（丙方）：广州大学

项目经办人：刘亚兰



项目编号：GDNKY-2024ZQQZ-K04

猪禽种业全国重点实验室开放课题合同书

项目名称：QSAR 辅助抗氧化脲酶抑制饲料添加剂设计及其在猪禽养殖中的应用

承担单位（盖章）：华南农业大学



项目负责人：闫希亮

项目执行期：2025 年 1 月 至 2025 年 12 月

项目联系人：闫希亮 联系电话：13884988366

广东省农业科学院

二〇二四年制

填 表 说 明

- 一、合同书各项内容要求实事求是，仔细查看备注说明，逐项认真填写。表达要明确、严谨，字迹要清楚易辨，外来语要同时用原文和中文表达。
- 二、合同书各项内容应与申请书保持一致，不得随意更改。
- 三、开放课题费用采用报销制，经费不直接拨付到承担单位，由合作单位参与人员协助经费报销等流程，但所有经费均由课题负责人支配，院内合作单位不得开支。
- 四、合同书为 A4 纸本。各栏空格如不够时，可自行加页，纸张大小一致。
- 五、请根据邮件通知填写封面上的“项目编号”，请认真检查需要签字盖章处，签章页盖章须为所在单位公章或合同章，学院盖章无效。

一、基本信息

(一) 项目负责人信息表

| | | | | | | |
|-------|-------|--------------------|------|----------------------------|------|------------|
| 项目负责人 | 姓名 | 闫希亮 | 性别 | 男 | 出生日期 | 1991.03.03 |
| | 职务/职称 | 副教授/专业副主任 | 所在单位 | 华南农业大学动物科学学院 | | |
| | 身份证号 | 370181199103036515 | | | | |
| | 移动电话 | 13884988366 | 电子邮箱 | yanxiliang1991@scau.edu.cn | | |

(二) 项目组成员 (须含有至少一名广东省农业科学院人员)

| 序号 | 姓名 | 出生年月 | 职称 | 学历 | 任务分工 | 所在单位 | 签名 |
|----|-----|---------|-----|----|-----------|-----------------|-----|
| 1 | 闫希亮 | 1991.03 | 副教授 | 博士 | 项目设计与结果分析 | 华南农业大学动物科学学院 | 闫希亮 |
| 2 | 李娟 | 1984.04 | 研究员 | 博士 | 项目设计 | 广东省农业科学院动物卫生研究所 | 李娟 |
| 3 | 罗晶晶 | 1996.06 | 无 | 硕士 | 体外发酵试验 | 华南农业大学动物科学学院 | 罗晶晶 |
| 4 | 陈静怡 | 1999.09 | 无 | 本科 | 呼吸代谢箱试验 | 华南农业大学动物科学学院 | 陈静怡 |
| | | | | | | | |
| | | | | | | | |
| | | | | | | | |

二、主要研究任务、合作内容

1. 研究目标

本项目的目标是开发一种高效、经济的绿色饲料添加剂，并通过在养殖中的应用，显著提高养殖效率、降低臭气等环境污染。具体目标包括：

(1) 饲料添加剂开发目标：获得 DPPH 自由基清除能力和脲酶活性抑制能力强的饲料添加剂产品，为绿色健康应用奠定基础。

(2) 环境效益目标：通过开发的绿色饲料添加剂产品，从源头降低臭气排放量。实现养殖源头氨气减排 50% 以上，为实现绿色养殖提供新技术手段。

(3) 效益目标：提高动物免疫能力，提升养殖健康度和生产效率。

2. 研究内容

(1) DPPH 自由基清除能力和脲酶活性抑制能力数据集构建。高质量的 DPPH 自由基清除能力和脲酶活性抑制分子数据，是接下来构建可靠人工智能预测模型的前提条件。因此，本项目将先从 ChEMBL、Binding 数据库和文献中获取具有脲酶活性 IC_{50} 值和具有 DPPH 自由基 IC_{50} 值的分子数据集，通过异常值处理、缺失值填充、数据标准化等操作，最终得到高质量分子数据集。

(2) 多任务机器学习预测模型开发。开发可同时准确预测 DPPH 自由基清除能力和脲酶活性抑制能力的机器学习模型是本项目的关键目标。因此，在该部分研究中，我们将根据研究内容一所得的 DPPH 自由基清除能力数据集和脲酶活性抑制数据集，运用 QSAR 技术，深入解析化合物结构组成与其 DPPH 自由基清除能力和脲酶抑制能力之间的定量关系。首先计算能够表征分子的二维描述符，并将其作为预测模型的输入；然后将 DPPH 自由基清除能力和脲酶活性抑制能力作为预测模型的输出，并将全部数据集分为训练集和测试集，进一步通过模型调参，最终精准解析输入和输出之间的定量关系。

(3) 新型天然化合物筛选和试验验证。为得到效果更好的新型天然化合物，在该部分研究中，我们将通过虚拟筛选和实验验证检验我们所构建预测模型的实用性。具体来讲，通过 ZINC20 数据库构建虚拟天然产物数据集，并将其应用于 DPPH 自由基的 IC_{50} 预测模型和脲酶活性的 IC_{50} 预测模型进行预测，筛选出 2-5 种天然产物。进一步，通过体外、体内试验验证，验证其 DPPH 自由基清除能力和脲酶活性抑制能力。

3. 合作内容

(1) 实验室访问 1~2 次，科研交流 1~2 次，学术报告 1~2 次，联合培养硕士生 1 人，合作发表论文 1~2 篇。

三、将提供的成果和合作形式

| 成果形式 | 发明专利 | | 论文 | | 技术突破与应用 | | |
|------|------|----|-----|------|---------|-----|-----|
| 成果类型 | 国内 | 国外 | SCI | 中文核心 | 新技术 | 新材料 | 新品种 |
| 成果数量 | 0 | 0 | 2 | 0 | 1 | 0 | 0 |

其他成果形式及数量:

无

| 合作方式 | 学术交流 | | 研究生联合培养 | | 联合攻关或产出成果 | | |
|------|------|----|---------|----|-----------|------|------|
| 合作类型 | 报告 | 交流 | 硕士 | 博士 | 共享资源 | 承担项目 | 产出成果 |
| 合作数量 | 1 | 1 | 1 | 0 | 0 | 0 | 1 |

其他合作方式及数量:

无

四、项目经费支出预算（单位：万元）

| 支出科目 | 经费 | 用途说明 |
|-----------------|------|--|
| 1. 业务费 | 9.0 | 主要用于机器学习服务器租用、饲养试验相关耗材购买、动物体微生物相关指标测序等 |
| 2. 直接人力资源成本/劳务费 | 1.0 | 项目组研究生成员劳务费发放 |
| 合计 | 10.0 | |

备注：不得以任何形式开支设备费和间接费（管理费、绩效等）

财务负责人： .

吴晗

财务专用章



五、项目承担与参与单位分工

| 项目承担/参与单位(盖章) | 工作分工 | 经费分摊(万元) |
|--|-----------|----------|
|  华南农业大学 | 机器学习和实验验证 | 10.0 |
|  广东省农业科学院 植物研究所 | 项目设计 | 0.0 |
| | | |

备注：参与单位不分摊经费

六、项目进度和阶段目标

| | |
|---------------------------|-------------|
| 项目执行期：2025.1.1-2025.12.31 | |
| 2025.1.1-2025.3.31 | 数据收集 |
| 2025.4.1-2025.6.30 | 机器学习建模和虚拟筛选 |
| 2025.7.1-2025.9.30 | 体外验证和论文撰写 |
| 2025.10.1-2025.12.31 | 体内验证和论文撰写 |

六、签约各方

| | |
|--|--|
| 管理单位（甲方）：  | (盖章) |
| 法定代表人（或法人代理）：  | (签章) |
| 联系人（经办人）： | |
| E-mail及电话： | |
| 2025 年 5 月 15 日 | |
| 承担单位（乙方）： | |
| 法定代表人（或法人代理）：  |  (盖章) |
| 项目负责人：  | (签章) |
| E-mail及电话：   | |
| 2025 年 5 月 7 日 | |
| 猪禽种业全国重点实验室（丙方）：  | (盖章) |
| 实验室主任：  | (签章) |
| 联系人（经办人）： | |
| E-mail及电话： | |
| 2025 年 5 月 15 日 | |

国家自然科学基金资助项目批准通知

(预算制项目)

贾建博 先生/女士:

根据《国家自然科学基金条例》、相关项目管理办法规定和专家评审意见,国家自然科学基金委员会(以下简称自然科学基金委)决定资助您申请的项目。项目批准号: 22276042, 项目名称: 氧化应激效应导向的微污染水健康风险评价及智能预测研究, 直接费用: 54.00万元, 项目起止年月: 2023年01月至 2026年12月, 有关项目的评审意见及修改意见附后。

请您尽快登录科学基金网络信息系统(<https://isisn.nsf.gov.cn>), **认真阅读《国家自然科学基金资助项目计划书填报说明》并按要求填写《国家自然科学基金资助项目计划书》(以下简称计划书)**。对于有修改意见的项目,请您按修改意见及时调整计划书相关内容;如您对修改意见有异议,须在电子版计划书报送截止日期前向相关科学处提出。

请您将电子版计划书通过科学基金网络信息系统(<https://isisn.nsf.gov.cn>)提交,由依托单位审核后提交至自然科学基金委。自然科学基金委审核未通过者,将退回的电子版计划书修改后再行提交;审核通过者,打印纸质版计划书(一式两份,双面打印)并在项目负责人承诺栏签字,由依托单位科研、财务管理等部门审核、签章并在承诺栏加盖依托单位公章,且将申请书纸质签字盖章页订在其中一份计划书之后,一并报送至自然科学基金委项目材料接收工作组。纸质版计划书应当保证与审核通过的电子版计划书内容一致。**自然科学基金委将对申请书纸质签字盖章页进行审核,对存在问题的,允许依托单位进行一次修改或补齐。**

向自然科学基金委提交电子版计划书、报送纸质版计划书并补交申请书纸质签字盖章页截止时间节点如下:

1. **2022年10月8日16点:** 提交电子版计划书的截止时间;
2. **2022年10月14日16点:** 提交修改后电子版计划书的截止时间;
3. **2022年10月19日:** 报送纸质版计划书(一式两份,其中一份包含申请书纸质签字盖章页)的截止时间。
4. **2022年10月28日:** 报送修改后的申请书纸质签字盖章页的截止时间。

请按照以上规定及时提交电子版计划书，并报送纸质版计划书和申请书纸质签字盖章页，逾期不报计划书或申请书纸质签字盖章页且未说明理由的，视为自动放弃接受资助；未按要求修改或逾期提交申请书纸质签字盖章页者，将视情况给予暂缓拨付经费等处理。

附件：项目评审意见及修改意见表

国家自然科学基金委员会
2022年9月7日

附件：项目评审意见及修改意见表

| | | | | | |
|---|----------------------------|-------|---------------------|-------|-------|
| 项目批准号 | 22276042 | 项目负责人 | 贾建博 | 申请代码1 | B0607 |
| 项目名称 | 氧化应激效应导向的微污染水健康风险评价及智能预测研究 | | | | |
| 资助类别 | 面上项目 | 亚类说明 | | | |
| 附注说明 | | | | | |
| 依托单位 | 广州大学 | | | | |
| 直接费用 | 54.00 万元 | 起止年月 | 2023年01月 至 2026年12月 | | |
| <p>通讯评审意见：</p> <p><1>具体评价意见：</p> <p>一、该申请项目是否面向国家需求并试图解决技术瓶颈背后的基础问题？请结合应用需求详细阐述判断理由。</p> <p>饮用水水源微污染问题严重。项目面向“饮用水安全”的国家需求，针对现有的生物效应测试检测灵敏度不足，且微污染水健康效应分析和预测无法避开关键致毒组分鉴定过程的瓶颈，拟效应导向关键毒性组分，解决饮用水及其水源微污染物潜在健康风险的科学问题。</p> <p>二、请评述申请项目所提出的科学问题与预期成果的科学价值。</p> <p>本项目1) 开发基于细胞氧化应激的微污染饮用水水源水健康风险评价体系，实现无浓缩饮用水水源水诱导细胞效应的快速高通量检测，并解析水样致毒的关键分子机制；2) 完成饮用水源水中污染物全组分的质谱分析，结合水样生物效应大数据，建立不少于200个包括不同水源地水样的化学组分和细胞效应的数据库；3) 建立深度神经网络模型，实现基于微污染物图谱数据的饮用水源水健康风险的智能预测，并鉴别关键致毒污染物图谱特征。本项目的实施将为微污染饮用水水源水健康风险的科学评估与智能预测提供理论基础和方法学支撑。</p> <p>三、请评述申请人的研究基础及研究方案的创新性和可行性。</p> <p>申请人长期从事典型环境污染物体内、体外生物效应及致毒机制的研究，在易感群体动物、细胞建模、痕量污染物生物效应评价及分子机制等方面有很好的工作积累。项目构思新颖，技术路线合理，研究方案具体可行。</p> <p>四、其他建议</p> <p><2>具体评价意见：</p> <p>一、该申请项目是否面向国家需求并试图解决技术瓶颈背后的基础问题？请结合应用需求详细阐述判断理由。</p> <p>饮用水安全关乎国计民生，是我国实现生态文明建设与可持续发展宏伟目标的中大战略需求。现行基于化学指标和生物效应指标的水质评价方法存在局限性。该项目拟研究氧化应激效应导向的微污染水健康风险评价及智能预测研究。项目的完成将为微污染饮用水水源健康风险的科学评估与智能预测提供理论基础和方法支撑。</p> <p>二、请评述申请项目所提出的科学问题与预期成果的科学价值。</p> <p>微污染饮用水及其水源安全性系统评估在环境健康领域一直受到广泛关注。现有的生物效应测试检测灵敏度不足，微污染水健康效应分析和预测无法避开关键致毒组分鉴定过程，该项依据这一研究瓶颈开展研究。项目的实施有望为微污染饮用水水源健康风险的评估和智能预测提供理论基础。</p> <p>三、请评述申请人的研究基础及研究方案的创新性和可行性。</p> <p>申请人长期从事典型环境污染物体内外生物效应和致毒机制研究，在相关领域已有较好的研究基础。研究方案整体较合理，具有一定的可行性。</p> <p>四、其他建议</p> | | | | | |

<3>具体评价意见:

一、该申请项目是否面向国家需求并试图解决技术瓶颈背后的基础问题? 请结合应用需求详细阐述判断理由。

该项目申请书针对我国饮用水安全开展相关研究。饮用水安全的好坏与我国人民的身体健康乃至生命安全密切相关。但是, 目前有关饮用水中为污染物的评价方法仍然存在很多局限性。针对该局限性, 申请人拟开展方法学的创新, 为微污染饮用水水源健康风险的科学评估与智能预测提供基础和方法支撑, 从而保障我国饮用水的安全, 符合需求牵引、突破瓶颈的科学属性。

二、请评述申请项目所提出的科学问题与预期成果的科学价值。

该项目申请书主要针对两个科学问题开展研究, 第一, 是现有的生物监测方法灵敏度不高, 不能直接反映微污染物的健康风险水平; 第二。目前的微污染物风险分析, 无法避开致毒组分的鉴定。因此申请人。提出了新的微污染物评价方法, 通过建立高灵敏的生物监测方法, 系统评价饮用水及水源中为污染物潜在的健康风险, 发展高效微污染物风险评估策略方法, 对于指导微污染饮用水水源地的污染控制将具有重要意义。

三、请评述申请人的研究基础及研究方案的创新性和可行性。

申请人长期从事典型环境污染物体内和体外生物效应及机制的研究, 在EST等期刊上发表多篇学术论文。在具体的研究工作基础上, 比如申请人建立了脂肪变肝细胞模型, 开展了实际环境水样的化学分析, 基于电子电离质谱的有机化合物毒性预测, 建立了污染物致毒的分子机制研究平台。以上表明, 申请人在相关研究上具有较强的工作积累。申请书研究方案详实、思路清晰、具有较强的可行性。

四、其他建议

建议优先资助

修改意见:

化学科学部

2022年9月7日



| | |
|--------|--------------------|
| 项目批准号 | 22276042 |
| 申请代码 | B0607 |
| 归口管理部门 | |
| 依托单位代码 | 51000608A0268-0524 |



22276042 1002708

国家自然科学基金 资助项目计划书 (预算制项目)

资助类别: 面上项目

亚类说明:

附注说明:

项目名称: 氧化应激效应导向的微污染水健康风险评价及智能预测研究

直接费用: 54万元

执行年限: 2023.01-2026.12

负责人: 贾建博

通讯地址: 广州市大学城外环西路230号

邮政编码: 510006

电话: 17722856901

电子邮件: jb_jia@gzhu.edu.cn

依托单位: 广州大学

联系人: 王玉林

电话: 020-39366266

填表日期: 2022年09月15日

国家自然科学基金委员会制

Version: 1.002.708



国家自然科学基金资助项目计划书填报说明 （预算制项目）

- 一、项目负责人收到《国家自然科学基金资助项目批准通知》（以下简称《批准通知》）后，请认真阅读本填报说明，参照国家自然科学基金相关项目管理办​​法和新修订的《国家自然科学基金资助项目资金管理办法》（以下简称《资金管理办法》，请查阅国家自然科学基金委员会官方网站首页“政策法规”栏目），按《批准通知》的要求认真填写和提交《国家自然科学基金资助项目计划书》（以下简称《计划书》）。
- 二、填写《计划书》时要科学严谨、实事求是、表述清晰、准确。《计划书》经国家自然科学基金委员会相关项目管理部门审核批准后，将作为项目研究计划执行、检查和验收的依据。
- 三、《计划书》各部分填写要求如下：
 - （一）简表：由系统自动生成。
 - （二）摘要及关键词：各类获资助项目都应当填写中、英文摘要及关键词。
 - （三）项目组主要成员：计划书中列出姓名的项目组主要成员由系统自动生成，与申请书原成员保持一致，不可随意调整。如果《批准通知》所附“项目评审意见及修改意见表”中“修改意见”栏目有调整项目组成员相关要求的，待项目开始执行后，按照项目成员变更程序另行办理。
 - （四）资金预算表：根据批准的项目资助额度，按规定调整项目预算，并按照《国家自然科学基金项目计划书预算表编制说明》填报资金预算表和预算说明书。
 - （五）正文：
 1. 面上项目、地区科学基金项目：如果《批准通知》所附“项目评审意见及修改意见表”中“修改意见”栏目没有修改要求的，只需选择“研究内容和研究目标按照申请书执行”即可；如果《批准通知》中上述栏目明确要求调整研究期限或研究内容等的，须选择“根据研究方案修改意见更改”并填报相关修改内容。
 2. 重点项目、重点国际（地区）合作研究项目、重大项目、国家重大科研仪器研制项目、原创探索计划项目：须选择“根据研究方案修改意见更改”，根据《批准通知》的要求填写研究（研制）内容，不得自行降低、更改研究目标（或仪器研制的技术性能与主要技术指标、验收技术指标等）或缩减研究（研制）内容。此外，还要突出以下几点：
 - （1）研究的难点和在实施过程中可能遇到的问题（或仪器研制风险），拟采用的研究（研制）方案和技术路线；
 - （2）项目主要参与者分工，合作研究单位（如有）之间的关系与分工，重大项目还需说明课题之间的关联；
 - （3）详细的年度研究（研制）计划。
 3. 创新研究群体项目：须选择“根据研究方案修改意见更改”，按下列提纲撰写：
 - （1）研究方向；



- (2) 结合国内外研究现状，说明研究工作的学术思想和科学意义（限两个页面）；
 - (3) 研究内容、研究方案及预期目标（限两个页面）；
 - (4) 年度研究计划；
 - (5) 研究队伍的组成情况。
4. 基础科学中心项目：须选择“根据研究方案修改意见更改”，根据《批准通知》的要求和现场考察专家组的意见和建议，进一步完善并细化研究计划，按下列提纲撰写：
 - (1) 五年拟开展的研究工作（包括主要研究方向、关键科学问题与研究内容）；
 - (2) 研究方案（包括骨干成员之间的分工及合作方式、学科交叉融合研究计划等）；
 - (3) 年度研究计划；
 - (4) 五年预期目标和可能取得的重大突破等；
 - (5) 研究队伍的组成情况。
5. 对于其他类型项目，参照面上项目的方式进行选择和填写。

简表

| | | | | | | | | |
|---------|-----------|----------------------------|-----|---|------|-------------------------|-----|---------------|
| 项目负责人信息 | 姓 名 | 贾建博 | 性 别 | 男 | 出生年月 | 1989年02月 | 民 族 | 汉族 |
| | 学 位 | 博士 | | | 职 称 | 讲师 | | |
| | 是否在站博士后 | 否 | | | 电子邮件 | jb_jia@gzhu.edu.cn | | |
| | 电 话 | 17722856901 | | | 个人网页 | | | |
| | 工 作 单 位 | 广州大学 | | | | | | |
| | 所 在 院 系 所 | 大湾区环境研究院 | | | | | | |
| 依托单位信息 | 名 称 | 广州大学 | | | | | 代 码 | 51000608A0268 |
| | 联 系 人 | 王玉林 | | | 电子邮件 | wyl@gzhu.edu.cn | | |
| | 电 话 | 020-39366266 | | | 网站地址 | http://kyc.gzhu.edu.cn/ | | |
| 合作单位信息 | 单 位 名 称 | | | | | | | |
| | | | | | | | | |
| 项目基本信息 | 项 目 名 称 | 氧化应激效应导向的微污染水健康风险评价及智能预测研究 | | | | | | |
| | 资 助 类 别 | 面上项目 | | | | 亚 类 说 明 | | |
| | 附 注 说 明 | | | | | | | |
| | 申 请 代 码 | B0607:环境毒理与健康 | | | | | | |
| | 基 地 类 别 | | | | | | | |
| | 执 行 年 限 | 2023.01-2026.12 | | | | | | |
| | 直 接 费 用 | 54万元 | | | | | | |

项目摘要

中文摘要:

饮用水水源微污染问题严重。效应导向分析方法在微污染水样健康风险评价中表现出优异的应用潜力。然而，现有的生物效应测试检测灵敏度不足，且微污染水健康效应分析和预测无法避开关键致毒组分鉴定过程。为突破以上研究瓶颈，本项目将重点完成：1) 开发基于细胞氧化应激的微污染饮用水水源健康风险评价体系，实现无浓缩饮用水源水诱导细胞效应的快速高通量检测，并解析水样致毒的关键分子机制；2) 完成饮用水源水中污染物全组分的质谱分析，结合水样生物效应大数据，建立不少于200个包括不同水源地水样的化学组分和细胞效应的数据库；3) 建立深度神经网络模型，实现基于微污染物谱图数据的饮用水源水健康风险的智能预测，并鉴别关键致毒污染物图谱特征。本项目的实施将为微污染饮用水水源健康风险的科学评估与智能预测提供理论基础和方法学支撑。

Abstract:

Nowadays, micro-contamination is a common problem affecting the quality of water from drinking water sources. Effect-directed analytical methods have shown excellent potential for their application in health risk assessment of micro-contaminated water samples. However, the lack of sensitivity of the existing biological effect testing methods and the inability of the health effect analysis and prediction of environmental micro-polluted water bodies to bypass the identification process of key toxicogenic components are the main research bottlenecks in the current effect-directed systematic assessment and monitoring of the safety of micro-polluted water sources. In order to break through the above research bottlenecks, this proposed project will focus on 1) developing high-sensitive in vitro cellular models to achieve rapid detection of drinking water-induced cellular effects using unconcentrated real micro-polluted water samples, and analyzing the key molecular mechanisms of toxicity of water samples in key areas, 2) completing full-range mass spectrometry analysis of pollutants in water from micro-polluted drinking water sources, and utilizing data on the biological effects of water samples to establish a database of no less than 200 chemical components and cellular effect data including water samples from different water sources, and 3) establish a deep neural network model to achieve intelligent prediction of health risks of drinking water sources based on micro-pollutant profile data and identify key toxicogenic pollutant profile features. The successful implementation of this project will provide the theoretical basis and methodological support for scientific assessment and intelligent prediction of health risks of micro-polluted drinking water sources.

关键词(用分号分开): 效应导向分析; 微污染水; 健康风险; 氧化应激; 智能预测

Keywords(用分号分开): Effect-directed analysis; Micro-polluted water; Health risk; Oxidative stress; Intelligent prediction

项目组主要成员

| 编号 | 姓名 | 出生年月 | 性别 | 职称 | 学位 | 单位名称 | 电话 | 证件号码 | 项目分工 | 每年工作时间 (月) |
|-----|-----|---------|----|----|----|------|--------------|--------------------|-----------|---------------|
| 1 | 贾建博 | 1989.02 | 男 | 讲师 | 博士 | 广州大学 | 17722856901 | 372923198902181118 | 项目负责人 | 10 |
| 2 | 闫希亮 | 1991.03 | 男 | 讲师 | 博士 | 广州大学 | 13884988366 | 370181199103036515 | 深度学习建模与预测 | 6 |
| 3 | 刘国红 | 1989.06 | 女 | 无 | 博士 | 广州大学 | 15053173015 | 371402198906066125 | 细胞效应及机制 | 6 |
| 4 | 孙海南 | 1986.08 | 男 | 无 | 博士 | 广州大学 | 15615536923 | 370304198608304216 | 化学分析与细胞效应 | 6 |
| 5 | 刘荣涛 | 1990.08 | 男 | 无 | 博士 | 广州大学 | 020-39386084 | 610425199008114111 | 细胞效应及机制 | 6 |
| 总人数 | | | | 高级 | 中级 | 初级 | | 博士后 | 博士生 | 硕士生 |
| 8 | | | | | 2 | | | 3 | | 3 |



国家自然科学基金预算制项目预算表

项目批准号：22276042

项目负责人：贾建博

金额单位：万元

| 序号 | 科目名称 | 金额 |
|----|----------------|---------|
| 1 | 一、基金资助项目直接费用合计 | 54.0000 |
| 2 | 1、设备费 | 4.5000 |
| 3 | 其中：设备购置费 | 4.5000 |
| 4 | 2、业务费 | 37.5000 |
| 5 | 3、劳务费 | 12.0000 |
| 6 | 二、其他来源资金 | 0.0000 |
| 7 | 三、合计 | 54.0000 |

注：请按照项目研究实际需要合理填写各科目预算金额。

预算说明书

（请按照《国家自然科学基金项目申请书预算表编制说明》等的有关要求，按照政策相符性、目标相关性和经济合理性原则，实事求是编制项目预算。填报时，直接费用应按设备费、业务费、劳务费三个类别填报，每个类别结合科研任务按支出用途进行说明。对单价≥50万元的设备详细说明，对单价<50万元的设备费用分类说明，**对合作研究单位资质及资金外拨情况、自筹资金进行必要说明。**）

项目直接经费预算为**54.00万元**，具体预算如下：

（一）设备费4.50万元

购置用于计算模拟和数据存储的服务器1台，单价4.50万元/台。设备费合计4.50万元。

（二）业务费37.50万元

1. 材料费16.70万元

根据实验方案，主要消耗性指出包括各种细胞和分子生物学试剂耗材和化学分析用试剂耗材等，具体预算如下：

细胞培养用试剂耗材（共计4.72万元）：细胞培养基+PBS：60元/瓶×30瓶/年×4年，计0.72万元；血清5000元/瓶×2瓶/年×4年，计4.00万元。

分子生物学试剂耗材（共计9.78万元）：荧光素酶反应底物5000元/套×6套，计3.00万元；各类细胞效应抑制剂3000元/支×3支，计0.90万元；细胞存活率检测试剂盒4000元/套×3套，计1.20万元；RNA提取、反转录及PCR相关试剂1.00万元；抗体及western blot相关试剂1.10万元；细胞因子检测试剂盒5000元/套×3套，计1.50万元；分子生物学实验过程中用到的离心管、移液管、酶联板等，计1.08万元。

化学分析用试剂耗材（共计2.20万元）：拟购买甲醇、DMSO等化学试剂用于水样有机污染物分析的前处理及分析测试，约300元/箱×20箱，计0.60万元；ICP-MS用液氩2500元/罐×2罐，计0.50万元；用于LC-MS分析的液质小瓶500元/盒×10盒，计0.50万元；0.22 μm孔径各种需求的滤膜0.30万元；封口膜、自封袋、锡箔纸等其他相关耗材0.30万元。

2. 测试化验加工费10.80万元

本项目需用UPLC-MS/MS对有机污染物全组分进行非靶标分析，400元/小时，120小时，计4.80万元；细胞全转录组分析分析5000元/样，6样，计3.00万元；代谢组分析1.00万元/样，3样，计3.00万元。测试化验加工费合计10.80万元。

3. 出版/文献/信息传播/知识产权事务费5.00万元

预计文章出版版面费2.00万元，文献检索费用2.00万元，图书计文献购买计1.00万元。合计5.00万元。

4. 会议/差旅/国际合作交流费5.00万元

为更好的把握课题前沿进展和研究动态，需参加国内外相关学术会议。项目执行期间，拟参加国内高水平学术会议4人次，每次按3天计，会议注册费2000元/人/次，交通费2000元/人/次，住宿费320元/人/天，伙食补贴100元/人/天，交通补贴80元/人/天，每人次参会费用合计5500元，共计2.20万元。拟参加国际学术会议1人次，按5天计，国际往返交通费12000元/人/次，会议注册费4000元/人/次，住宿费600元/人/天，杂费200元/人/天，合计2.00万元。此外，需赴各水源地采集水样，预计由此产生的各类交通费用约0.80万元。差旅费合计5.00万元。

（三）劳务费12.00万元**1. 研究生劳务费9.60万元**

用于项目中成员中没有工资性收入的研究生。硕士研究生按800元/人/月计，项目执行期间共120人月，合计9.60万元。

2. 专家咨询费2.40万元

用于项目执行过程中邀请本领域专家对项目进行指导。咨询费标准平均800元/人/次，计30人次，合计2.40万元。



报告正文

研究内容和研究目标按照申请书执行。



国家自然科学基金项目负责人、依托单位承诺书

国家自然科学基金项目负责人承诺书

本人郑重承诺：我接受国家自然科学基金的资助，严格遵守中共中央办公厅、国务院办公厅《关于进一步加强科研诚信建设的若干意见》《关于进一步弘扬科学家精神加强作风和学风建设的意见》《关于加强科技伦理治理的意见》等规定，及国家自然科学基金委员会关于资助项目管理、项目资金管理等各项规章，在《计划书》填写及项目执行过程中：

（一）按照《批准通知》《国家自然科学基金资助项目计划书填报说明》的要求填写《计划书》，未自行降低、更改目标任务或约定要求，或缩减研究（研制）内容；

（二）树立“红线”意识，严格履行科研合同义务，按照《计划书》负责实施本项目（批准号：22276042），切实保证研究工作时间，按时报送有关材料，及时报告重大情况变动，不违规将科研任务转包、分包他人，不以项目实施周期外或不相关成果充抵交差；

（三）遵守科研诚信、科技伦理规范和学术道德，认真开展研究工作，对资助项目发表的论著和取得的研究成果按规定进行标注，不在非本项目资助的成果或其他无关成果上标注本项目批准号，反对无实质学术贡献者“挂名”，不在成果署名、知识产权归属等方面侵占他人合法权益，并如实报告本人及项目组成员发生的违背科研诚信要求的任何行为；

（四）尊重科研规律，弘扬科学家精神，严谨求实，追求卓越，反对浮夸浮躁、投机取巧，不人为夸大学术或技术价值，不传播未经科学验证的现象和观点；

（五）将项目资金全部用于与本项目研究工作相关的支出，并结合科研活动需要，科学合理安排项目资金支出进度；

（六）做好项目组成员的教育和管理，确保遵守以上相关要求。

如违背上述承诺，本人愿接受国家自然科学基金委员会和相关部门做出的各项处理决定。

项目负责人（签字）

顾建峰

2022年9月22日

依托单位科研管理部门：

依托单位财务管理部门：

负责人（签章）

2022年10月14日

负责人（签章）：

2022年10月12日

国家自然科学基金项目依托单位承诺书

我单位同意承担上述国家自然科学基金项目，将保证项目负责人及其研究队伍的稳定和研究项目实施所需的条件，严格遵守国家自然科学基金委员会有关资助项目管理、项目资金管理、科研诚信管理和科技伦理管理等各项规定，并督促实施。

依托单位（公章）

2022年10月12日



国家自然科学基金资助项目签批审核表

本栏自由自然科学基金委填写

科学处审查意见：

同意按计划执行

庄乾坤

负责人（签章）：

年 月 日

2023-01-05

科学部审查意见：

同意科学处意见

杨俊林

负责人（签章）：

年 月 日

2023-01-05



项目名称：氧化应激效应导向的微污染水健康风险评价及智能预测研究

资助类型：面上项目

申请代码：B0607. 环境毒理与健康

国家自然科学基金项目申请人和参与者承诺书

为了维护国家自然科学基金项目评审公平、公正，共同营造风清气正的科研生态，本人**在此郑重承诺**：严格遵守《中华人民共和国科学技术进步法》《国家自然科学基金条例》《关于进一步加强科研诚信建设的若干意见》《关于进一步弘扬科学家精神加强作风和学风建设的意见》以及科技部、自然科学基金委关于科研诚信建设有关规定和要求；申请材料信息真实准确，不含任何涉密信息或敏感信息，不含任何违反法律法规或违反科研伦理规范的内容；在国家自然科学基金项目申请、评审和执行全过程中，恪守职业规范和科学道德，遵守评审规则和工作纪律，杜绝以下行为：

- (一) 抄袭、剽窃他人申请书、论文等科研成果或者伪造、篡改研究数据、研究结论；
- (二) 购买、代写申请书；购买、代写、代投论文，虚构同行评议专家及评议意见；购买实验数据；
- (三) 违反成果发表规范、署名规范、引用规范，擅自标注或虚假标注获得科技计划等资助；
- (四) 在项目申请书中以高指标通过评审，在项目计划书中故意篡改降低相应指标；
- (五) 以任何形式打听或散布尚未公布的评审专家名单及其他评审过程中的保密信息；
- (六) 本人或委托他人通过各种方式和途径联系有关专家进行请托、游说，违规到评审会议驻地窥探、游说、询问等干扰评审或可能影响评审公正性的行为；
- (七) 向工作人员、评审专家等提供任何形式的礼品、礼金、有价证券、支付凭证、商业预付卡、电子红包，或提供宴请、旅游、娱乐健身等任何可能影响评审公正性的活动；
- (八) 违反财经纪律和相关管理规定的行为；
- (九) 其他弄虚作假行为。

如违背上述承诺，本人愿接受国家自然科学基金委员会和相关部门做出的各项处理决定，包括但不限于撤销科学基金资助项目，追回项目资助经费，向社会通报违规情况，取消一定期限国家自然科学基金项目申请资格，记入科研诚信严重失信行为数据库以及接受相应的党纪政务处分等。

申请人签字：贺建坤

| 编号 | 参与者姓名 / 工作单位名称（应与加盖公章一致）/ 证件号码 | 签字 |
|----|--------------------------------|-----|
| 1 | 闫希亮 / 广州大学 / 3*****5 | 闫希亮 |
| 2 | 刘国红 / 广州大学 / 3*****5 | 刘国红 |
| 3 | 孙海南 / 广州大学 / 3*****6 | 孙海南 |
| 4 | 刘荣涛 / 广州大学 / 6*****1 | 刘荣涛 |
| 5 | | |
| 6 | | |
| 7 | | |
| 8 | | |
| 9 | | |



项目名称： 氧化应激效应导向的微污染水健康风险评价及智能预测研究
资助类型： 面上项目
申请代码： B0607. 环境毒理与健康

国家自然科学基金项目申请单位承诺书

为了维护国家自然科学基金项目评审公平、公正，共同营造风清气正的科研生态，**本单位郑重承诺**：申请材料中不存在违背《中华人民共和国科学技术进步法》《国家自然科学基金条例》《关于进一步加强科研诚信建设的若干意见》《关于进一步弘扬科学家精神加强作风和学风建设的意见》以及科技部、自然科学基金委关于科研诚信建设有关规定和要求的情况；申请材料符合《中华人民共和国保守国家秘密法》和《科学技术保密规定》等有关法律法规和规章制度要求，不含任何涉密信息或敏感信息；申请材料不含任何违反法律法规或违反科研伦理规范的内容；申请人符合相应项目的申请资格；在项目申请和评审活动全过程中，遵守有关评审规则和工作纪律，杜绝以下行为：

（一）以任何形式打听或公布未公开的项目评审信息、评审专家信息及其他评审过程中的保密信息，干扰评审专家的评审工作；

（二）组织或协助申请人/参与者向工作人员、评审专家等给予任何形式的礼品、礼金、有价证券、支付凭证、商业预付卡、电子红包等；宴请工作人员、评审专家，或组织任何可能影响科学基金评审公正性的活动；

（三）支持、放任或对申请人/参与者抄袭、剽窃、重复申报、提供虚假信息（含身份和学术信息）等不当手段申报国家自然科学基金项目疏于管理；

（四）支持或协助申请人/参与者采取“打招呼”“围会”等方式影响科学基金项目评审；

（五）其他违反财经纪律和相关管理规定的行为。

如违背上述承诺，本单位愿接受自然科学基金委和相关部门做出的各项处理决定，包括但不限于停拨或核减经费、追回项目已拨经费、取消本单位一定期限国家自然科学基金项目申请资格、记入科研诚信严重失信行为数据库以及主要责任人接受相应党纪政务处分等。



依托单位公章：

日期：2022年10月12日

合作研究单位公章：

日期： 年 月 日

合作研究单位公章：

日期： 年 月 日

检索证明

根据委托人提供的论文材料，委托人华南农业大学动物科学学院 闫希亮 1 篇论文收录情况如下表。

| 序号 | 论文名称 | 发表刊物及发表的年月卷期/页码等 | 作者排名 | 论文等级 | 作者文中单位 | 收录情况 | 影响因子 | 中科院大类分区 |
|----|--|--|------|------|--------|------|--|----------------------------------|
| 1 | Construction of a web-based nanomaterial database by big data curation and modeling friendly nanostructure annotations | NATURE COMMUNICATIONS 出版年: 2020 出版日期: MAY 20 卷期: 11 1 页码: - 文献号: 2519 文献类型: Article | 第一作者 | T2 类 | 广州大学 | SCI | IF2-year=14.919 IF5-year=15.805 (2020) | 综合性期刊 1 区 Top 期刊: 是 (2020) |

说明: 论文等级和中科院大类分区按《华南农业大学学术论文评价方案(试行)》划分。

报告免责声明: 如未盖章, 报告无效



检索证明

根据委托人提供的论文材料，委托人华南农业大学动物科学学院 闫希亮 1 篇论文收录情况如下表。

| 序号 | 论文名称 | 发表刊物及发表的年月卷期/页码等 | 作者排名 | 论文等级 | 作者文中单位 | 收录情况 | 影响因子 | 中科院大类分区 |
|----|--|--|------|------|--------|------|--|-------------------------------|
| 1 | Converting Nanotoxicity Data to Information Using Artificial Intelligence and Simulation | CHEMICAL REVIEWS 出版年: 2023 出版日期: JUN 1 卷期: 123 13 页码: 8575-8637 文献类型: Review | 第一作者 | T2 类 | 广州大学 | SCI | IF2-year=51.5 IF5-year=63.6 (2023) | 化学 1 区 Top 期刊: 是 (2023) |

说明: 论文等级和中科院大类分区按《华南农业大学学术论文评价方案(试行)》划分。

报告免责声明: 如未盖章, 报告无效

检索员: 尹银怀

华南农业大学图书馆

2025-07-14

SCAULIB202519141

检索证明

根据委托人提供的论文材料，委托人华南农业大学动物科学学院 闫希亮 1 篇论文收录情况如下表。

| 序号 | 论文名称 | 发表刊物及发表的年月卷期/页码等 | 作者排名 | 论文等级 | 作者文中单位 | 收录情况 | 影响因子 | 中科院大类分区 |
|----|---|---|------|------|--------|------|--|-------------------------------|
| 1 | Prediction of Nano-Bio Interactions through Convolutional Neural Network Analysis of Nanostructure Images | ACS SUSTAINABLE CHEMISTRY & ENGINEERING 出版年: 2020 出版日期: DEC 28 卷期: 8 51 页码: 19096-19104 文献类型: Article | 第一作者 | A 类 | 广州大学 | SCI | IF2-year=8.198 IF5-year=8.471 (2020) | 化学 2 区 Top 期刊: 是 (2020) |

说明: 论文等级和中科院大类分区按《华南农业大学学术论文评价方案(试行)》划分。

报告免责声明: 如未盖章, 报告无效

检索员: 尹银怀

华南农业大学图书馆

2025-07-14

检索证明

根据委托人提供的论文材料，委托人华南农业大学动物科学学院 闫希亮 1 篇论文收录情况如下表。

| 序号 | 论文名称 | 发表刊物及发表的年月卷期/页码等 | 作者排名 | 论文等级 | 作者文中单位 | 收录情况 | 影响因子 | 中科院大类分区 |
|----|--|--|------|------|------------------|------|--|--------------------------------|
| 1 | The Glutamatergic System Regulates Feather Pecking Behaviors in Laying Hens Through the Gut-Brain Axis | ANIMALS 出版年：2025 出版日期：APR 30 卷期：15 9 页码：- 文献号：1297 文献类型：Article | 第一作者 | A 类 | 华南农业大学 动物科学学院 | SCI | IF2-year=2.7 IF5-year=3.2 (2024) | 农林科学 2 区 Top 期刊：否 (2025) |

说明：论文等级和中科院大类分区按《华南农业大学学术论文评价方案（试行）》划分。

报告免责声明：如未盖章，报告无效

检索员：尹银怀

华南农业大学图书馆

2025-07-14

检索证明

根据委托人提供的论文材料，委托人华南农业大学动物科学学院 闫希亮 1 篇论文收录情况如下表。

| 序号 | 论文名称 | 发表刊物及发表的年月卷期/页码等 | 作者排名 | 论文等级 | 作者文中单位 | 收录情况 | 影响因子 | 中科院大类分区 |
|----|--|---|------|------|--------|------|--|------------------------------------|
| 1 | Linking electron ionization mass spectra of organic chemicals to toxicity endpoints through machine learning and experimentation | JOURNAL OF HAZARDOUS MATERIALS 出版年：2022 出版日期：JUN 5 卷期：431 页码：- 文献号：128558 文献类型：Article | 通讯作者 | T2 类 | 广州大学 | SCI | IF2-year=13.6 IF5-year=12.7 (2022) | 环境科学与生态学 1 区 Top 期刊：是 (2022) |

说明：论文等级和中科院大类分区按《华南农业大学学术论文评价方案（试行）》划分。

报告免责声明：如未盖章，报告无效



检索证明

根据委托人提供的论文材料，委托人华南农业大学动物科学学院 闫希亮 1 篇论文收录情况如下表。

| 序号 | 论文名称 | 发表刊物及发表的年月卷期/页码等 | 作者排名 | 论文等级 | 作者文中单位 | 收录情况 | 影响因子 | 中科院大类分区 |
|----|---|--|--------------|------|--------|------|--|-------------------------------------|
| 1 | Implementing comprehensive machine learning models of multispecies toxicity assessment to improve regulation of organic compounds | JOURNAL OF HAZARDOUS MATERIALS 出版年: 2023 出版日期: SEP 15 卷期: 458 页码: - 文献号: 131942 文献类型: Article | 共同通讯作者(倒数第一) | T2 类 | 广州大学 | SCI | IF2-year=12.2 IF5-year=11.9 (2023) | 环境科学与生态学 1 区 Top 期刊: 是 (2023) |

说明: 论文等级和中科院大类分区按《华南农业大学学术论文评价方案(试行)》划分。

报告免责声明: 如未盖章, 报告无效



检索证明

根据委托人提供的论文材料，委托人华南农业大学动物科学学院 闫希亮 1 篇论文收录情况如下表。

| 序号 | 论文名称 | 发表刊物及发表的年月卷期/页码等 | 作者排名 | 论文等级 | 作者文中单位 | 收录情况 | 影响因子 | 中科院大类分区 |
|----|--|---|------|------|------------------|------|--|------------------------------------|
| 1 | Unraveling the ecotoxicity of micro(nano)plastics loaded with environmental pollutants using ensemble machine learning | JOURNAL OF HAZARDOUS MATERIALS 出版年：2025 出版日期：SEP 5 卷期：495 页码：- 文献号：138911 文献类型：Article | 通讯作者 | T2 类 | 华南农业大学 动物科学学院 | SCI | IF2-year=11.3 IF5-year=12.4 (2024) | 环境科学与生态学 1 区 Top 期刊：是 (2025) |

说明：论文等级和中科院大类分区按《华南农业大学学术论文评价方案（试行）》划分。

报告免责声明：如未盖章，报告无效



检索证明

根据委托人提供的论文材料，委托人华南农业大学动物科学学院 闫希亮 1 篇论文收录情况如下表。

| 序号 | 论文名称 | 发表刊物及发表的年月卷期/页码等 | 作者排名 | 论文等级 | 作者文中单位 | 收录情况 | 影响因子 | 中科院大类分区 |
|----|---|---|--------|------|--------------|------|--|------------------------------|
| 1 | Advanced Mass-Spectra-Based Machine Learning for Predicting the Toxicity of Traditional Chinese Medicines | ANALYTICAL CHEMISTRY 出版年：2024 出版日期：DEC 20 卷期：97 1 页码：783-792 文献类型：Article | 共同通讯作者 | T2 类 | 华南农业大学动物科学学院 | SCI | IF2-year=6.7 IF5-year=6.6 (2024) | 化学 1 区 Top 期刊：是 (2025) |

说明：论文等级和中科院大类分区按《华南农业大学学术论文评价方案（试行）》划分。

报告免责声明：如未盖章，报告无效



检索证明

根据委托人提供的论文材料，委托人华南农业大学动物科学学院 闫希亮 1 篇论文收录情况如下表。

| 序号 | 论文名称 | 发表刊物及发表的年月卷期/页码等 | 作者排名 | 论文等级 | 作者文中单位 | 收录情况 | 影响因子 | 中科院大类分区 |
|----|--|---|---------------|------|--------|------|--------------------------------------|----------------------------------|
| 1 | ILTox: A Curated Toxicity Database for Machine Learning and Design of Environmentally Friendly Ionic Liquids | ENVIRONMENTAL SCIENCE & TECHNOLOGY LETTERS 出版年: 2023 出版日期: MAR 21 卷期: 10 11 页码: 983-988 文献类型: Article | 共同通讯作者 (倒数第一) | A 类 | 广州大学 | SCI | IF2-year=8.9 IF5-year=10.0 (2023) | 环境科学与生态学 2 区 Top 期刊: 否 (2023) |

说明：论文等级和中科院大类分区按《华南农业大学学术论文评价方案（试行）》划分。

报告免责声明：如未盖章，报告无效



检索证明

根据委托人提供的论文材料，委托人华南农业大学动物科学学院 闫希亮 1 篇论文收录情况如下表。

| 序号 | 论文名称 | 发表刊物及发表的年月卷期/页码等 | 作者排名 | 论文等级 | 作者文中单位 | 收录情况 | 影响因子 | 中科院大类分区 |
|----|--|--|------|------|------------------|------|--|--------------------------------|
| 1 | De novo Design of Biocompatible Nanomaterials Using Quasi-SMILES and Recurrent Neural Networks | ACS APPLIED MATERIALS & INTERFACES 出版年：2024 出版日期：NOV 20 卷期：16 48 页码：66367-66376 文献类型：Article | 通讯作者 | A 类 | 华南农业大学 动物科学学院 | SCI | IF2-year=8.2 IF5-year=8.5 (2024) | 材料科学 2 区 Top 期刊：否 (2025) |

说明：论文等级和中科院大类分区按《华南农业大学学术论文评价方案（试行）》划分。

报告免责声明：如未盖章，报告无效



检索证明

根据委托人提供的论文材料，委托人华南农业大学动物科学学院 闫希亮 1 篇论文收录情况如下表。

| 序号 | 论文名称 | 发表刊物及发表的年月卷期/页码等 | 作者排名 | 论文等级 | 作者文中单位 | 收录情况 | 影响因子 | 中科院大类分区 |
|----|--|--|------|------|------------------|------|--|--------------------------------|
| 1 | Effect of different strategies for modifying graphene on the adsorption and gas sensing of trimethylamine: Insights from DFT study | INTERNATIONAL JOURNAL OF HYDROGEN ENERGY 出版年：2024 出版日期：APR 3 卷期：61 页码：1330-1339 文献类型：Article | 通讯作者 | A 类 | 华南农业大学 动物科学学院 | SCI | IF2-year=8.3 IF5-year=7.7 (2024) | 材料科学 2 区 Top 期刊：否 (2025) |

说明：论文等级和中科院大类分区按《华南农业大学学术论文评价方案（试行）》划分。

报告免责声明：如未盖章，报告无效

检索员：尹银怀
华南农业大学图书馆

2025-07-14

检索证明

根据委托人提供的论文材料，委托人华南农业大学动物科学学院 闫希亮 1 篇论文收录情况如下表。

| 序号 | 论文名称 | 发表刊物及发表的年月卷期/页码等 | 作者排名 | 论文等级 | 作者文中单位 | 收录情况 | 影响因子 | 中科院大类分区 |
|----|--|---|------|------|------------------|------|--|-------------------------------|
| 1 | MnN4 embedded zeolite-templated carbon for methylamine and trimethylamine sensing: Insights from DFT study | JOURNAL OF MOLECULAR LIQUIDS 出版年: 2024 出版日期: MAR 1 卷期: 397 页码: - 文献号: 124090 文献类型: Article | 通讯作者 | A 类 | 华南农业大学 动物科学学院 | SCI | IF2-year=5.2 IF5-year=5.1 (2024) | 化学 2 区 Top 期刊: 否 (2025) |

说明: 论文等级和中科院大类分区按《华南农业大学学术论文评价方案(试行)》划分。

报告免责声明: 如未盖章, 报告无效



检索证明

根据委托人提供的论文材料，委托人华南农业大学动物科学学院 闫希亮 1 篇论文收录情况如下表。

| 序号 | 论文名称 | 发表刊物及发表的年月卷期/页码等 | 作者排名 | 论文等级 | 作者文中单位 | 收录情况 | 影响因子 | 中科院大类分区 |
|----|---|---|--------------|------|--------------|------|--|--------------------------------|
| 1 | DFT perspective of gas sensing properties of metal oxide nanocages toward trimethylamine: Effects of humidity, temperature and electric field | MATERIALS TODAY SUSTAINABILITY 出版年：2024 出版日期：MAR 卷期：25 页码：- 文献号：100668 文献类型：Article | 共同通讯作者（倒数第一） | B 类 | 华南农业大学动物科学学院 | SCI | IF2-year=7.9 IF5-year=8.0 (2024) | 材料科学 3 区 Top 期刊：否 (2025) |

说明：论文等级和中科院大类分区按《华南农业大学学术论文评价方案（试行）》划分。

报告免责声明：如未盖章，报告无效



检索证明

根据委托人提供的论文材料，委托人华南农业大学动物科学学院 闫希亮 1 篇论文收录情况如下表。

| 序号 | 论文名称 | 发表刊物及发表的年月卷期/页码等 | 作者排名 | 论文等级 | 作者文中单位 | 收录情况 | 影响因子 | 中科院大类分区 |
|----|---|---|--------------|------|--------|------|--|------------------------------------|
| 1 | Reaching the Full Potential of Machine Learning in Mitigating Environmental Impacts of Functional Materials | REVIEWS OF ENVIRONMENTAL CONTAMINATION AND TOXICOLOGY 出版年：2022 出版日期：DEC 卷期：260 1 页码：- 文献号：21 文献类型：Review | 共同通讯作者（倒数第一） | A 类 | 广州大学 | SCI | IF2-year=6.0 IF5-year=7.1 (2022) | 环境科学与生态学 2 区 Top 期刊：否 (2022) |

说明：论文等级和中科院大类分区按《华南农业大学学术论文评价方案（试行）》划分。

报告免责声明：如未盖章，报告无效



检索证明

根据委托人提供的论文材料，委托人华南农业大学动物科学学院 闫希亮 1 篇论文收录情况如下表。

| 序号 | 论文名称 | 发表刊物及发表的年月卷期/页码等 | 作者排名 | 论文等级 | 作者文中单位 | 收录情况 | 影响因子 | 中科院大类分区 |
|----|---|---|--------|------|--------|------|--|------------------------------------|
| 1 | Comprehensive Interrogation on Acetylcholinesterase Inhibition by Ionic Liquids Using Machine Learning and Molecular Modeling | ENVIRONMENTAL SCIENCE & TECHNOLOGY 出版年：2021 出版日期：NOV 2 卷期：55 21 页码：14720-14731 文献类型：Article | 共同通讯作者 | T2 类 | 广州大学 | SCI | IF2-year=11.357 IF5-year=12.154 (2021) | 环境科学与生态学 1 区 Top 期刊：是 (2021) |

说明：论文等级和中科院大类分区按《华南农业大学学术论文评价方案（试行）》划分。

报告免责声明：如未盖章，报告无效



检索证明

根据委托人提供的论文材料，委托人华南农业大学动物科学学院 闫希亮 1 篇论文收录情况如下表。

| 序号 | 论文名称 | 发表刊物及发表的年月卷期/页码等 | 作者排名 | 论文等级 | 作者文中单位 | 收录情况 | 影响因子 | 中科院大类分区 |
|----|---|--|--------|------|--------|------|--|------------------------------------|
| 1 | Predicting cytotoxicity of binary pollutants towards a human cell panel in environmental water by experimentation and deep learning methods | CHEMOSPHERE 出版年：2022 出版日期：JAN 卷期：287 页码：- 文献号：132324 文献类型：Article | 共同通讯作者 | A 类 | 广州大学 | SCI | IF2-year=8.8 IF5-year=8.3 (2022) | 环境科学与生态学 2 区 Top 期刊：是 (2022) |

说明：论文等级和中科院大类分区按《华南农业大学学术论文评价方案（试行）》划分。

报告免责声明：如未盖章，报告无效

检索员：尹银怀
华南农业大学图书馆
2025-07-14

检索证明

根据委托人提供的论文材料，委托人华南农业大学动物科学学院 闫希亮 1 篇论文收录情况如下表。

| 序号 | 论文名称 | 发表刊物及发表的年月卷期/页码等 | 作者排名 | 论文等级 | 作者文中单位 | 收录情况 | 影响因子 | 中科院大类分区 |
|----|---|--|--------|------|--------|------|--|-------------------------------|
| 1 | Developmental toxicity of fenbuconazole in zebrafish: Effects on mitochondrial respiration and locomotor behavior | TOXICOLOGY 出版年: 2022 出版日期: MAR 30 卷期: 470 页码: - 文献号: 153137 文献类型: Article | 共同通讯作者 | B 类 | 广州大学 | SCI | IF2-year=4.5 IF5-year=4.6 (2022) | 医学 3 区 Top 期刊: 否 (2022) |

说明: 论文等级和中科院大类分区按《华南农业大学学术论文评价方案(试行)》划分。

报告免责声明: 如未盖章, 报告无效

检索员: 尹银怀

华南农业大学图书馆

2025-07-14

SCAULIB202519138

检索证明

根据委托人提供的论文材料，委托人华南农业大学动物科学学院 闫希亮 1 篇论文收录情况如下表。

| 序号 | 论文名称 | 发表刊物及发表的年月卷期/页码等 | 作者排名 | 论文等级 | 作者文中单位 | 收录情况 | 影响因子 | 中科院大类分区 |
|----|--|---|--------|------|--------|------|--|-------------------------------------|
| 1 | Paper Unraveling the joint toxicity of transition-metal dichalcogenides and per- and polyfluoroalkyl substances in aqueous mediums by experimentation, machine learning and molecular dynamics | JOURNAL OF HAZARDOUS MATERIALS 出版年: 2023 出版日期: FEB 5 卷期: 443 页码: - 文献号: 130303 文献类型: Article | 共同通讯作者 | T2 类 | 广州大学 | SCI | IF2-year=12.2 IF5-year=11.9 (2023) | 环境科学与生态学 1 区 Top 期刊: 是 (2023) |

说明: 论文等级和中科院大类分区按《华南农业大学学术论文评价方案(试行)》划分。

报告免责声明: 如未盖章, 报告无效

检索员: 尹银怀
华南农业大学图书馆

2025-07-15

ARTICLE



<https://doi.org/10.1038/s41467-020-16413-3>

OPEN

Construction of a web-based nanomaterial database by big data curation and modeling friendly nanostructure annotations

Xiliang Yan ^{1,2}, Alexander Sedykh^{2,3}, Wenyi Wang², Bing Yan ^{1,4}✉ & Hao Zhu ^{2,5}✉

Modern nanotechnology research has generated numerous experimental data for various nanomaterials. However, the few nanomaterial databases available are not suitable for modeling studies due to the way they are curated. Here, we report the construction of a large nanomaterial database containing annotated nanostructures suited for modeling research. The database, which is publicly available through <http://www.pubvinas.com/>, contains 705 unique nanomaterials covering 11 material types. Each nanomaterial has up to six physico-chemical properties and/or bioactivities, resulting in more than ten endpoints in the database. All the nanostructures are annotated and transformed into protein data bank files, which are downloadable by researchers worldwide. Furthermore, the nanostructure annotation procedure generates 2142 nanodescriptors for all nanomaterials for machine learning purposes, which are also available through the portal. This database provides a public resource for data-driven nanoinformatics modeling research aimed at rational nanomaterial design and other areas of modern computational nanotechnology.

¹Institute of Environmental Research at Greater Bay, Key Laboratory for Water Quality and Conservation of the Pearl River Delta, Ministry of Education, Guangzhou University, Guangzhou 510006, China. ²The Rutgers Center for Computational and Integrative Biology, Camden, NJ 08102, USA. ³Sciome, Research Triangle Park, North Carolina 27709, USA. ⁴School of Environmental Science and Engineering, Shandong University, Jinan 250100, China. ⁵Department of Chemistry, Rutgers University, Camden, NJ 08102, USA. ✉email: drbingyan@yahoo.com; hao.zhu99@rutgers.edu

The global market value of nanotechnology is expected to reach \$90.5 billion by 2021¹ as commercial and consumer nano-products continue to rise^{2–4}. Increased production, use and environmental accumulation of these nanomaterials present important toxicology concerns^{5–7}. A variety of in vitro and in vivo assays evaluating their potential environmental and human health effects have generated vast quantities of experimental data^{8,9}, requiring data extraction, analysis, and sharing for guiding the safe design of next-generation nanomaterials^{10,11}. This urgency is echoed in the recent Nanoinformatics Roadmap 2030 in USA and Europe, aimed at promoting the capture, preservation, and dissemination of publicly available data on nanomaterials. The Roadmap, which outlined the importance of coordinating research efforts and charting the challenges in nanoinformatics as a set of milestones, envisages the flow of data from experimentalists into structured databases that can be used by computational modelers to predict nanomaterial properties, exposure and hazard values that will support regulatory actions¹².

Two large databases for chemicals and proteins have already impacted different areas of science. As a small molecule database, PubChem provides structural annotation (e.g., chemical structures, SMILES, and InChi key), physicochemical properties (e.g., logP and molecular weight) and available bioactivities (e.g., EC50 and IC50)¹³. Since its launch in 2004, PubChem has served various scientific communities including cheminformatics, chemical biology, medicinal chemistry, and drug discovery. Another crucial database for scientific community is the Protein Data Bank (PDB)¹⁴, which provides three-dimensional structures of biological macromolecules, (e.g., proteins and nucleic acids) as PDB files for broad researchers in fields like molecular biology, structural biology, and computational biology. However, a comparable nanomaterial database is not available. The key to building such a database of nanomaterials is nanostructure annotation—a computer-friendly format for encoding information.

Several nanomaterial databases serving specific areas are available (Table 1)^{15–19}. For example, the cancer Nanotechnology Laboratory (caNanoLab) database (<https://cananolab.nci.nih.gov/>) built by the National Cancer Institute in 2007¹⁵ is designed to

expedite and validate the use of nanotechnology in biomedicine. However, it is not fully accessible to the public because it contains proprietary data. While these nanomaterial databases, which are shown in Table 1, share published data and have been used for modeling studies^{16,20,21}, they are limited by the way they are curated. Although, new file formats (e.g., JSON¹⁷ and ISA-TAB-Nano²²) are also specially designed in several nanomaterial databases, such as eNanomapper and NANOREG, to store and manage the curated nanomaterial data. Nanomaterial entities (e.g., composition, physicochemical properties, and biological activities of the nanomaterials) in these databases exist as text outputs extracted directly from publications, ignoring nanostructure annotations that are critical for modeling studies. As a result, variables (e.g., physicochemical properties) used in previous modeling studies were mostly experimentally generated. Without nanostructure annotations, diverse structural information for predictive modeling and other research such as nanostructure analysis and visualization cannot be performed.

Here, we report a publicly available nanomaterial database that contains annotated nanostructures of diverse nanomaterials suitable for immediate modeling research. The database, constructed from thousands of scientific papers, currently contains 705 unique nanomaterials, 1365 physicochemical property (e.g., logP, zeta potential, and hydrodynamic diameter) and 2386 bioactivity (e.g., cell viability, cellular uptake, and ROS) data points. All experimentally obtained information on the structure of the nanomaterials, such as form, size, shape, and surface ligand were annotated and stored as PDB files, which are downloadable from the web portal (<http://www.pubvinas.com/>). The PDB files can be used to generate nanodescriptors, which were created in-house to quantitatively represent nanostructure diversity. Using these nanodescriptors, we developed predictive models for three critical property/bioactivity endpoints of various nanomaterials using machine learning (*k*-nearest neighbor) and deep learning (deep neural network) approaches. This is the largest and the only nanomaterial database that contains nanostructure annotations to support nanomaterial modeling and rational nanomaterial design. Furthermore, the predictive models developed from this database can be used to predict three critical properties and

| Table 1 Nanomaterial databases. | | | |
|---|-------------|--|-----------|
| Database | Data points | Usage | Reference |
| caNanoLab https://cananolab.nci.nih.gov/ | 1308 | Expedite and validate the use of nanotechnology in biomedicine | 15 |
| S ² NANO http://portal.s2nano.org/ | 6854 | Develop and commercialize safe and sustainable nano-products | 16 |
| eNanomapper http://www.enanomapper.net/ | 5528 | Develop a computational framework for nanotoxicity data management | 17 |
| Nanomaterial registry http://nanohub.org/ | 2031 | Help understanding the fundamental properties of nanomaterials | 18 |
| Nanoparticle information library http://nanoparticlelibrary.net/ | 88 | Capture the information about nanomaterial physicochemical characteristics | 19 |
| NanoMILE https://ssl.biomax.de/nanomile/cgi/login_bioxm_portal.cgi | 120 | Contain characterization data and high throughput screening toxicity data of nanomaterials | — |
| DaNa Knowledge Base https://www.nanopartikel.info/en/ | — | Help understanding the impacts of nanomaterials for humans and the environment | — |
| NanoDatabank http://nanoinfo.org/nanodatabank/ | >1000 | Design with simplicity of nanomaterial data storing and sharing | — |
| NBI Knowledgebase http://nbi.oregonstate.edu/ | 200 | Help understanding the mechanism of nanomaterial exposure effects in biological systems | — |
| Nanowerk https://www.nanowerk.com/ | 4000 | Help the nanotechnology community to research nanomaterials | — |

The low curation of existing nanomaterials's databases is limiting their application in modeling studies. Here the authors report a publicly available nanomaterial database that contains annotated nanostructures of diverse nanomaterials immediately available for modeling research studies.

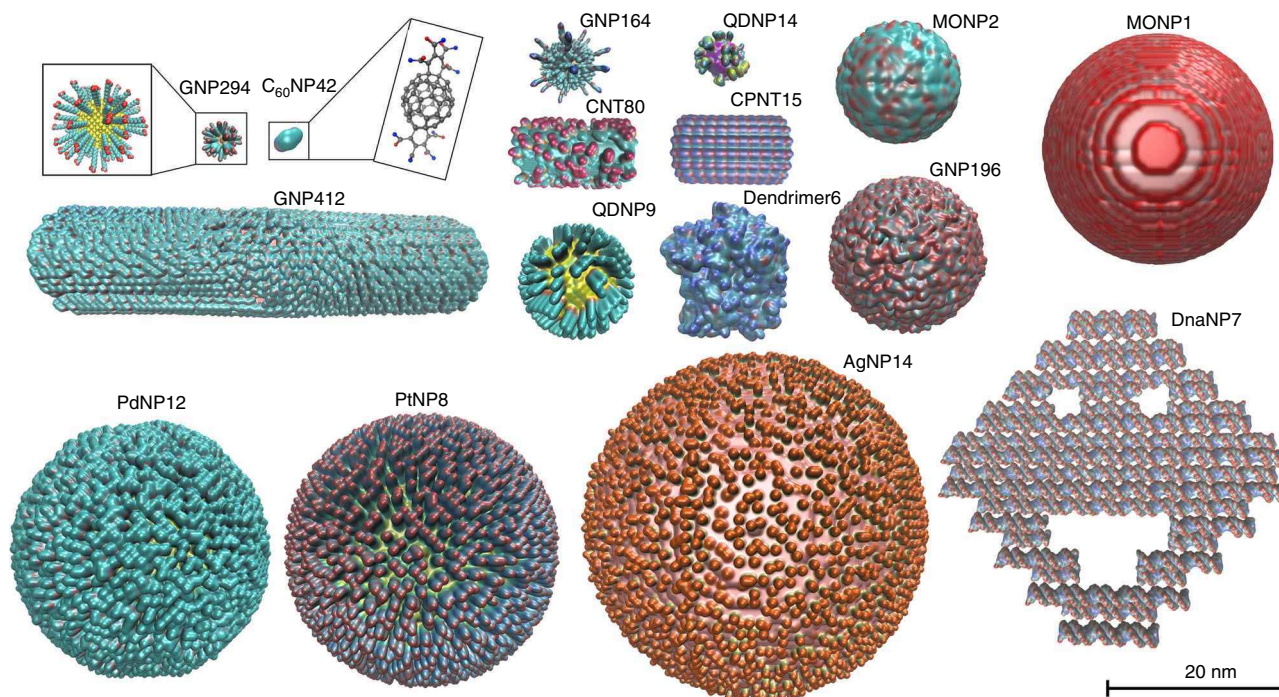


Fig. 1 Visualization of 16 representative nanomaterials in the database. The database contains 705 nanomaterials that vary in material type, size, shape, and surface ligand. Most nanomaterials were spherical but rod-like and irregular ones were also annotated and included in the database. Different surface chemistries of the nanostructures were rendered in different colors by QuickSurf drawing method in VMD, offering direct impressions of the nanomaterials. Emboldened text represents text identifiers that can be used to search for the nanomaterial in the database.

bioactivity (i.e., logP, zeta potentials, and cellular uptake) of new nanomaterials.

Results

Construction of the nanomaterial database. A total of 705 nanomaterials, comprising 414 gold nanoparticles (GNPs), 17 silver nanoparticles (AgNPs), 12 platinum nanoparticles (PtNPs), 12 palladium nanoparticles (PdNPs), 80 carbon nanotubes (CNTs), 48 buckminsterfullerenes (C_{60}), 34 quantum dots (QDs), 32 metal oxides nanoparticles (MONPs), 21 DNA origami nanoparticles (DnaNPs), 11 dendrimers, and 24 cyclic peptide nanotubes (CPNTs), were annotated for the database. Figure 1 shows 16 representative nanostructures covering all nanomaterial types in the database and are rendered by visual molecular dynamics (VMD) using the QuickSurf method²³. This method uses positions of atoms and the Monte Carlo simulation for generating the volumetric density maps and isosurface that simulate electron density and solvent accessible surface for the input nanostructures. For example, GNP164 represents the 164th gold nanoparticle in the database that has a core diameter of 5 nm (Fig. 1, see Supplementary Data for other structure information). The nanostructures varied in material type, size, shape, and surface ligand. For example, C_{60} NP42 and AgNP14 are 1 nm and 40 nm, respectively. Although most nanomaterials are spherical, the database also contains rod-like (e.g., GNP412, CNT80, and CPNT15) and irregular (e.g., Dendrimer6 and DnaNP7) nanomaterials. Different surface chemistries of the nanostructures were rendered with different colors. For example, the nanoparticle PdNP12 (logP = 2.52) with hydrophobic surface ligands are shown as cyan while the nanoparticle PtNP8 (logP = -1.47) with hydrophilic surface ligands are rendered purple. Other structural details can also be observed, for example, the long surface ligand chains on GNP164 are shown as tentacles. These detailed 3D plots of nanomaterials in the database provide direct

impressions of the relevant surface chemistry and physicochemical properties.

Figure 2 is an overview of the data curated in this study (see Supplementary Data for details), including physicochemical properties (logP and zeta potential), bioactivities (cell viability, reactive oxidative stress (ROS), and cellular uptake), along with the nanomaterial types and structure information (surface ligands and size). Although majority of the nanomaterials are GNPs, there are 291 other types of nanomaterials (Fig. 2a). The functions of nanomaterials are affected by surface small molecules (e.g., drugs and peptides), which determine their diverse applications (e.g., drug delivery and tumor diagnosis). As shown in Fig. 2b, the number of surface ligands ranged from 1 (such as C_{60} nanomaterials) to more than 6000 (such as GNP12). This is because ligand density is highly affected by the properties of the surface ligands. For example, similar sized GNP (~5.8 nm) can have around 200 ligands per particle for positively charged ligands (e.g., GNP130) and negatively charged ligands (e.g., GNP138). Meanwhile, ligands without charges can pack up to over 700 surface ligands per GNP (e.g., GNP152). Among the 705 nanomaterials, one contained up to four different ligands (GNP392) and there were in total 314 unique surface ligands. The spherical nanomaterials in the database also had a wide size distribution (Fig. 2c). At the lower end, there are GNPs with diameter less than 10 nm that are suitable for biomedical applications^{24,25}. Some spherical nanoparticles have sizes ranging from 10 to 45 nm.

The nanomaterials in this database are also biologically diverse (Fig. 2d–h). The logP values of the nanomaterials, which describe the hydrophobicity of relevant nanomaterials, ranged from -2.68 to 2.72. Zeta potential—the charge at the interface between the nanomaterial surface and its liquid medium—of nanomaterials in this database was tested in three solutions (water, aqueous buffer, and serum) and they ranged from -93.73 mV to 86.80 mV (Fig. 2e). Cell viability showed a spread from 2% to 118.05%

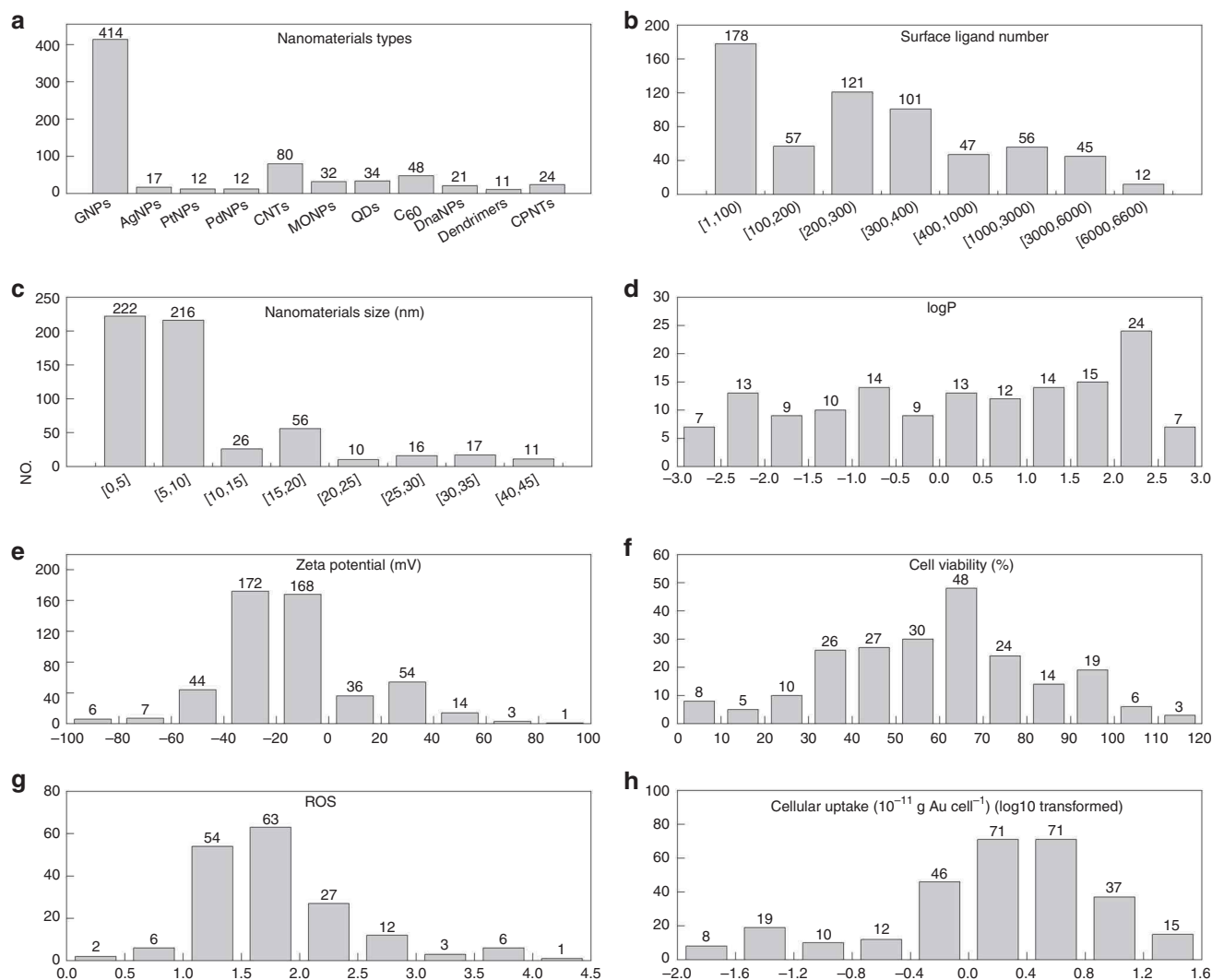


Fig. 2 Overview of the nanomaterial database. **a–h** Distributions of nanomaterials accounting to **a** nanomaterial type, **b** surface ligand number, **c** nanomaterial size, **d** logP, **e** zeta potential, **f** cell viability, **g** reactive oxidative stress (ROS) and **h** cellular uptake. Nanomaterials in the database show chemical, structural, and biological diversity. The numbers in the brackets of **b**, **c** represent the range of the surface ligand number and nanomaterial size, respectively.

(Fig. 2f), indicating the various nanomaterials induced varying degrees of cytotoxicity. ROS level, which is used to evaluate cellular oxidative stress, linked to cancer, diabetes, and aging, also ranged widely from 0.44 to 4.10 (Fig. 2g). For nanomaterials, cellular uptake is usually a prerequisite for their applications in drug delivery, bioimaging and, etc.²⁶. In this database, cellular uptake capacity of all nanomaterials varied from $-1.87 \text{ g cell}^{-1}$ to 1.36 g cell^{-1} with a log10-transformation (Fig. 2h).

Analysis of nanostructure diversity. After annotating and saving the structures of all 705 nanomaterials in our database as PDB files, we calculated 680 nanodescriptors using the Virtual Nanostructure Simulations (VINAS) toolbox²⁷—an in-house cheminformatics program designed to calculate descriptors based on the annotated nanomaterial structures. The current descriptors calculated by VINAS are based on Delaunay tessellation, which is a fast way to transform the nano surface geometry into quantitative values as nanodescriptors. Using the 680 calculated nanodescriptors, we performed principal component analysis (PCA) and used the top three principal components, which account for 79% of the total descriptor variance, to show the occupation of all nanomaterials in a 3D chemical space

(Fig. 3a). All the nanomaterials were structurally diverse and occupied most of this chemical space. Compared to other nanomaterials, MONPs occupied a larger area because the relevant VINAS nanodescriptor values, which are based on atomic properties, varied significantly according to the unique atoms (e.g., Zn, Co, and Ce) that make up each MONPs.

Chemical structure is the key to determine a molecule's physicochemical properties and biological activities. The content that structurally similar molecules should exhibit similar bioactivities is the fundamental hypothesis of all quantitative structure-activity relationship (QSAR) and other relevant modeling studies^{28,29}. To quantitatively study the structural similarity among nanomaterials, we calculated the pairwise Euclidean distance for all nanomaterials. All nanodescriptor results were normalized to a range between 0 and 1 before calculation. A total of 248,160 distances were generated among each two of the 705 nanomaterials. The distribution of values ranged from 0.004 to 17.31 with an average of 5.3 (Fig. 3b). Two substances are typically considered similar if their normalized Euclidean distance is less than 0.5^{30,31}. In this database, some nanomaterials that belong to different nanomaterial types, are also structurally similar. For example, the Euclidean distances between PtNP1 and

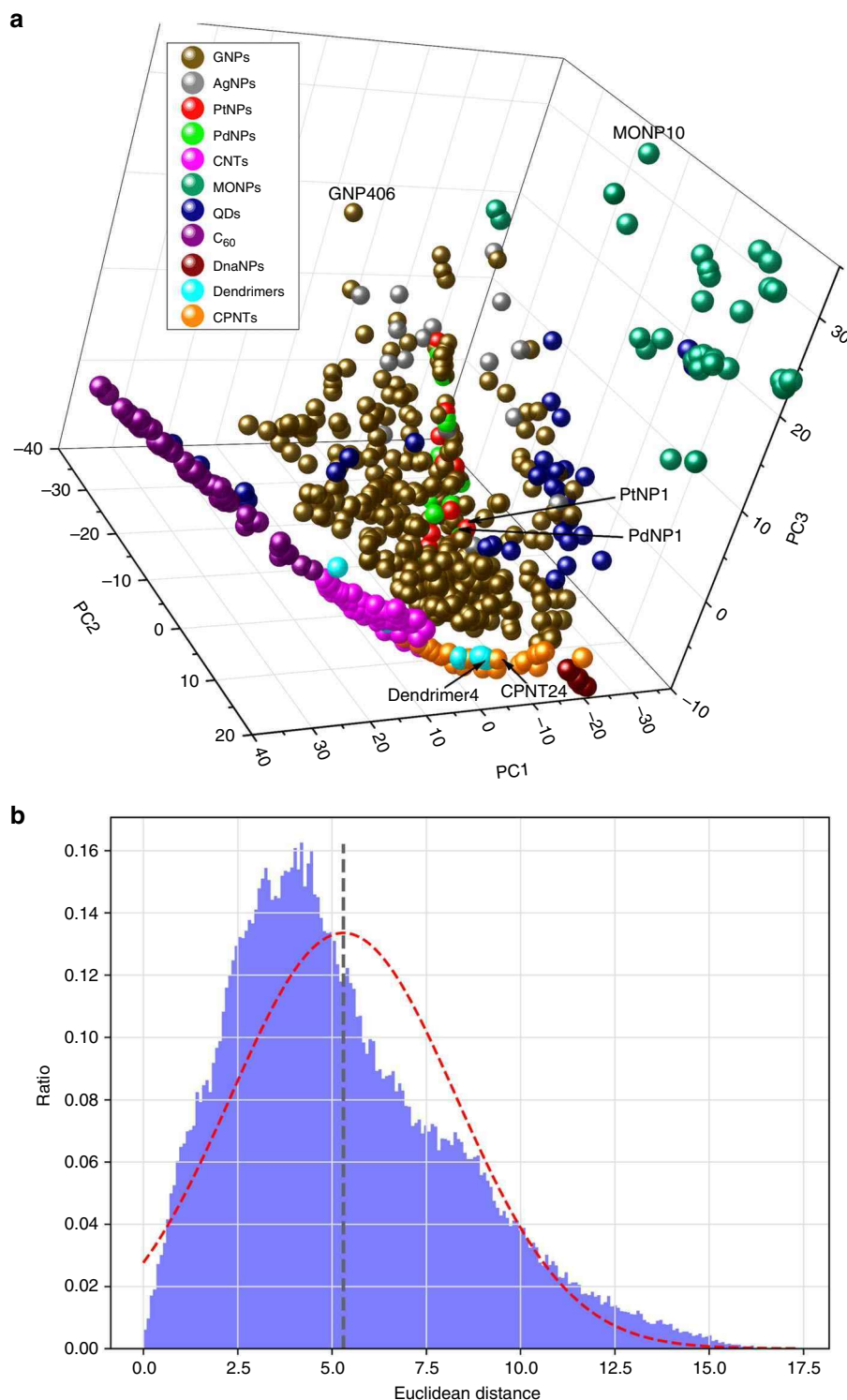


Fig. 3 Nanostructure diversity analysis. **a** Nanomaterial chemical space shown by principal component analysis (PCA) of all 705 nanomaterials. The three principal components (PC1, PC2, PC3) account for 43%, 23, and 13% of the total descriptor variance, respectively. Different colors were associated with different nanomaterial classifications. Six nanomaterials are shown with their identifiers (i.e., PtNP1, PdNP1, Dendrimer4, CPNT24, GNP406, and MONP10). **b** Distribution of the 248,160 Euclidean distances calculated from each pair of nanomaterials in the database. The distribution ranged from 0.004 to 17.31 with an average of 5.3 (black dashed line). Normalized distribution curve is shown as red dotted line.

PdNP1, and between Dendrimer4 and CPNT24 are 0.037 and 0.14, respectively. PtNP1 and PdNP1 with Euclidean distance near zero are considered structurally similar because they are about the same size (6 nm and 5.8 nm, respectively) and have the same surface ligand at the similar density (371 and 365 ligands

per particle, respectively). Although Dendrimer4 is irregular and CPNT24 is rod-like, they are considered structurally similar because they have similar sizes (2 nm and 1.41 nm * 1.44 nm) and atomic compositions (C, N, O, and H). Some structural outliers such as GNP406 and MONP10 were also seen. GNP406 is

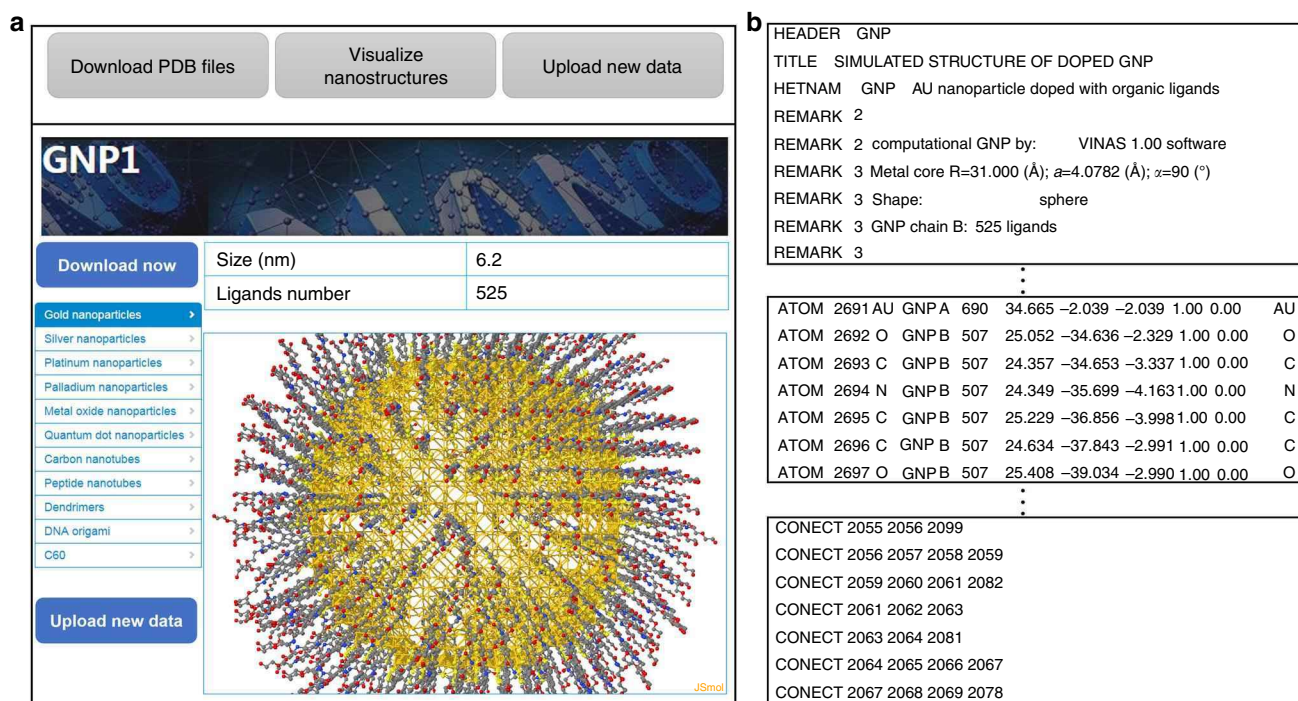


Fig. 4 PubVINAS online portal. **a** Screenshot of PubVINAS. The bars on the above and left of the picture show the user functions (e.g., download/upload data, visualize data, select data based on classifications, and, etc.) and **b** example PDB file as an output shown as three parts: (1) the basic information, (2) atom type and coordinates, and (3) the connections between atoms.

structurally different because it is a rod-like gold nanoparticle (most are spherical) that is also relatively large at 30 nm × 33 nm. MONP10, which is a La₂O₃ metal oxide nanoparticle around 24.6 nm in diameter, is structurally different because of the unique properties of the Lanthanum (La) atom.

Nanomaterial database portal. To share the structural annotated data, we developed an online database portal (<http://www.pubvinas.com/>) that currently can be used to download the PDB files, visualize the nanostructures and upload new data (Fig. 4a). A full-time computer systems administrator will be responsible for maintaining the portal. Each PDB file of the nanomaterials can be downloaded by clicking the dropdown bars with their corresponding classification (e.g., gold nanoparticles, silver nanoparticles, and platinum nanoparticles). Users can view the nanostructure online from the corresponding PDB file and open the downloaded PDB file using well-known cheminformatics software (e.g., VMD, RasMol, and MOE). An example PDB file is shown in Fig. 4b. The first part of the file contains the basic information on the structure of the nanomaterial (e.g., the form, shape and size); the second part contains information about the atoms (e.g., atom type and coordinates); and the third part includes information on the bond/connection between atoms. Users may also share their new data (e.g., new nanomaterials synthesized and/or tested against new bioassays) by uploading them as a text file (Fig. 4a). After reviewing the upload files, the system administrator will generate the PDB files and add the new dataset to the nanomaterial database. We expect to add more functions, such as an online toolbox to calculate nanodescriptors and several trained models, in the future to predict the properties of new nanomaterials.

Predictive nano property/bioactivity modeling. Using data from the database, we used *k*-Nearest Neighbor (*k*NN), a traditional machine learning approach, and deep neural network (DNN), a

representative deep learning algorithm, to build computational models that will identify quantitative relationships between the annotated nanostructures and target activities. Two properties and one bioactivity (i.e., logP, zeta potential tested in water at pH = 7, and cellular uptake capacity in A549 cells) were selected for modeling. The logP dataset contains 147 unique nanomaterials, including 123 GNPs, 12 PtNPs and 12 PdNPs. The zeta potential dataset contains 213 unique nanomaterials, including 148 GNPs, 6 AgNPs, 12 PtNPs, 12 PdNPs, 8 MONPs, 24 QDNPs, and 3 Dendrimers. The cellular uptake dataset contains 71 GNPs, which were tested against A549 cells. Each model was developed using the *k*NN and DNN approach with VINAS nanodescriptors calculated from the associated nanomaterials in the dataset. The performance of the model was evaluated by both the 5-fold cross-validation and external prediction methods common in modeling studies^{32,33}. For each endpoint, the available data were randomly split into a training set (80% of the data) for developing the model, and a test set (20% of the data) for external validation of the model. The training set was further split into five subsets. The model was developed using four of the five subsets and the remaining subset was used for validation. This procedure was repeated five times until all subsets were used for validation once.

The correlations between experimental and predicted values of the six resulting models based on *k*NN and DNN are shown in Fig. 5, which also includes the root mean square error (RMSE) and correlation coefficients (*R*²). Overall, both *R*² and RMSE for 5-fold cross validation (*R*²_{5CV} and RMSE_{5CV}) and external prediction (*R*²_{val} and RMSE_{val}) are at the same order of magnitude, indicating the 5-fold cross-validation process and external prediction yielded similar results. All correlation coefficients (both *R*²_{5CV} and *R*²_{val}) were above 0.5, indicating that all six models successfully predicted the relationships between the annotated the nanostructures and target activities³⁴. When comparing *R*²_{5CV} and *R*²_{val}, *k*NN models (Fig. 5a, c, e) showed better predictability than DNN models (Fig. 5b, d, f). Although DNN is a popular modeling tool and has demonstrated

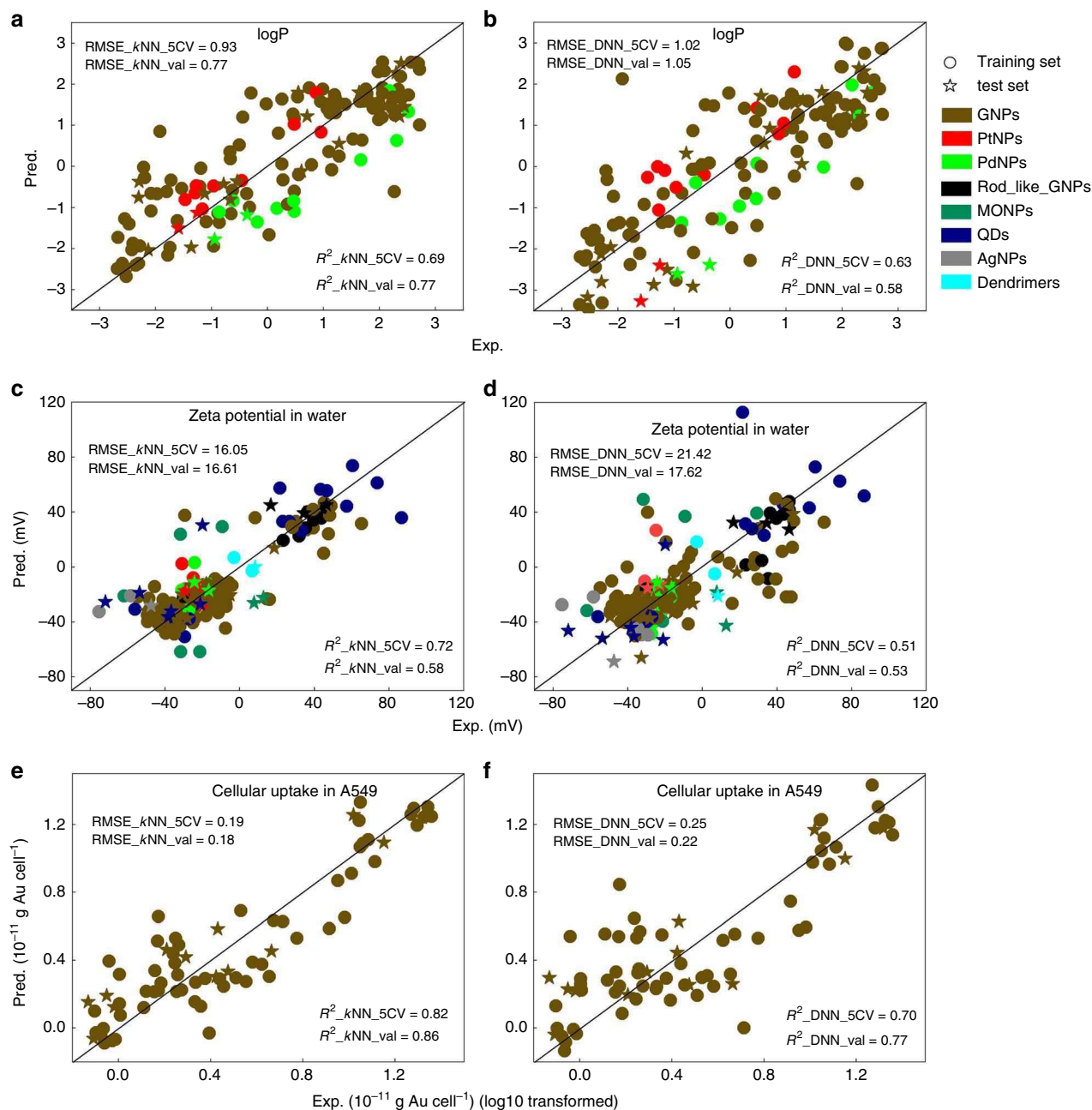


Fig. 5 Correlations between experimental (Exp) and predicted (Pred) values. kNN (**a**, **c**, **e**) and DNN (**b**, **d**, **f**) models are developed for predicting logP (**a**, **b**), zeta potential (**c**, **d**) and cellular uptake (**e**, **f**). logP dataset contains 147 unique nanomaterials, including 123 GNPs, 12 PtNPs and 12 PdNPs. Zeta potential dataset contains 213 unique nanomaterials, including 148 GNPs, 6 AgNPs, 12 PtNPs, 12 PdNPs, 8 MONPs, 24 QDNPs, and 3 Dendrimers. Cellular uptake dataset contains 71 GNPs, which were tested against A549 cells. Root mean square error (RMSE) and correlation coefficients (R^2) are also shown. RMSE_5CV and R^2_{5CV} represent the RMSE and R^2 values for 5-fold cross validation, while RMSE_val and R^2_{val} represent the values for external prediction. R^2_{CV} and R^2_{val} above 0.5 indicate high correlation between Exp and Pred values.

high predictability in recent modeling challenges in drug discovery^{35,36}, it performed differently in other studies^{37,38}. Here, the lower predictability of DNN models is likely due to overfitting caused by too many neurons in the layers compared to the size of the input data. Both kNN (Fig. 5e) and DNN (Fig. 5f) cellular uptake models performed better (i.e., higher R^2 values) than the logP and zeta potential models.

The resulted models, especially the kNN models, can be used to predict new nanomaterials directly from their structures and assist rational nanomaterial design. Because the cellular uptake dataset consists of only one type of nanomaterial (GNP) so that the applicability of the resulted cellular uptake model can be

reliably applied to predict new GNPs. The logP and zeta potential datasets consist of various types of nanomaterials collected from different sources. The two models can be used to predict the properties of a wide range of nanomaterials. In addition, based on the same nanostructure annotation method, machine learning models were recently built to predict the inflammatory responses and cytotoxicity of various carbon nanoparticles³⁹. Once a new nanomaterial is virtually designed using computer, its properties will be assessed using the developed models before chemical synthesis. This procedure will greatly save resources by prioritizing new nanomaterials with desired properties and/or cellular uptake potentials.

Discussion

In summary, we constructed a universal nanomaterials database containing structure annotations suitable for direct computational modeling. The database currently contains 705 unique nanomaterials with multiple biological testing results. Structures of these nanomaterials were annotated and stored as PDB files that are retrievable from online portal. The new data being uploaded in the future will rapidly expand the database. We also developed several machine learning models using three property and bioactivity datasets in this database and showed the models had highly accurate predictability based on cross-validation and external validation results (i.e., $R^2 > 0.5$). The resulted models can be used to predict two critical properties and one bioactivity of new nanomaterials directly from their nanostructures. Some materials such as alloy nanomaterials⁴⁰, polymeric micelles⁴¹, mesoporous nanomaterials⁴², and metal-organic frameworks (MOFs)-based nanomaterials⁴³ were tentatively not included in the database because their nanostructures were poorly defined and the related publications currently lack quality control information on their synthesis. Other nanomaterials that were annotated still lack representative data in some target endpoints, for example, cellular uptake potentials. For the database to be more useful, there is still a need to generate more biological data of diverse nanomaterials.

Methods

Experimental data curation. The database was compiled from in-house data (297 unique nanomaterials) and external data (408 unique nanomaterials). The in-house data were collected from our previously published studies (these references were provided in Supplementary References). The external data was collected by manual literature searching. This process resulted in more than 1000 papers with nanomaterial data for further examination. The data were included into the database with the following conditions satisfied: (1) the material (e.g., core atoms) and size information were provided in this paper; (2) the surface ligand structures can be annotated and transferred into SMILES; (3) the nano-bioactivity or physico-chemical property data were provided with detailed experimental information. There are 69 publications that were identified to contain useful data by fulfilling all criterions (these references were provided in Supplementary References). Each publication was manually examined, and relevant structure information (e.g., core, size, and surface ligands), experimental data, and testing details were extracted from the corresponding papers. For raw data with size and shape information of a set of nanoparticles instead of a single molecular entity, the same core was set for all the nanoparticles in this data source. Data were also obtained directly from figures of published papers using PlotDigitizer. The surface ligand structures were converted to SMILES, which were shown in Supplementary Data.

Nanostructure annotation. For nanoparticles, the core atoms were first put together as a nano core based on the particle size information. Then the associated surface ligands were randomly placed on the core surface. For GNPs, AgNPs, PtNPs, PdNPs, MONPs, and QDs, the core of the corresponding nanostructure was generated by replicating the unit cell of the most thermodynamically stable crystal structures and then deleting atoms outside the input diameter data. The lattice parameters (e.g., unit cell lengths and angles) were obtained from the Materials Project (<https://materialsproject.org/>). For CNTs, the python toolkit scikit-nano (<https://scikit-nano.org/>) was applied to construct the carbon core (pristine CNTs). All the surface ligands were optimized before being grafted to the nano core. As for C₆₀, the SMILES obtained from the paper⁴⁴ were directly converted to PDB file. The PDB files of DnNPs were either collected from the corresponding papers^{45–48} or generated by the Legogen⁴⁹. The PDB files of dendrimers were collected from corresponding papers^{50–53}. For CPNTs, the nanostructures were generated by an in-house program written in C++⁵⁴. In this procedure, the amino acids were firstly connected as various cyclic peptides through peptide bonds and then these cyclic peptides were stacked as CPNTs through H-bonds.

Nanodescriptor generation. At first, 126 tetrahedron fragments were generated for each nanostructure based on our previous study, which were calculated by combining the Delaunay tessellation and atom types²⁷. In our previous study, the value of a nanodescriptor was calculated as the value of each tetrahedron electronegativity multiplied by its occurrences in the nanostructure. As described above, the range of nanomaterial size has a wide distribution in the current database. As a result, there will be a large difference of the tetrahedron occurrences between the large nanomaterials and small nanomaterials. In order to resolve this issue, property-based descriptors were also calculated in this study. The procedure can be described as follows: (1) The occurrence of each tetrahedron was converted

to frequency (the occurrence of each tetrahedron divided by the total number of all the tetrahedrons in each nanostructure). (2) More atomic properties were introduced, which included the calculated radii (R_{cal}), the covalent radii (R_{cov}), the empirical radii (R_{emp}), the atom mass (M), the boiling point (T_{bol}), the density (ρ), the electron affinity (E_{ea}), the electronegativity (χ), the heat of fusion (ΔH_{fus}), the heat of vaporization (ΔH_{vap}), the first ionization energy (IE_1), the second ionization energy (IE_2), the melting point (T_{mel}), the molar volume (V_{mol}), the specific heat (Q), the thermal conductivity (λ) and the valence (q). Then, these 17 property values of each tetrahedron were multiplied respectively by the corresponding tetrahedron frequency, as described in our previous study²⁷. As a result, 17 descriptor matrices were generated that each descriptor matrix contained 126 individual descriptors (the tetrahedron fragments integrated with atomic properties). The calculated nanodescriptors for all nanomaterials are available from the web portal. After removing descriptors with limited information (e.g., with consistent values over all nanomaterials), total 680 nanodescriptors were used for modeling purpose. The nanostructure annotations and nanodescriptor generations were described in details in our previous papers^{27,55}.

Computational modeling. The datasets were split into training sets (80% of the original datasets) and test sets (20% of the original datasets). The training sets were used to build models, and the associated test sets were used to evaluate the developed models. The performance of each model was indicated by 5-fold cross validation within the training set and the external validation by predicting the test set. In this study, two different machine learning approaches were used to develop the computational models. The k -nearest neighbor (k NN) method used the weighted average of nearest neighbors as its prediction and employed a variable selection procedure to define neighbors^{27,55}, which was developed in-house (also available at <http://chembench.mml.unc.edu/>). The deep neural network (DNN) is a multi-layer feed-forward neural network, which was implemented using Keras 2.2.4 (<https://keras.io/>) python deep learning library, with the TensorFlow backend. The DNN architecture used in this study included a sequence of five dense layers (three hidden layers), which were fully connected neural layers. Three hidden layers contained 512, 128, and 64 nodes, respectively. The relu was used as activation function to perform non-linear transformations. The dropout function, set as 0.2, was used to prevent overfitting of the resulting models. The rmsprop and mean squared error (MSE) were used as optimizer and loss function to compile the DNN model in this study. The learning rate was set as the default value of the rmsprop optimizer. Each DNN model was trained for 300 epochs.

Data availability

All experimental data can be accessed from the Supplementary Data or from the Experimental data page of the web portal (<http://www.pubvinas.com/>).

Received: 9 January 2020; Accepted: 22 April 2020;

Published online: 20 May 2020

References

- McWilliams, A. *The Maturing Nanotechnology Market: Products and Applications* (BCC Research, Wellesley, MA, 2016).
- Quadros, M. E. & Marr, L. C. Silver nanoparticles and total aerosols emitted by nanotechnology-related consumer spray products. *Environ. Sci. Technol.* **45**, 10713–10719 (2011).
- Stamm, H., Gibson, N. & Anklaam, E. Detection of nanomaterials in food and consumer products: bridging the gap from legislation to enforcement. *Food Addit. Contam.* **29**, 1175–1182 (2012).
- Vance, M. E. et al. Nanotechnology in the real world: redeveloping the nanomaterial consumer products inventory. *Beilstein J. Nanotechnol.* **6**, 1769–1780 (2015).
- Valsami-Jones, E. & Lynch, I. How safe are nanomaterials? *Science* **350**, 388–389 (2015).
- Cao, M., Li, J., Tang, J., Chen, C. & Zhao, Y. Gold nanomaterials in consumer cosmetics nanoproducts: analyses, characterization, and dermal safety assessment. *Small* **12**, 5488–5496 (2016).
- Djurišić, A. B. et al. Toxicity of metal oxide nanoparticles: Mechanisms, characterization, and avoiding experimental artefacts. *Small* **11**, 26–44 (2015).
- Zhang, Y. et al. Perturbation of physiological systems by nanoparticles. *Chem. Soc. Rev.* **43**, 3762–3809 (2014).
- Sharifi, S. et al. Toxicity of nanomaterials. *Chem. Soc. Rev.* **41**, 2323–2343 (2018).
- Maojo, V. et al. Nanoinformatics: a new area of research in nanomedicine. *Int. J. Nanomed.* **7**, 3867–3890 (2012).
- Hendren, C. O., Powers, C. M., Hoover, M. D. & Harper, S. L. The nanomaterial data curation initiative: a collaborative approach to assessing, evaluating, and advancing the state of the field. *Beilstein J. Nanotechnol.* **6**, 1752–1762 (2015).

12. Haase, A. & Klaessig, F. *EU US Roadmap Nanoinformatics 2030* (EU NanoSafety Cluster, 2018).
13. Kim, S. et al. PubChem substance and compound databases. *Nucleic Acids Res.* **44**, D1202–D1213 (2016).
14. Rose, P. W. et al. The RCSB protein data bank: Integrative view of protein, gene and 3D structural information. *Nucleic Acids Res.* **45**, D271–D281 (2017).
15. Gaheen, S. et al. CaNanoLab: data sharing to expedite the use of nanotechnology in biomedicine. *Comput. Sci. Disco.* **6**, 014010 (2013).
16. Trinh, T. X., Ha, M. K., Choi, J. S., Byun, H. G. & Yoon, T. H. Curation of datasets, assessment of their quality and completeness, and nanoSAR classification model development for metallic nanoparticles. *Environ. Sci. Nano* **5**, 1902–1910 (2018).
17. Jeliaskova, N. et al. The eNanoMapper database for nanomaterial safety information. *Beilstein J. Nanotechnol.* **6**, 1609–1634 (2015).
18. Mills, K. C., Murry, D., Guzan, K. A. & Ostraat, M. L. Nanomaterial registry: database that captures the minimal information about nanomaterial physico-chemical characteristics. *J. Nanopart. Res.* **16**, 2219 (2014).
19. Miller, A. L., Hoover, M. D., Mitchell, D. M. & Stapleton, B. P. The Nanoparticle Information Library (NIL): A prototype for linking and sharing emerging data. *J. Occup. Environ. Hyg.* **4**, D131–D134 (2007).
20. Ha, M. K. et al. Toxicity classification of oxide nanomaterials: effects of data gap filling and pchem score-based screening approaches. *Sci. Rep.* **8**, 1–11 (2018).
21. Choi, J. S., Trinh, T. X., Yoon, T. H., Kim, J. & Byun, H. G. Quasi-QSAR for predicting the cell viability of human lung and skin cells exposed to different metal oxide nanomaterials. *Chemosphere* **217**, 243–249 (2019).
22. Thomas, D. G. et al. ISA-TAB-Nano: a specification for sharing nanomaterial research data in spreadsheet-based format. *BMC Biotechnol.* **13**, 2 (2013).
23. Krone, M., Stone, J., Ertl, T. & Schulten, K. Fast visualization of Gaussian density surfaces for molecular dynamics and particle system trajectories. *EuroVis(Short Papers)* <https://doi.org/10.2312/PE/EuroVisShort/EuroVisShort2012/067-071> (2012).
24. Khlebtsov, N. & Dykman, L. Biodistribution and toxicity of engineered gold nanoparticles: a review of in vitro and in vivo studies. *Chem. Soc. Rev.* **40**, 1647–1671 (2011).
25. Huo, S. et al. Ultrasmall gold nanoparticles as carriers for nucleus-based gene therapy due to size-dependent nuclear entry. *ACS Nano* **8**, 5852–5862 (2014).
26. Depan, D. & Misra, R. D. K. Hybrid nanoparticle architecture for cellular uptake and bioimaging: direct crystallization of a polymer immobilized with magnetic nanoparticles on carbon nanotubes. *Nanoscale* **4**, 6325–6335 (2012).
27. Yan, X. et al. *In silico* profiling nanoparticles: predictive nanomodeling using universal nanodescriptors and various machine learning approaches. *Nanoscale* **11**, 8352–8362 (2019).
28. Cherkasov, A. et al. QSAR modeling: where have you been? Where are you going to? *J. Med. Chem.* **57**, 4977–5010 (2014).
29. Zhu, H. Big data and artificial intelligence modeling for drug discovery. *Annu. Rev. Pharmacol. Toxicol.* **60**, 573–589 (2020).
30. Dragos, H., Gilles, M. & Alexandre, V. Predicting the predictability: a unified approach to the applicability domain problem of qsar models. *J. Chem. Inf. Model.* **49**, 1762–1776 (2009).
31. Shen, M. et al. Quantitative structure-activity relationship analysis of functionalized amino acid anticonvulsant agents using k nearest neighbor and simulated annealing PLS methods. *J. Med. Chem.* **45**, 2811–2823 (2002).
32. Wang, W., Kim, M. T., Sedykh, A. & Zhu, H. Developing enhanced blood-brain barrier permeability models: integrating external bio-assay data in QSAR modeling. *Pharm. Res.* **32**, 3055–3065 (2015).
33. Kim, M. T. et al. Mechanism profiling of hepatotoxicity caused by oxidative stress using antioxidant response element reporter gene assay models and big data. *Environ. Health Perspect.* **124**, 634–641 (2016).
34. Eriksson, L. et al. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. *Environ. Health Perspect.* **111**, 1361–1375 (2003).
35. Mayr, A. et al. Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chem. Sci.* **9**, 5441–5451 (2018).
36. Feng, C. et al. Gene expression data based deep learning model for accurate prediction of drug-induced liver injury in advance. *J. Chem. Inf. Model.* **59**, 3240–3250 (2019).
37. Russo, D. P., Zorn, K. M., Clark, A. M., Zhu, H. & Ekins, S. Comparing multiple machine learning algorithms and metrics for estrogen receptor binding prediction. *Mol. Pharm.* **15**, 4361–4370 (2018).
38. Rodríguez-Pérez, R., Miyao, T., Jasial, S., Vogt, M. & Bajorath, J. Prediction of compound profiling matrices using machine learning. *ACS Omega* **3**, 4713–4723 (2018).
39. Liu, G. et al. Analysis of model PM2.5-induced inflammation and cytotoxicity by the combination of a virtual carbon nanoparticle library and computational modeling. *Ecotoxicol. Environ. Saf.* **191**, 110216 (2020).
40. Liu, X., Wang, D. & Li, Y. Synthesis and catalytic properties of bimetallic nanomaterials with various architectures. *Nano Today* **7**, 448–466 (2012).
41. Movassaghian, S., Merkel, O. M. & Torchilin, V. P. Applications of polymer micelles for imaging and drug delivery. *Wiley Interdiscip. Rev. Nanomed. Nanobiotechnol.* **7**, 691–707 (2015).
42. Tang, F., Li, L. & Chen, D. Mesoporous silica nanoparticles: synthesis, biocompatibility and drug delivery. *Adv. Mater.* **24**, 1504–1534 (2012).
43. Dang, S., Zhu, Q. L. & Xu, Q. Nanomaterials derived from metal-organic frameworks. *Nat. Rev. Mater.* **3**, 1–14 (2017).
44. Toropova, A. P., Toropov, A. A., Benfenati, E., Leszczynska, D. & Leszczynski, J. QSAR modeling of measured binding affinity for fullerene-based HIV-1 PR inhibitors by CORAL. *J. Math. Chem.* **48**, 959–987 (2010).
45. Bai, X., Martin, T. G., Scheres, S. H. W. & Dietz, H. Cryo-EM structure of a 3D DNA-origami object. *Proc. Natl Acad. Sci. USA* **109**, 20012–20017 (2012).
46. Nguyen, N. et al. The absence of tertiary interactions in a self-assembled DNA crystal structure. *J. Mol. Recognit.* **25**, 234–237 (2012).
47. Dong, Y., Chen, S., Zhang, S. & Sodroski, J. Folding DNA into a lipid-conjugated nanobarrel for controlled reconstitution of membrane proteins. *Angew. Chem.* **130**, 2094–2098 (2018).
48. Pan, K. et al. Lattice-free prediction of three-dimensional structure of programmed DNA assemblies. *Nat. Commun.* **5**, 5578 (2014).
49. Slone, S. M. *Building DNA Brick Structures with LegoGen*. Theoretical and Computational Research at the Interface of Physics, Biology, and Nanotechnology, <http://bionano.physics.illinois.edu/tutorials/using-legogen-build-dna-brick-structures> (2016).
50. Maingi, V., Jain, V., Bharatam, P. V. & Maiti, P. K. Dendrimer building toolkit: Model building and characterization of various dendrimer architectures. *J. Comput. Chem.* **33**, 1997–2011 (2012).
51. Schillreiff, P., Mundiña-Weilenmann, C., Romero, E. L. & Morilla, M. J. Selective cytotoxicity of PAMAM G5 core-PAMAM G2.5 shell tecto-dendrimers on melanoma cells. *Int. J. Nanomed.* **7**, 4121–4133 (2012).
52. Maiti, P. K., Çağın, T., Wang, G. & Goddard, W. A. Structure of PAMAM dendrimers: generations 1 through 11. *Macromolecules* **37**, 6236–6254 (2004).
53. Naha, P. C., Davoren, M., Lyng, F. M. & Byrne, H. J. Reactive oxygen species (ROS) induced cytokine production and cytotoxicity of PAMAM dendrimers in J774A.1 cells. *Toxicol. Appl. Pharmacol.* **246**, 91–99 (2010).
54. Yan, X., Fan, J., Yu, Y., Xu, J. & Zhang, M. Transport behavior of a single Ca²⁺, K⁺, and Na⁺ in a water-filled transmembrane cyclic peptide nanotube. *J. Chem. Inf. Model.* **55**, 998–1011 (2015).
55. Wang, W. et al. Predicting nano-bio interactions by integrating nanoparticle libraries and quantitative nanostructure activity relationship modeling. *ACS Nano* **11**, 12641–12649 (2017).

Acknowledgements

X.Y. and B.Y. were supported by the National Key R&D Program of China (2016YFA0203103), the National Natural Science Foundation of China (91543204 and 91643204), and the introduced innovative R&D team project under the “The Pearl River Talent Recruitment Program” of Guangdong Province (2019ZT08L387). W.W. and H. Z. were partially supported by the National Institute of Environmental Health Sciences (grant number R01ES031080, R15ES023148, and P30ES005022). We thank A. L. Chun of Science StoryLab for editorial service.

Author contributions

H.Z. and B.Y. conceived and designed the study. H.Z. designed the project strategy. X.Y. curated the experimental data, constructed the web portal, simulated the virtual nanomaterials, calculated nanodescriptors, built the models, and performed validation. A.S. designed, wrote and tested codes for constructing the virtual nanomaterials and guided several nanodescriptors calculation. W.W. helped analyze the results. X.Y., B.Y., and H.Z. wrote the paper. All authors have read and approved this paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41467-020-16413-3>.

Correspondence and requests for materials should be addressed to B.Y. or H.Z.

Peer review information *Nature Communications* thanks Christine Ogilvie Hendren and David Winkler for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020

Converting Nanotoxicity Data to Information Using Artificial Intelligence and Simulation

Xiliang Yan, Tongtao Yue, David A. Winkler, Yongguang Yin, Hao Zhu, Guibin Jiang, and Bing Yan*



Cite This: *Chem. Rev.* 2023, 123, 8575–8637



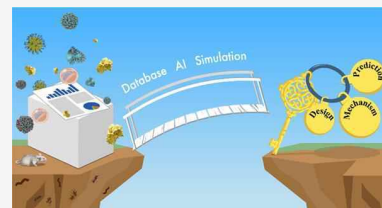
Read Online

ACCESS |

Metrics & More

Article Recommendations

ABSTRACT: Decades of nanotoxicology research have generated extensive and diverse data sets. However, data is not equal to information. The question is how to extract critical information buried in vast data streams. Here we show that artificial intelligence (AI) and molecular simulation play key roles in transforming nanotoxicity data into critical information, i.e., constructing the quantitative nanostructure (physicochemical properties)–toxicity relationships, and elucidating the toxicity-related molecular mechanisms. For AI and molecular simulation to realize their full impacts in this mission, several obstacles must be overcome. These include the paucity of high-quality nanomaterials (NMs) and standardized nanotoxicity data, the lack of model-friendly databases, the scarcity of specific and universal nanodescriptors, and the inability to simulate NMs at realistic spatial and temporal scales. This review provides a comprehensive and representative, but not exhaustive, summary of the current capability gaps and tools required to fill these formidable gaps. Specifically, we discuss the applications of AI and molecular simulation, which can address the large-scale data challenge for nanotoxicology research. The need for model-friendly nanotoxicity databases, powerful nanodescriptors, new modeling approaches, molecular mechanism analysis, and design of the next-generation NMs are also critically discussed. Finally, we provide a perspective on future trends and challenges.



CONTENTS

| | | | |
|---|------|---|------|
| 1. Introduction | 8576 | 3.2. Encoding NM Properties into Nanodescriptors | 8587 |
| 2. Nanotoxicity Data and Model-Friendly Databases | 8577 | 3.2.1. The Significance of Feature Engineering | 8587 |
| 2.1. Challenge of Nanotoxicity Data Availability | 8577 | 3.2.2. Experimental Properties as Nanodescriptors | 8588 |
| 2.2. Generation of Nanotoxicity “Big Data” | 8578 | 3.2.3. Nanodescriptors from Quantum Chemical Calculations | 8588 |
| 2.2.1. The Sources of Nanotoxicity Data | 8578 | 3.2.4. Structural Nanodescriptors from 1D Text Representations | 8588 |
| 2.2.2. High-Throughput Synthesis of Diverse NMs | 8579 | 3.2.5. Nanodescriptors from Periodic Table Properties, Liquid Drop Model, and Metal–Ligand Binding Theory | 8589 |
| 2.2.3. Rigorous Characterization of NMs to Guarantee Data Quality | 8581 | 3.2.6. Nanodescriptors from Full Nanostructures | 8589 |
| 2.2.4. Quality Assays for Toxicity Testing of NMs | 8581 | 3.2.7. Nanodescriptors from MD Simulations | 8590 |
| 2.2.5. Improving Nanotoxicity Data Quality by Standardization | 8581 | 3.2.8. Latent Nanodescriptors from Deep Learning | 8591 |
| 2.2.6. Improving Data Availability with Nanotoxicity Quantification | 8582 | 3.2.9. Developing Universal Nanodescriptors | 8592 |
| 2.3. Collection and Curation of Nanotoxicity Data | 8582 | 3.3. Application of ML to Nanotoxicity Modeling | 8593 |
| 2.4. Currently Available Nanotoxicity Databases | 8583 | 3.3.1. Nanodescriptor Preprocessing and Feature Selection | 8593 |
| 2.5. Construction of Model-Friendly Databases under the FAIR Principles | 8584 | | |
| 3. Unraveling Quantitative Nanostructure–Toxicity Relationships Using Artificial Intelligence | 8585 | | |
| 3.1. Application of AI to Nanotoxicology | 8586 | | |
| 3.1.1. A Brief History of AI | 8586 | | |
| 3.1.2. Traditional Machine Learning and Deep Learning | 8586 | | |
| 3.1.3. The Basic Steps in ML | 8587 | | |
| 3.1.4. Application of ML to Nanotoxicology | 8587 | | |

Received: February 5, 2023

Published: June 1, 2023



| | |
|---|------|
| 3.3.2. Modeling Physicochemical Properties of NMs | 8594 |
| 3.3.3. Modeling Interactions between NMs and Biomolecules | 8595 |
| 3.3.4. Modeling Cellular Responses Induced by NMs | 8595 |
| 3.3.5. Modeling Ecological Risks of NMs | 8596 |
| 3.3.6. Modeling Toxicity of NMs in Mammals | 8597 |
| 3.3.7. Applicability Domain of ML Models | 8597 |
| 3.4. Applications of ML to Mechanism Elucidation and NM Design | 8597 |
| 3.4.1. Applications of ML to Mechanism Analysis in Nanotoxicology | 8597 |
| 3.4.2. Design of Bespoke NMs with Targeted Properties | 8599 |
| 3.5. Summary of Applying AI in Nanotoxicology | 8601 |
| 4. Elucidating Nanotoxicity Mechanisms by Molecular Simulations | 8601 |
| 4.1. The Role of Molecular Simulation in Nanotoxicology | 8602 |
| 4.1.1. A Brief Introduction of Molecular Simulations | 8602 |
| 4.1.2. The Commonly Used Molecular Simulation Methods for Nanotoxicology | 8603 |
| 4.1.3. The Workflow of Molecular Simulations | 8603 |
| 4.1.4. Nanotoxicology Studies Using Molecular Simulations | 8604 |
| 4.2. Chemical Reactivities of NMs | 8605 |
| 4.2.1. NM-Induced ROS Generation | 8606 |
| 4.2.2. Chemical Transformation and Degradation of NMs | 8606 |
| 4.3. NM Interactions with Cell Membranes | 8608 |
| 4.3.1. Pathways of NM–Cell Membrane Interactions from MD | 8609 |
| 4.3.2. Effects of NM Physicochemical Properties on Cell Membrane Interactions | 8609 |
| 4.3.3. Synergistic Cell Entry of Multiple NMs | 8611 |
| 4.3.4. Influence of the Cellular Microenvironment on NM–Cell Interactions | 8613 |
| 4.4. MD Simulations of NM Interactions with Functional Proteins | 8613 |
| 4.4.1. Molecular Insights into Protein Adsorption by NMs | 8613 |
| 4.4.2. Protein Denaturation and Dysfunction Induced by NMs | 8614 |
| 4.5. ML Approaches to Analyze and Enhance Molecular Simulations | 8615 |
| 4.6. Summary of Applying Molecular Simulations in Nanotoxicology | 8616 |
| 5. Concluding Remarks and Future Perspectives | 8616 |
| Author Information | 8617 |
| Corresponding Author | 8617 |
| Authors | 8617 |
| Notes | 8618 |
| Biographies | 8618 |
| Acknowledgments | 8618 |
| Abbreviations | 8618 |
| References | 8619 |

1. INTRODUCTION

In the past three decades, nanotechnology applications have developed rapidly in diverse areas such as medicine, energy,

materials, and environment.^{1,2} According to a report from Emergen Research (<https://www.emergenresearch.com/>), the global nanotechnology market is expected to reach 290 billion USD in 2028. Currently, engineered nanomaterials (NMs) are used in more than 5,000 commercial products, including medical equipment, textiles, fuel additives, cosmetics, and plastics (<https://nanodb.dk/>).³ Because these products aim to enhance the quality of human life, people will be inevitably exposed to the NMs during production, transportation, daily use, and disposal. Whether by inhalation, oral uptake, skin absorption, or medical use, NMs may cause adverse biological effects by crossing biological barriers (e.g., skin,⁴ air–blood (lung),⁵ reproductive system,⁶ or blood–brain⁷) and can accumulate in different tissues and organs. Concerns about the health and safety of NMs have catalyzed a new research field called nanotoxicology.^{8,9} For two decades, nanotoxicology research has aimed to address adverse effects of NMs to facilitate the safe and sustainable applications of NMs and nanotechnology via regulatory agencies.^{10,11} As of December 2022, at least 170,000 nanotoxicology-related papers have been published (data were retrieved from the Web of Science with the keywords “nano*” and “toxic*”). This area of research has generated extensive data sets but also created a new challenge to extract critical information from potentially vast data streams. Recently, “big data” (the definition varies markedly for different disciplines) is changing the world and rapidly becoming a crucial driving force for almost every field, for example, drug discovery,^{12,13} materials design,^{14,15} and medical diagnosis and theranostics.^{16,17} Nanotoxicology studies have not yet fully benefited from big data so far. This raises a fundamental question of how we can effectively convert the nanotoxicity data into critical information, i.e., generating mechanism information on nanotoxicity and predicting potential adverse effects of new NMs.

The primary challenge comes from the ways current nanotoxicity data are obtained.^{18–20} Most data are generated from traditional one-at-a-time experiments under vastly different laboratory conditions.^{8,21,22} In addition, substandard or incomplete characterization and toxicity testing of NMs significantly compromises the quality of nanotoxicity data used for modeling and simulation. Moreover, the labor-intensive data collection, annotation, curation, and integration create additional challenges.²³ Although the construction of model-friendly databases will improve the availability of nanotoxicity data, the lack of NM annotation, standardization, and ontologies is a critical technical challenge.^{18,24} The European Union has funded several Horizon 2020 projects to address the challenges of faster synthesis and characterization and annotation enhancements such as NInChI (International Chemical Identifier for NMs).^{25–27}

Another challenge is unraveling the quantitative nanostructure (or physicochemical properties of NMs)–toxicity relationships (QNTR) in high multidimensional feature spaces.^{26,28} The success of artificial intelligence (AI) in the safety assessment of small organic molecules has demonstrated their ability to identify specific toxicity features from a myriad of molecular properties.^{29–32} However, examples of the applications of AI to nanotoxicology are much less popular due to the paucity of model-friendly databases and the lack of suitable nanodescriptors.^{19,33} Nanodescriptors, the mathematical entities that encode important physicochemical information about NMs, should apply to all types of nanostructures and properties such as composition, size, shape, and surface chemistry. They

should ideally reflect the influence of the external environment (e.g., the protein corona—the layer of proteins that naturally bind to NMs in biological fluids that modulates their biological effects).³⁴ Unfortunately, most available nanodescriptors were originally developed for small molecules and are not specific to NMs.^{25,35} Thus, complex NM structures and their interactions with the environment and biological systems are not sufficiently represented.³⁶ Moreover, machine learning (ML, a subset of AI algorithms) deployment in nanotoxicology has to deal with feature selection, model optimization, and interpretation.^{14,37,38}

Elucidation of the molecular mechanisms underlying nanotoxicity remains a significant challenge. Molecular simulations have made valuable contributions to model the interactions of small molecules with the target biomolecules (e.g., proteins and lipids) and ultimately provide insights into the toxicity mechanism.^{39,40} However, the application of molecular simulation to nanotoxicology is compromised by the complexity of NMs, which exist as distributions rather than discrete entities, are much larger than organic molecules, and have the added complication of the biologically relevant entity being the NM plus corona. Specifying complex NM systems in specific environments and performing simulations over biologically relevant time and length scales demands large computational resources.^{41,42} Although coarse-graining can allow simulation over larger time and length scales,⁴³ some atomistic details are inevitably sacrificed, and the simulation precision reduced. The other obstacle is that the actual complexity of the biological medium cannot be fully captured in molecular simulations.⁴⁴ For instance, widely used lipid bilayer models often do not accurately represent biological membranes.

The accumulation of large quantities of nanotoxicity data creates a risk of being overwhelmed by these data and missing crucial information in nanotoxicology.^{26,45} Here, we address these questions by first establishing ways to integrate big data into model-friendly nanotoxicity databases (Section 2). We then summarize methods and progress in applying AI and ML to unravel NMs property–toxicity relationships (Section 3). The application of molecular simulation to elucidate the molecular mechanism underlying nanotoxicity will be presented (Section 4) before concluding remarks and future perspectives (Section 5). Data is not equivalent to information, and with model-friendly databases, AI, and molecular simulation, we can bridge the gap between massive nanotoxicity data and critical information (Figure 1).

2. NANOTOXICITY DATA AND MODEL-FRIENDLY DATABASES

AI approaches and molecular simulation methods can extract crucial information from nanotoxicity data that can be used to create safe and sustainable (safe-by-design) nanotechnology. Most AI methods, especially ML, are data-driven, so sufficient high-quality and chemically diverse data is a prerequisite. Although molecular simulation methods do not necessarily require large volumes of data, the simulation conditions, including force field and system parameters, are also generated from high-quality experimental data. However, current nanotoxicity data are primarily unsuitable for direct use by AI modeling methods and molecular simulations. This is due to the low quality of data and the need for NM ontologies, characterization standards, nanotoxicity quantification, and the curation of model-friendly databases and related issues. In this section, we first discuss the problems with current nanotoxicity

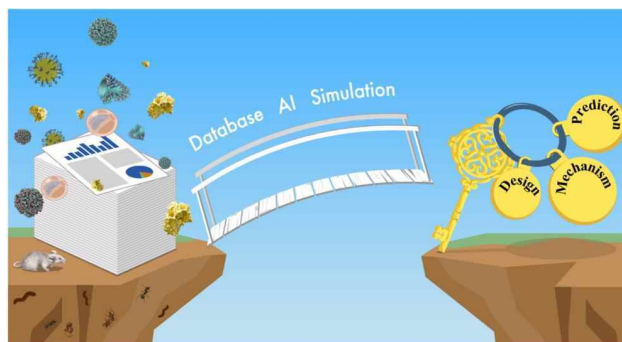


Figure 1. Bridging the gap between nanotoxicity data and critical information through the database, AI, and molecular simulation. Nanotoxicity databases provide an integrated platform for improved data availability. AI approaches excel at unraveling complex relationships between the physicochemical properties of NMs and their toxicities. Molecular simulation methods aim to elucidate the molecular mechanism underlying nanotoxicity.

data and then focus on integrating nanotoxicity data into model-friendly databases.

2.1. Challenge of Nanotoxicity Data Availability

Data has become an important engine driving everyday activities, such as education, the economy, and healthcare. Data science has been very influential in transforming data into useful information. For example, millions of small molecules and associated bioactivities data allow ML studies to significantly reduce costs and research times in drug discovery.^{46,47} Currently, it is difficult for nanotoxicology to generate such a large volume of data as in the field of small molecule drug discovery due to the challenges from NM synthesis, characterization, toxicity testing, and even data collection. Most nanotoxicity-related ML models were trained on relatively small data sets (mainly less than 100 data points; see the summary of Section 3.5). However, big data does not mean a mere increase in the volume of data. A recent study discussed the term of “big data” in nanoscience.⁴⁸ The authors believed that big data did not prescribe an absolute number for set size but referred to the point when a data set was sufficiently large that the model accuracy achieved no longer depended on the set size.⁴⁸ On the other hand, the data science community has also confirmed the adage “garbage in, garbage out”. This has been proven again and again in drug discovery practices. An increased volume of data *per se* may not yield better models; relatively higher quality data is more valuable to train applicable models. With reliable quality control, small data sets can yield equivalent performance to those trained on a more extensive data set.⁴⁸ Many researchers are calling for a shift in focus from big data to good data.⁴⁹ Therefore, we must focus more on improving nanotoxicity data quality and diversity rather than just increasing the volume. For nanotoxicology, efficient use of high-quality data can avoid duplication of efforts, leverage the advances of others, and ultimately promote safe and sustainable nanotechnology.

Nanotoxicity data are associated with the synthesis, characterization, and toxicity testing of NMs. Unlike small molecules, it is technically more challenging to synthesize many NMs rapidly. The quality of NMs and their toxicity data are also affected by other issues, such as the lack of standard experimental protocols and the need for uniform reporting formats.^{50,51} These result in only a fraction of data generated in few research laboratories

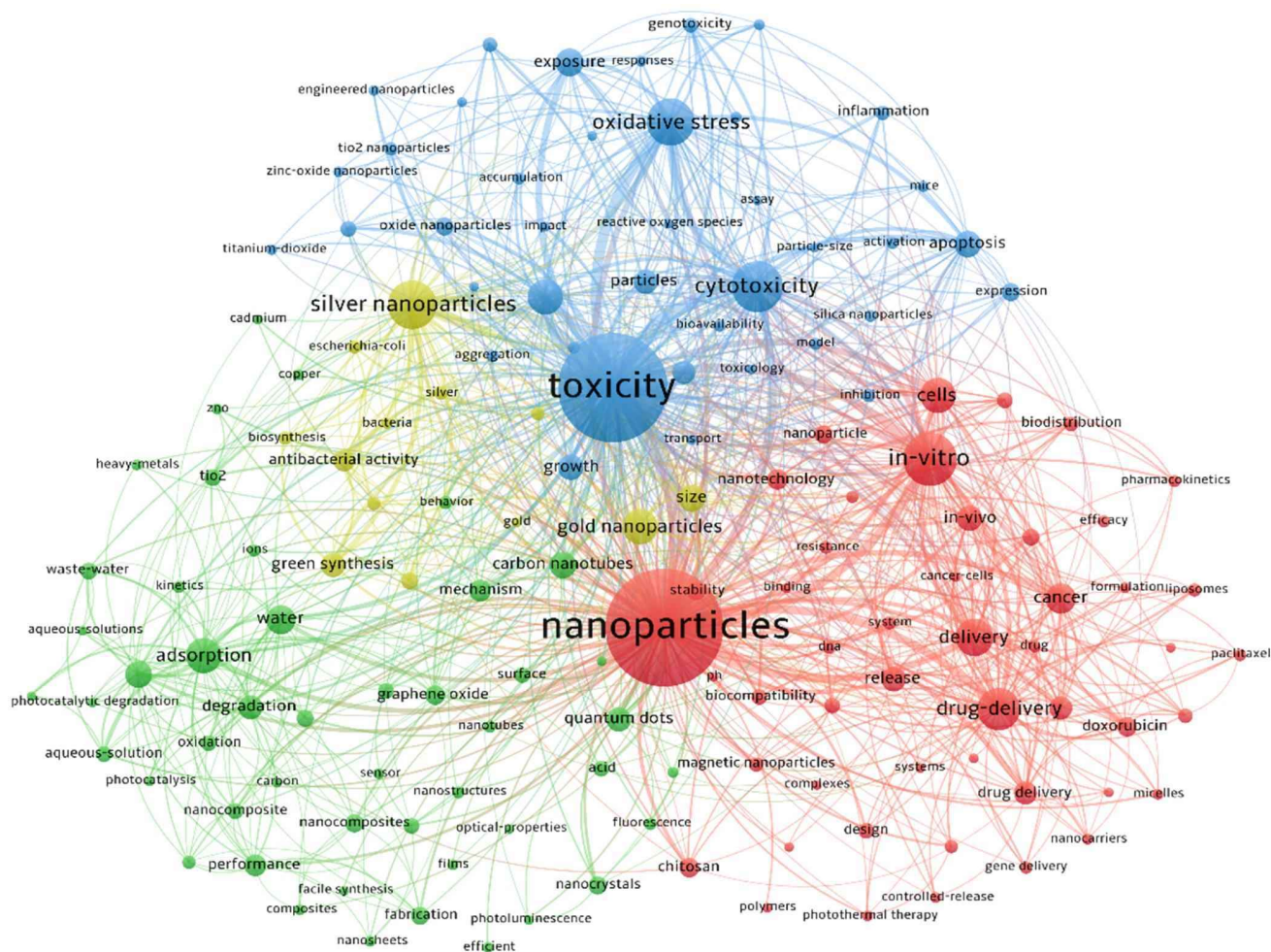


Figure 2. Top keywords of nanotoxicology and their co-occurrence network. Top keywords were extracted from over 90,000 previously published articles (as of December, 2022) related to “nano* and toxic*” in the Web of Science Core Collection Database. Keywords are grouped into color-coded categories using VOSviewer, depending on their intrinsic relationship.

being useful for computational modeling.¹⁸ One outcome is that the literature contains reports of multiple ML models trained on the same data sets generated from few laboratories,⁵² and only few nanotoxicity data sets were modeled by ML during the past decade.¹⁹ In other words, the quality of most current nanotoxicity data is insufficient for models, and those data have not been effectively exploited. Thus, the limited availability of model-friendly nanotoxicity data is due to the limitations in how NMs are synthesized, characterized, and tested, as how the data are extracted from diverse literature sources.

2.2. Generation of Nanotoxicity “Big Data”

2.2.1. The Sources of Nanotoxicity Data. Nanotoxicity data come from two sources: experiments and simulations. Experiments are labor- and time-intensive and involve manipulating chemicals and biological samples. Experimental nanotoxicity data may capture crucial nanobio interactions and faithfully characterize NMs in relevant environmental or biological media. Nanobio interactions refer to interactions between NMs and multiple biological systems such as lipids, proteins, microorganisms, cells, fish, plants, and higher animals. NMs may produce various toxicological effects such as cytotoxicity, genotoxicity, inflammation, carcinogenicity, and reproductive toxicity in biological systems exposed to them.

Figure 2 shows that current nanotoxicity data are mainly generated from *in vitro* studies of cytotoxicity, oxidative stress, and apoptosis. Compared to *in vivo* assays, *in vitro* tests are easier for data quality control. *In vitro* tests are also relatively cheaper, faster, and more efficient, allowing more NMs to be evaluated with the same resource.

The physicochemical properties of NMs play essential roles in determining their behavior in environmental and biological media and are the basis for understanding the mechanism of nanobio interactions. Among these, particle size is one of the most relevant factors in nanotoxicology research (Figure 2). Nanotoxicity data can also provide useful information about the environmental fate and pharmacological behavior of NMs, such as adsorption, transformation, and degradation. In the critical application of nanotechnology to healthcare, the nanomedicine field requires sufficient toxicology information on possible adverse effects of NMs to ensure their safe use.⁵³ Therefore, a significant portion of nanotoxicity data comes from nanomedicine studies, especially for cancer therapy (the red cluster in Figure 2).

Alternatively, molecular simulation involves physics-based methods such as density functional theory (DFT), molecular docking, and molecular dynamics (MD) simulations. Molecular

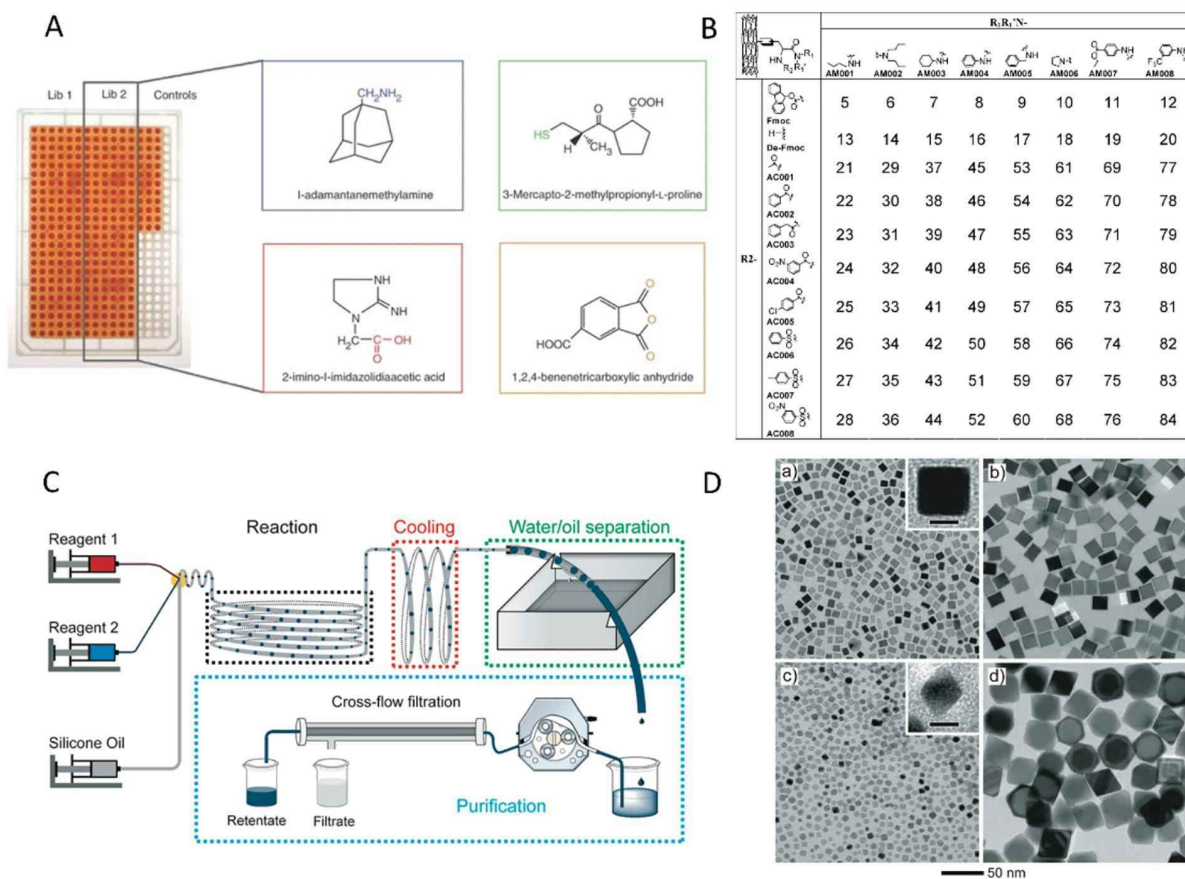


Figure 3. High-throughput synthesis of diverse NPs. (A) Parallel synthesis of NP collections on a large scale. These NP collections were parallelly synthesized by chemical coupling reactions after sufficient optimization of reaction conditions. Reproduced with permission from ref 70. Copyright 2005 Springer Nature. (B) Synthesis of nanocombinatorial NP libraries. Based on computer-aided design and starting material optimization, combinatorial chemistry can quickly create high-diversity NP libraries by combinatorial modifying of many NP structural elements. Therefore, the combinatorial library differs from synthesizing large NP collections based on easy coupling reactions. Reproduced with permission from ref 79. Copyright 2008 American Chemical Society. (C) Schematic illustration of the microfluidic reactor. Reproduced with permission from ref 80. Copyright 2018 American Chemical Society. (D) TEM images of the Pd NPs synthesized from the microfluidic reactor. Reproduced with permission from ref 80. Copyright 2018 American Chemical Society.

simulation is mainly used for generating nanodescriptors and elucidating molecular mechanisms underlying nanotoxicity. NM–protein binding affinities from molecular docking simulations can sometimes be used to predict toxicity quantitatively. For example, the inhibition of human immunodeficiency virus type 1 aspartic protease by various fullerene derivatives was evaluated, and this simulation data set has been used to train ML models.^{54–57} In other fields, such as heterogeneous catalysis, simulation data has been used to train ML models since structured data can be generated relatively easily.^{58–63} In a recent review, predictions based on simulation data are regarded as the bottom-up approach while learning from experimental data is viewed as the top-down approach.⁶⁴ Considering the time-consuming and laborious processes of generating experimental data, simulation data may be a valuable method to augment training data for ML models in nanotoxicology research. In the following, we will mainly focus on the current challenges in acquiring experimental nanotoxicity data.

2.2.2. High-Throughput Synthesis of Diverse NMs.

Although automation and robotics have greatly enhanced the synthesis and characterization of small molecules and materials,^{65,66} NM synthesis still mainly relies on traditional one-at-a-time synthesis methods and low-throughput testing

assays. As well as being time-consuming and laborious, these methods are prone to batch effects, causing huge discrepancies between laboratories on crucial data such as physicochemical properties and toxicity end points. In many cases, experimental results from different laboratories are contradictory or self-contradictory due to the lack of uniform experimental protocols and standards.^{8,21,22} Furthermore, a ML model can make accurate predictions when trained on large and chemically diverse data sets.⁶⁷ However, currently it is very difficult to generate NM libraries with sufficient structural diversity to support ML modeling in a short time by traditional methods.

The issues with batch-to-batch variations have driven the rapid development of new technologies that can synthesize NMs in a high-throughput and reproducible way.^{68,69} Parallel synthesis methods can generate NM collections on a relatively large scale. This approach involves rapid conjugation of small molecules^{70–72} to NMs, or polymer synthesis^{73–76} in a parallel format. Using parallel synthesis, a collection of 146 magnetic nanoparticles (NPs) decorated with different small organic molecules was successfully created (Figure 3A).⁷⁰ These NPs were synthesized by reactions of diverse small molecules with primary amino groups on the core surface. The functionalized NPs provided candidates with high binding specificities for

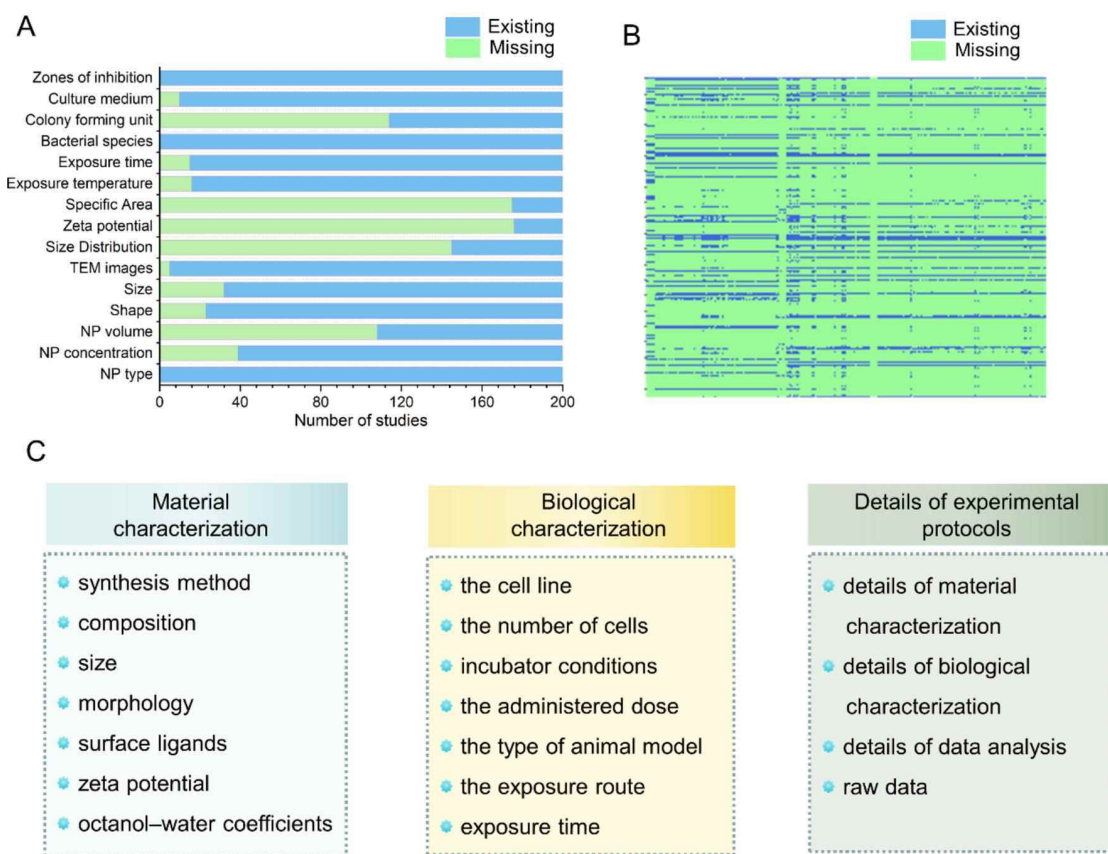


Figure 4. Low-quality data are resulting from missing values. (A) Some critical data (e.g., NP concentration, size distribution, and zeta potential) was not present in the nano-antimicrobial assays. (B) Many responses in this profile were shown as missing data in the bioprofile of 2203 Tox21 compounds against 300 PubChem assays. (C) A recommendation on the critical data required for nanotoxicity assays.

mammalian pancreatic cancer, endothelial, and macrophage-like cells. Similarly, a collection of 50 surface-modified NPs with different core materials was generated by parallel synthesis and modeled by ML methods.^{71,77,78} These NPs were synthesized by a dehydration condensation reaction of commonly used molecules/functional groups (e.g., $-\text{COOH}$ and fluorescein isothiocyanate) with polymers (e.g., cross-linked dextran and poly(vinyl alcohol)) coated on the core surface. Although parallel synthesis has generated several widely used nanotoxicity data sets for ML modeling, this approach still has several limitations, such as the lack of diversity in final NM products due to the simple chemical coupling reactions used.

Using combinatorial chemistry, large libraries can be constructed quickly from a few compounds with high diversity. This approach allows modification of many NM structural elements, such as the type of organic functional groups on surface ligands, core materials, size, hydrophobicity, electric charge, and redox property. Libraries were made by combining all of these diversity elements^{81–87} or by changing a single physicochemical property while other properties were fixed.^{88–90} The application of the nanocombinatorial chemistry strategy was originally reported in 2008 when a library comprising 80 multiwalled carbon nanotubes (MWCNTs) modified by small molecules was synthesized and screened against different biological assays (Figure 3B).⁷⁹ In the next few years, several gold NP (GNP) libraries were also synthesized by the nanocombinatorial approach.^{82,88–91} In a recent study, a comprehensive library of 36 NPs was created by combining

three core materials (Au, Pd, and Pt), two sizes (26 and 6 nm), and six types of small molecules with different hydrophobicities.⁸⁷ After rigorous characterization and biological testing, high-quality data from these NP libraries have been widely used for ML modeling and molecular simulations.^{24,92–94}

Microfluidic reactors provide a novel approach to NP synthesis, as they have the advantage of small reagent volumes, rapid reaction time, and improved control over mass and heat transfer compared to conventional synthesis methods.⁹⁵ This technique has enabled size-, shape-, and composition-controlled synthesis of various NPs. As shown in Figure 3C,⁸⁰ a droplet-reactor system was developed to automatically synthesize noble-metal NPs with well-controlled shapes and uniform sizes (Figure 3D).⁸⁰ For more complex NMs or those that do not lend themselves to microfluidic reactors, robotic high-throughput batch synthesis is also a possibility for generating broad, reproducible libraries of NMs.⁹⁶ Furthermore, the AI-based decision-making strategies have recently been used to create self-optimizing or autonomous NP synthesis platforms.^{97,98} Quantum dots (QDs),^{98–101} metal–organic nanocapsules,¹⁰² perovskite nanocrystals,^{103,104} GNPs,^{105,106} silver NPs (SNPs),¹⁰⁷ and CNTs¹⁰⁸ have been synthesized this way. AI algorithms can efficiently explore experimental space involving reaction time, temperature, pressure, and catalysts to find the optimum conditions. For instance, using circular dichroism to monitor the target quantity, a reinforcement-learning-based AI platform successfully discovered inorganic optically active CsPbBr₃ perovskite nanocrystals from a 2-dimensional param-

eter space within 250 experimental loops.¹⁰⁴ Recently, an AI-guided microfluidic reactor was developed to design high-quality inorganic lead halide perovskite QDs with desired peak emission energy.¹⁰¹ Such self-optimizing microfluidic reactors hold considerable promise for further accelerating the synthesis of NMs. Technologies such as parallel synthesis, nanocombinatorial chemistry, and AI-assisted microfluidic reactors will make a significant progress toward better-quality NMs compared to the traditional one-at-a-time approach.

2.2.3. Rigorous Characterization of NMs to Guarantee Data Quality. The completeness of nanotoxicity data is an essential factor when considering data quality. However, it cannot be optimistic when considering the completeness of current nanotoxicity data.^{109,110} For example, 200 antibacterial-related papers were retrieved from a large pool of research papers and used to investigate nanotoxicity data completeness. As shown in Figure 4A, critical physicochemical properties (e.g., size, shape, and zeta potential) and experimental conditions (e.g., the exposure time, dosage, and temperature) are not fully presented in these studies. Data incompleteness (or missing data) is a relatively common problem in almost all research areas. For example, a bioprofile of 2203 Tox21 compounds against 300 PubChem assays is shown in Figure 4B.¹¹¹ There are more than 600,000 data points, but most assays contain missing data. Moreover, the amount of missing data increased with the complexity of toxicity end points.⁴⁶ Although imputation methods have been designed to solve this problem, information loss cannot be eliminated, which reduces the prediction power of ML models.^{112,113} To improve this situation in the future, it is necessary to provide a detailed characterization of NMs' physicochemical properties and sufficient information about the testing protocols (the so-called metadata). A previous study listed a minimum set of information on the properties of NMs used for toxicity evaluation, and this information covered some fundamental physicochemical properties such as NM size, zeta potential, and water solubility.⁸ Recently, two EU-funded projects, NANoREG and GRACIOUS, also listed a minimum information checklist for the physicochemical properties of NMs.¹⁰⁹ Furthermore, some essential data about the toxicity or bioactivity testing methods were also included in each study.

In a recent study, a preliminary standard for reporting nanobio data was proposed.⁵¹ It suggested that nanotoxicity data should include three main components: materials characterization, biological characterization, and details of experimental protocols (Figure 4C). Adequate characterization of NMs is essential for assuring the quality of NMs and reliable nanotoxicity data.¹¹⁴ Citing Whitesides, "When things are large, they are what they are. When they are small, it's a different game: they are what our measurements make them."¹¹⁵ Several physicochemical properties, including composition, size, morphology, surface ligands, zeta potential, and octanol–water partition coefficients, are known to be important for nanotoxicity and should be rigorously characterized. Due to their crucial role in determining nanotoxicity, NM size and size distribution should be comprehensively measured. Furthermore, researchers should measure the NM size immediately after its synthesis and characterize the changes in size during storage or after interaction with biological media.¹¹⁶ The cell seeding details (e.g., the cell line, the number of cells, and the incubator conditions) should be provided when reporting cell culture assays. For *in vivo* experiments, the animal exposure details (e.g., the administered dose, type of animal model, exposure route, and time) should be recorded.¹¹⁷ Besides reporting the

experimental details of material and biological characterization, the raw data and details of data analysis should also be provided.

2.2.4. Quality Assays for Toxicity Testing of NMs. While synthesizing high-quality NMs quickly is essential, the materials must also be characterized rapidly and reliably. Efficient methods for fast toxicity testing and hazard ranking of synthesized NMs are equally important.^{118,119} High-throughput screening (HTS), developed by the pharmaceutical industry several decades ago, is an essential tool for testing thousands to millions of chemicals for biological activities in short periods. For example, while the number of compounds in PubChem increased from 19.3 million to 110.6 million between 2008 and 2021, the number of bioassays rose from 1197 to 1.93 million during the same period.^{120,121} Tox21 is a significant HTS effort in toxicology that aims to develop more efficient approaches for fast evaluation of the toxicity of compounds. As of 2018, the Tox21 program generated over 120 million data entries for approximately 8500 chemicals. These toxicity data generated from HTS have been widely used to train ML models.

Although the scale of nanotoxicity data is not comparable to small molecule toxicity data generated by Tox21, some HTS assays have proven valuable for generating nanotoxicity data for modeling. These HTS assays employ isolated molecular targets,^{85,91} cell cultures,^{70,122} and even whole organisms^{123,124} and are mainly performed on 24-, 96-, 384-, or 1536-well plates to rapidly generate NMs dose–response data. For instance, 384-well plates were used to evaluate the cytotoxicity of 50 NPs at four different doses in four cell types, using four physiological assays such as apoptosis, mitochondrial potential, and ATP content.⁷¹ Another critical HTS study came from screening the potential hazards of metal oxide NPs (MONPs) according to the multitier oxidative stresses.¹²⁵

In addition to the screening of cells, whole organisms (e.g., *Caenorhabditis elegans*,¹²⁶ *Daphnia magna*,¹²⁷ and Zebrafish embryo¹²⁸) have been commonly used in HTS platforms. In recent years, zebrafish has proven to be an excellent *in vivo* model because of its high fecundity, embryo transparency, fast and well-characterized development, low cost, gene manipulation accessibility, and short reproduction time.^{128,129} In summary, these screening assays not only provide copious nanotoxicity data but also allow the creation of ML models that consider more experimental variables (e.g., exposure times, dose concentrations, and biological responses) in addition to the physicochemical properties of NMs in the modeling procedure and mechanism extrapolations.¹¹⁹

2.2.5. Improving Nanotoxicity Data Quality by Standardization. The lack of standardization in assessing NM's toxicity impedes the effective use of nanotoxicity data. It causes inconsistencies in toxicity end points and data reporting formats. Several researchers have suggested the establishment of experimental standards.^{18,50,109} High-quality data can only be generated with reliable experimental protocols and universal data reporting formats.

Several cytotoxicity indicators, such as tetrazolium, LDH (lactate dehydrogenase) activity, and CellTiter-Glo assays, are widely used to determine the effects on cell viability induced by NMs, resulting in highly variable values. Meta-analysis of 93 peer-reviewed research articles showed that LDH assays usually result in lower cell viability values compared to tetrazolium-based assays.¹³⁰ This is because the former is a cell death assay and the latter is a live cell assay. Inconsistent data from different testing methods may result in misleading or erroneous conclusions when using these data for computational modeling.

Furthermore, inappropriate experimental methods may cause both false positive and false negative results. For example, previous studies have demonstrated that MTT assays will produce erroneous results when used to evaluate the viability of cells treated with carbon nanotubes.^{131,132} Furthermore, the exposure concentrations are often expressed in different units such as mg/L, mM, and pmol/cell. This requires additional work to convert the results into a single consistent unit for computational modeling.^{133,134} To address this challenge, there is an urgent need to standardize the testing and reporting protocols for physicochemical properties and toxicity end points. The standardized protocols for the toxicology assessment of NMs have been reflected in a variety of guidelines, including the International Organization for Standardization, the Organization for Economic Cooperation and Development (OECD), and the American Society for Testing and Materials.^{53,135–137} Recently, the standardized test methods, guidelines for assessing physicochemical properties, and *in vitro* nanotoxicity methods have been summarized.⁵³ These proposed standards have taken an essential step toward improved data comparability and availability from different sources. However, proposing a standard is insufficient to improve the quality of nanotoxicity data alone. To be useful, standards should be widely agreed on and strictly enforced by all stakeholders involved in manufacturing engineered NMs, researchers, regulatory agencies, and publishers.

2.2.6. Improving Data Availability with Nanotoxicity Quantification. We generally expected any machine learning models to provide a definite toxicity value or the risk level of the predicted NMs. This makes data labeling one of the most important steps of ML model development. Data labeling accounted for much of the time during data preparation, especially for imaging data.^{49,138} For nanotoxicology, data labeling can be regarded as assigning a category or a numerical value to the nanotoxicity end point. Herein, we call this process nanotoxicity quantification. Nanotoxicity quantification helps supervised ML models learn the relationships between the structural features (or physicochemical properties) of NMs and their corresponding toxicity labels.

Typically, nanotoxicity quantification involves several critical steps, such as defining the toxicity end point, determining the toxicity test method, unifying the quantitative indicator, and calculating the toxicity value. Over the past two decades, many biological systems (e.g., biomacromolecules, cells, plants, and mammals) have been used to evaluate the different adverse effects of NMs, such as cell uptake, cytotoxicity, immunological response, oxidative stress, and genotoxicity. Herein, we take the cytotoxicity and cell uptake as examples to clarify the details of nanotoxicity quantification. Different cell line models derived from organs are widely used for nanotoxicology to evaluate the toxicity and potency of NMs in the human body. The viability of cells can be determined by methods like the CellTiter-Glo assay, which would provide the relative number of living cells under different treatments. The generated cell viability (the relative percentage) can be directly used as a target quantity for ML¹³⁰ or converted into the half-maximal inhibitory concentration (IC₅₀).¹³⁴ The cell uptake can be quantitatively determined by the concentration of NMs in the cells. However, these concentration units (e.g., number·cell⁻¹, nm²·cell⁻¹, and g·cell⁻¹) always varied with different laboratories. Therefore, the units should be unified before being fed into the ML models. Here, we emphasize the importance of nanotoxicity quantification and provide two examples for reference. Still, the most

important thing is to formulate standards that can be recognized and widely accepted for each toxicity end point. Such standards are also helpful for assembling nanotoxicity data of different sources into machine-readable data sets.

2.3. Collection and Curation of Nanotoxicity Data

Nanotoxicity data generated from toxicity studies are usually stored as documents, such as scientific publications, dissertations, patents, and conference reports. Thus, currently, text documents are the major sources of nanotoxicity data. Compiling information from these unstructured data sources mainly involves document retrieval and data extraction (i.e., text mining). Using different combinations of suitable keywords, a useful collection of documents can be acquired from academic research databases such as Web of Science and Google Scholar. Extracting nanotoxicity data from text documents involves complex tasks, such as chemical structure recognition, image processing, and extracting experimental conditions and toxicity entities. Although the need for unstructured data processing has resulted in the development of novel text mining technologies,^{139–141} most current nanotoxicity data still came from very time- and resource-intensive manual extraction. For example, when extracting information on cellular toxicity of cadmium-containing QDs, it is estimated that more than 1,200 person-hours of work is needed.¹³⁴ Currently, the use of text mining techniques is hampered by the disparate and heterogeneous data sources that provide data in different formats (e.g., patents, technical reports, and scientific publications). Furthermore, the lack of standardization in data reporting also limits the machine's readability, for example, no standard language, terms, tables, or data storage location (in the main text or Supporting Information).⁵⁰ Until recently, copyright and paywall issues also limited large-scale text mining. However, with increasing open access literature, this problem is abating.

Previous efforts have been dedicated to recovering machine-readable data sets from the massive scientific literature. These data sets involved the toxicity of NMs toward multiple biological systems, such as biomacromolecules,¹⁴² microorganisms,¹⁴³ cells,^{130,133,134,144,145} plants,¹⁴⁶ and mammals.¹⁴⁷ Among this, cell viability was the most widely used toxicity end point due to the HTS approach's advantage and unified target quantity. For instance, a recent study extracted over 3000 cell viability data points from 517 publications.¹³³ Recently, the rapid development of large language models, such as BERT¹⁴⁸ and GPT-3,¹⁴⁹ also provided promising tools for data extraction. Massive quantities of textual data are utilized for training these models, enabling them to discern patterns and relationships between entities within language. Researchers have used these large language models to automatically generate machine-readable data sets for ML applications in material science.^{150–153} For instance, a materials-aware language model, MatSciBERT, was developed to perform several tasks including material entity recognition, relation classification, and abstract classification.¹⁵¹ These successful examples would provide guiding principles for automatically extracting nanotoxicity data from massive scientific literature.

Before assimilating into existing data sets or online databases, nanotoxicity data must be curated because the raw data often contains errors. A statistical analysis of the World of Molecular Bioactivity database demonstrated an average of approximately two errors per medicinal chemistry publication. The overall error rate for compounds was as high as 8%, including incorrect chemical structures, replicates, and wrong biological activ-

Table 1. Selected Publicly Available Nanotoxicity Databases

| Database | Description | Web Access |
|--------------|---|---|
| eNanoMapper | Proposes a computational infrastructure for nanotoxicity data sharing and management based on standardized ontologies. Contains data from multiple nanosafety projects and scientific literatures. | https://search.data.enanomapper.net/ |
| NBI | Serves as a repository for annotated data of NM physicochemical properties and their biological interactions. Contains computational tools to predict the potential hazards of unsynthesized NMs. | http://nbi.oregonstate.edu/ |
| Nanowerk | A leading nanotechnology and nanoscience portal that delivers useful, entertaining, and cutting-edge information from all nanothings. Contains information about more than 5,800 manufactured NMs from over 175 suppliers worldwide. | https://www.nanowerk.com/ |
| caNanoLab | A data sharing portal designed to facilitate the use of nanotechnology in biomedicine. | https://cananolab.nci.nih.gov/ |
| S2NANO | Provides measurement protocols, curated data sets, predictive models, and decision support tools for safe and sustainable nanotechnology. | http://portals.s2nano.org/ |
| NanoCommons | An infrastructure developed as part of the H2020 NanoCommons and NanoSolveIT projects. Contains physicochemical characterization of NMs and their toxicological and omics data. | https://ssl.biomax.de/nanocommons/ |
| NIL | Provides information about health and safety-associated properties of NMs to help workers, customers, and researchers. | http://nanoparticlelibrary.net/ |
| DaNa | A web-resource funded by German Federal Ministry of Education and Research to provide information about the risk of NMs to humans and environment. | https://nanopartikel.info/en/ |
| NanoMILE | Intends to establish a fundamental understanding of the mechanism of NMs interacting with environment and living systems. Contains data on NM structure and transformation information and their high-throughput screening toxicity. | http://nanomile.eu-vri.eu/ |
| ESL | Includes the exposure scenario for uses of NMs by workers and consumers, together with the exposure measurements data. | http://marina.iom-world.co.uk/ |
| nanoHUB | Provides a variety of resources for computational nanotechnology research, education, and collaboration. | https://nanohub.org/ |
| Nanodatabase | Includes information about nanoparticle manufacturers and their risk category. | https://nanodb.dk/ |
| PubVINAS | Contains electronic files of nanostructures and their physicochemical property and toxicity data. | http://www.pubvinas.com/ |

ities.¹⁵⁴ Poor quality training data can cause bias models and may lead to significant errors. According to a recent report from Gartner research (<https://www.gartner.com/>), poor data quality incurred huge costs each year.¹⁵⁵ Noncurated data have been characterized as five I's (i.e., incomplete, incompatible, inaccurate, imprecise, and irreproducible). Therefore, a workflow for curating chemical and biological data must deal with incorrect chemical structure annotations and errors in experimental tests.¹⁵⁶ The central role of data curation in nanoinformatics (i.e., methods and software tools for modeling NMs and their properties and interactions with biological entities) is well recognized. It is the focus of the nanomaterial data curation initiative.¹⁵⁷ A general workflow of nanotoxicity data curation was also proposed by the nanomaterial data curation initiative that includes four steps: assessing the data quality and completeness; contacting the authors for any missing and dubious data; formatting data into a database; reviewing and releasing the curated data in the public domain. To reiterate, data curation is critical to ensure data quality that supports and facilitates the application of AI to nanotoxicology. Previous studies also demonstrated that rigorous curation of nanotoxicity data would improve model performance.^{109,110,112,113} Clearly, creating high-quality data requires joint efforts from experimentalists who generate the data and modelers who curate and use the data.

2.4. Currently Available Nanotoxicity Databases

Databases store and manage data in a unified format, allowing stakeholders to search and use them freely. For example, PubChem and Protein Data Bank (PDB) are widely used in the scientific research community and have significantly impacted broad areas of science.^{120,158} In PubChem, structural information on chemicals, their physicochemical properties, and biological activities are stored in several electronic formats such as structure-data files (SDF) and comma-separated value

(CSV) formats. Since its launch in 2004, PubChem has continuously served the toxicology, medicinal chemistry, cheminformatics, and bioinformatics communities.^{46,159} The Protein Data Bank database provides the three-dimensional structures (e.g., atomic coordinates and sequences of peptide chains) of large biological molecules (e.g., nucleic acids and proteins) stored in PDB file format. This structure information is crucial in structural biology, bioinformatics, and computational biology.^{160,161}

Seminal efforts to create nanotoxicity databases have been made over the past decade. Table 1 summarizes the databases that store biological data for NMs. Almost all of them contain at least some basic nanotoxicity information, such as the physicochemical properties of NMs and the toxicity testing results. As early as 2006, the U.S. National Cancer Institute created a data portal, the cancer Nanotechnology Laboratory (caNanoLab), to facilitate sharing of nanomedicine data and advance nanotechnology's use in biomedicine.¹⁶² caNanoLab allows submission and retrieval of NM characterization data, *in vitro* and *in vivo* assays, and associated nanotechnology protocols and publications in a secure environment. In Europe, more than 50 nanoEHS (Nano Environmental Health and Safety)-related projects have been funded over the last ten years or so, such as the funding programs FP6 (2002–2006), FP7 (2007–2013), and Horizon 2020 (2014–2020).¹⁶³ These projects resulted in a comprehensive platform for coordinating nanosafety research in Europe, the EU-NanoSafety Cluster. Under this platform, several significant nanotoxicity databases have also been developed, such as eNanoMapper,¹⁶⁴ caLIBRate, GRACIOUS,¹⁶⁵ and NanoReg2. Recently, most of these nanotoxicity databases have been merged into the eNanoMapper database and have adopted the harmonized infrastructure for data collection, curation, and sharing. Currently, the eNanoMapper database consists of three main parts, eNanoMapper ontology,

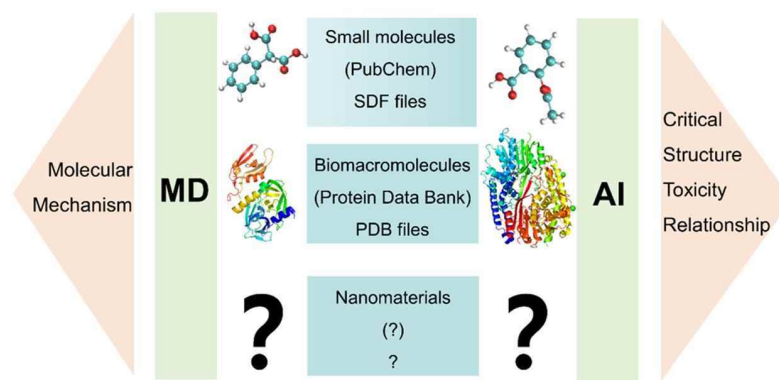


Figure 5. Critical role of nanostructure annotation for computational modeling. Nanostructure annotation refers to storing physicochemical properties and structural features of NMs in machine-readable files, such as SDF and PDB files. AI modeling relies on nanostructure annotation to generate relevant nanodescriptors. Molecular simulation of nanobio interactions also relies on the accurate annotation of three-dimensional nanostructures. Reproduced with permission from ref 33. Copyright 2022 Springer Nature.

nanotoxicity data, and modeling tools.^{164,166–168} Apart from these web-based databases, several specialized nanotoxicity databases have been published in the Nature journal *Scientific Data* (<https://www.nature.com/sdata/>), such as the scanning transmission electron microscopy database, the high-throughput imaging database of NM effects, and the NaKnowBase database on NM actions in environment or biological media.^{169–172}

Although these nanotoxicity databases have made promising progress in sharing published data, most nanotoxicity databases were rarely used by the nanotoxicology community.²⁰ In other words, current nanotoxicity databases are unsuitable for computational modeling. One of the main reasons is the way that they were curated. The NM records in these databases, such as size, composition, surface ligands, and nanotoxicity, exist as text outputs extracted directly from the literature and ignore nanostructure annotations crucial for computational modeling (Figure 5). Without nanostructure annotations, diverse theoretical nanodescriptors cannot be generated for ML modeling (see Section 3)³³ and nanobio interfacial systems cannot be constructed for molecular simulations (see Section 4).⁴¹ Other analyses such as nanostructure studies and visualization cannot be performed. Therefore, it is urgent to develop well-structured and publicly accessible nanotoxicity databases that can be easily interrogated—so-called *model-friendly*.

2.5. Construction of Model-Friendly Databases under the FAIR Principles

In a recent proof-of-concept study, a nanotoxicity database (PubVINAs) was constructed to improve data availability for computational modeling.²⁴ The core of the PubVINAs database is the electronic files that contain nanostructure information such as material type, size, atomic coordinates, and chemical bonds. These machine-readable files in PubVINAs can be used directly to train ML models and for molecular simulations. These files potentiate data sharing and management by simplifying the data deposition requirements. They can also be used to visualize three-dimensional nanostructures, providing direct structure features of the relevant surface chemistry and physicochemical properties of NMs. Currently, there are still several critical technical challenges that limit the further expansion and usage of nanotoxicity databases, including the lack of standardized reporting systems or harmonized ontologies, the paucity of unique identifiers for NMs, the scarcity of a standardized web service for data retrieval, and

varying levels of data quality and completeness due to source variations.^{18,26} Other challenges, such as intellectual property protection, lack of project funding, and standardized data deposition, also hamper the integration of nanotoxicity data into model-friendly databases.¹⁷³ Compared with well-known databases such as PubChem and Protein Data Bank, the current nanotoxicity databases are still nascent.

In 2016, a concise and measurable set of FAIR principles was defined and published, aiming to improve the findability, accessibility, interoperability, and reuse of scientific data.¹⁷⁴ The FAIR principles emphasize enhancing the ability of computational systems to automatically find and use data with little or no human interventions. Recent studies suggested that the FAIR principles could resolve the issues with nanotoxicity databases and substantially improve the reuse of nanotoxicity data.²³ Under the FAIR principles, concerted efforts from individuals and groups are urgently needed to progress on data integration and enhance the reuse of nanotoxicity data. Actions to be taken should include, but not be limited to, the following:

- (1) establishing a universal scheme and infrastructure for nanotoxicity data sharing and integration,¹⁷⁵
- (2) assigning a unique identifier for each NM, like the CAS (Chemical Abstracts Service) registry number, in which the NInChI is a good candidate for this identifier,²⁷
- (3) building user-friendly platforms for data searches and retrieval by ontological terms,
- (4) creating quantifiable methods or systems for evaluating data quality and completeness,¹¹⁰
- (5) reinforcing financial support and international collaborations,
- (6) strengthening data-sharing policies to support the longevity and sustainability of nanotoxicity databases,
- (7) simplifying the data deposition formats with standardized electronic files, such as the PDB file.²⁴

Furthermore, an ideal nanotoxicity data cycle needs to resolve the challenges in data flow, including initial generation, collection, curation, storage, management, and computational modeling. As these challenges are addressed, the resulting well-structured nanotoxicity databases will accelerate the use of ML to nanotoxicology research. Persistent efforts from all stakeholders (e.g., researchers, industry, funding agencies, regulatory authorities, and professional data publishers) are needed to consolidate and accelerate this process.

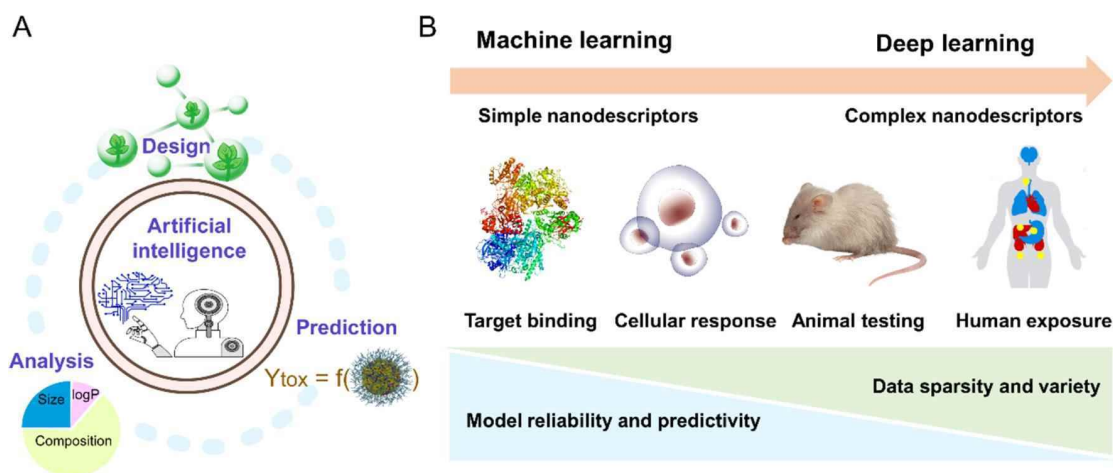


Figure 6. Application of AI approaches to unravel the QNTR. (A) AI has three essential roles in nanotoxicology, i.e., prediction of nanotoxicity, design of novel NMs, and analysis of toxicity mechanism. (B) The applications of AI in nanotoxicology are typically data-dependent. With the advancement of nanotoxicology research, the sparsity and variety of nanotoxicity data have increased and brought multiple challenges for AI modeling.

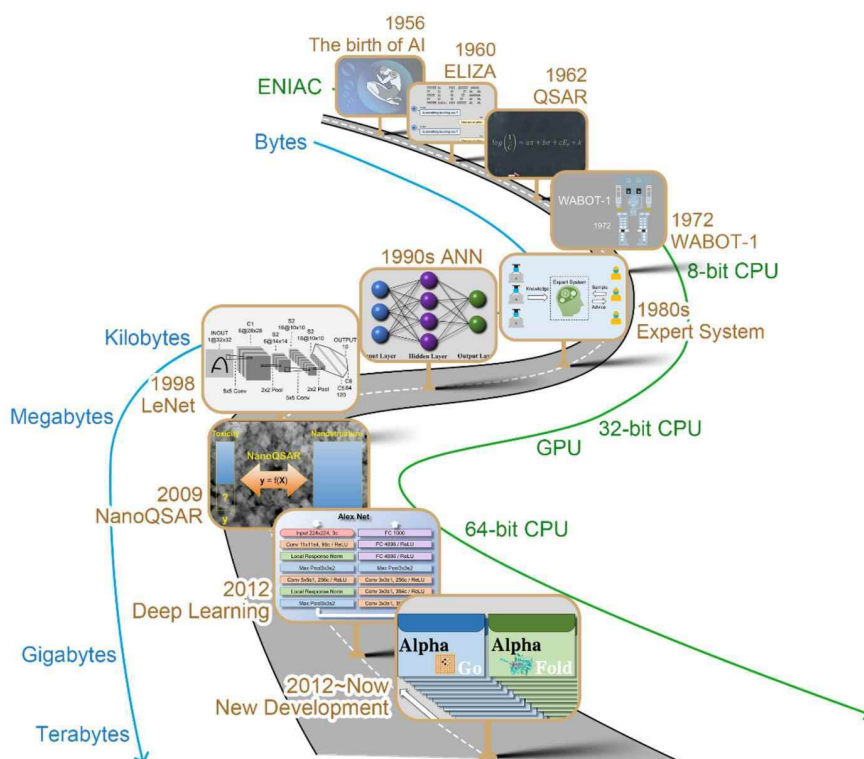


Figure 7. A brief history of AI. This field experienced several booms and busts over the past 60+ years. With the development of computer power and the rise of big data in the past decade, AI techniques are profoundly influencing all fields of human society, such as medical health and drug discovery. LeNet: Reproduced with permission from ref 185. Copyright 1998 Institute of Electrical and Electronics Engineers. NanoQSAR: Reproduced with permission from ref 192. Copyright 2009 John Wiley and Sons.

3. UNRAVELING QUANTITATIVE NANOSTRUCTURE–TOXICITY RELATIONSHIPS USING ARTIFICIAL INTELLIGENCE

The complex interplay between NM properties makes identifying the features driving toxicity challenging. AI, particularly ML methods, provides relatively simple and effective ways to map the complex relationships between NM properties and the biological responses they invoke. For nanotoxicology, AI must achieve the following goals: extract the critical QNTR information and predict the potential toxicity of new NMs;

design NMs with better biocompatibility or desirable activities; and elucidate underlying toxicity mechanisms (Figure 6A). The application of ML approaches has been advanced to a new stage with the newly developed deep learning approaches (Figure 6B). In the early stage of ML studies of nanotoxicity, multiple linear regression (MLR) models were used to model the physicochemical properties or simple toxicity end points of a few NMs (typically less than 20).^{176,177} Even with few simple nanodescriptors (e.g., size and composition of NMs), those basic ML models could provide acceptable model performance

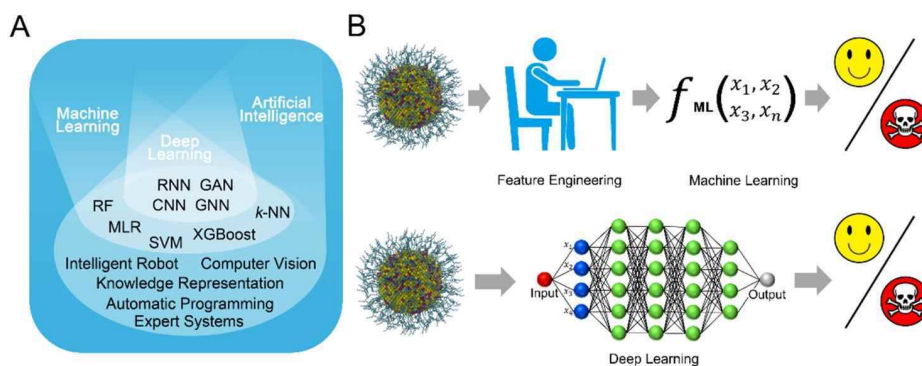


Figure 8. Machine learning and deep learning. (A) The relationships between AI, ML, and deep learning. ML and deep learning are subsets of AI, and they are also the main approaches to realize AI at present. (B) One of the main differences between traditional ML and deep learning. Traditional ML needs to generate diverse nanodescriptors before constructing prediction models, while deep learning can directly extract features from raw data. RNN, Recurrent Neural Networks; GAN, Generative Adversarial Networks; CNN, Convolutional Neural Networks; GNN, Graph Neural Networks; RF, Random Forest; MLR, Multiple Linear Regression; SVM, Support Vector Machine; XGBoost, eXtreme Gradient Boosting; *k*NN, *k*-Nearest Neighbors.

(Figure 6B). However, NM behavior in the environment and in biological media involving adsorption of biomolecules and pollutants, nanosafety at the ecosystem level, and human exposure risk assessment are substantially more complex. Most nanotoxicity end points cannot be successfully modeled by simple nanodescriptors or linear ML methods (Figure 6B). Critical issues also exist in dealing with new types of nanotoxicity data (e.g., biological images and omics data)^{127,178} and encoding complex nanostructures as numerical variables.^{179,180} Contemporary nanotoxicity data commonly requires advanced ML methods and novel nanodescriptors.

In this section, we discuss the application of AI to nanotoxicology following the basic AI workflow. We first briefly review the history and basic concepts of AI methods and the status of their application to nanotoxicology. We then discuss the significance and challenges of nanospecific descriptors (nanodescriptors), and how to develop new descriptors with more relevant features. We also summarize recent developments in AI applied to nanotoxicity modeling at the level of NM–biomolecule interactions, cellular responses, aquatic toxicity, and mammalian toxicity. The domains of applicability of ML models, their interpretation, and use to guide NM design were also discussed in this section.

3.1. Application of AI to Nanotoxicology

3.1.1. A Brief History of AI. AI is defined by Kurzweil et al.,¹⁸¹ as “the art of creating machines that perform functions that require intelligence when performed by people”. AI research aims to create algorithms that think like humans and simulate human behaviors, including perception, learning, reasoning, and decision-making. The term AI is widely credited to John McCarthy at the Dartmouth Conference in 1956.¹⁸² This research area experienced several boom and bust cycles in the past 60 years (Figure 7). Initially (1956–1974), AI was driven by pure optimization, and logical reasoning and heuristic searching were introduced for solving mathematical problems. The symbolic method allowed the computer to communicate in natural language (e.g., dialogue system “ELIZA”¹⁸³). However, most achievements were considered as “toy problems”, and AI research fell out of favor (1974–1980). Expert systems drove the second boom of AI (1980–1987). One successful example is XCON, which was used to automatically select the computer system components based on the customer’s requirements. This

expert system saved one company 40 million dollars annually by 1986.¹⁸⁴ However, expert systems cannot automatically learn from data. During this period, the backpropagation algorithm was also proposed, and LeCun developed the first modern convolutional neural network, LeNet, in 1998.¹⁸⁵ At that time, due to the lack of massive training data sets and high-performance computers, deep neural networks had no advantage compared to simpler shallow neural networks. The rise of big data and the spectacular increase in computer power resulted in another boom in AI starting in the 2010s. Here, AI research focused on deriving features and information from massive data sets from a wide range of fields, such as drug design,^{186,187} materials discovery,^{14,188} and medical diagnosis.^{189,190} In 2021, deep learning solved a grand challenge in biology—predicting the 3D structure of proteins from the sequence. DeepMind’s open-sourced AlphaFold2 was trained on 170,000 protein structures from the Research Collaboratory for Structural Biology (www.rcsb.org) using novel neural network architectures.¹⁶¹ It could predict protein structures to near experimental accuracy based solely on their amino acid sequences. A Massachusetts Institute of Technology (MIT) research team reported another important AI platform for screening potential antibiotics from more than 107 million molecules.¹⁸⁶ One of the hits exhibited broad-spectrum antibiotic activity, including strains resistant to all known antibiotics. These achievements exemplified the importance of AI in accelerating scientific discoveries using big data. Despite the enormous advances, the current field is still mostly restricted to narrow AI (excellence at a single task). However, there are promising embryonic results of general AI (e.g., GPT-3¹⁴⁹ and LaMDA¹⁹¹), and human-superior (super AI) intelligence in machines is mooted but still relatively far from ultimate success.

3.1.2. Traditional Machine Learning and Deep Learning. AI is the superset of tasks that demonstrate characteristics of human intelligence, while ML is a subset of AI that assesses data, analyzes trends, and generates insights. Deep learning is a subset of ML. There is a broad set of techniques in AI, such as computer vision,¹⁹³ expert systems,¹⁹⁴ natural language processing,¹⁹⁵ automation and robotics,¹⁹⁶ evolutionary algorithms, and ML.¹⁹⁷ Currently, ML¹⁹⁷ and deep learning¹⁹⁸ are the most widely used types of AI for chemical and biological problems (Figure 8A). ML algorithms allow computers to learn

automatically from data without explicit programming. The key difference between ML and deep learning is the algorithm's complexity. For example, traditional shallow neural networks typically have a single hidden layer in which the computation is performed. Deep neural networks have multiple layers and many more neurons, relying on regularization methods to control the complexity of models and avoid overfitting. Differences also arise from the data types used to train them (Figure 8B). Traditional ML methods need domain knowledge to transform raw data into suitable features (i.e., featurization) before creating a predictive model. In contrast, deep learning models can perform automatic feature generation from raw data. Conspicuously, deep learning requires very large data sets, and its prediction accuracy increases with training data set size. Thus, deep learning models trained on small data sets commonly perform worse than traditional ML models.^{24,32,199} Furthermore, the application of deep learning methods to nanotoxicology has also been hampered by their lack of interpretability. There are many types of ML methods such as *k*-nearest neighbors (*k*NN),²⁰⁰ random forests (RF)²⁰¹ and other tree-based methods, support vector machines (SVM),²⁰² and their sparse Bayesian version relevance vector machines (RVM), shallow neural networks, and their Bayesian regularized version. Deep learning methods include graph neural networks (GNN),²⁰³ convolutional neural networks (CNN),²⁰⁴ and recurrent neural networks (RNN).²⁰⁵ Generally, ML models can be classified as four main types—supervised, unsupervised, semisupervised, and reinforcement learning—depending on whether the data are labeled or unlabeled. Supervised ML models can be classified or regression based on whether the data are classes or continuous. ML is ideally suited for the prediction of adverse effects of NMs and can be trained on diverse types of data such as images and transcriptomics data.

3.1.3. The Basic Steps in ML. The development of ML models involves six steps: data collection, feature generation, feature selection, training a suitable model, model validation, model interpretation, and experimental validation.¹⁴ Training data are collated from experiments, databases, publications, and patents. Data usually require preprocessing to handle replicates, errors, missing values, and unit and scale conversion (e.g., logarithmic transformation). Feature engineering or featurization uses domain knowledge to measure or calculate relevant features for modeled materials. For molecular and material science, these features are called descriptors that encode structures and properties of molecules and materials. Then the most relevant subset of features is selected during the model training. This can be done in many ways, such as forward addition/backward elimination. Efficient methods include sparsity-inducing L1 regression such as LASSO (Least Absolute Shrinkage and Selection Operator)²⁰⁶ and MLREM (Multiple Linear Regression with Expectation Maximization).⁷⁸ A suitable ML method needs to be selected. An ML model aims to predict unseen (new) data, which is the golden criterion for evaluating model performance. This is achieved by predicting a test set, which is partitioned from the initial data set and not used for model training. Ideally, the resulting model should be interpreted for molecular or mechanistic insight, although this is difficult in many cases. Interpretability depends on the selected features (some can be directly mapped onto structures, while many are arcane) and the correct application of feature importance methods.^{35,147} The model interpretation aims to identify which features primarily influence the target activity/toxicity. Model predictions and interpretations provide a

rationale for designing NMs with lower toxicity and higher desired activities/properties.

3.1.4. Application of ML to Nanotoxicology. Compared to other fields (e.g., drug discovery), the application of ML in nanotoxicology is still in its preliminary stage. Although QSAR modeling has been applied to predict the bioactivity of organic compounds for drug discovery since 1962,²⁰⁷ applying ML to nanotoxicity or nanobioactivity prediction was not realized until about a decade ago. In the early stages of nanotoxicology (e.g., before 2009), the commonly used modeling approaches were linear regression for physicochemical properties (e.g., water solubility,^{208,209} octanol–water partition coefficient,¹⁷⁶ and Young's modules²¹⁰) of NMs. In 2009, QNTR (or nano-QSAR) was presented.¹⁹² In 2010, ML approaches were used to predict the toxicity of NMs for the first time.⁷⁷ However, modeling toxicity or bioactivity of NMs was hindered by the diversity, complexity, and heterogeneity of nanostructures and the paucity of quality data sets. By December 2022, less than 400 ML modeling of nanotoxicology papers were published, according to the Web of Science. The application of ML to nanotoxicology thus lags behind other related fields due to lack of suitable nanotoxicity data. Although a large volume of unstructured data has been reported, it is not yet ideal for full exploitation by ML approaches.^{20,211} Consequently, most existing ML models were trained on small data sets, limiting their applicability, and their reliability and utility need to be validated by other data sets or external experiment testing. Furthermore, current ML models mainly focus on simple toxicity end points (e.g., cytotoxicity) under laboratory conditions and are far from directly evaluating essential human health risks of NMs.

3.2. Encoding NM Properties into Nanodescriptors

3.2.1. The Significance of Feature Engineering. Many NM features can be measured or calculated, but not all features are relevant to the biological properties or toxicities being modeled and predicted. The process of transforming NM structures/properties into mathematical features suitable for training ML models is called feature engineering or featurization.²¹² It is important to include relevant features and avoid irrelevant ones.²¹³ Expert knowledge can help select important features correlated with target properties. In many cases, it is essential to conduct feature selection to ensure that the features used to train models are sufficient to capture important physicochemical properties of NMs. There are several useful algorithms for choosing the most relevant features from a larger feature pool,^{214–216} such as weighted penalized logistic regression,²¹⁷ the evolutionary feature weighting approach,²¹⁸ RF-based methods,²¹⁹ and the interaction-weight-based algorithm.²²⁰ These methods remove irrelevant features (i.e., noise), avoid the risk of subjective bias in the choice of features (i.e., human errors), and result in models with better predictivity and higher interpretability.²²¹

Feature engineering, which is to select the correct features to represent NMs, plays a major role in determining the performance of a ML model. Many studies have shown that feature engineering has a much greater impact on the predictive accuracy of models than the ML algorithm.^{30,222} Mathematical features encoding NM properties comprise several different families. For instance, whole molecule/material/NM (global) properties (e.g., size, shape, zeta potential, molecular weight, lipophilicity) or properties of substructures (e.g., molecular fragments, surface functionality) are both useful. The core of

feature engineering is to find features (descriptors) that best represent the structures and properties of a molecule or NM. Descriptors can be calculated or experimentally measured, but the latter cannot be used when predicting the properties of new NMs not yet synthesized. There are commercial and open-source software packages, such as RDKit (www.rdkit.org), MOE,²²³ and Dragon²²⁴ for descriptor calculations. However, these packages were mostly designed for organic small molecules rather than NMs. In fact, the lack of nanospecific molecular descriptors is generally considered one of the major roadblocks to the application of ML in nanotoxicology.²⁵

Compared with organic small molecules, it is quite a challenge to generate meaningful nanodescriptors due to the complexity of NM morphology. They are not well-defined monodisperse entities like organic molecules but rather large particles that are mixtures of sizes, shapes, compositions, and surface chemistries. The incomplete characterization of nanostructures and properties is another barrier to generating optimal features. Another reason for nanodescriptor burden is that structures and physicochemical properties of NMs are greatly affected by environmental or biological media (e.g., agglomeration, dissolution, protein/natural organic matter adsorption). Despite these difficulties, useful nanodescriptors have been developed to encoding the structural and physicochemical properties of NMs, as discussed in the following sections.

3.2.2. Experimental Properties as Nanodescriptors.

Nanostructures can be represented by size, shape, composition, and surface area. These features have been widely used as nanodescriptors and can be acquired from transmission electron microscopy (TEM), atomic force microscopy (AFM), scanning electron microscopy (SEM), and dynamic light scattering (DLS). Accordingly, a set of software packages (e.g., Pebbles,²²⁵ ImageJ,²²⁶ and NanoXtract²²⁷) or ML algorithms^{228–230} were also developed to quantitatively analyze and extract the morphological features of NMs from the microscope images. For example, morphological nanodescriptors from the TEM images (i.e., size, surface area, curvature, and agglomeration) were sufficient to predict the toxicity profiles of some NMs.^{230,231}

Other physicochemical properties such as $\log P$,²³² zeta potential,¹³⁰ solubility,²³³ magnetic relaxivities,²³⁴ isoelectric point,²³⁵ and surface charge⁷⁸ can also be important nanodescriptors for ML modeling. Recently, nine physicochemical properties and cellular uptake were used to model and predict cytotoxicity induced by silica- and carbon-based NPs.²³² As physicochemical properties of NPs change when bound to biological corona molecules, the corona features can also be used as descriptors. For example, protein corona fingerprints (i.e., the relative abundance of adsorbed proteins) were used to predict the cellular interaction of GNPs and SNPs.⁷² Similarly, corona features were used to predict the *in vivo* fate (e.g., half-life and accumulation) of GNPs in a recent study.²³⁶ In addition, experimental conditions such as cell species and assay type can also be used as experimental descriptors. Using these nanodescriptors, ML models may elucidate the effects of experimental conditions on nanotoxicity.^{130,133,134} Using 24 qualitative and quantitative features, a RF model mapped the complex relationship between quantum dot physicochemical properties and experimental conditions (exposure concentration and time, cell lines) and measured cytotoxicity.¹³⁴ Recently, a Bayesian-network-based web application was developed to predict the cytotoxicity of QDs and elucidate toxicity mechanisms.¹³³

Although experimental nanodescriptors can represent properties of NMs, their generation is time-consuming and laborious. They are highly susceptible to changes in different experimental conditions so they can vary with and between laboratories. Conspicuously, it is not feasible to measure the experimental properties of new NMs that are not yet synthesized. In other words, for NM design, virtual screening cannot be performed using experimental nanodescriptors.

3.2.3. Nanodescriptors from Quantum Chemical Calculations. Compared to the disadvantages of experimental nanodescriptors, theoretically derived nanodescriptors can be generated faster and at lower cost. Computed nanodescriptors can also be generated for new NMs before they are synthesized. Nanodescriptors derived from quantum chemical calculations were some of the earliest theoretical descriptors used in QNTR studies. There are many open-source and commercial software packages, such as VASP (Vienna Ab initio Simulation Package) and Gaussian, that contain semiempirical or *ab initio* methods.^{237,238} They can provide information about geometric and electronic properties, such as energies of the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO). These have been widely used as features for nanotoxicity modeling.^{177,239–242} For example, an MLR model was established to describe the toxicity of 17 MONPs to *Escherichia coli* using the calculated quantum chemical descriptors. The ΔH_{Me+} (i.e., the enthalpy of formation of a gaseous cation having the same oxidation state as that in the metal oxide structure) was found to be useful to predict bacterial toxicity.¹⁷⁷ Similarly, quantum chemical descriptors and various supervised ML methods were also used to model the cytotoxicity of metal-doped TiO₂ NPs.²⁴² More recently, three types of descriptors from quantum chemical calculations, experimental characterization, and the periodic table properties were used to successfully train a ML model for inflammatory response of 30 MONPs.²⁴³

For large (i.e., size >5 nm) and polydisperse materials like NMs, quantum chemical calculations require large computing resources. In some studies, NMs were represented by unit cells or small-sized clusters to make the quantum chemical calculations tractable.^{241,244} However, quantum chemical nanodescriptors are only useful for idealized, pristine monodisperse NMs and are not applicable to account for distributions of sizes, shapes, and surface modifications. On the other hand, although quantum chemical descriptors can provide electronic properties of a molecule, the need for high computational resources limits their application in machine learning.²⁴⁵ Since many descriptors are derived from semiempirical methods and molecular fragments, accurate descriptors from *ab initio* methods are not a prerequisite for constructing a useful machine learning model. We should find a balance between improving model accuracy and the required computational resources when using quantum chemical descriptors. A previous study discussed the necessity of calculating quantum chemical descriptors at the DFT level.²⁴⁶ It was demonstrated that using semiempirical descriptors allowed QSAR models to obtain similar accuracy to DFT-based models. Therefore, we do not necessarily need to calculate *ab initio*-based descriptors if the machine learning models have achieved a high accuracy with semiempirical methods or even molecular fragments. Of course, the *ab initio*-based descriptors are desired if we do want to explore the effects of accurate electronic properties on chemical toxicity.

3.2.4. Structural Nanodescriptors from 1D Text Representations. Nanodescriptors derived from SMILES

(Simplified Molecular Input Line Entry System) strings were also one of the earliest computed features used in nanotoxicity or nanobioactivity modeling. In 2010, SMILES-based nano-descriptors for organic ligands decorating the surfaces of 109 fluorescent magnetic NPs with similar cores were used to train a ML model.⁷⁷ In other QNTR studies, SMILES strings were also employed to represent the surface^{78,247} and core^{248–250} composition of NPs (Figure 9A) and simple nanostructures

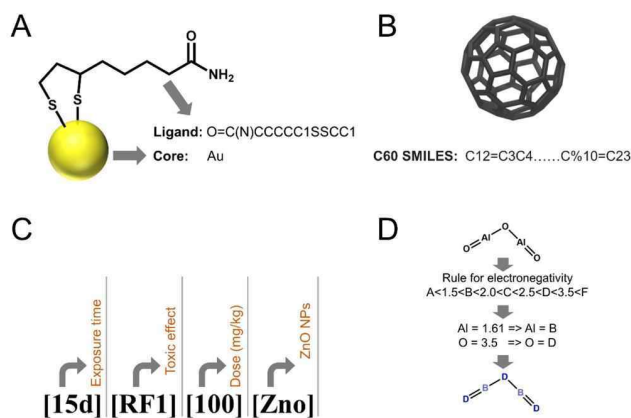


Figure 9. Structural nanodescriptors calculated from SMILES, quasi-SMILES, and SiRMS. (A, B) The SMILES can be used to represent the surface ligands of NPs and the simple nanostructures. (C) The quasi-SMILES can encode the information about the molecular structures, experimental conditions, and physicochemical properties. (D) In the SiRMS system, any molecule can be represented as different fragments of fixed composition or structure. Reproduced with permission from ref 253. Copyright 2014 Royal Society of Chemistry.

such as fullerenes (Figure 9B).^{57,251} SMILES-based nano-descriptors can be calculated using commercial and open-source software packages.^{223,224} SMILES-based nanodescriptors have also been used to model the toxicity of ligand-grafted MWCNTs,^{85,252} GNPs,^{91,247} and MONPs.^{77,78} For example, Dragon and MOE descriptors encoding surface ligand features were used to model protein binding, acute cytotoxicity, and immune response induced by 83 MWCNTs.²⁵² However, the SMILES-based nanodescriptors cannot distinguish NPs with the same surfaces but different cores, or different sizes and shapes.

To encode more NM features, a novel sequence of symbols called quasi-SMILES was developed.²⁵⁴ It extends traditional SMILES by encoding information on molecular structures, experimental conditions (e.g., cell type and exposure time), and physicochemical properties (e.g., surface chemistry).^{254,255} Thus, string characters and numbers are merged to generate quasi-SMILES (Figure 9C).²⁵⁵ Using Monte Carlo optimization in the CORAL (CORrelations And Logic) software package,²⁵⁶ the correlation weights of each character in quasi-SMILES constitute efficient descriptors for nanotoxicity modeling.^{257,258} Quasi-SMILES provides an example of combining molecular structures and other relevant features such as experimental conditions and physicochemical properties to generate useful NM features (Figure 9C). This idea has been extended further recently with the development of NInChI text representations of NMs.²⁷

The SiRMS (Simplex Representation of Molecular Structure) system represents a molecule as fragments of fixed composition or structure. Decomposition of molecules can also be performed on the basis of atomic properties, such as electronegativity,

partial charge, and lipophilicity.²⁵⁹ As shown in Figure 9D, the molecule Al_2O_3 can be represented as B_2D_3 on the basis of atomic electronegativity and further decomposed as different fragments using the rules of SiRMS representation.²⁵³ Although like SMILES, SiRMS-derived descriptors were originally developed for small organic molecules,²⁵⁹ they were later adapted for mixtures of organic molecules²⁶⁰ and polymers.²⁵⁰ The usefulness of SiRMS-derived descriptors for nanostructures was exemplified by the successful development of ML models for toxicity of MONPs.²⁵³ The SiRMS-based nanodescriptors can distinguish NMs not only on the basis of structures but also by atomic properties (e.g., electronegativity and partial charge).

3.2.5. Nanodescriptors from Periodic Table Properties, Liquid Drop Model, and Metal–Ligand Binding Theory. Periodic-table-based nanodescriptors are calculated from the molecular formula of a nanostructure and elemental properties obtained from the periodic table. Similar to nanodescriptors calculated from SMILES, quasi-SMILES, and SiRMS, these nanodescriptors require insignificant computing resources. Seven periodic-table-based nanodescriptors were used to train a ML model for cytotoxicity of MONPs.²⁶¹ Sixteen new nanodescriptors (second generation periodic nanodescriptors) were also applied to a QNTR study.²⁶² These nanodescriptors encoded more periodic table properties (e.g., valence electron of metal) and have proven to be efficient in nanotoxicity modeling.

In liquids, the physicochemical properties of NMs are highly affected by water, organic molecules, or biomacromolecules. For this reason, a liquid drop model (LDM) was proposed to simplify the behavior of NMs.²⁵³ In the LDM, an NM or an agglomerate is assumed as a liquid drop, where molecules or atoms are densely packed. Based on this assumption, LDM-based descriptors can describe several important physicochemical properties such as the Wigner–Seitz radius²⁶³ and aggregation behavior of NMs in liquids.

Metal–ligand binding (MLB) theory has also been used to describe the affinity of metal ions for biochemical ligands. Two MLB-based nanodescriptors, covalent index and cation polarizing power, were successful in modeling the toxicity of MONPs toward *E. coli* and HaCaT cells.²⁵³

3.2.6. Nanodescriptors from Full Nanostructures. A problem in generating theoretical and other classic computational nanodescriptors is the use of simplified nanostructures, disregarding or simplifying information on NM size, shape, ligand type, density, and position. Clearly, it is rational to develop nanodescriptors based on complete nanostructures. To this end, a recent study used 3D structures of MONPs.²⁶⁴ The full sphere nanostructures were constructed by replicating the unit cells of the most likely (thermodynamically stable) crystal structures and removing atoms outside of a set radius. In this study, nanodescriptors were derived from a core-surface model and computed parameters such as potential energy and coordination number. These descriptors were effective in modeling the biological activities of MONPs.^{264–266} However, the proposed method only studied monodisperse, uncoated, spherical NPs, and the resulting nanodescriptors were also limited for representing MONPs.

In the PubVINAS database, a Python program was developed to construct virtual nanostructures by importing the information about core materials, shape, size, chemical structure, and the number of ligands.²⁴ It can generate virtual NPs with user input material types, shapes, sizes, and surface ligands. Nanodescriptors for 34 GNPs were generated by simulating the

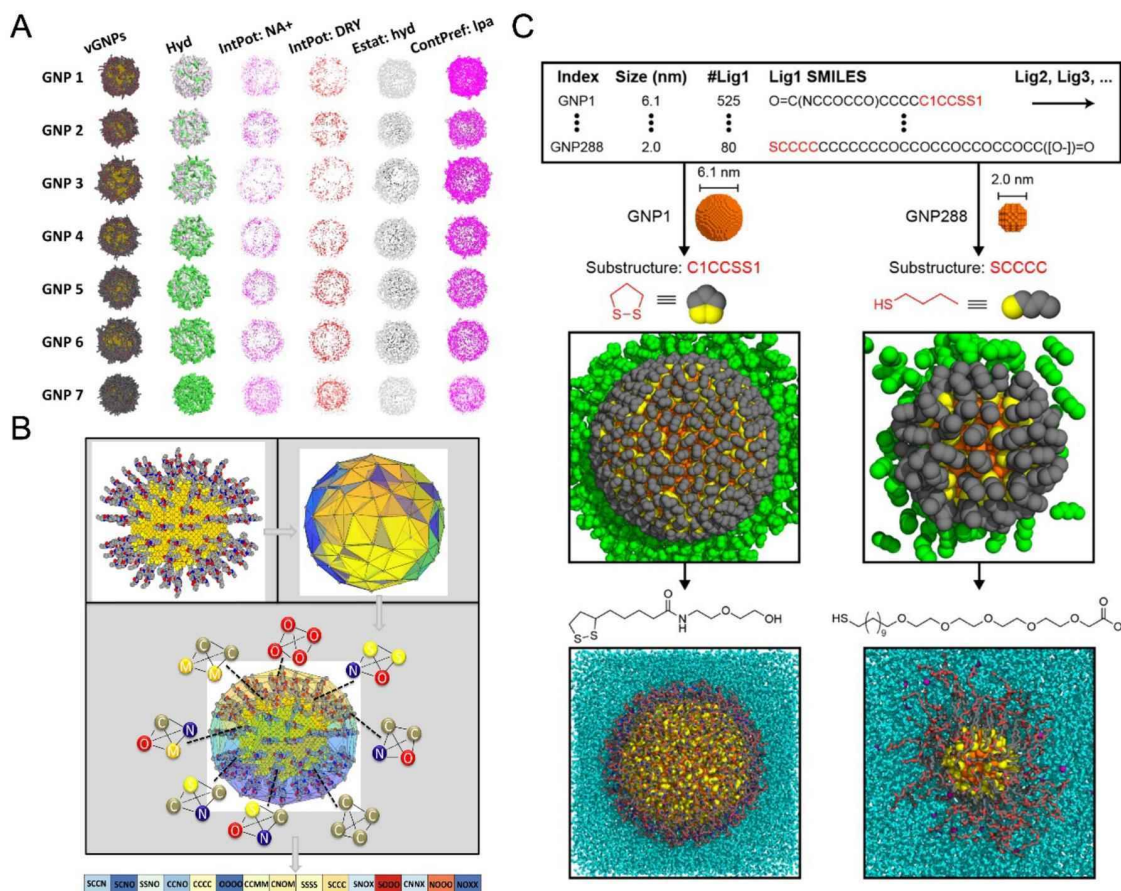


Figure 10. Descriptors calculated from full nanostructures. (A) Nanodescriptors generated from surface simulation of virtual nanostructures. Reproduced with permission from ref 86. Copyright 2017 American Chemical Society. (B) Novel tetrahedron nanodescriptors calculated from atomic electronegativity and the Delaunay tessellation of virtual nanostructures. Reproduced with permission from ref 35. Copyright 2019 Royal Society of Chemistry. (C) Nanodescriptors generated from molecular dynamics simulations. Reproduced with permission from ref 92. Copyright 2022 American Chemical Society.

surface of virtual nanostructures (Figure 10A) and were successfully used to model their bioactivities and to design new GNPs.⁸⁶ However, the calculations are resource intensive and rely on commercial software. Novel tetrahedral nanodescriptors based on atomic electronegativity and the Delaunay tessellation of virtual nanostructures have also been reported (Figure 10B).³⁵ Each nanodescriptor was the sum of the electronegativities of the four atoms within each tile multiplied by the tetrahedron number from the Delaunay tessellation. These tetrahedron nanodescriptors provide a way that features can be extracted from full nanostructures. They have significant potentials for modeling nanotoxicity or nanobioactivity.^{24,35,87,267}

3.2.7. Nanodescriptors from MD Simulations. Nanodescriptors derived from MD simulations are based on complete nanostructures and have the potential to elucidate interactions between NMs and liquid media (e.g., water) or biological systems (e.g., proteins). Features that can be derived from MD include the following: solvent-accessible surface area, NM–solvent interaction energy, NM–protein interactions, and ligand–water hydrogen bonds.^{41,86,92} Descriptors derived from MD simulations are sometimes called four-dimensional descriptors and have been widely used for computational aided drug discovery.^{268–270} For example, MD fingerprints of organic molecules in different environments (water, membranes, and

protein pockets) were used to train models of P-glycoprotein substrates.²⁶⁹ Recently, a set of theoretical nanodescriptors from MD simulations of 154 ligand-coated GNPs in aqueous solution were used to model their logP, zeta potential, and cellular uptake (Figure 10C).⁹²

Like quantum chemical calculations, NM size is a major factor limiting the use of MD simulations. Although nanostructures can be represented by a CG model when NM sizes are too large for whole structure simulations, small differences at the atomic scale are hard to discriminate using CG modeling. MD simulations are also highly dependent on accurate force fields, not always available for complex NMs. Furthermore, MD is impractical for modeling the protein corona compositions because of the size and complexity of the NM system and biological environment. For example, human plasma contains over 3700 proteins and many other small and large molecules. Each type of NM typically has its own unique corona, which is dynamic and changes with time and biological compartments.

In the above discussion, we introduced several representative experimental nanodescriptors and types of commonly used theoretical nanodescriptors. Typically, the experimental nanodescriptors can represent the properties of NMs in a realistic environment, while the theoretical nanodescriptors can be calculated in a fast and high-throughput way. Regarding model accuracy, it is hard to say which nanodescriptor is better, since

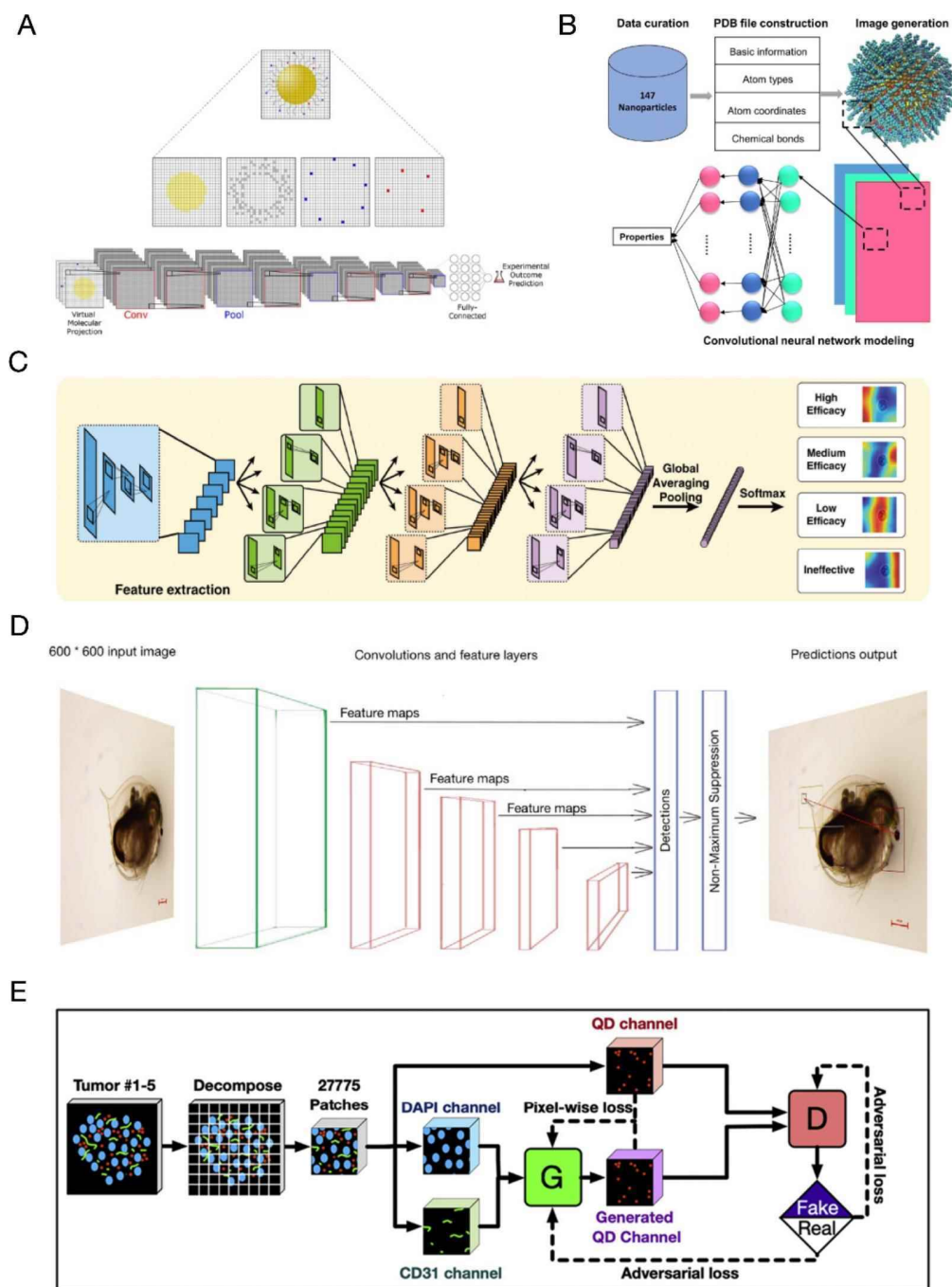


Figure 11. Application of end-to-end deep learning methods in nanotoxicology. (A) The CNN model is used for directly extracting features from virtual molecular projections. Reproduced with permission from ref 179. Copyright 2020 American Chemical Society. (B) The CNN model is used for directly extracting features from nanostructure images. Reproduced with permission from ref 180. Copyright 2020 American Chemical Society. (C) The CNN model is used for directly extracting features from cell images. Reproduced with permission from ref 178 (Creative Commons Attribution 4.0 International License: <http://creativecommons.org/licenses/by/4.0/>). Copyright 2018. (D) The CNN model is used for directly extracting features from *Daphnia* images. Reproduced with permission from ref 127 (Creative Commons Attribution 4.0 International License: <http://creativecommons.org/licenses/by/4.0/>). Copyright 2020. (E) The GAN model is used for generating the intratumoral distribution of QDs. Reproduced with permission from ref 279. Copyright 2021 Elsevier.

both experimental nanodescriptors and theoretical nanodescriptors were successfully used for nanotoxicity prediction in previous studies.^{35,86,92,142,147} Currently, experimental nanodescriptors are still the most widely used ones for nanotoxicity prediction due to the complexity of nanostructures (see the summary of Section 3.5). However, given the critical role of theoretical descriptors in the property prediction of other

molecules and materials,^{30,222,271} we should pay more attention to developing universal theoretical nanodescriptors in the future.

3.2.8. Latent Nanodescriptors from Deep Learning.

Important advances in deep learning have generated a paradigm shift in how nanodescriptors can be generated from ML models. CNN can accept simple representations of molecules or materials such as graphs or text strings (SMILES) and generate

latent descriptors within convolutional layers that can be used to train deep learning models, which was also called end-to-end deep learning.^{272–274} For example, a CNN model was used to generate latent features from multidimensional arrays of atomic coordinates obtained from nanostructures by molecular projections. The resultant ML model could reliably recapitulate experimental properties of NMs (Figure 11A).¹⁷⁹ Inspired by face recognition technology, a novel nanostructure annotation method was developed to automatically convert nanostructures into uniform images used to train models for physicochemical properties and biological activities of NMs (Figure 11B).¹⁸⁰ The end-to-end deep learning was also applied to model biological images generated from high-throughput microscopes.^{275,276} Trained on millions of images obtained from single-cell flow cytometry, a deep learning system could identify changes in cell states and evaluate the efficacy of nanomedicines (Figure 11C).¹⁷⁸ Moreover, CNN was useful for detecting malformations in *Daphnia magna* exposed to SNPs and TiO₂ NPs from thousands of microscopic images (Figure 11D).¹²⁷ Another class of end-to-end deep learning frameworks widely used for biological images analysis is the GAN.^{277,278} A recent study used a GAN model to map the intratumoral distribution of QDs after intravenous injection (Figure 11E).²⁷⁹ These deep learning models are increasingly capable of deciphering critical phenomena from biological images and show great promise in providing scientific insights.

In summary, end-to-end deep learning models can learn representations from original input data in an automated way, avoiding the complicated feature engineering process. Furthermore, such learning methods are flexible to uncover patterns difficult for experts to see.^{280,281} However, end-to-end deep learning generally involves complex neural network architectures that have the risk of poor generalization and “black box”. Therefore, a useful end-to-end deep learning model should be trained on enough high-quality data to improve the generalization. Recently, the development of a novel model interpretation method has opened the possibility of solving the end-to-end deep learning black-box problem.^{282–284} For instance, the class activation mapping was used to visualize the class-specific discriminative regions of cells treated with NMs.¹⁷⁸ As these obstacles are resolved, we do believe that the end-to-end deep learning would be a promising tool for automatically extracting nanotoxicity-related representations.

3.2.9. Developing Universal Nanodescriptors. Table 2 summarizes the advantages and disadvantages of commonly used nanodescriptors described above. Despite their success in published studies, there is no universal nanodescriptor (i.e., nanodescriptor that can be generated for any type of NMs) due to the complexity of nanostructures. First, correctly formatted machine-readable files (e.g., PDB files and SDF files) are required for generating nanodescriptors from physics-based simulations (e.g., DFT and MD). Therefore, accurate annotation of nanostructures is a prerequisite for calculating a set of theoretical nanodescriptors. Thus, rigorous characterization of NMs is essential for generating both experimental and theoretical nanodescriptors (Figure 12A). Second, the following issues need to be addressed to improve existing nanodescriptors. Real-world biological effects are induced by a polydisperse distribution of NM sizes, shapes, etc., not a monodisperse one. However, most studies use the average properties of similar NMs but a slight structure difference due to synthesis to train QNTR models.^{92,179,180,266} Thus, predictions will be more meaningful if the distributions of NM properties can be used

Table 2. Advantages and Disadvantages of Several Commonly Used Nanodescriptors

| Type | Examples | Advantages | Disadvantages |
|--------------------------------------|--|--|---|
| Experimental nanodescriptors | Size, logP, zeta potential | Directly relate to the characters of NMs under real conditions | Time-consuming, laborious, poor repeatability |
| Quantum chemical nanodescriptors | HUMO, LUMO, ΔH_{Me} | Describe the properties of NMs at the electronic scale | Require high computing resource and prior knowledge, ignore the true NM size |
| SMILES-derived nanodescriptors | Properties of surface ligands, such as partial charge, surface area, solubility, atomic number, molecular weight, and molecular connectivity | Easy to obtain | Information on three-dimensional nanostructures is not included |
| Periodic-table-based nanodescriptors | Cationic charge, periodic number of elements, atomic electronegativity | Easy to obtain | Information on three-dimensional nanostructures is not included; only applicable to MONPs |
| Full particle nanodescriptors | Total surface area, fragments of full nanostructure, lattice energy | Represent the three-dimensional features of full nanostructures | Require accurate annotations of nanostructures |
| MD-based nanodescriptors | Solvent accessible surface area, NM–solvent interaction energy, ligand–water H-bonds | Character of the interactions between NMs and the solvents at the atomic scale | Require high computing resources and precise force field parameters |

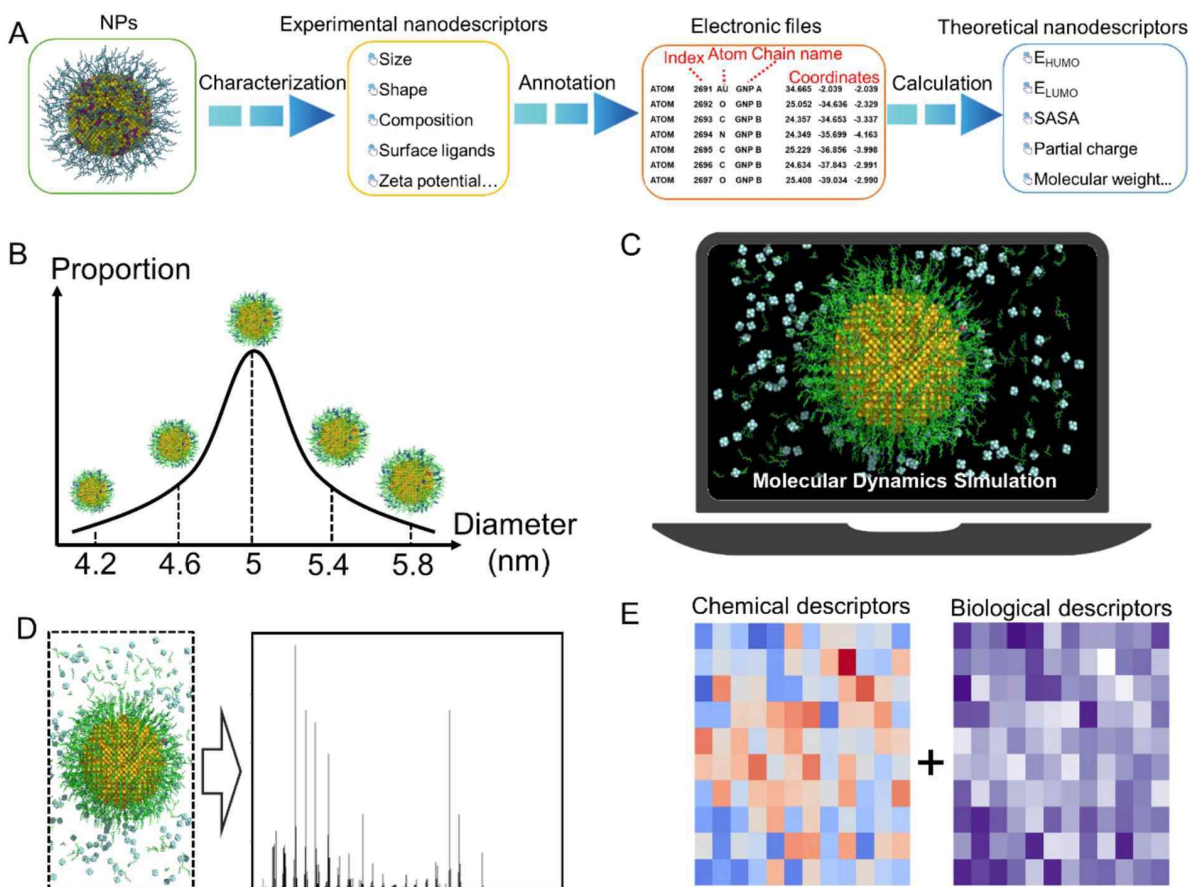


Figure 12. Prospects for developing universal nanodescriptors. (A) Nanostructure annotation is a prerequisite for calculating theoretical nanodescriptors. (B) Nanodescriptors representing a property distribution of multiple NMs. (C) MD-derived nanodescriptors to describe the nanocomposites with known structures. (D) Spectral nanodescriptors to describe nanocomposites with unknown structures. (E) The combination of structural and biological nanodescriptors used to distinguish structurally similar NMs.

instead of average values (Figure 12B).³⁴ Additionally, once released into the environment or entering biological systems, the NMs will inevitably interact with numerous pollutants or biomolecules and form nanocomposites. Characterization of the nanocomposites instead of original NMs can increase the applicability of the resulting models for environmental fate or toxicological effects induced by NM exposure. Inspired by descriptors for protein–protein complexes, novel structural descriptors for NM–protein complexes were developed and used to predict the sites of interaction of NMs with proteins.²⁸⁵ Nanocomposite descriptors can be calculated from MD simulations if we know the identity of the biological corona (Figure 12C). However, if corona information is unavailable, spectral features can be used as a fingerprint descriptor for ML model training (Figure 12D). There is also evidence that specific types of NM surfaces attract a consistent corona in a given medium, and the combined biologically relevant entity generates a consistent biological response. Thus, there is a complex but modellable relationship between NM surface chemistry and biological response. In a recent study, the electron ionization mass spectra of organic chemicals could be applied to model their toxicity end points without knowing the chemical structures.¹⁹⁹ Sometimes, structurally similar chemicals or NMs may have dissimilar biological or toxicological results, so biological similarity rather than structural or chemical similarity is also important for modeling (Figure 12E).^{286,287} As these

challenges are addressed, universal nanodescriptors are potentially feasible in the future.

3.3. Application of ML to Nanotoxicity Modeling

After obtaining the biological response data (i.e., dependent variables) for NMs and their corresponding nanodescriptors (i.e., independent variables), nanotoxicity models can be generated using ML methods. Here, we discuss the steps involved in using ML to model nanotoxicity data from a variety of biological assays.

3.3.1. Nanodescriptor Preprocessing and Feature Selection. After gathering toxicity data and generating nanodescriptors, a series of data preprocessing steps should be applied before model developments. These include data gap filling (e.g., imputation), data class balancing, and relevant feature selection. High heterogeneity of nanotoxicity data collected from different sources and missing data entities in data sets is ubiquitous and a major challenge in developing reliable ML models.^{19,37,45} Except for the experimental generation of new data, there are several mathematical ways to account for missing data entities: removing NMs with missing values; replacing missing values with the imputed results; and using read-across methods that replace missing values from similar NMs.^{288,289} Some recent ML algorithms such as XGBoost and category boost can automatically handle missing values without an imputation preprocessing.^{290,291} Another important issue is balancing the data set when one class has

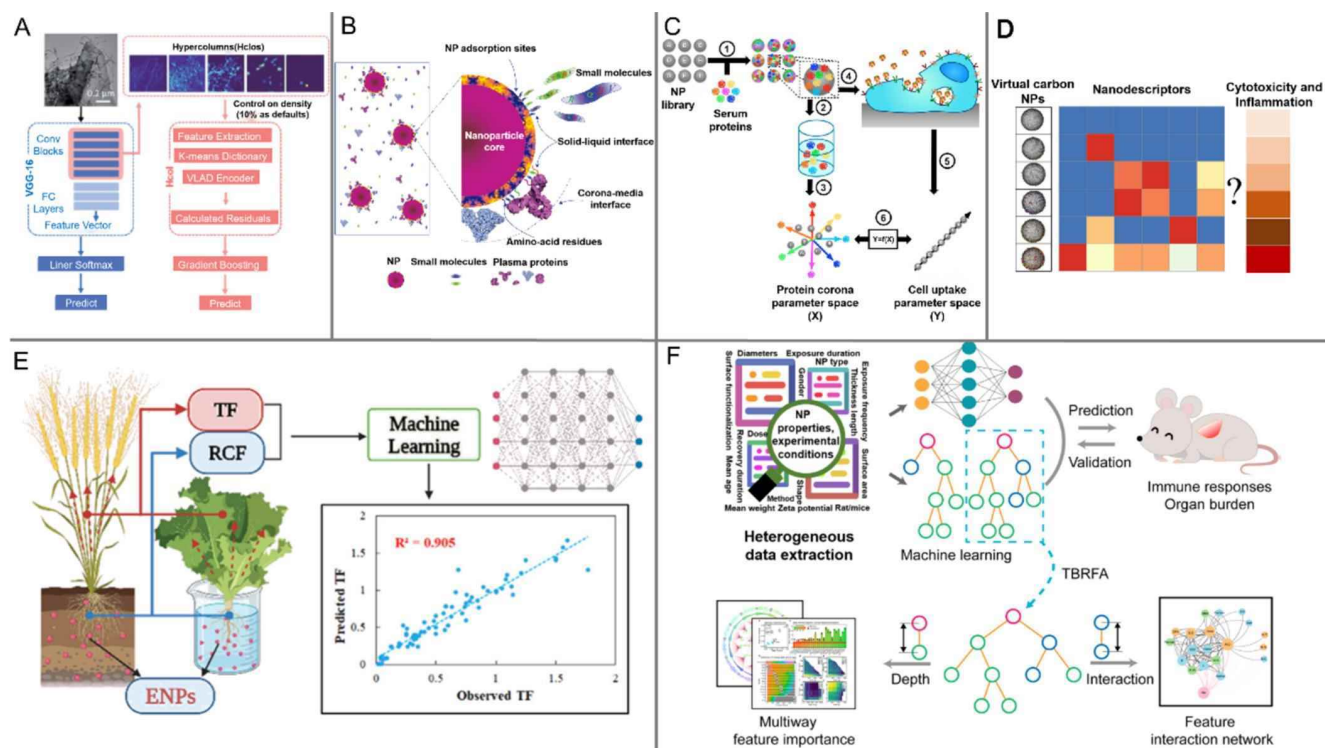


Figure 13. Modeling of NM's properties and bioactivities using AI approaches. (A) A transfer learning approach was used to identify carbon nanotubes/nanofibers from TEM images. Reproduced with permission from ref 306 (Creative Commons Attribution 4.0 International License: <http://creativecommons.org/licenses/by/4.0/>). Copyright 2021. (B) Competitive adsorption of proteins and small molecules onto the NM surface. Reproduced with permission from ref 315. Copyright 2010 Springer Nature. (C) Prediction of cellular interactions of GNPs and SNPs using protein corona fingerprinting and ML. Reproduced with permission from ref 72. Copyright 2014 American Chemical Society. (D) Modeling of cytotoxicity and inflammation of carbon NPs using virtual NP library and ML. Reproduced with permission from ref 267. Copyright 2020 Elsevier. (E) Prediction of plant uptake and translocation of NMs. Reproduced with permission from ref 316. TF, translocation factor; RCF, root concentration factor; ENPs, engineered nanoparticles. Copyright 2021 American Chemical Society. (F) Prediction of immune responses induced by NMs. TBRFA, tree-based random forest feature importance and feature interaction network analysis framework. Reproduced with permission from ref 147 (Creative Commons Attribution 4.0 International License: <http://creativecommons.org/licenses/by/4.0/>). Copyright 2021.

much more examples than the other class. Biased data is a common problem for modeling studies, especially for drug discovery, since most chemicals are inactive against the target.^{46,48} Biased data can be a big challenge for developing an accurate classifier as most ML algorithms are designed based on the assumption of a uniform distribution of examples for each class. Problems induced by biased data can be tackled by choosing proper evaluation metrics (e.g., F-score or G-means), resampling to balanced sample classes, or using a balanced bagging classifier.^{292,293}

Feature selection refers to reducing the number of input descriptors before ML model development. This process is important, especially in situations where the number of features is larger than the number of data points used to train models. The presence of redundant variables increases the overall complexity of the model, reduces the model performance and interpretability, and induces overfitting. There are three main types of feature selection techniques, filter methods, wrapper methods, and embedded methods.^{216,294} Filter methods include removing constant descriptors, descriptors with low variance, or highly correlated descriptors. These methods select features based on their inherent properties without considering labels (e.g., the property being modeled). Wrapper methods select features that result in the best model performance, so they are context-dependent or labeled methods. Some ML algorithms (e.g., RF, kNN, and LASSO) include wrapper feature selection

functions. Embedded feature selection methods based on penalty terms such as MLREM and L1 regularization are particularly effective for feature selections as they set less relevant features identically to zero and remove them from modeling entirely.²⁹⁵ Previously, some dimensionality reduction techniques, such as principal component analysis (PCA),²⁹⁶ t-distributed stochastic neighbor embedding (tSNE),²⁹⁷ and uniform manifold approximation and projection (UMAP),²⁹⁸ have also been used to reduce the ratio between the dimension of feature space and the number of unique training samples. These techniques were able to remove irrelevant and redundant features and thus reduce overfitting. In a recent study, it was demonstrated that the PCA-based feature reduction could greatly improve the accuracy and generalization performance of machine learning models.²⁹⁹

3.3.2. Modeling Physicochemical Properties of NMs.

Subtle changes in the physicochemical properties of NMs can influence their toxicity.^{300,301} The first application of ML to nanotoxicology is modeling the physicochemical properties of NMs.^{176,208} The main NM physicochemical properties that influence the toxicity include morphology, surface charge, hydrophobicity, solubility, and dispersion ability. The morphology of NMs is usually characterized by TEM,³⁰² mass spectrometry,³⁰³ and small-angle X-ray scattering.³⁰⁴ Manual analysis of these data is time-consuming and error-prone, so *in silico* methods have been used for unbiased and high-throughput

analyses. ML or deep learning algorithms have been shown to objectively extract morphological information (e.g., size, shape, surface area, and structure deflection) from TEM images, requiring little or even no manual intervention.^{228,229,305–309} For example, a transfer learning model pretrained on a large data set of images (i.e., ImageNet) can automatically classify complex TEM images of carbon nanotubes/nanofibers with good accuracy (Figure 13A).³⁰⁶ It is well-known that the hydrophobicity of NMs plays an important role in determining their environmental fate, biological partitioning, and toxicity.^{310,311} Thus, ML methods have been successfully used to model the hydrophobicity of NMs.^{86–88,92,312} Recently, MD-derived descriptors and the RF method were used to model the hydrophobicity of 110 GNPs. It was reported that ligand stability in water was the most important factor influencing the hydrophobicity of GNPs.⁹² The zeta potential and electrophoretic mobility determine the ability of NMs to form agglomerates. They have been successfully modeled by various ML approaches.^{35,180,313,314} Recently, ML models that include properties of the medium (ionic strength and conductivity) of NMs accurately modeled electrophoretic mobility.³¹³

3.3.3. Modeling Interactions between NMs and Biomolecules. The concept of a protein corona was first proposed by Dawson and co-workers in 2007,³¹⁷ and other NM–macromolecule coronas (e.g., environmental corona and eco-corona) have also been studied recently.^{318,319} For instance, inhaled NMs exhibit strong interactions with the lung surfactant systems.^{320,321} At the same time, released NMs can adsorb a large number of ecological molecules, such as natural organic matter and extracellular polymeric substances in the aquatic environment.^{322,323} As the NM plus corona constitutes the “biologically relevant entity”, there is a growing effort to elucidate the physicochemical and biological identities of coronas.^{324–326} As early as 2010, a linear free energy relationship model was developed to model the adsorption of various small molecules onto carbon nanotubes.³¹⁵ The model suggested that the macromolecule adsorption ability of NMs was mainly affected by a few critical properties of the adsorbate molecules, such as molecular volume, polarity, and polarizability (Figure 13B). It is well-known that the structure, stability, and dynamic properties of the corona can significantly alter biological responses to NMs.^{327,328} In this section, we will systematically summarize recent advances in the use of ML to model interactions of NMs with biomolecules such as proteins, lipids, and nucleic acids.

3.3.3.1. Modeling of NM–Protein Interactions. Studies of NM–protein interactions can be divided into two categories, i.e., analysis of a single protein interacting with one NM and identification of an entire corona population bound to an NM. Single-protein model systems can provide detailed atomic-level information for the disease mechanisms and aid in developing therapeutic nanomedicines. Currently, there are four data sets available to model the interactions of NMs with a single protein.^{56,79,85,91} These data sets were used to train a series of ML models to quantitatively describe the relationships between the nanostructures and protein binding affinities or enzyme inhibitions.^{35,57,85,91,252} These data sets were small (i.e., less than 100 data points), and each data set was generated by a single laboratory. To learn more about how nanostructures modulate protein function, larger and more diverse data sets must be generated and curated for modeling purposes.

It is well-known that thousands of proteins can bind to the NM's surface in biological fluids. The populations on the NM

surface change over time and in different biological environments. Characterizing the adsorbed proteins can provide critical information on the factors influencing NM biodistribution, clearance, and toxicity.^{329,330} To model entire corona populations, ML methods should fulfill two main functions: identify the quantitative relationships between the physicochemical properties of NMs and the binding affinities of various proteins^{142,331,332} and reveal how the protein corona populations influence NM responses to different biological systems.^{72,333–335} Relationships between the physicochemical properties of NMs and the adsorbed protein corona composition have been modeled by ML and meta-analysis. Researchers found that NMs' surface modification was important to determine protein corona formation.¹⁴² ML models trained on protein corona fingerprints have been shown to adequately model the cell association abilities of surface-modified GNPs and SNPs (Figure 13C).⁷² Models incorporating protein corona data were 50% more accurate than models trained on physicochemical properties of NMs alone. Clearly, protein corona fingerprints encode additional biologically relevant information than physicochemical properties.

3.3.3.2. Modeling Interactions of NMs with Other Biomolecules. Phospholipids are amphiphilic molecules with one negatively charged phosphate group (head) and two hydrophobic fatty acid chains (tail). They occur naturally in all living organisms as the major components of the pulmonary systems and cell membranes. As with NM–protein interactions, two questions can be asked for NM interactions with other biomolecules: what are the NM-induced structural changes in these biomolecules, and how does the adsorbed biomolecular corona affect NM's biological responses? Experimentally, interactions between NMs and cell membranes have been investigated using biophysical approaches to evaluate the cellular translocation of NMs and the effects of NMs on the integrity of the cell membrane.^{336–338} NM–DNA interactions can be detected from complex formations, binding-induced DNA conformational changes, and DNA degradations.^{339,340} Conversely, adsorption of phospholipids or nucleic acids onto NM surfaces alters their dispersibility, surface charge, and *in vitro* and *in vivo* behaviors.^{341–344} Previous studies have shown that DNA sequence recognition predicted by ML is useful to improve the biosensing and imaging capabilities of CNTs.^{345,346} However, to our best knowledge, there are no ML models applicable to predict NM–phospholipid/nucleic acid interaction for nanotoxicology research purposes due to the paucity of training data.

3.3.4. Modeling Cellular Responses Induced by NMs. Understanding NM–cell interactions is critical for nanotoxicology and nanomedicine. Cell lines are often used for toxicity mechanism studies. Of the top ten most researched topics in over 90,000 publications, four involve NM–cell interactions (cytotoxicity, *in vitro*, oxidative stress, and cells). Cell-based assays allow toxicity testing of multiple NMs at different concentrations, thus reducing the interexperimental variation and making substantial savings in time and cost. Results from *in vitro* assessment of nanotoxicity can identify potential toxicity mechanisms via different cell responses. Toxicity end points of *in vitro* testing include cellular uptake, oxidative stress, cell morphology changes, inflammatory responses, genotoxicity, and cell viability.^{347,348}

With the increasing use of NMs as drug carriers, the question of the amount of NMs taken up by each cell is becoming increasingly important. It is crucial to assess how intracellular

NMs interact with normal cells for nanosafety. NMs first interact with the cell membrane and then enter the cell through various known endocytic pathways, such as clathrin-mediated endocytosis,³⁴⁹ caveolae-dependent endocytosis,³⁵⁰ micropinocytosis,³⁵¹ and phagocytosis.³⁵² Upon entering the cells, the NMs may localize in different cell compartments, such as endosomes, lysosomes, or the cytoplasm. Internalized NMs can be characterized quantitatively by several analytical techniques. For example, metal-containing NMs can be quantified by ICP-MS^{353,354} and carbon NPs by electrophoresis.^{355,356} To date, several available cellular uptake data sets have been used to train ML models for estimating the quantity of NM per cell.^{70,72,82,86,87,357} Most ML models focus on finding quantitative relationships between the structural features of NMs and their cellular uptake through theoretical nano-descriptors like SMILES-based descriptors^{247,358–367} and full particle properties.^{35,86,87,180} As with other types of nanotoxicity data, the lack of standardized experimental protocols limits the comparability of cellular uptake data. For example, currently available data sets differ in experimental conditions, such as exposure times and doses.

Another important end point modeled by ML is cell viability following NM exposure. Because it is widely used and has a consistent unit of measurement, it is possible to assemble large-scale data sets through comprehensive and meticulous data curation from large numbers of nanotoxicology publications. This effort has generated several cytotoxicity data sets consisting of several thousand data points for different types of NMs.^{130,133,134,144,145,368} In addition, several smaller data sets generated from a few laboratories were used to train ML models for NP-induced cytotoxicity.^{235,239,240,242,369–373} For instance, ML models trained by cytotoxicity data on 30 MONPs revealed that toxic ions released in the lysosome were one of the most important factors determining toxicity to immune cells.³⁷⁴ By contrast, modeling results from larger and more diverse data sets provide more comprehensive insights into latent relationships between the physicochemical properties of NMs and their cytotoxicity. For example, a rule-based ML model trained on 4111 data points revealed that the cytotoxicity of NMs was determined primarily by the core material, surface chemistry, synthesis method and cell types exposed.¹⁴⁴

Other NM–cell end points such as oxidative stress, inflammatory response, and genotoxicity are also modeled by ML. Oxidative stress reflects an imbalance between the production and accumulation of reactive oxygen species (ROS) in cells or tissues and the biological system's ability to detoxify these reactive products. Typically, NM-induced cytotoxicity is regarded as a multilayered event in which the generation of antioxidant defense (tier 1) precedes the activation of proinflammatory (tier 2) and cytotoxicity (tier 3) responses at higher levels of oxidative stress.³⁷⁵ Existing oxidative stress data sets are mainly obtained from individual laboratories because it is difficult to merge these data generated from different groups under different experimental conditions and methods. Even with limited training data, useful information has been extracted by various ML models.^{35,92,125,372,376} For instance, previous studies have suggested that the ability of MONPs to induce oxidative stress was highly related to their band gaps.^{125,376} Increases in oxidative stress activate the expression of inflammatory genes, such as interleukin-6 (IL-6) and interleukin-8 (IL-8) genes. Previous studies demonstrated that the models based on ML methods and a virtual carbon NP library clarified the relationships between the critical compo-

nents of PM_{2.5} and the induced inflammatory responses (Figure 13D).²⁶⁷ In addition, virtual screening results from ML models can predict the inflammatory potential of MONPs not yet synthesized.²⁴³ An imbalance between ROS generation and the antioxidant system leads to damage to cells, proteins, lipids, and DNA. An important toxicity end point, the genotoxic effect of NPs, was investigated by QNTR modeling.³⁷⁷ The results revealed that the genotoxicity was strongly correlated to the formation heat, molecular weight, and surface area of the MONP cluster.

3.3.5. Modeling Ecological Risks of NMs. NMs that are persistent in the environment may lead to ecological risks.^{378–380} For example, NMs are toxic to microorganisms similar to their toxicity to human cells by disrupting the microbial membrane and generating intracellular ROS that causes DNA and protein damage.^{381–384} Currently, ML models of NM toxicity to microorganisms can be classified into three main types: cytotoxicity of NPs to the bacteria,^{177,289,385–391} antibacterial effects,^{392,393} and effects on microbial communities.^{143,394} MLR combined with a genetic algorithm was used to model the cytotoxicity of MONPs to *Escherichia coli*.¹⁷⁷ This work demonstrated that cytotoxicity was strongly related to the enthalpy of formation of a gaseous cation having the same oxidation state. As one of the earliest data sets on NM toxicity to bacteria, this data set has been used in many modeling studies.^{261,387,391,395–397} On the other hand, the cytotoxicity of NMs to bacteria can also be described in terms of their antibacterial capacity. An ML regression model was reported to study the antibacterial capacity of NMs and showed that the NM size, the exposure dose, and the bacterium species are the most important features modulating antibacterial activity.³⁹³ A RF regression model has also mapped relationships between the properties of NMs and hazards to the soil microbial community from a macro perspective.¹⁴³ Such ML models can guide the rational use and disposal of NMs to minimize the impacts on soil microbiota.

Compared to terrestrial ecosystems, NMs are more mobile in aquatic ecosystems. NMs may enter aquatic systems in many different ways. Zebrafish (*Danio rerio*) are useful as *in vivo* model organisms for nanotoxicity studies.^{128,129} HTS data from the zebrafish model has been used to develop AI-based platforms to handle high volume data sets and hazard ranking of NPs.^{398–401} Recently, various ML algorithms were used to create QNTR tools for modeling the toxicity of metal NPs and MONPs to embryonic zebrafish.⁴⁰² The ML models suggested that nanotoxicity was highly associated with the core material, and NP concentration and surface charge. SVM and image-based descriptors have also been used to automatically assign Zebrafish embryo phenotype after exposure to SNPs.⁴⁰¹ This approach has shown that ML models can effectively utilize image-based HTS assay data. Recently, deep learning was used to extract features from microscopic images of *Daphnia magna* and detect malformations induced by engineered NMs.¹²⁷

The use and disposal of commercial nanoproducts, such as nanofertilizers and nanopesticides, also introduce NMs into agricultural soils. Their uptake by plants, especially by food crops, poses potential risks to food safety. It is unclear how much NMs are taken up by plant roots and translated to plant shoots. Using a back-propagation neural network model, researchers were able to model the root concentration factor and translocation factor of NPs from their fundamental physicochemical properties and key experimental conditions such as plant species and exposure time (Figure 13E).³¹⁶ Similarly,

meta-analysis and ML were used to investigate the impacts of nanoplastics on terrestrial plants.¹⁴⁶ This study revealed that toxicity metrics most affected adverse effects, followed by plant species, nanoplastic mass concentration, and nanoplastic size. Thus, ML is valuable for the ecological risk assessment of NMs and can also extract new knowledge from the algorithm decision-making process. Recently, an SVM model trained on infrared spectroscopy data to predict NP composition in *Cicer arietinum* samples⁴⁰³ was shown to be useful for monitoring of the accumulation and distribution of NPs in environmental systems, such as soil and plants.

3.3.6. Modeling Toxicity of NMs in Mammals. Understanding the relationships between the physicochemical properties of NMs and their *in vivo* behavior in mammals provides a basis for assessing human toxicity. Although *in vitro* and aquatic animal models like Zebrafish provide high-throughput, low-cost methods to study the biological effects of NMs, these data do not correlate well with human responses. Therefore, nanomedicines must undergo rigorous mammalian testing for toxicity before being tested in humans. Mice or rats are the most common primary animal models for studying human diseases. Mice share 95% of their genes with humans, suggesting that data in mice can be useful for understanding analogous effects in humans.

Various ML models have been developed for the *in vivo* behavior of NMs in mice or rats. These include modeling of reproductive⁴⁰⁴ and pulmonary toxicity,^{147,405,406} genotoxicity,⁴⁰⁷ tissue-specific oxidative stress,⁴⁰⁸ metabolic pathways,⁴⁰⁹ delivery efficiency,⁴¹⁰ and *in vivo* fate.^{236,279,411,412} For example, an RF model was developed to identify the most important factors determining reproductive toxicity of NPs, determined by the sperm count, percentage of motile sperm, and sperm abnormalities in rats.⁴⁰⁴ Feature importance analysis revealed that the reproductive toxicity was strongly related to the NP type and toxicity end points used. More recently, an interpretable RF model was used to study the pulmonary immune response and organ burden of NPs in mice and rats. This model could recapitulate the training set data with $R^2 > 0.9$ and half of the test set data with $R^2 > 0.75$ (Figure 13F).¹⁴⁷ Their proposed method utilized multiway importance analysis to reduce the bias created by small data sets and employed feature interaction networks to clarify the joint effects of multiple features. As important factors in clinical decision-making for nanomedicines, the pharmacokinetic properties of NPs can be estimated by *in silico* ML methods before experimental testing to triage the experiments to the most promising nanomedicines. To personalize NP treatment for micrometastases, 3D optical light-sheet microscopy imaging and ML techniques were used to analyze interactions between the micrometastases and NPs. The resultant ML model could predict NP delivery based on the specific pathophysiology of a micrometastasis.⁴¹¹

3.3.7. Applicability Domain of ML Models. As stated by the Organization for Economic Co-operation and Development (OECD) guiding principles, an ML toxicity model should have a defined applicability domain (AD), the region of materials, and the biology space in which the training set exists. No ML model can reliably predict the properties or bioactivity of every NM. Model predictions become increasingly unreliable when predictions are further from the domain of applicability. The AD plays a critical role in defining the region in which model predictions are reliable. Many methods have been developed to determine the AD of ML models. These approaches can be classified into range-based methods, distance-based approaches, and probability-density distribution-based strategies.⁴¹³ Range-

based methods define the AD as an n -dimensional hypercube whose boundaries are constrained by each descriptor's maximum and minimum values. Distance-based approaches use the "distance-to-centroid" principle, comparing distances between predictions and the training set using a predefined threshold. Commonly used distance metrics included the Mahalanobis distance, Euclidean distance, leverage approach, and Tanimoto similarity method.⁴¹⁴ The probability density distribution strategies estimate the probability density function for given data.⁴¹⁵ They are implemented by calculating the probability density of the data set and then identifying the smallest region that encompasses some predefined fraction of the total probability density. The smallest (highest density) region is used as the probability cutoff to determine the AD.

Previous studies have also employed the Williams plot to determine the AD of QNTR models.^{177,289,357,358,360,367,416,417} This is a plot of standardized cross-validated residuals versus leverage values. It provides immediate and simple graphical detection of both the response and structural outliers in a model. Other novel AD approaches include the local outlier factors approach,⁴¹⁸ distribution of SMILES attributes,^{257,419} and probability-oriented distance-based method (AD_{probDist}).^{242,420,421} The latter combines the average Euclidean distance and standardized residual to determine a model's AD. Unlike the range-, distance-, and leverage-based approaches, the AD_{probDist} method works with relatively small data sets and identifies unreliable predictions for newly screened NMs without experimental data.⁴²¹ However, no universal AD approach has been adopted by stakeholders and modelers in industries and regulatory authorities.⁴²² More conservative and restrictive approaches are sometimes an advantage because of the precautionary principle ("better safe than sorry").⁴²¹

3.4. Applications of ML to Mechanism Elucidation and NM Design

3.4.1. Applications of ML to Mechanism Analysis in Nanotoxicology. Another important OECD requirement for an ML model is interpretability, determining how a model makes a prediction. Knowledge gained from model interpretation can lead to testable theories and hypotheses, further advancing scientific understanding. Specifically, the interpretation of an ML model can lead to a better understanding of the mechanisms involved in nanotoxicity. It provides a theoretical rationale for making reasonable toxicity assessments and designing sustainable and safe-by-design NMs. Although most ML algorithms can make useful model predictions, it is a grand challenge to improve the transparency of predictions, especially for DNN.^{423–427} Here, we focus on the recent progress in molecular mechanism analysis for nanotoxicology using ML and other *in silico* methods.

3.4.1.1. Modeling Nanotoxicity with Interpretable ML Methods. Typically, approaches to explain ML models are of two main types, ante-hoc (intrinsic) and posthoc methods.^{424–426} Intrinsic interpretability refers to models trained on simple, chemically relatable structural features. Models that clearly explain how they produce predictions are also called glass-box or white-box ML models. The simplest ML algorithm to develop intrinsic nanotoxicity models is MLR.^{241,358,364,416,428–431} This method uses property and structural features as regressors to construct a linear relationship equation that predicts nanotoxicity.⁴³² Such models are easily interpretable because the signs and magnitudes of the regression coefficients show how each feature influences nanotoxicity. For

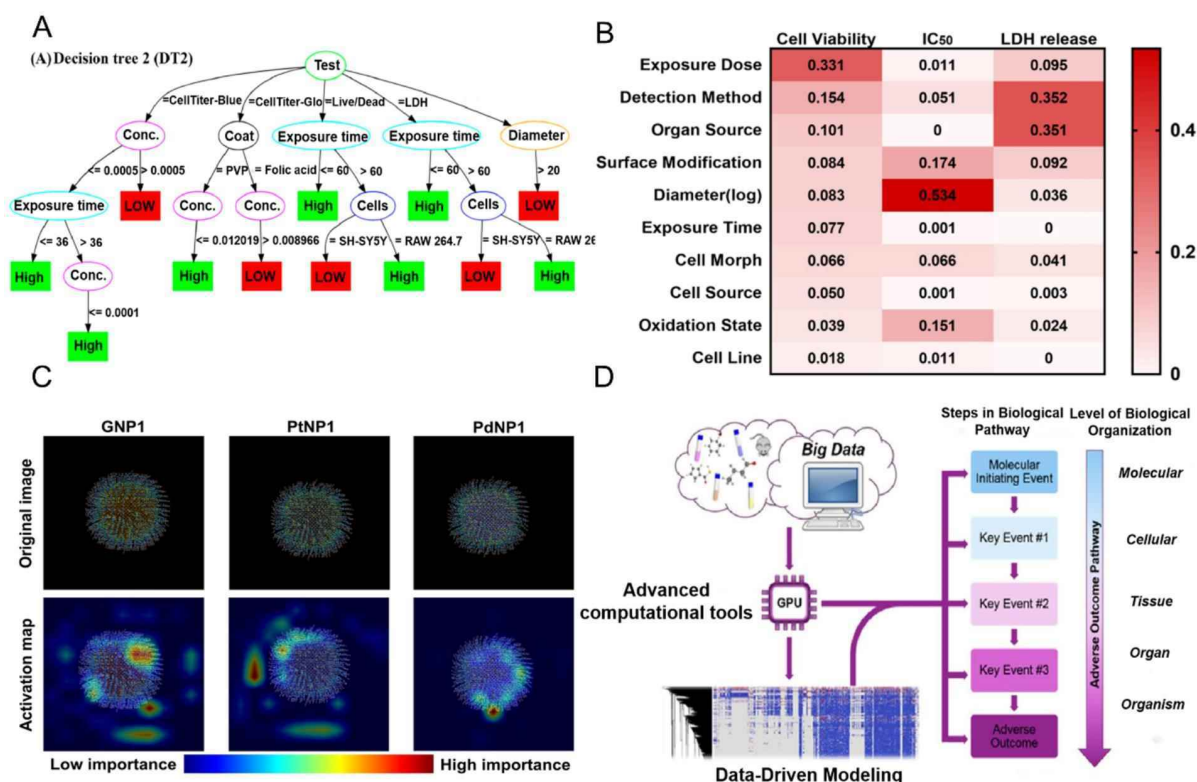


Figure 14. Applications of ML in molecular mechanism analysis of nanotoxicology. (A) Model interpretation of the decision tree. Reproduced with permission from ref 130. Copyright 2019 American Chemical Society. (B) Feature importance from the RF model. Reproduced with permission from ref 145. Copyright 2021 Elsevier. (C) Class activation map of the CNN model. Reproduced with permission from ref 180. Copyright 2020 American Chemical Society. (D) Mechanism-driven modeling based on an adverse outcome pathway. Reproduced with permission from ref 159. Copyright 2019 American Chemical Society.

example, an MLR model trained on quantum-chemical property features identified electronegativity (χ_{mix}) as the most relevant influence on the cytotoxicity of TiO₂-based hybrid NPs.^{242,433} This result suggested that increased cytotoxicity may be attributed to electron generation and ROS formation.²⁴² Other researchers have also used interpretable ML methods such as naïve Bayes,^{133,407,434–436} decision trees,^{130,402,437–440} and random forests^{147,267,405,429,441–444} to predict the potential hazards of NMs. The visualization capacity of decision trees helps explain how experimental conditions and NM properties affect cytotoxicity (Figure 14A).¹³⁰ These examples highlight the role of interpretable ML methods in building accurate and explainable models that relate intrinsic properties of NMs to their toxicity and other biological effects, further guiding the development of safe-by-design NMs.

There are classes of models whose predictions cannot be interpreted directly. These are often called “black-box” models and require additional algorithms to identify the important features in the model (posthoc analysis) and extract meaningful insights from them. Postanalysis methods can take many forms, such as permuted feature importance,⁴⁴⁵ partial dependence plots,⁴⁴⁶ local interpretable model-agnostic explanations (LIME),⁴⁴⁷ and Shapley additive explanations (SHAP).⁴⁴⁸ The most common posthoc analysis methods used in nanotoxicology applications are permuted feature importance^{134,145,368,420,443,449} and SHAP.^{92,410,450,451} The permuted feature importance is a model-agnostic global explanation method that provides insights into ML model behavior. It estimates and ranks feature importance by evaluating how the

prediction error increases when a feature is unavailable. In a recent study, the feature importance score of RF models was used to determine the relative impact of experimental conditions and NMs’ properties on their cytotoxicity (Figure 14B).¹⁴⁵ Ultimately, the exposure dose, NP size, and test method were identified as the most important features inducing cell viability, IC₅₀, and LDH release, respectively. SHAP values are assigned to each feature to represent the deviation from the average prediction by the feature in one sample. A positive SHAP value indicates the feature increases the probability of nanotoxicity and vice versa. For instance, high SHAP values suggested that the number of GNP–water hydrogen bonds normalized by the number of ligands plays a dominant role in determining the cellular uptake of GNPs.⁹² In addition, several model-specific explanation methods such as activation maximization and class activation mapping (CAM) were designed for DNN and other black-box models.^{282,283} These are visualization techniques that aid in understanding the behavior of DNN intuitively and simply. However, these methods can only produce coarse-grained visualizations and lack the ability to quantify the contributions of each feature to the prediction results.^{178–180} As shown in Figure 14C, the CAM suggested that the outer layers of nanostructures were the most important contributors to the hydrophobicity of NPs.¹⁸⁰ These examples highlight the feasibility and utility with which posthoc explanation methods can extract information from black-box models.

3.4.1.2. Mechanism-Driven Modeling for Predictive Nanotoxicology. In traditional QNTR models, a diverse set of structural and physicochemical features of NMs were used to

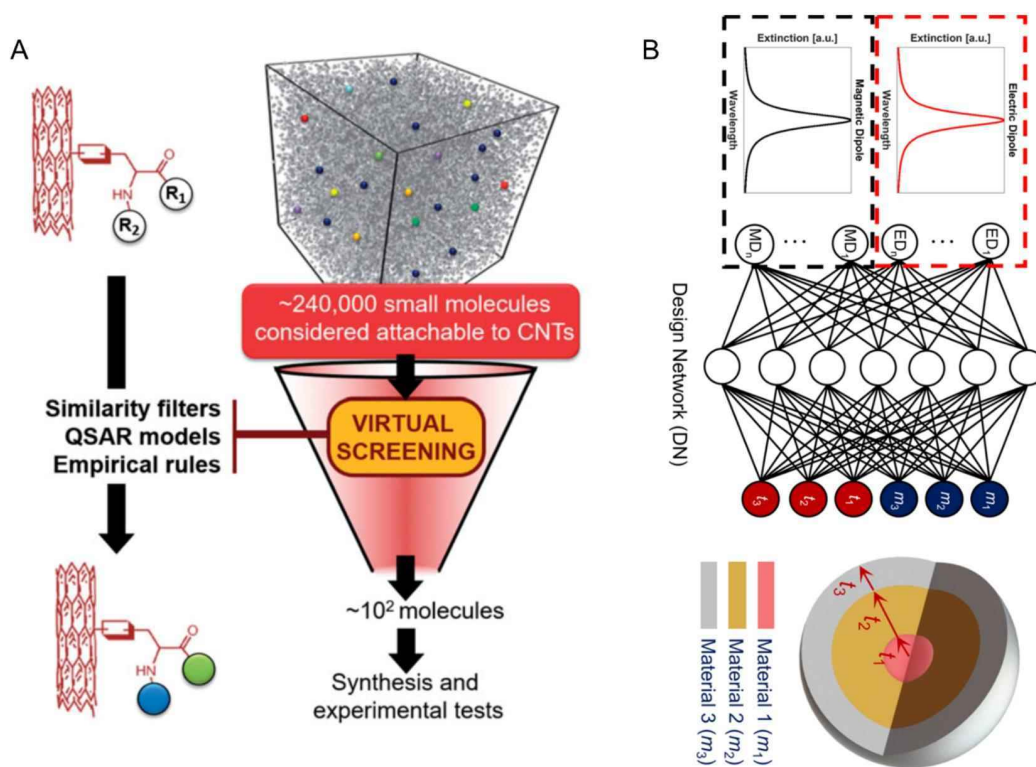


Figure 15. NM design with targeted properties using AI approaches. (A) Virtual screening of desired CNTs. Reproduced with permission from ref 252. Copyright 2016 Taylor & Francis. (B) Inverse design of core-shell NPs with targeted properties (from top to bottom). Reproduced with permission from ref 477. Copyright 2019 American Chemical Society.

train models of their biological responses, such as cellular uptake and cytotoxicity. Although applicable to making predictions, it is still challenging to elucidate the underlying toxicity mechanism with the limited number of features identified from interpretable ML models, especially when the model is quite nonlinear. However, it is known that NMs with similar physicochemical properties may sometimes induce different or even opposite biological responses, meaning empirical interpretations of nanotoxicity mechanisms based on physicochemical features can be unreliable.^{287,452} Due to the complexity of organisms, the relevance of predicting *in vitro* toxicity to eventual adverse effects at the organism level is often absent.

Since its establishment in 2010,⁴⁵³ the adverse outcome pathway (AOP) framework has provided mechanism-driven extrapolations of toxicity of new materials, including NMs.^{159,454–456} The AOP framework is a theoretical concept that describes a cascade of measurable key events linking a molecular initiating event (MIE) to an adverse outcome (AO) through intermediate key events (Figure 14D). Currently, AOP-based mechanism-driven modeling is being developed for various types of toxicities such as endocrine disruption,⁴⁵⁴ neurotoxicity,⁴⁵⁷ growth impairment,⁴⁵⁸ hepatotoxicity,⁴⁵⁹ acute inhalation toxicity,⁴⁶⁰ and skin sensitization.⁴⁶¹ The AOP strategy was implemented into a computational framework that depicts a pathway from structural alerts in which oxidative stress is a key event and hepatotoxicity is the adverse outcome.⁴⁵⁹ This oxidative stress hepatotoxicity model can predict the hepatotoxicity potential of new compounds and infer mechanisms involved in toxicity. Although the concept of AOPs in nanotoxicology is still in a preliminary stage, there were previous studies to construct AOPs with direct relevance for NMs.^{455,456,462–464} Recently, an AOP-informed QNTR model

was established to quantify the influence of the structural properties of MWCNTs on lung tissue inflammatory responses observed at the transcriptomics level.⁴⁵⁶ The results served as a proof-of-concept for further development of a framework that combines the QNTR and AOP for improved understanding and prediction of the adverse effects of NMs on human health and form the basis for comprehensive and realistic risk assessments.

3.4.2. Design of Bespoke NMs with Targeted Properties. Robust ML nanotoxicology models can be useful in rationalizing NMs with specified properties. Previous studies have shown that ML models enable property optimization of materials in multidimensional parameter space to provide improved beneficial properties while filtering out detrimental properties. In this section, we focus on the main strategies and progress employing ML to design NMs with bespoke properties.

3.4.2.1. Virtual Screening of NMs Using ML Models. Virtual screening, coined in the late 1990s,⁴⁶⁵ has become an increasingly important tool for identifying novel molecules with desired properties.^{466–468} Virtual screening allows the elimination of large numbers of “unfit” materials from the virtual library and the identification of a few top-performing candidates for further experimental validation.⁴⁶⁹ Compared with randomly selecting molecules for experimental testing, the virtual screening technique can save significant resources to avoid testing unsuitable molecules. Typically, virtual screening methods can be classified into structure-based and ligand-based methods.^{470–472} In ligand-based techniques, pharmacophores (patterns of chemical features in 3D space) or QSAR models are used to screen possible compounds of interest from chemical libraries. The structure-based methods rely mainly on molecular simulation techniques, such as docking and MD, to model the interactions of biological targets with a library of

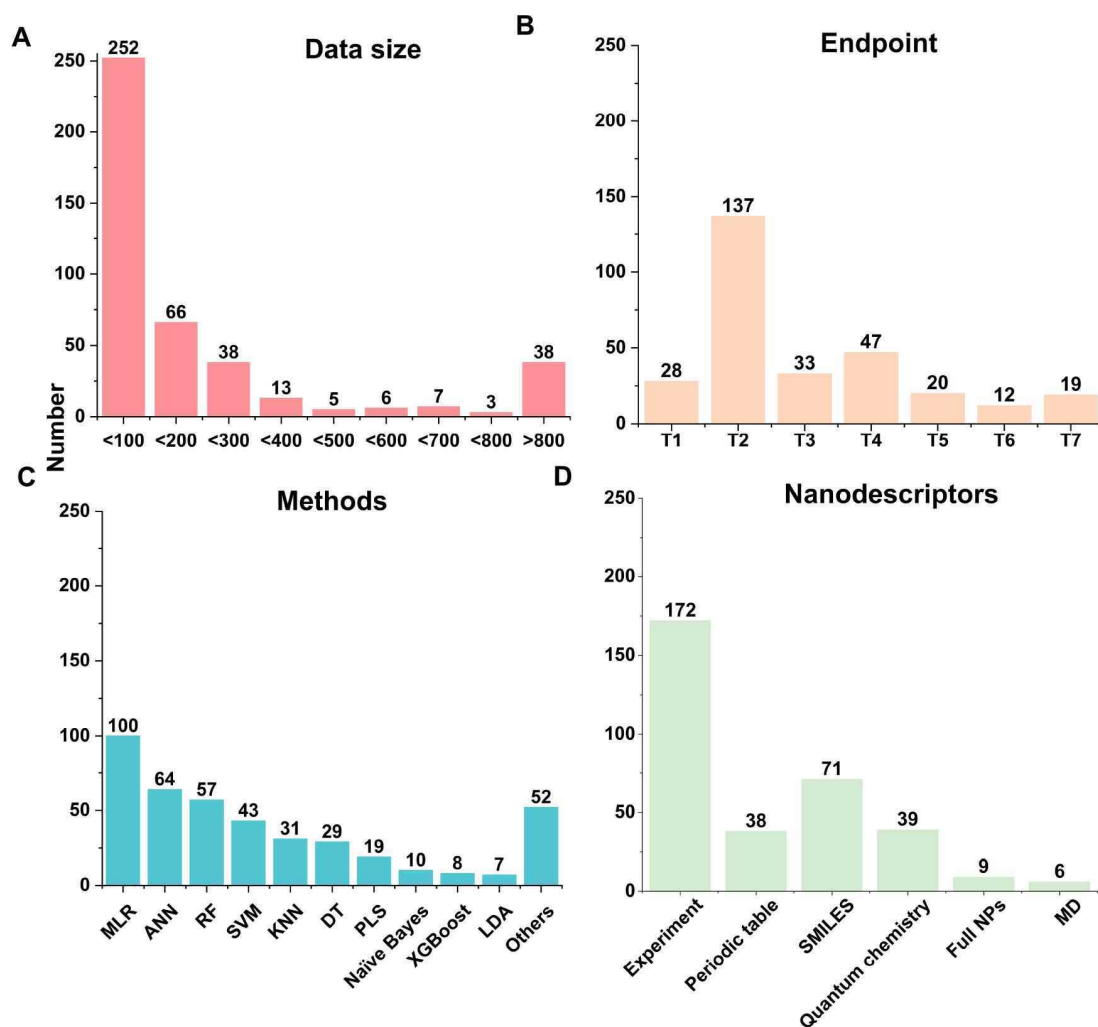


Figure 16. A summary of the applications of AI in nanotoxicology from more than 240 studies. (A) Nanotoxicity data size is used to train ML models. (B) Statistic analysis of biological assays used in previous ML models. T1, physicochemical property. T2, cellular response. T3, NM-biomolecule interaction. T4, NM–microorganism interaction. T5, aquatic toxicity. T6, NM–plant interaction. T7, mammalian toxicity. (C) ML algorithms used in previous studies. (D) Nanodescriptors generated in previous studies.

compounds. Conspicuously, it is increasingly difficult to deploy conventional virtual screening methods to meet the explosive growth of synthesizable molecules (a theoretical estimate is 10^{80} possible small molecules). The rapid development of ML-based virtual screening is expected to address this challenge.^{473–475} In a recent study, deep docking, a computational framework that iteratively trains DNN to accelerate molecular docking, was reported.⁴⁷³ This provides a 100-fold acceleration of structure-based virtual screening by discarding low-scoring molecules predicted from deep learning models. ML-driven virtual screening can also guide precision NM design.^{86,122,252,362,476} By using virtual screening and ML, multiple NMs with desired properties have been identified from material libraries.^{86,252} As shown in Figure 15A, ML models trained on 83 surface-modified CNTs were used to virtually screen a library of 240,000 ligands that could potentially be ligated to CNTs.²⁵² The selected CNTs were mostly shown desired activities in experimental testing. Such a screening strategy was proven to be efficient for designing NMs with the desired biological and safety profiles.

3.4.2.2. Inverse Design of NMs with Targeted Properties. Despite the success of virtual screening in designing molecules and materials, this strategy still has challenges. Creating a virtual

library usually requires substantial prior chemical knowledge and thus provides a limited number of candidates. Furthermore, screening a large virtual library can be time-consuming and sometimes inefficient as the number of resources required is disproportionately large compared to a small number of hits discovered. Recent advances in inverse design allow *de novo* design and prediction of new structures with improved properties.^{478–481} In contrast to the traditional design process that leads from chemical structures to properties, inverse design starts with the desired properties and searches for ideal molecular structures.^{480,481} This method is implemented by a range of generative models such as RNN,⁴⁸² GAN,⁴⁷⁸ variational autoencoders,^{483,484} and reinforcement learning.⁴⁸⁵ For instance, an automated materials discovery platform was established using the variational autoencoder algorithm to generate metal–organic framework (MOF) structures with specified properties.⁴⁸⁴ This platform discovered a variety of novel MOFs competitive with some of the best MOFs/zeolites previously reported. However, so far, there are few reports on the application of the inverse design of NMs^{477,486,487} and no clear examples in nanotoxicology.¹⁹ In a recent study, a simultaneous inverse design of core–shell NPs was achieved

using DNN, which could automatically learn the mapping from extinction spectra to design parameters (Figure 15B).⁴⁷⁷ The proposed deep learning-assisted inverse design was shown to be effective in identifying novel core–shell NPs with desired optical properties. Given the advantages of inverse design, it shows considerable promise for designing sustainable NMs more efficiently.

3.5. Summary of Applying AI in Nanotoxicology

So far, we have reviewed progress in applying ML to nanotoxicology such as the generation of novel nanodescriptors, deployment of advanced ML algorithms, prediction of nanotoxicity, and design of novel NMs. To better capture current progress, we provide a systematic analysis of all 246 ML-based nanotoxicology studies, which do not include review papers, in Figure 16.

There is still a considerable way for ML to accurately predict nanotoxicity broadly. As reported, most ML nanotoxicity models were trained on relatively small data sets (i.e., less than 100 data points) with limited NM diversity and small domains of applicability (Figure 16A). Only about 20 ML models were experimentally validated. As a result, most published ML models did not apply to designing new NMs. Nearly half of the reported ML models predict the NM–cell interactions, such as cell viability, cellular uptake, and oxidative stress under laboratory conditions (Figure 16B). These models, therefore, cannot elucidate the real-world impacts of NMs on humans and the environment. Therefore, more comprehensive and systematic studies are necessary using actual exposure conditions. With the progress of ML approach developments, if sufficient comprehensive real-world data can be obtained, the existing ML methods may be capable of generating robust and predictive models. Aside from obtaining additional nanotoxicity data from experiments, integrating different *in silico* assessment tools (e.g., biomass models and physiologically based pharmacokinetic modeling) can also address these challenges.³⁶ For instance, the biomass model can mimic the environmental fate of NMs, providing the basis for ecosystem-level safety assessment.³⁶ A recent study demonstrated that physiologically based pharmacokinetic modeling could be integrated with ML methods to predict the delivery efficiency of NPs to different tumors in animals, thus providing one potential solution to *in vitro*-to-*in vivo* translation and even animal-to-human extrapolation.⁴¹² Methods that combine physiochemical features with *in vitro* data can also improve the ability to predict *in vivo* responses.^{286,287}

The quantity, quality, and diversity of training data and the efficacy and interpretability of nanodescriptors are the critical factors in determining the robustness, quality, predictive power, and mechanistic utility of nanotoxicity models. Among the 246 QNTR studies reviewed, more than 400 ML models were constructed using over 30 different ML algorithms (Figure 16C). Tree-based algorithms (e.g., RF and decision tree) were the most popular, followed by MLR, mainly due to their good predictive accuracy and easy interpretability. However, the current popular deep learning (mainly neural networks) methods require large training data, so they are hindered by the available small nanotoxicity data sets. Furthermore, the black box and time-consuming nature of deep learning also limit its application to nanotoxicology.^{488,489} Some deep learning methods, such as CNN, show great promise in dealing with complex data formats such as TEM images of NMs,³⁰⁶ cell images,^{127,178} and mass spectrometry data.²³⁶ These data sets usually contain thousands^{127,306} to hundreds of thousands^{178,236}

of data points generated from high-throughput analysis or screening (Figure 16B).

The main differences in ML algorithm performance for a given training set are between linear and nonlinear methods. However, the increase in model complexity has the risk of overfitting and generating an unphysical dependence on descriptors. This can lead to misleading or erroneous predictions and hinder the interpretability and trustworthiness of the model. There are several ways to prevent overfitting, such as increasing the amount of training data, using regularization techniques, performing feature selection, and using ensemble models.

Ensemble learning is a ML technique that seeks to achieve better predictive performance by combining the predictions from multiple individual models.^{490,491} The strategies to implement ensemble models mainly include bagging, boosting, and stacking. The idea behind ensemble models is that, by combining the predictions of several different models, the strengths of each model can be leveraged. In contrast, the weaknesses of each model can be mitigated. Ensemble model predictions from different algorithms can sometimes improve performance, as can meta-models trained on a series of weak learners.^{62,492} Our previous studies demonstrated that the ensemble results could perform better than individual models.^{32,35} Ensemble models are generally most effective when the individual models have diverse and complementary strengths and weaknesses. However, it should also be noted that the ensemble models can be more complex and harder to interpret than individual models. Some examples to guide the application of ensemble learning in chemical toxicity prediction can be found in previous studies.^{493–496} In addition, most ML models still rely on experimentally measured features (incapable of predicting the properties of materials not yet synthesized, or for virtual screening) and traditional small-molecule-specific theoretical descriptors (Figure 16D). Given the weaknesses in current nanodescriptors, there is an urgent need to develop more advanced nanospecific descriptors in the future. Ideally, nanodescriptors should encode not only the entire nanostructure but features of the complex environments (e.g., the adsorbed protein corona) of NMs in biological media.

4. ELUCIDATING NANOTOXICITY MECHANISMS BY MOLECULAR SIMULATIONS

Quantitative relationships between NM structures, physicochemical properties, and nanotoxicity can be robustly established using AI methods, principally ML. Key parameters (e.g., size, shape, hydrophobicity, and charge) contributing to nanotoxicity can be identified. However, AI cannot tell us how NMs interact with the target biomolecules to generate the observed toxicities. Most adverse outcomes of NMs start from their interactions with biomolecules, such as lipids, proteins, and nucleic acids. These nanobio interactions involve complex molecular events, such as binding, conformational change, and dynamic adsorption/desorption equilibria. In some cases, the toxic effects involve chemical reactions catalyzed by or directly caused by NMs. Molecular simulation techniques are becoming indispensable for revealing molecular mechanisms underlying toxicity induced by NMs.

Nanobio interactions such as protein binding and denaturation, cell interaction and internalization, and generation of ROS are common biological responses induced by NMs. Combining AI and molecular simulation techniques has a strong potential to fill gaps in nanotoxicity data and provide critical predictions on

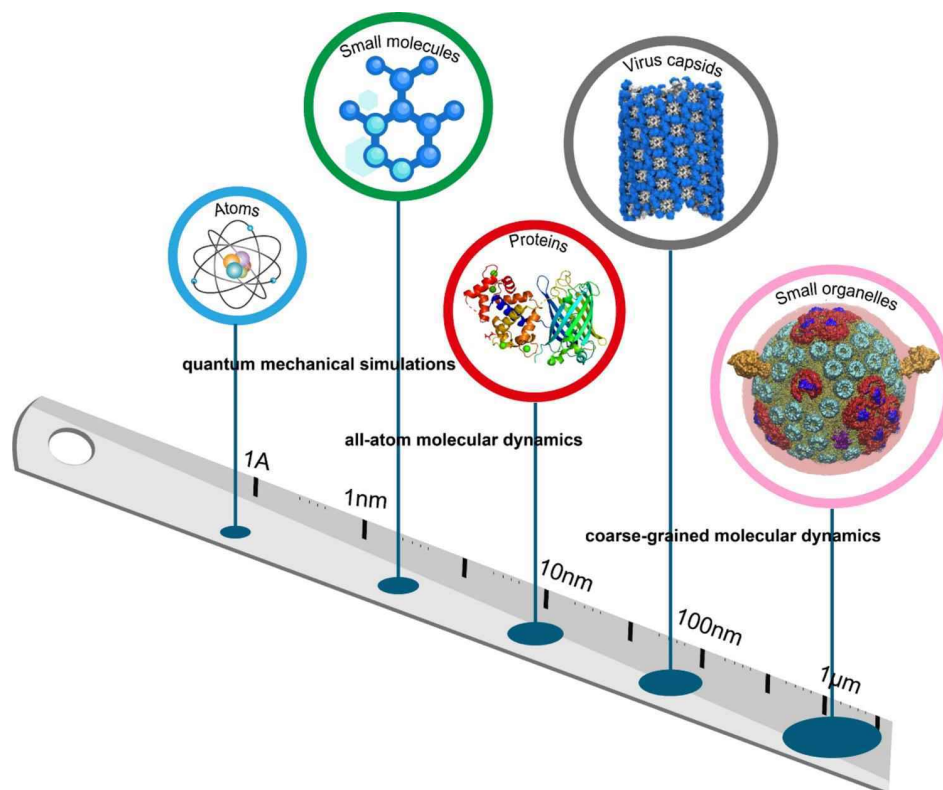


Figure 17. Molecular simulation methods. QM- and MM-based simulation methods are currently associated with varying levels of description according to the scale of physical sizes. QM simulations treat electrons as the fundamental interactive particles of the system. MM simulations treat individual atoms or groups of atoms as the fundamental interactive particles of the system. The decreased computational complexity granted by progressive coarse-graining makes it possible to access greater length scales.

the properties of new materials and mechanisms of toxicity. This section focuses on applying molecular simulation methods to elucidate mechanisms underlying nanotoxicity. We will emphasize the practices and challenges of generating high-quality mechanistic information and how simulation results can train interpretable AI models of relationships between nanostructure and nanotoxicity.

4.1. The Role of Molecular Simulation in Nanotoxicology

Molecular simulation has become an indispensable tool in contemporary nanotoxicology studies. The primary role of molecular simulation is to elucidate molecular mechanisms underlying nanotoxicity. Chemical reactions and physical interactions are the two main processes driving NM-induced toxicity. A range of molecular simulation methods is available to investigate these processes. Often, mechanistic information from the simulation is complementary to that from experiments, and it can be used to train predictive and interpretable AI models. Better information can be obtained by increasing the complexity of the simulation system.

4.1.1. A Brief Introduction of Molecular Simulations.

Molecular simulation is a type of computational chemistry, an important branch of chemistry that uses a computer to simulate physical or chemical processes at the molecular, atomistic, or electronic level. The two main types of molecular simulation methods are based on quantum mechanics (QM) and molecular mechanics (MM), respectively (Figure 17). The QM-based methods consider electrons as the fundamental particles of the system and treat them as quantum objects. In contrast, the MM-based methods focus on interactions between individual atoms

or groups of atoms in the system. These interactions are described by Newtonian mechanics. In many cases, the size of the system to be simulated is too large for QM calculations. Although MM-based methods cannot calculate electronic properties, large molecular systems can be modeled successfully. Generally, empirical and conceptually simple force fields describe intra- and intermolecular interactions, which contribute additively to the total energy of the system. In nanotoxicology studies, QM-based methods can reveal chemical processes involved in nanotoxicity, such as chemical transformations and ROS generation by NMs. MM-based methods are useful for elucidating molecular-level events, such as membrane damage, protein adsorption, and denaturation induced by NMs. For processes during which bonds are formed and broken, hybrid QM/MM methods can be used in which the chemically relevant parts are simulated by QM with the rest described by MM.

After the first development of valence bond theory by Heitler and London in 1927,⁴⁹⁷ advances in computer technology and improved wave functions, methods, and simplifying approximations allowed numerical solutions of wave equations for complex atomic systems to become common. Semiempirical atomic orbital methods based on the Hartree–Fock method were developed in the 1950s,⁴⁹⁸ employing parameters from empirical data and allowing much larger molecular systems to be studied, albeit with less rigor than with *ab initio* QM methods. Because the full Hartree–Fock methods for large molecules without the approximations are too costly, semiempirical quantum chemistry has become essential. Efficient QM packages such as GAUSSIAN, VASP, ABINIT, and Q-Chem

were developed for *ab initio* calculations in a similar time frame as empirical MM-based methods.^{499–502} Classical MD simulations based on MM were developed in the early 1950s. MD is used to examine the dynamics (temporal evolution) of atomic-level phenomena that cannot be observed directly. In biophysics and structural biology, it is frequently applied to study the motion of macromolecules such as proteins and nucleic acids. The technique can also aid the interpretation of results of biophysical experiments and model molecule/molecule or molecule/particle interactions.

4.1.2. The Commonly Used Molecular Simulation Methods for Nanotoxicology. Molecular simulation aids elucidation of molecular mechanisms underlying nanotoxicity and can also predict potential toxicity induced by NMs. Choosing an appropriate simulation method is crucial for mechanism elucidation. QM-based methods are best for chemical reaction processes involved in nanotoxicity. Some NMs react with the surrounding media to generate new groups on the NM surface, especially relevant for NM degradation processes. The structure and properties of NMs can be modified by the environment, and their fate and toxicity are altered. Therefore, using only the pristine nanostructure to predict toxicity may generate biased information. DFT is one of the most popular simulation methods for studying chemical interactions between NMs and biology. However, due to the complexity of real biology and the limit of computational capacity, it is very challenging for DFT methods to explicitly describe interactions between NMs and real biological systems.⁵⁰³ In particular, the current computational capacity allows DFT methods to treat systems containing a limited number of atoms. As such, the environmental effect induced by complex interactions between NMs and larger molecules cannot be simulated explicitly by DFT. Current efforts are made to allow DFT calculations to treat larger systems more accurately. MD is a proper method to simulate physical interactions between NMs and complex biological molecules, which should be considered in DFT calculations.⁵⁰⁴

Analysis of charge densities and frontier orbitals can identify possible reaction sites. Calculating electron transfer and changes in the energy levels can determine the optimal pathway and mechanism of chemical transformation. NMs can catalyze the generation of ROS that attack lipids, proteins, and DNA and induce oxidative stress, inflammation, and apoptosis. DFT is effective in predicting the chemical reactivity of NMs with certain atomic structures. For example, the capability of NMs to catalyze ROS, their redox activity, is an important property of nanotoxicity. DFT can calculate the LUMO and HOMO energies to determine the redox potential.^{505,506} Furthermore, the HOMO and LUMO distributions can be identified to predict chemically active sites or facets.⁵⁰⁷ Guided by DFT calculations, NMs with specific element doping or surfaces can be synthesized in experiments.⁵⁰⁸ The predicted activity can be validated by measuring the amount of product under certain conditions.⁵⁰⁹ With the chemical structure of the reaction product determined experimentally, DFT calculations can be further performed to reveal the reaction pathway.⁵¹⁰ The other important outcome of NM interactions with biology is the chemical transformation in atom rearrangement or surface functionalization. The chemical affinity of small molecules on NMs with certain atomic structures can be calculated using DFT calculations.⁵¹¹ Such predicted transformations can be experimentally validated using high-resolution characterization methods, such as XRD, FTIR, etc.⁵¹²

Clearly, *in silico* design of NMs with defined structures is much cheaper and faster than traditional trial-and-error, one-at-a-time methods. High-throughput DFT methods can easily generate very large virtual data sets. ML models can be trained based on these data to yield mechanistic information. Previously, the generated DFT data has trained many useful ML models for screening desired materials, such as CO₂ electrocatalysts,¹⁸⁸ new stable double perovskites,⁵¹³ thermally robust inorganic phosphor host,⁵¹⁴ superoxide dismutase nanozymes,⁵¹⁵ and ultraincompressible, superhard materials.⁵¹⁶ More importantly, these newly designed materials were also successfully synthesized and confirmed experimentally. For instance, in a recent study, the DFT-calculated Gibbs reaction-free energy was used to train ML models for screening superoxide dismutase-like nanozyme.⁵¹⁵ Finally, a novel MnPS₃ microneedle patch was discovered and experimentally confirmed to exhibit higher ability on free radical scavenging and hair regeneration.

MM-based methods can reveal physical mechanisms underlying nanotoxicity, while QM-based approaches can deal with chemical processes. Finding the optimal binding site is a prerequisite for discovering mechanisms involved in NM–protein interactions. Here, molecular docking is an effective method to predict putative binding sites for NMs and can provide considerable insight into how NMs bind to a protein in 3D. Most docking methods search high-dimensional spaces and use a separate scoring function to rank candidate poses. Residues contacting NMs can be analyzed to determine their contributions to binding affinity.

Some of the simpler and faster docking packages assume rigid proteins and NMs, while some allow conformational flexibility for the macromolecular target, albeit at much increased computational cost and complexity. Docking solutions are often used as initial structures for MD simulations, optimizing the interactions and accounting better for conformational flexibility. NMs often induce conformational changes in the cell membrane (damage) when they incorporate into them, a necessary process if NMs are to enter cells. MD simulation is the most commonly used and successful way of simulating NM–cell membrane interactions.^{40,517,518} Classical atomistic MD methods can typically simulate several thousand atoms over a few tens of nanometers dimensions for ≤1000 ns. As with QM-based methods, atomistic MD methods become too computationally expensive when the system size grows, often a problem for NMs. By combining groups of atoms into larger “coarser” particles and modeling their interactions using simpler and softer energy terms, the degrees of freedom are reduced as are the costs of simulating nanobio interactions. This CG method allows simulations over much larger time and length scales. Enhanced sampling methods, such as replica exchange, parallel tempering, and umbrella sampling, are commonly used to improve conformational sampling.⁵¹⁹ Common MD simulation packages include GROMACS, NAMD, AMBER, and LAMMPS.^{520–522}

4.1.3. The Workflow of Molecular Simulations. An appropriate simulation method is a prerequisite for conducting an effective computational nanotoxicology study. As discussed above, QM-based methods, such as DFT, are necessary if the critical process inducing nanotoxicity involves electron transfer and changes in chemical bonding. Otherwise, MM-based methods, such as MD, are more appropriate for describing the physical processes of NM interactions with biomolecules. In some cases, chemical and physical processes should be considered to comprehensively understand nanotoxicology. As such, the *ab initio* MD method becomes a better choice. In an *ab*

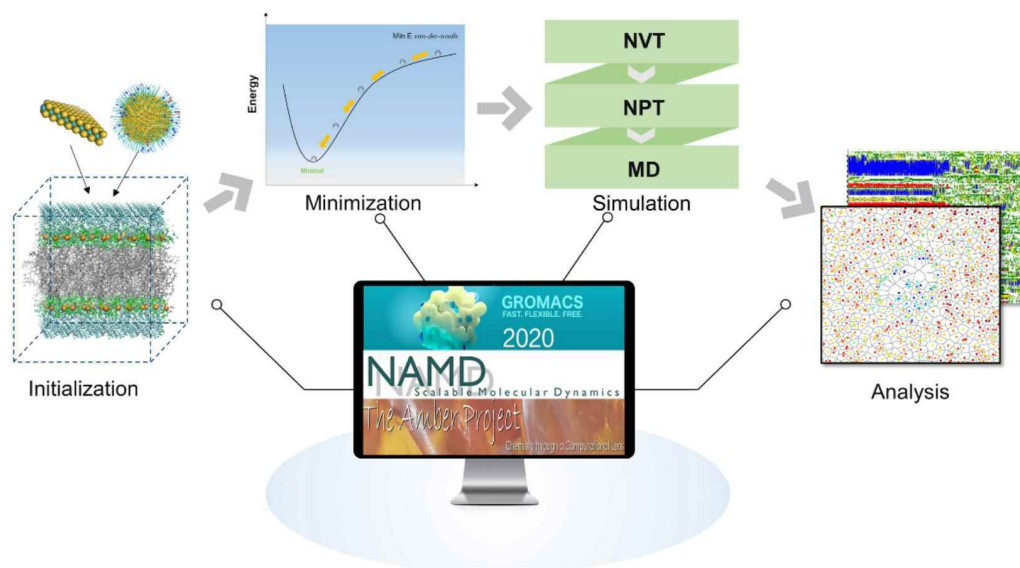


Figure 18. Workflow of MD simulation. A typical MD simulation includes initialization, energy minimization, equilibration, and analysis. Both all-atom and CG MD simulations can be performed using multiple software packages, such as GROMACS, NAMD, and AMBER.

initio MD simulation, finite-temperature dynamical trajectories are generated using forces obtained directly from electronic structure calculations performed “on the fly” as the simulation proceeds. It permits chemical bond-breaking and -forming events and accounts for electronic polarization effects.

The first step in DFT calculations is creating the 3D NM model that contains structural information. Due to the computational cost, NMs are usually modeled by a periodic slab or a cluster (<100 atoms). A reduced model is typically built around the NM moiety responsible for chemical activities. It should contain essential atoms or functional groups and/or molecules physically or chemically bound to the NM surface. In principle, reaction kinetics can be derived for a given chemical reaction by calculating reaction-free energies and activation barriers as a function of coverage and surface structure. The most probable reactive/catalytic centers for a given NM can be identified from the HOMO and LUMO. The LUMO energy is a measure of the reduction potential of the NM, while that of the HOMO reflects the oxidation potential, both important for the reactivity of NMs. The mechanism and reaction pathway can be elucidated when the reaction’s initial and final states are defined. The energy of each reaction step can be calculated to determine the most probable rate-determining step. Thus, DFT is useful for elucidating the mechanism of a specific chemical reaction and can optimize the nanostructure or suggest how to modulate reactivity.

Typical MD simulations involve four main steps: initialization, energy minimization, equilibration, and analysis (Figure 18). Before performing MD simulations to reveal mechanisms of NM interactions with specific biomolecules, 3D model structures of NMs and target molecules are again required. In general, NMs of specific core material and surface properties can be constructed using multiple software packages, such as Materials Studio, NanoModeler,⁵²³ and VMD.⁵²⁴ The lipid bilayer model is useful for investigating NM–cell membrane interactions. CHARMM-GUI is an excellent web-based platform to interactively build complex membrane models with desired components for molecular simulations.⁵²⁵ Structures of many proteins can be downloaded from the Protein Data Bank

(rcsb.org). For those without experimental structures, the deep learning package Alphafold can make surprisingly accurate predictions of their protein structures from amino acid sequences.¹⁶¹ The models of NMs and biomolecules are placed in a water box where ions neutralize the system charge or represent a specific ionic strength. The system energy is minimized to remove bad contacts; then, the equilibration and production runs are performed under specified temperatures, pressure, and cutoff radius for interaction conditions. Periodic boundary conditions are also defined to avoid edge effects. External mechanical, electrical, and/or magnetic driving forces can also be applied as required.

4.1.4. Nanotoxicology Studies Using Molecular Simulations. Physical mechanisms, such as membrane damage, protein denaturation, and DNA cleavage, have been relatively well understood at the molecular level using MD. For example, two useful insights can be obtained from MD simulations of NM–cell membrane interactions (Figure 19). One is the process by which an NM crosses a biological membrane to achieve cell uptake. The other related insight is how NMs perturb the mechanical and structural integrity of membranes. During simulations, the NM position and energy of NM–membrane interactions can be predicted. A mechanical force can be applied on the NM to pull or push it through the membrane at a constant velocity. NM translocation through the membrane can be analyzed by calculating the membrane-resistance force. Several parameters such as the membrane thickness, lipid density, and order parameters around the site of NM interactions can be calculated to describe the membrane ultrastructural perturbation by NMs. Membrane fluidity and permeability modulated by NMs can be characterized by calculating the mean square displacement of lipid molecules. Membrane compressibility and bending rigidity also reflect the mechanical perturbations induced by NMs. Although our understanding of these critical issues has been improved by simulations, the simplified lipid bilayer models used currently need to closely approach the complexity of actual cell membranes. Besides the lateral heterogeneity, two leaflets of the cell membrane are usually asymmetric due to the preferential

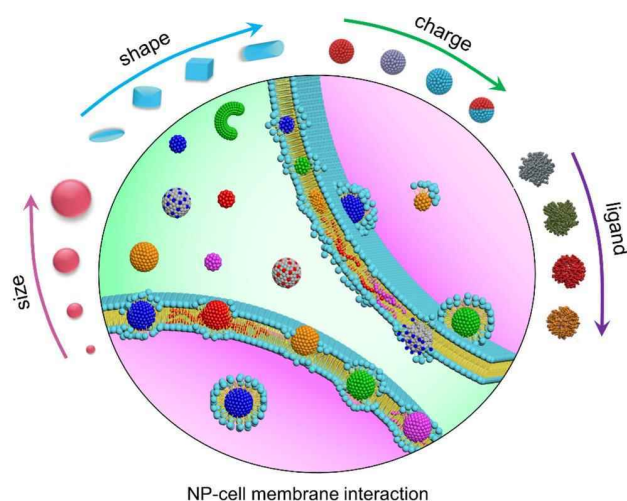


Figure 19. Schematic illustration of multiple pathways of NP–cell membrane interactions for NPs of various physicochemical properties. Reproduced with permission from ref 526. Copyright 2023 Royal Society of Chemistry.

selection of different lipid components. Although proteins are important components in NM–cell interactions, they have received much less attention than bilayers.

With the massive increase in computational power, simulations of protein folding and unfolding by atomistic MD have become feasible. MD is a powerful tool for probing the binding of proteins to NMs. Given the heterogeneity of protein structures, exploring the entire protein surface to find the optimal site with the lowest free energy is challenging for unbiased MD simulations. However, molecular docking has enabled more reliable prediction of protein binding sites by calculating the static binding affinity and score (Figure 20A).^{527,528} The docking results can provide guidance and initial starting geometries for MD simulations that describe the dynamic conformational changes on binding (Figure 20B).^{529,530} Analysis of the interactions can reveal mechanistic details. Metadynamics simulations with enhanced sampling can also construct the binding free energy landscape (Figure 20C).⁵³¹ Once the relationship between the structure and function of specific proteins is known, changes in the secondary and tertiary structures during interactions with NMs can be

predicted to reveal mechanisms of protein inhibition by NMs. For enzymes, it is clearly more important that attention is focused on local structural changes in the active site region and what these may mean for the binding of substrates. For example, the probability of residues in contact with NMs can be calculated as a measure of the accessible area change induced by NMs. Although functional inhibition of some proteins by NMs can be inferred from the relationship between structure and function, the structural change revealed by MD simulations is generally not quantitatively related to the loss or perturbation of chemical reactivity. While QM-based methods are more effective in evaluating the catalytic activity of specific chemical structures, the effects of structural changes in the environment are usually not tractable due to limited computation capacity.

Some aspects of nanotoxicology are relatively well understood due to their success in simulating NM interactions with specific biomolecules and interfaces. However, knowledge of equally important chemical mechanisms in nanotoxicology has received much less attention. NMs transform in the environmental or biological systems, resulting in changes in the NM's morphology, size, and surface properties. NMs can also generate ROS leading to oxidative protein carbonylation, lipid peroxidation, DNA/RNA breakage, and membrane structure destruction. In cells, this ultimately leads to necrosis, apoptosis, or mutagenesis. These processes involve chemical or catalytic reactions between the NM surface and the relevant target molecule. Direct experimental observations of the detailed changes in NM surface structure and amounts of reaction intermediates and products are quite sparse. A few specific active sites often dominate the catalytic activities of NMs. NM crystal faces with well-defined surface sites have been widely studied to clarify the contributions of each type of surface site to chemical reactions.⁵³² Clearly, this method is very limited for highly inhomogeneous atomic configurations, such as alloy NMs.⁵³³ Although DFT calculations on NMs can provide detailed kinetic information, their computational cost is still too high to acquire statistically meaningful kinetic data for NMs of various sizes, shapes, and compositions. However, new developments in deep QM promise to greatly expand the size of systems that can be studied by QM in the near future.⁵³⁴

4.2. Chemical Reactivities of NMs

Some types of NMs can have relatively high chemical reactivities important for their toxicities. Chemical reactions between NMs

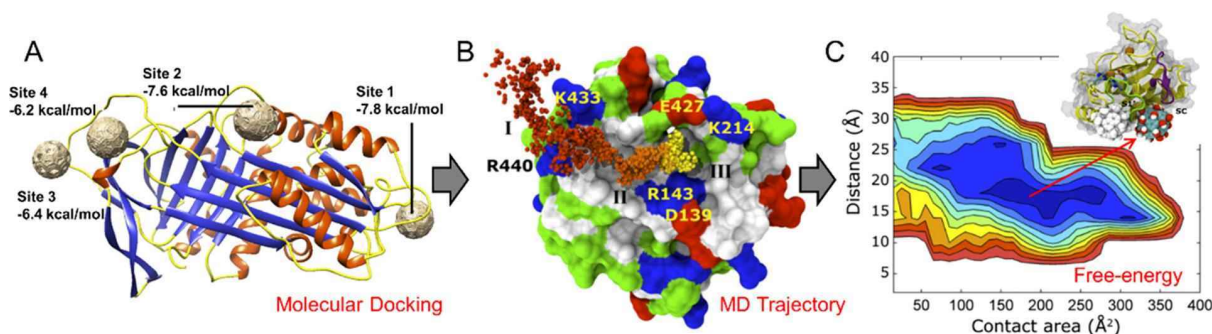


Figure 20. Molecular docking and dynamics simulations of NP interactions with proteins. (A) Molecular docking of C60 with protein, from which the possible binding sites with the corresponding affinity can be determined. Reproduced with permission from ref 528 (Creative Commons Attribution 4.0 International License: <http://creativecommons.org/licenses/by/4.0/>). Copyright 2017. (B) The trajectory of NPs moving around the protein acquired by MD simulation. Reproduced with permission from ref 529 (Creative Commons Attribution 4.0 International License: <http://creativecommons.org/licenses/by/4.0/>). Copyright 2012. (C) Free-energy landscape to find the optimal binding site. Reproduced with permission from ref 529 (Creative Commons Attribution 4.0 International License: <http://creativecommons.org/licenses/by/4.0/>). Copyright 2012.

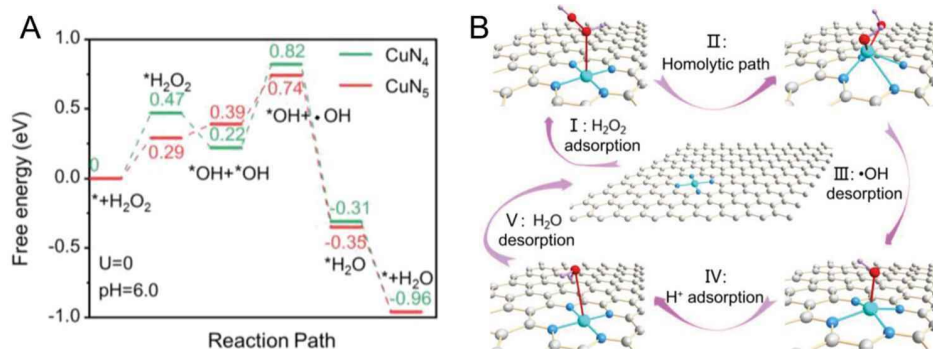


Figure 21. DFT calculations of Fenton activity. (A) Gibbs free energy diagram of the Fenton process on Cu-N₅ and Cu-N₄. Reproduced with permission from ref 541. Copyright 2020 John Wiley & Sons. (B) Schematic of the proposed reaction mechanism on Cu-N₄. Reproduced with permission from ref 541. Copyright 2020 John Wiley & Sons.

and surrounding media can generate ROS that induce oxidative stress. Functional groups can also be added to alter the surface chemistry of NMs and, consequently, their toxicities and interactions with biomolecules. These processes often feature electron transfer and chemical bonding that can be well described by QM-based methods. Therefore, chemical reaction mechanisms for NMs revealed by QM calculations can provide complementary information for training interpretable AI models.

4.2.1. NM-Induced ROS Generation. Reactive oxygen species (ROS) are superoxide anions ($O_2^{\bullet-}$), hydroxyl radicals (OH^\bullet), and hydrogen peroxide (H_2O_2). NM-induced ROS generation is one of the most frequent mechanisms responsible for NM-induced toxicity. For example, it promotes inflammatory responses in macrophages and neutrophils. Metal-based NMs elicit nanotoxicity via free radicals resulting from the Fenton reaction, whereas carbon NM-mediated ROS generation plays a role in mitochondrial damage.⁵³⁵ The critical factors in NM-induced ROS include the following: pro-oxidant functional groups on the reactive surface of NM, active redox cycling on the surface of transition-metal-based NMs, and NM–cell interactions.^{535,536} Surface-bound radicals such as SiO^\bullet and SiO_2^\bullet are involved in the formation of OH^\bullet and $O_2^{\bullet-}$.⁵³⁷ Structural defects on the NM surface create additional reactive groups in which electron donors or active acceptor sites interact with molecular O_2 to form $O_2^{\bullet-}$.⁵³⁸ Free radicals can also be generated as free entities in water.⁵³⁷ For example, the dissolution of NMs and subsequent release of metal ions can enhance the ROS response.⁵³⁹ Although the level of the components of ROS can be measured experimentally, our understanding of the exact mechanisms of ROS generation induced by different NPs is far from complete.

QM-based methods can elucidate mechanisms underlying ROS generation by NMs. Using band energy levels, previous studies have revealed how specific ROS species could be generated from different NMs in the aqueous phase.¹²⁵ The ability of NMs to induce ROS generation and oxidative stress is strongly correlated with the overlap between the energy of the NM conduction band and biological redox potential. In a seminal paper, DFT was used to calculate the lattice and HOMO and LUMO energies (E_{HOMO} , E_{LUMO}) that describe the redox properties of 17 different MONPs.¹⁷⁷ Using DFT, E_{HOMO} and E_{LUMO} of graphene NMs decorated with other functional groups were also calculated.⁵⁴⁰ By comparing the biological redox potential and E_{LUMO} of graphene-based NMs, unfunctionalized graphene and carboxylated graphene were predicted to induce

higher oxide stress than the other graphenes. Photochemical pathways for ROS generation can be identified by comparing E_{HOMO} and E_{LUMO} of NPs, the excitation energy of O_2^\bullet , and the redox potential of ROS. The Gibbs free energy along the reaction path can be calculated using DFT for a well-defined coordination structure (Figure 21A), allowing the reaction mechanism underlying ROS generation by specific NMs to be elucidated (Figure 21B).⁵⁴¹ While QM-based methods show promise in explaining the mechanisms of ROS generation by NMs, current published work is scant. A large amount of data can potentially be generated on NM-induced ROS generation by constructing NM libraries with controlled physicochemical properties for high-throughput DFT calculations.

4.2.2. Chemical Transformation and Degradation of NMs. Environmental and *in vivo* processes work to degrade NMs by altering their physicochemical properties and activities. Understanding these processes is important for predicting NM's environmental fate and toxicity.⁵⁴² Both amplification and attenuation of NM toxicity after transformation can occur.³²⁷ Clearly, the prediction of the toxicity or fate of NMs based solely on their pristine structures is inadequate. Although considerable efforts have been made to explore transformations of NMs and how these affect their stability and fate in natural systems,⁵⁴³ less attention has been paid to the toxic effects of environmentally or biologically transformed NMs. Inconsistent results were often observed when different materials and experimental conditions were used.

One typical pathway for NM transformation is a surface modification by the surrounding environment to form new functional groups. Using DFT, the chemical reaction mechanism was revealed by calculating the relaxed structure, molecular orbital energies, energy gap, and adsorption energy. For example, graphene can be modified by different functional groups.⁵⁴⁴ The chemisorption energy of species on graphene can be calculated as

$$E_{\text{chemisorption}} = (E_{\text{graphene with chemisorbed species}} - E_{\text{graphene before adsorption}} - E_{\text{species}}) / N$$

where $E_{\text{graphene with chemisorbed species}}$ and $E_{\text{graphene before adsorption}}$ are the total energies of the original and transformed graphene, E_{species} is the energy of chemical species, and N is the total number of adsorbed species. This value does not include the chemical species' activation energy (energy required to dissociate covalent bonds). This process allows step-by-step functionalization to be studied at an atomic structure level. The

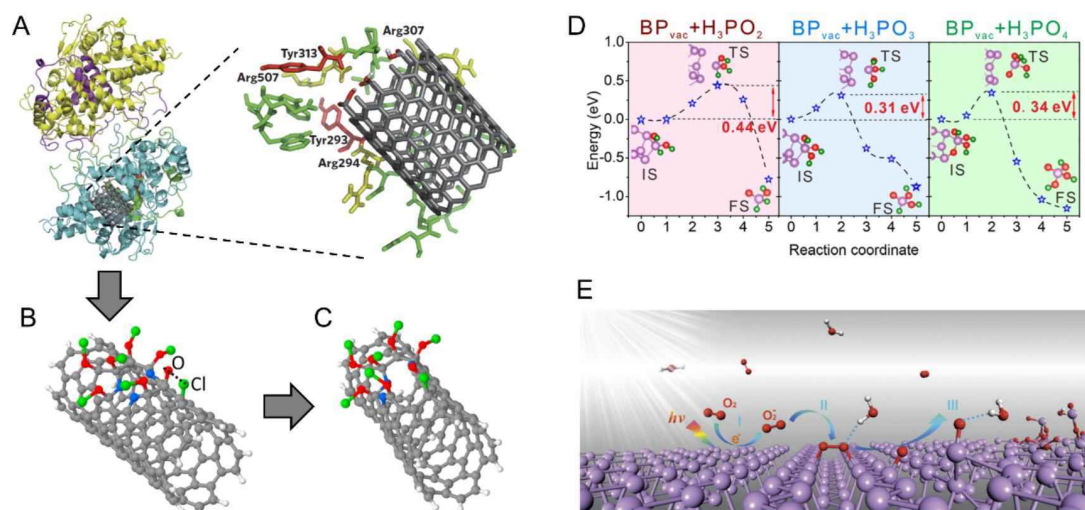


Figure 22. Degradation of carbon- and phosphorus-based NMs revealed by MD simulation and DFT calculation. (A) MD simulations of carbon nanotube interactions with human myeloperoxidase as the first step for nanotube degradation. Reproduced with permission from ref 556. Copyright 2010 Springer Nature. (B) Low-energy atomic configurations for ClO molecules chemisorbed on single-walled carbon nanotubes. Reproduced with permission from ref 557. Copyright 2021 Elsevier. (C) Cleavage of the C–C bonds, creating hole regions on the graphitic walls. Reproduced with permission from ref 557. Copyright 2021 Elsevier. (D) Reaction route and minimum energy pathway for the generation of H₃PO_x ($x = 2, 3, 4$) from the reconstructed zigzag edge of BP. Reproduced with permission from ref 558. Copyright 2018 American Chemical Society. (E) Schematic of the proposed pathway of stepwise BP degradation. Reproduced with permission from ref 559. Copyright 2016 John Wiley and Sons.

effects of the out-of-plane distortion and in-plane compression or expansion in the chemical reaction can be elucidated by comparing the total energy.

Another important route of NM's transformation is the rearrangement of atoms in the lattice and changes to the valent state. NM activity will change as the morphology or crystal surface is altered via atomic rearrangement. For example, the morphology of zinc oxide NPs can be altered through interactions with phosphate, from uniform nanosized, spherical particles to amorphous porous particles of much larger size.⁵⁴⁵ How such morphological transformation occurs via atom rearrangement is yet to be elucidated. DFT has been used to calculate the electronic properties and energy of NMs as a function of the exposed surface area.⁵⁴⁶ Chemical transformations between adjacent NMs can lead to hybrid systems with dramatic changes in their properties. Chemical reactions between two metal NPs studied by mass spectrometry and molecular simulations have revealed alloy clusters and have shed light on the detailed pathway and the mechanism of such reactions. Molecular docking simulations have shown that van der Waals forces between the clusters Au₂₅(SR)₁₈ and Ag₂₅(SR)₁₈ play an essential role in the initial stages of the reaction.⁵⁴⁷ DFT calculations were further employed to understand the energetics of single metal atom substitution into the various unique symmetry sites in the cluster and the overall single metal atom substitution reaction. Results showed that metal atoms could spontaneously exchange by forming an adduct between two clusters. Such structure-conserving reactions between NMs suggest new possibilities for the transformation of NMs in nature. Although the stability of a specific NM surface morphology can be estimated by measuring the electronic properties and energies, our understanding of how atoms are rearranged at the NM surface is still poor. Simulating the dynamic process of atom rearrangement, with the accompanied change of the valent state, is challenging by the conventional DFT method but may yield deep QM methods in the future. Unlike NMs with high surface energy and reactivity,

molecular drugs are relatively inert in their chemical structures under normal conditions. Nevertheless, drugs undergo changes when processed by enzymes (e.g., cytochrome P450) in different organs, especially the liver.⁵⁴⁸ Since the first discovery of biotransformation of drugs in the mid-19th century, considerable efforts have been made in understanding the metabolism and transformation of drugs in the body.⁵⁴⁹ In most cases, the transformation of drugs leads to pharmacological inactivation. With the biotransformation well regulated, drug metabolism can also be utilized for pharmacological activation, where pharmacologically active metabolites are generated.⁵⁵⁰ Thus, challenges and strategies can be shared between NMs and molecular drugs to reach a better understanding of the NM's biotransformation process.

Degradation is an important NM transformation for fate and toxicity. Metal ions can be released during degradation, particularly for metal-based NPs, such as ZnO, Ag, and CuO NPs, which are believed to be the dominant source of enhanced toxicity.^{551–553} Environmental conditions, such as pH and ionic strength, modulate metal ion release from NMs.⁵⁵⁴ Oxidation of NMs is enhanced under low pH. At the same time, high oxygen concentrations promote ion release from them.⁵⁵⁵ Despite increasing evidence that metal ions are released from NMs under diverse conditions, relatively little is known about the dissolution process.

Carbon-based NMs can also be degraded. For example, CNTs are catalytically biodegraded in neutrophils and macrophages by myeloperoxidase. The physical binding site of carbon nanotubes to the enzyme and their interaction mechanisms were revealed by MD simulations (Figure 22A).⁵⁵⁶ This enzyme generates hypochlorite and reactive radical intermediates to catalyze degradation. DFT calculations showed the formation of hypochlorite islands on the nanotube surface that promote ClO dissociation to strongly bound O species and Cl atoms (Figure 22B).⁵⁵⁷ These reactions result in the cleavage of the underlying C–C bonds, creating a hole in the graphitic walls (Figure 22C). Compared to carbon-based NMs, phosphorus-

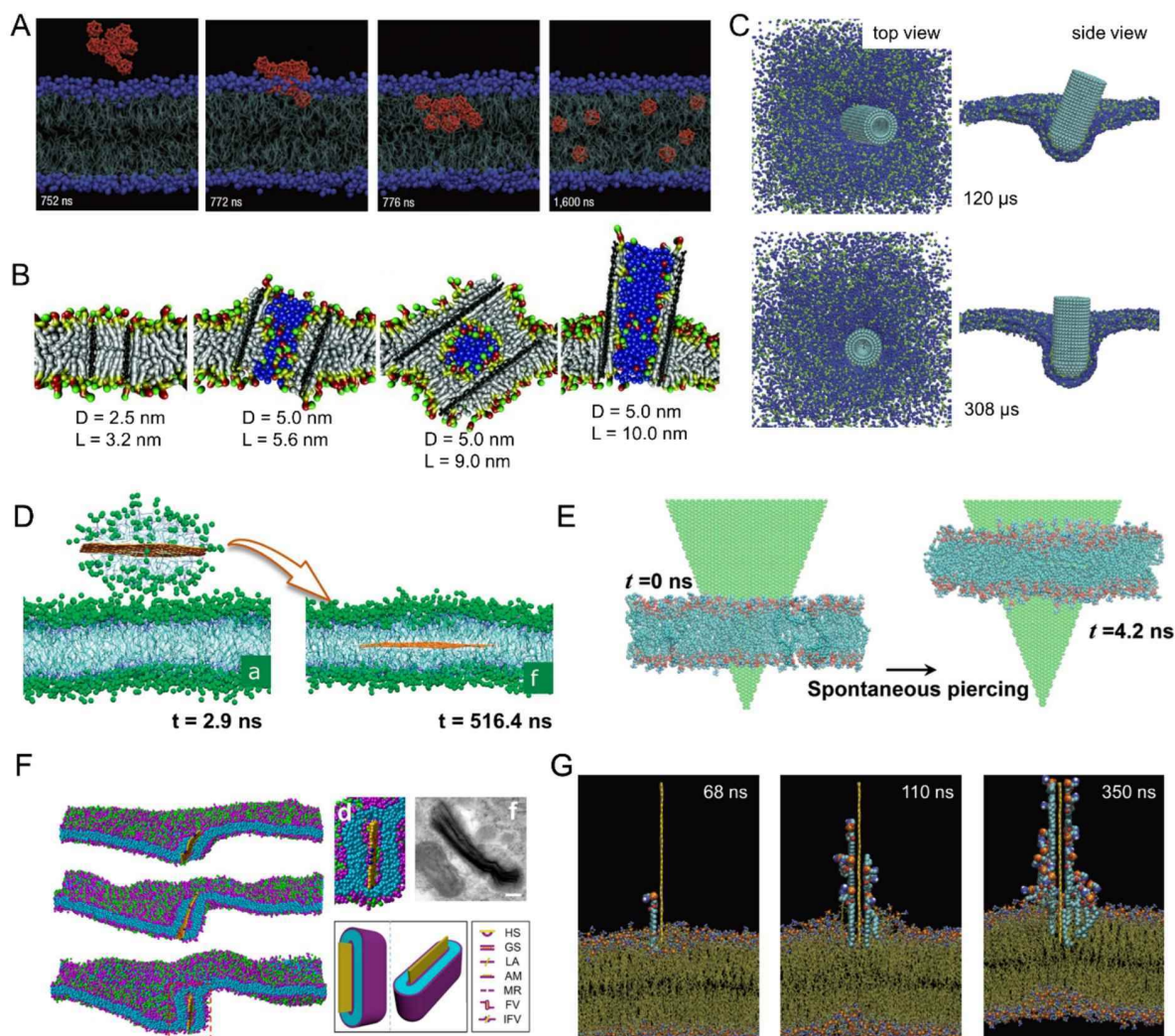


Figure 23. Multiple pathways of cell membrane interactions with carbon-based NMs revealed by MD simulations. (A) Spontaneous translocation and disaggregation of fullerene in the lipid bilayer membrane. Reproduced with permission from ref 572. Copyright 2008 Springer Nature. (B) MD simulations of carbon nanotube insertion into the membrane and induced perturbation. Reproduced with permission from ref 576. Copyright 2013 John Wiley & Sons. (C) MD simulations of cell membrane wrapping on carbon nanotubes. Reproduced with permission from ref 577. Copyright 2011 Springer Nature. (D) Spontaneous insertion of a graphene nanosheet into the membrane to form the sandwich-like structure predicted by MD simulation. Reproduced with permission from ref 578. Copyright 2010 American Chemical Society. (E) Spontaneous vertical translocation of graphene through the membrane revealed by MD simulations. Reproduced with permission from ref 585. Copyright 2013 National Academy of Science. (F) Cell membrane wrapping on graphene nanosheets revealed by MD simulation and visualized by TEM imaging. Reproduced with permission from ref 569. Copyright 2016 American Chemical Society. (G) Destructive extraction of membrane components by suspended graphene revealed by MD simulations. Reproduced with permission from ref 584. Copyright 2013 from Springer Nature.

based NMs, such as black phosphorus (BP), are less stable and are readily degraded in the environment. The degradation mechanism of BP has been well elucidated by DFT.^{558,559} Based on the calculation of the free energy of BP oxidation and dissociation of PO_2^{3-} , PO_3^{3-} , and PO_4^{3-} (Figure 22D), the minimum free energy pathways of BP degradation were determined (Figure 22E). Degradation of NMs can generate diverse molecules and ions potentially more or less toxic than the original NMs. Defects can also be introduced through partial and selective degradation to alter the NMs' surface activity. This critical information acquired by a combination of experiments and computation is essential to predict nanotoxicology from the original state of NMs.

4.3. NM Interactions with Cell Membranes

NM–cell interactions induce cytotoxicity,^{560,561} a process initiated by the primary physical barrier against NM invasion, the cell plasma membrane.⁵⁶² As mentioned previously, NMs cross the membrane and enter cells via multiple pathways,⁵⁶³ such as passive translocation,⁵⁶⁴ endocytosis,⁵⁶⁵ and phagocytosis.⁵⁶⁶ Upon entering cells, NMs may interact with diverse functional molecules to interfere with cell functions. The structure and function of the cell membrane can be perturbed by interactions with NMs.⁵⁶⁷ MD simulations have been widely used for molecular-level studies of NM–cell membrane interactions.⁵¹⁸ So far, multiple pathways of NM–cell membrane interactions have been identified. The relevant path for specific NMs is determined by size, shape, hydro-

phobicity, charge, elasticity, membrane surface tension, phase separation state, and bending rigidity.

Most simulation studies on the NM–cell membrane interactions used bilayers consisting of five or fewer lipid components for simplification. Real cell membranes are composed of hundreds of lipids and proteins asymmetrically distributed across two leaflets. Although the simplified lipid bilayer model can represent major structural features and properties of cell membranes, the roles of proteins have often been neglected. For example, some receptors participate directly in the endocytosis of NMs, while current simulations only use a defined number of lipids to act as receptors. This can be useful for modeling NM–membrane adhesion as the significant driving force in membrane engulfment, but more complex molecular components should be considered in future studies.

Since modeling cell uptake of NMs, especially at the atomistic scale, is very time-consuming by traditional MD methods, enhanced sampling can maintain atomistic precision while saving time. A combination of AI and MD is also a potential solution to this problem. ML techniques can be crucial in faster energy evaluations and MD time evolution to reach longer simulated times.

4.3.1. Pathways of NM–Cell Membrane Interactions from MD. NM–cell membrane interactions determine how NMs enter cells, cause cell membrane damage, or are transformed inside/outside of the cell. Therefore, an important application of MD is elucidating the pathways of cell membrane interactions with diverse NMs. Specifying the basic structure and properties of the cell membrane is a prerequisite for identifying pathways of NM–cell membrane interactions. The molecular size of lipids and the process of self-assembly into bilayers determine the membrane thickness of $\sim 4\text{--}5$ nm. Plasma membranes can segregate into functional and dynamic nanoscale membrane domains due to differential interactions between cholesterol and lipids with saturated and unsaturated acyl chains. Another structural feature of the cell membrane is asymmetry. Phosphatidylcholines (PC), sphingomyelin (SM), and gangliosides (GM) distribute predominantly in the outer leaflet. In contrast, phosphatidylethanolamine (PE), phosphatidylserine (PS), and other negatively charged lipids are present primarily in the inner leaflet. This asymmetry, together with the ionic gradient between inside and outside the cell, induces a transmembrane potential that strikingly influences cell membrane functioning and its interactions with NMs. Spontaneous curvature is caused by membrane asymmetry and the molecular shape of lipids in the local membrane patch. Surface tension and bending rigidity are two other significant mechanical properties influencing NM–cell membrane interactions.^{568–570} In classical MD simulations, surface tension is usually controlled by adjusting the lateral pressure. Given the specific relationship between lipid density and surface tension, the N-varied DPD (Dissipative Particle Dynamics) method was developed to control surface tension by lipid addition and subtraction in the defined boundary region.^{570,571} Bending rigidity, another mechanical property of cell membrane is an intrinsic membrane property. It is enhanced by physical support from the extracellular matrix and the cytoskeleton. The membrane patch of the liquid-ordered phase is more resistant to NM invasion due to the higher bending rigidity related to the tighter arrangement of lipid molecules. These cell membrane structural and mechanical properties explain why NMs of different physicochemical properties interact with the membrane differently.

Despite using the simplicity of the membrane model, MD simulations can obtain useful information on the pathways of NM–cell membrane interactions. The amphiphilic nature of the lipid bilayer suggests that hydrophobic NMs of sizes relative to the membrane thickness may displace hydrophilic lipid headgroups and enter the membrane interior to interact favorably with hydrophobic lipid tails. For example, fullerene is known to perturb cell functions and cross the blood–brain barrier. MD simulations showed that they rapidly aggregate in water but disaggregate upon entering the membrane interior (Figure 23A).⁵⁷² As fullerene is smaller than the membrane thickness, both the ultrastructure and mechanical properties of the membrane are barely affected by fullerenes. Simulation results suggested that membrane damage is an unlikely mechanism for fullerene toxicity. To elucidate whether and how carbon nanotubes penetrate the cell membrane, the single chain means field theory was developed to describe the cell membrane translocation of carbon nanotubes.⁵⁷³ Although membrane insertion of carbon nanotubes is energetically favorable, further separation from the membrane requires higher energy due to attraction by the hydrophobic membrane core.⁵⁷⁴ Increasing membrane surface tension was found to reduce the barrier of membrane poration, thus allowing short carbon nanotubes to escape from the membrane.⁵⁷⁵ For carbon nanotubes embedded into the membrane, MD simulations illustrated lipid rearrangements induced by the tube, and the extent of membrane perturbations was dependent on the nanotube size (Figure 23B).⁵⁷⁶ These computational results suggested that experimentally observed cell uptake of carbon nanotubes is probably due to other energy-dependent pathways, such as endocytosis. CGMD simulations have been used to identify the pathway of cell entry by carbon nanotubes. Nanotubes enter cells through a process of tip recognition, rotation driven by asymmetric elastic strain at the tube–membrane interface, and near-vertical entry (Figure 23C).⁵⁷⁷

Graphene is a 2D carbon-based NM. MD simulations predicted a sandwich superstructure, with graphene laterally embedded between the lipid bilayer (Figure 23D).⁵⁷⁸ With the graphene sandwiched inside the cell membrane, further MD simulations showed that the transport of graphene varied significantly from Brownian to Levy and even directional dynamics.⁵⁷⁹ For larger graphene sheets with sharp edges or corners, their vertical insertion into the membrane at edge asperities and corner sites was more energetically favorable (Figure 23E).⁵⁸⁰ As a result, the normal lipid flip-flop behavior can be perturbed, as revealed by free energy changes calculated by MD simulations.⁵⁸⁰ Under certain conditions, graphene cell uptake can also occur via receptor-mediated endocytosis.⁵⁸¹ MD simulations revealed that the endocytosis of graphene occurred via flat vesiculation and graphene self-rotation (Figure 23F).⁵⁶⁹ Graphene suspended outside the cell membrane can destructively extract numerous lipids from the membrane and induce lipid depletion, membrane damage, and finally cytotoxicity (Figure 23G).^{581–584} These distinctive pathways of graphene interactions with cell membrane suggest that the NM–cell membrane interactions are strongly dependent on the NM's physicochemical properties and the microenvironment, as will be discussed in the following section.

4.3.2. Effects of NM Physicochemical Properties on Cell Membrane Interactions. Given the membrane's amphiphilic nature and the diversity of molecular structures of components in the membrane, multiple NM–cell membrane interactions occur. These include electrostatic, van der Waals,

hydrophobic forces, and ligand–receptor interactions.⁵⁸⁶ Physicochemical properties of NMs, therefore, play a role in NM–cell membrane interactions.^{517,518,587} Among various physicochemical properties, NM size is important in regulating NP–cell membrane interactions.³⁴⁸ In simulations, cell membrane translocation of NMs can be accomplished by exerting an external force on the NM to pull it across the membrane.⁵⁶⁴ By comparing the force barriers of membrane translocation of NPs of different sizes, it was found that smaller NPs are easier to translocate through the membrane.^{588–590} The NM–cell membrane interactions are accompanied by the rearrangement and exchange of lipids, causing membrane damage. In the membrane translocation process, more lipids are displaced or rearranged by larger NPs (Figure 24A).⁵⁹¹ Larger NPs also extract more lipids from the membrane, creating more significant lipid depletion and membrane damage.⁵⁸² Another NM cell uptake pathway, receptor-mediated endocytosis, is also

size-dependent.⁵⁹² Thermodynamically, cell membrane wrapping around NMs is a competition between the NM–membrane adhesion energy and the membrane bending and stretching energy.⁵⁹³ Thus, membranes will wrap around NMs if the adhesive energy is sufficiently strong to compensate for the cost of membrane bending and stretching (Figure 24B).⁵⁶⁵ MD has simulated size-dependent membrane wrapping on rigid NMs. Using a CG model to simulate receptor-mediated endocytosis of NPs of different sizes, the calculations suggested that endocytosis of 14 nm NPs was easier than those around 6 nm.⁵⁹⁴

As NMs can vary considerably in their shapes, significant efforts have been made to elucidate how shape affects NM–cell membrane interactions. The contact area between the NM and the membrane determines the translocation capability of a nonspherical NM across the cell membrane. The translocation is accomplished via NP rotation to reduce the membrane-resistance force (Figure 24C).⁵⁶⁴ Although the fundamental biophysics of endocytosis for spherical NPs is a good guide, asymmetrical NPs should exhibit unique wrapping modes (Figure 24D). Researchers found that the efficiency of cell membrane wrapping is higher for spherocylindrical NPs than for spheres. Endocytosis was suppressed for NPs with sharp edges.⁵⁹⁴ Additional MD simulations and free energy analyses revealed that NP size is the primary determinant of endocytosis. NP shape breaks the symmetry of the curvature energy landscape and drives the endocytic pathway and the entry angle.^{577,595–597}

While NM's size and shape are inherent properties that change little, the surface chemistry of NMs can be readily altered in environmental and biological media or deliberately to tune their properties and functions, e.g., for targeted drug delivery. Therefore, understanding how surface chemistry modulates NM interactions with cell membranes is essential for nanotoxicity assessment and safe-by-design protocols. NMs coated with subnanometer striations of alternating anionic and hydrophobic groups penetrate the cell membrane without bilayer disruption. In contrast, NMs coated with the same moieties but with a random distribution are trapped in endosomes.⁵⁹⁸ This is the early experimental evidence highlighting the importance of NM surface chemistry and surface nano topography in cell membrane penetration and disruption. DPD simulations have been employed to reveal molecular mechanisms underlying this intriguing phenomenon.⁵⁹⁹ Unbiased simulations and free energy analysis revealed that the translocation of striated NPs was facilitated because its rotational degrees of freedom were constrained by the anisotropic ligand pattern, preventing the free energy of the system from falling into a deeper minimum as the NP passed through the hydrophobic core of the membrane (Figure 25A). Besides the ligand arrangement, ligand properties such as length, rigidity, and terminal chemistry have also been studied by simulations. For example, MD simulations revealed cell membrane interactions with gold NPs grafted with alkyl chains functionalized by ammonium (positive) or carboxylate (negative) groups.⁶⁰⁰ Given the electronegativity of the cell membrane, anionic NPs failed to adhere to the membrane due to electrostatic repulsion. In contrast, cationic NPs penetrated the membrane by generating defective areas across the entire surface of the upper leaflet. It shows the role of surface charge in cell membrane penetration and damage. This effect is well-known in the effect of cationic antiseptic agents on bacterial membranes.⁶⁰¹

Studies in which the lipophilicity of cationic ligands grafted on GNPs was altered showed that increasing lipophilicity promoted

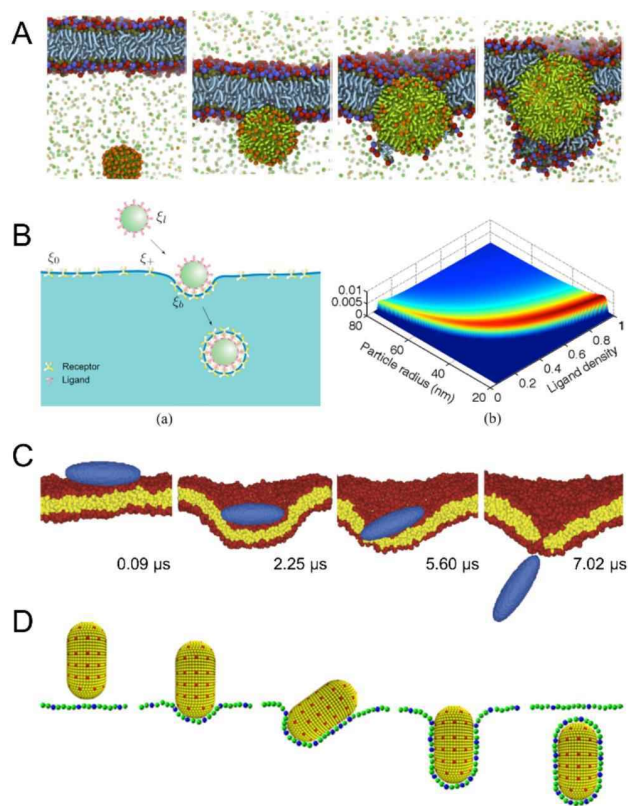


Figure 24. Size and shape effects on cell membrane interactions with NPs. (A) Simulated snapshots of cell membrane translocation of cationic NPs of different sizes. Smaller NPs are easier to translocate through the membrane. Reproduced with permission from ref 591 (Creative Commons Attribution 4.0 International License: <http://creativecommons.org/licenses/by/4.0/>). Copyright 2020. (B) Cell membrane wrapping on spherical NPs of different sizes and ligand density. Left panel: Schematic of the receptor-mediated endocytosis of NPs. Reproduced with permission from ref 565. Copyright 2015 American Chemical Society. Right panel: Interrelated effect of NP size and ligand density on the endocytosis time of an NP. Reproduced with permission from ref 565. Copyright 2015 American Chemical Society. (C) Cell membrane translocation of nonspherical NPs by MD simulations. Reproduced with permission from ref 564. Copyright 2010 Springer Nature. (D) Cell membrane wrapping on rod-like NPs revealed by MD simulations. Reproduced with permission from ref 597. Copyright 2013 American Chemical Society.

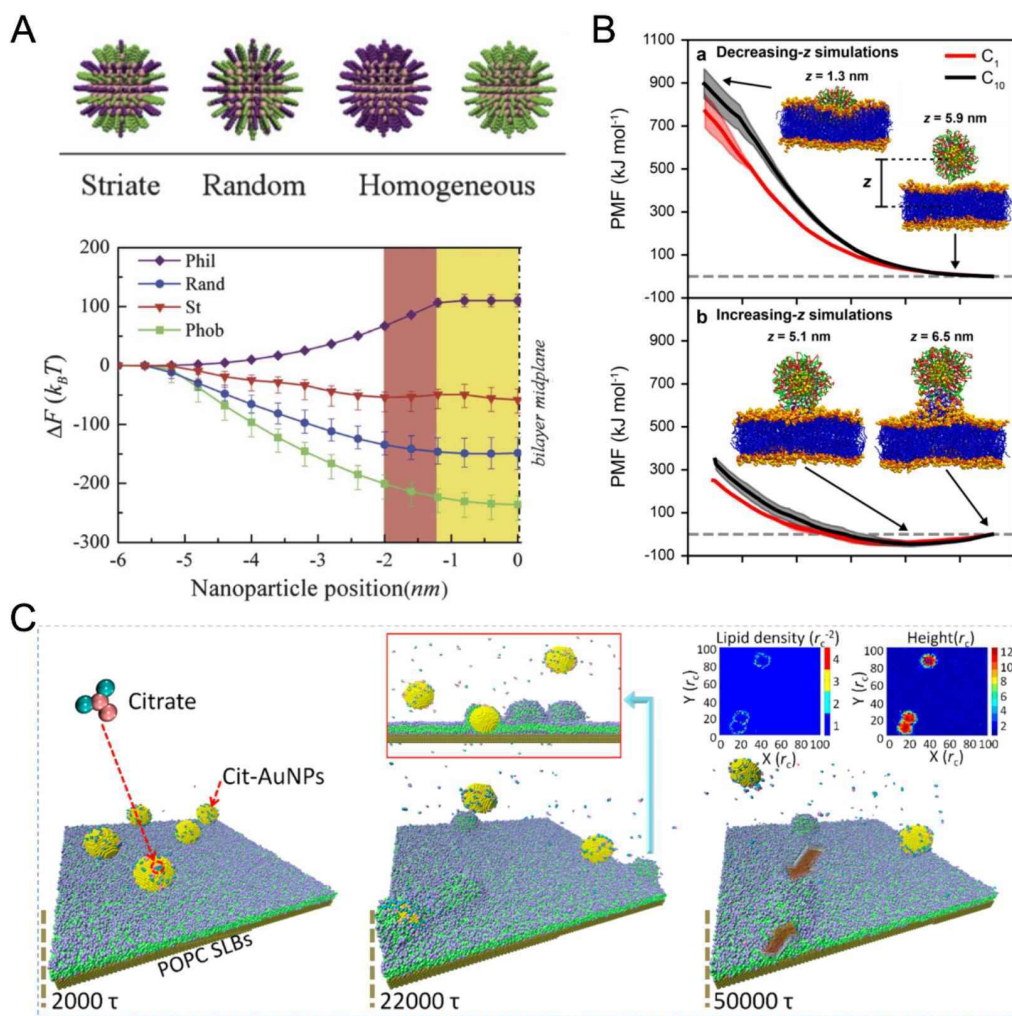


Figure 25. Effects of surface properties on NP–cell membrane interactions revealed by molecular simulation. (A) Surface structure regulated the translocation of NPs through a lipid bilayer membrane. The upper panel shows four NPs of the same core but different surface patterns. The lower panel shows the free energy changes associated with the penetration of four types of NPs as a function of the NP–membrane distance. Reproduced with permission from ref 599. Copyright 2012 Royal Society of Chemistry. (B) Irreversibility of membrane attachment of NPs regulated by ligand lipophilicity as revealed by free energy analysis. The upper figure shows the free energy change during NP approaching the membrane, and the lower curves are free energy changes as the attached NP leaves the membrane. Reproduced with permission from ref 602. Copyright 2021 American Chemical Society. (C) NP ligand (citrate) exchange and its effects on the NP–cell membrane interface revealed CGMD simulations. Reproduced with permission from ref 605. Copyright 2019 American Chemical Society.

ligand intercalation into the lipid bilayer and irreversible adsorption (Figure 25B).⁶⁰² Another DPD simulation on the effects of ligand length and bending rigidity revealed that NPs decorated with shorter and stiffer ligands were more likely to be wrapped up by cell membrane.⁶⁰³ MD simulations also revealed that structural and free energy changes in the grafted polymers accompanied the endocytosis of PEGylated NPs.⁶⁰⁴ For NPs with adsorbed citrate, MD simulations showed that these molecules could be competitively replaced by components of the cell membrane (Figure 25C). This interfacial exchange was found to alter cell membrane integrity and NP uptake efficiency.^{605,606}

4.3.3. Synergistic Cell Entry of Multiple NMs. Cell membrane interactions with multiple NMs are a cooperative process. The first experimental evidence was the cooperative budding of late domain mutated Mason-Pfizer monkey viruses, which lack individual budding activity.⁶⁰⁷ To elucidate the mechanism underlying such a cooperative effect, MD simu-

lations of cell membrane wrapping around multiple NPs were carried out.⁶⁰⁸ The results revealed that curvature-inducing NPs adsorbed on lipid bilayer membranes can experience attractive interactions that arise purely from membrane curvature (Figure 26A). Receptor-mediated endocytosis of multiple NPs was then simulated.⁶⁰⁹ The cooperative effect depended on the NP size, membrane tension and curvature, and NP concentration on the membranes.^{610,611} While smaller NPs generally cluster into a close-packed aggregate on the membrane and internalize as a whole, larger NPs tend to separate and internalize independently (Figure 26B).⁶⁰⁹ The simulated results were then experimentally validated by cryo-electron microscopy.⁶¹² To further elucidate the effect of shape on the cooperative wrapping of multiple NPs, the cell membrane wrapping of rod-like NPs was simulated, and results revealed that orientation-dependent inter-NP interactions drove the ordered arrangement of NPs on the membrane.⁶¹³ Furthermore, spherical, prolate, and oblate NPs

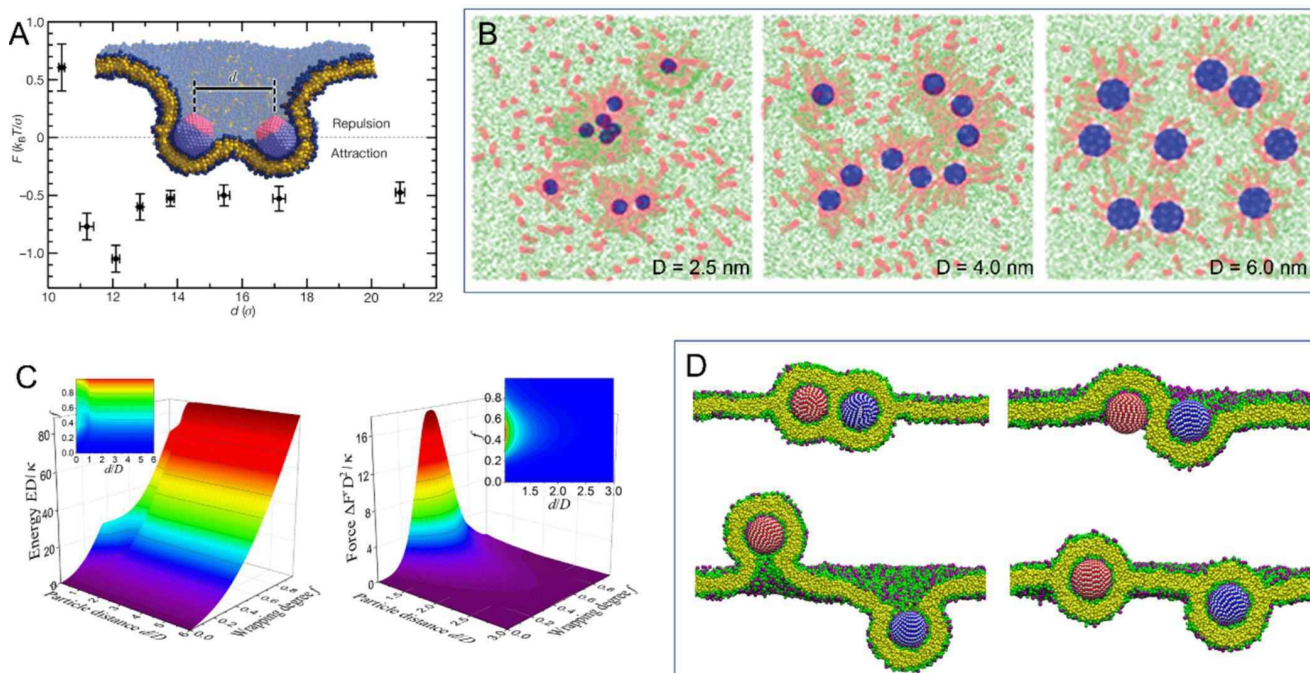


Figure 26. Cooperative effect in cell membrane interactions with multiple NPs revealed by molecular simulations. (A) Force versus distance for two NPs wrapped by a membrane. Negative forces signify attraction. Reproduced with permission from ref 608. Copyright 2007 Springer Nature. (B) Cooperative cell membrane wrapping on multiple NPs of different sizes. Smaller NPs (2.5 nm) form close-packed aggregates before internalization, and NPs of 4.0 nm aggregate into a pearl-chain-like arrangement on the membrane surface. In comparison, larger NPs (6.0 nm) are wrapped by membrane independently. Reproduced with permission from ref 609. Copyright 2012 American Chemical Society. (C) The membrane deformation energy as a function of distance and wrapping degree for two NPs at opposite membrane sides. Reproduced with permission from ref 615. Copyright 2019 Royal Society of Chemistry. (D) Simulated snapshots of cell membrane wrapping on multiple NPs at opposite sides. Reproduced with permission from ref 615. Copyright 2019 Royal Society of Chemistry.

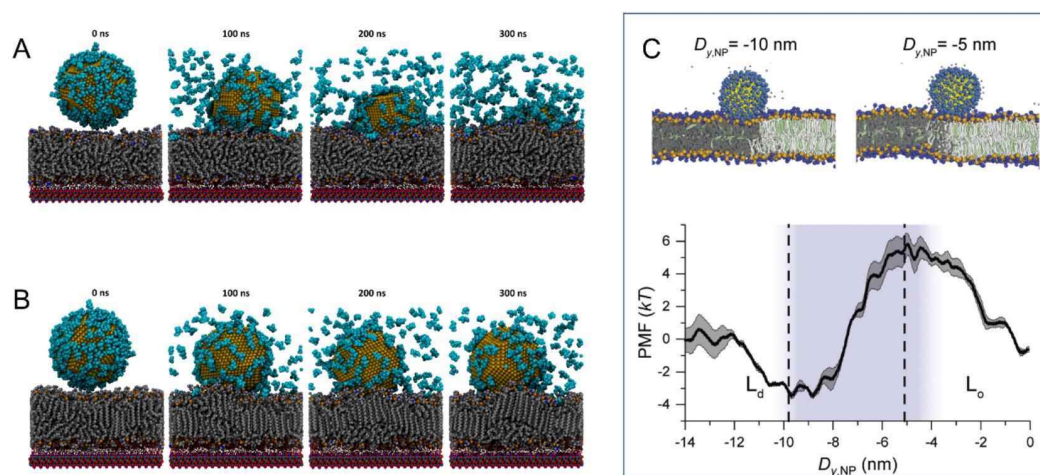


Figure 27. Influence of the phase state on NP–cell membrane interactions revealed by MD simulations. (A) Rapid internalization of NP into the membrane of the liquid-disordered phase as the capped citrates are displaced with surrounding lipids. Reproduced with permission from ref 606. Copyright 2022 American Chemical Society. (B) Failed internalization of the same NP into the membrane of the liquid-order phase. Reproduced with permission from ref 606. Copyright 2022 American Chemical Society. (C) Free energy curve for moving an adsorbed NP across the phase boundary of a two-phase bilayer. The right panel shows two snapshots of NPs located at the phase boundary. Reproduced with permission from ref 624. Copyright 2019 Royal Society of Chemistry.

were found to form multiple aggregate patterns in simulation studies.⁶¹⁴

The curvature-mediated interactions were either attractive or repulsive for multiple NPs binding to the same side of the membrane, depending on the competition between membrane bending, NP binding, and membrane protrusion. In contrast, the

interaction between two NPs binding to opposite sides of a membrane was always attractive (Figure 26C) because two NPs share the curved membrane and reduce the cost of membrane bending (Figure 26D).⁶¹⁵ Like multiple NPs with similar properties, cell entry of NPs of different types was also shown to be cooperative. For example, NPs functionalized with targeted

peptides were found to stimulate the cell uptake of coadministered unfunctionalized NPs.⁶¹⁶ MD simulations combining free energy analysis revealed that the cell membrane wrapping of functional NPs can cause bystander NPs trapped in the endocytic vesicles.⁶¹⁷ For similarly charged NPs, aggregation was simulated, and the results showed that ion bridging, lipid depletion, and membrane deformation were the major driving forces.⁶¹⁸ Additionally, a simulation study revealed that the translocation of NPs through the membrane proceeds cooperatively.⁶¹⁹

4.3.4. Influence of the Cellular Microenvironment on NM–Cell Interactions. As the primary barrier against NM invasion, the cell membrane is composed of multiple types of lipids and proteins asymmetrically distributed across two leaflets. Membranes are often segregated into domains due to interactions between saturated and unsaturated lipids. Patches usually exhibit other structural and mechanical properties, such as membrane thickness, bending rigidity, fluidity, and lipid packing. These properties influence the modes of cell membrane interactions with NPs. By contrast, domains of liquid-ordered phase state are tighter and more ordered of lipid packing, thus being more resistant to NP penetration.⁶²⁰ Recently, MD simulations, in combination with atomic force microscopy studies, probed NP interactions with supported lipid bilayers of pure fluid (DOPC, dipalmitoylphosphatidylcholine) and pure gel-phase (DPPC, dipalmitoylphosphatidylcholine) phospholipids.⁶⁰⁶ The fluid-phase DOPC membrane progressively internalized NPs as the capped citrates on NPs were displaced by the surrounding lipids (Figure 27A). However, the gel-phase DPPC membrane was more resistant to NP penetration, and only partial anchorage into the outer leaflet was achieved with fewer lipid rearrangements (Figure 27B). Similarly, increased cholesterol content hindered passive membrane insertion of amphiphilic NPs due to the reduced membrane lipid dynamics.⁶²¹ Recently, the preferential binding of NPs on membranes of various compositions was found to be an entropy-driven process.⁶²² The gel-phase membrane patch is stiffer and more resistant to bending, as analyzed by previous simulations.⁶²³ Thus, NPs are more challenging to be wrapped by the gel-phase membrane, due to the higher energy cost of membrane bending. MD simulations of NP interactions with phase-separated membranes showed a lower free energy barrier of NP adsorption on the fluid-phase membrane.⁶²⁴ A free energy minimum of the NP at the phase boundary was identified. This is attributed to the thickness difference between the two phases that enables favorable NP–lipid interactions without necessitating large deformation (Figure 27C).

Given the higher bending resistance for the raft domain to wrap NMs, it is not clear how the coexistence of membrane phase separation and protein accumulation in lipid domains dictate the interactions with NMs. Asymmetric distributions of different lipids and proteins across two leaflets can also induce nonzero spontaneous curvature.⁶²⁵ Besides, many proteins such as clathrin, caveolin, and actin have been known to participate in the NM cell uptake. However, molecular-level information on how NMs interact with these proteins is quite limited. Cell membrane wrapping on NMs with a clathrin assembly mimic has been simulated,^{626,627} showing that clathrin assembly influenced cell membrane wrapping of NMs by generating favorable curvature and altering the local membrane bending rigidity. Therefore, the size, geometry, and mobility of protein assemblage were thought to regulate the kinetics of cell membrane wrapping on NMs.

Compared to the lipid bilayer membrane interface, simulation studies have relatively overlooked the effect of the extracellular environment. Before reaching the target cell membrane surface, NMs must cross the extracellular matrix barrier, a three-dimensional network consisting of extracellular macromolecules and minerals that provide structural and biochemical support to cells. Notably, only 0.7% of the administrated NPs are reported to reach targeted tumor sites,⁶²⁸ meaning that numerous NPs are trapped at cell junctions. The short distance between cells may also allow NPs to interact simultaneously with two adjacent cells. Adhesion, bending, and protrusion of at least two membranes from these cells could generate complicated energy contributions that trap NPs at cell junctions, as revealed by MD simulations.⁶²⁹ NP transport in the modeled matrix was also simulated using the CG model of polymer networks.^{630,631} Combining MD simulation with theoretical analysis, NPs in a semiflexible matrix were shown to possess enhanced heterogeneous diffusion characterized by more evident hopping dynamics.⁶³² NP interactions with the lipid bilayer environment interfered with adjacent proteins' normal structure and functioning. For example, the activity of gramicidin A ion channels was indirectly reduced by anionic NPs by altering the surrounding bilayer's mechanical properties.⁴² Similarly, MD simulations showed that graphene could insert into the membrane to extract surrounding lipids, thus activating integrin by facilitating the separation of its two transmembrane domains.⁶³³ Previous simulations have provided valuable molecular information regarding NP interactions with cell membranes or membrane proteins. However, these data are acquired from many independent simulations performed under different conditions. Therefore, filtering and integrating these data based on the uniform standard is essential for comprehensively understanding this critical issue.

4.4. MD Simulations of NM Interactions with Functional Proteins

Proteins are abundant and diverse in biology and perform most biological functions. NMs inevitably interact with proteins upon entering biological systems. These interactions may interfere with the normal functions of proteins by altering the protein structures or disrupting protein–protein interactions. Due to their high surface energy, many NMs recruit surrounding proteins to form a protein corona. Thus, the original surface properties of NMs may or may not be modified by coronas. Therefore, understanding how NMs with diverse physicochemical properties interact with specific proteins under specified conditions is essential for understanding the nanotoxicology induced by NMs. Although numerous studies on nanotoxicology have considered the effects of proteins, the nanobio interface at the atomistic level is still rather arcane, as the binding sites, strengths, and subtle changes in protein structures remain to be elucidated. In this regard, MD simulation is useful for revealing molecular details of interactions between NPs and proteins.

4.4.1. Molecular Insights into Protein Adsorption by NMs. Based on the binding affinities and dissociation rates, the layers of NM-adsorbed proteins can be divided into the hard and soft corona. Recently, using an *in situ* fishing method, the soft and hard corona proteins were separately identified.⁶³⁴ An important question arises of how coronas form by competitive/cooperative protein–protein and protein–NM interactions. Abundant proteins bind first and are gradually displaced by less abundant but higher-affinity proteins. At the NM–protein

interface, where the strong interactions usually define the hard corona, it is essential to identify the site of protein binding to NMs. The adsorption conformation or orientation determines the surface exposed to alter NM surface properties. Earlier molecular-level studies on protein coronas focused on single serum protein adsorption on CNTs or GNPs.^{635,636} Efforts have been made to manipulate the composition and conformation of protein coronas via changing the physicochemical properties of NPs.⁶³⁷ With a change in NP hydrophobicity, adsorbed pulmonary surfactants on NPs were found to vary from monolayer to bilayer. Proteins adopted parallel and perpendicular orientations relative to the NP surface (Figure 28A).⁶³⁸

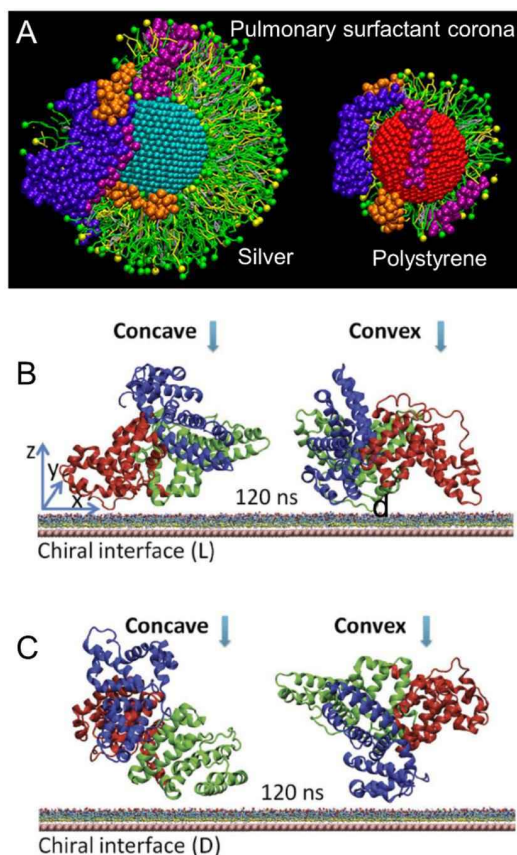


Figure 28. Protein adsorption on NPs dictated by the physicochemical properties. (A) Simulated snapshots of pulmonary surfactant corona formed on silver and polystyrene NPs. Reproduced with permission from ref 638. Copyright 2017 American Chemical Society. (B, C) Different orientations of BSA adsorption on gold NPs decorated with L- (B) and D-penicillamine (C) were revealed by MD simulations. Reproduced with permission from ref 639. Copyright 2018 John Wiley and Sons.

Surface chirality was also found to modulate the adsorption of serum proteins on NPs (Figure 28B and C). The importance of salt bridge interactions was revealed by MD simulation as a major driving force for the protein–NP interactions.⁶³⁹ Taking graphene and gold as representative surfaces, an MD simulation study showed that the surface hydroxyl groups on NPs played a role in regulating their interactions with proteins. This process could be used to manipulate the circulation of NPs in blood.⁶⁴⁰

The most frequently reported outcome of protein adsorption on NPs is reduced cytotoxicity. For example, DPD simulations revealed that hydrophobic NPs fail to insert into membranes

after serum protein adsorption.⁶⁴¹ The membrane damage induced by graphene nanosheets was also mitigated by protein corona. MD simulations revealed that the adsorbed proteins weakened the interactions between lipids and graphene surfaces due to smaller available surface area and unfavorable steric effects.⁶⁴² However, a consensus on the impact of protein corona on NP cytotoxicity has not been reached due to the complexity of the protein corona and the cell membrane surface. For example, some protein components, such as low-density lipoprotein and immunoglobulin G, in corona were found to recognize specific receptors on the cell membrane and mediate cell responses.⁶⁴³

Despite much research on protein corona, a much better understanding at the molecular level is required. Most previous simulations only considered a single protein adsorbed on NMs. As many types of proteins exist in serum, MD simulations of mixtures of various proteins and NMs showed that the protein layer on NMs was formed via the competitive adsorption of proteins (Figure 29A). This largely influenced the binding affinity and diffusivity of adsorbed proteins (Figure 29B).⁶⁴⁴ However, achieving a thermodynamically steady state for a complex multicomponent system is challenging as metastable states usually exist for individual proteins. Therefore, enhanced sampling methods should be used to allow the sampling of larger portions of the configuration space of complex systems in reasonable simulation times.⁶⁴⁵ Although the physicochemical properties and biological activities of NMs can be altered by protein corona, there is no consensus about whether the original NM properties are completely cloaked by coronas or still play a role in subsequent nanobio interactions. Earlier experiments have suggested that the NP core can play an important role in cell uptake and cytotoxicity under controlled size, shape, and surface coatings, suggesting coronas may not completely conceal the original NP properties.⁸⁷ Subsequent simulations supported this hypothesis, revealing that the adsorbed molecules could adopt conformations with specific groups exposed, modifying the NP surface properties and affecting subsequent cell interactions.⁹³ Moreover, adsorbed molecules undergo frequent exchanges whose on/off rates provide brief windows of exposure of the original NP surface for cell recognition.⁹⁴ Such cascading events involve competitive and cooperative interactions between different components at the nanobio interface. This aspect remains to be elucidated.

4.4.2. Protein Denaturation and Dysfunction Induced by NMs. Strong NM–protein interactions may block the protein binding/active sites, damage the protein structure, or disrupt the protein–protein interactions. Protein denaturation or misfolding induced by NMs is an important mechanism underlying nanotoxicity. Computational studies have investigated interactions between diverse NMs and functional proteins to optimize such interactions and improve bionano recognition. Given the surface heterogeneity of NMs and structural diversity of proteins, van der Waals, electrostatic, hydrogen bonding, and salt bridge interactions are likely to be dominant between NPs and proteins. NM–protein interactions are influenced by NM physicochemical properties, such as size, shape, hydrophobicity, and charge.

Multinano-second MD simulations on interactions between fullerene and antibodies showed that the hydrophobic interactions and the shape complementarity were major reasons for fullerene recognition by the antibody.⁶⁴⁶ Driven by electrostatic interactions, gold nanoclusters anchored with cationic peptides were found to rapidly bind to the negatively

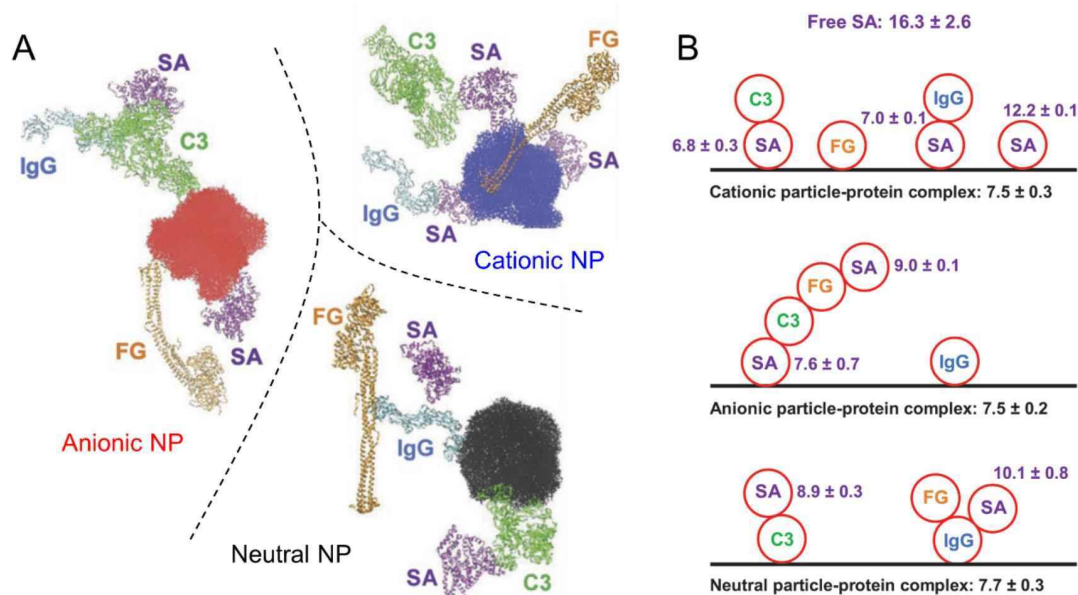


Figure 29. Competitive adsorption of multiple proteins on NPs. (A) Adsorption of four types of plasma proteins on neutral (black), cationic (blue), and anionic (red) PS NPs predicted by MD simulations. Reproduced with permission from ref 644. Copyright 2020 John Wiley and Sons. (B) Schematic of coadsorption of multiple proteins on NPs. Reproduced with permission from ref 644. Copyright 2020 John Wiley and Sons.

charged active site of the enzyme thioredoxin reductase 1 and inhibit the protein activity.⁶⁴⁷ MD simulations revealed that carbon nanotubes could insert into the hydrophobic core of proteins, blocking the active ligand binding sites and leading to loss of protein function.⁶⁴⁸ To explain why aggregated carbon nanotubes are effective in the treatment of methamphetamine addiction,⁶⁴⁹ MD simulations revealed that dopamine-related proteins adsorb onto carbon nanotubes and the aggregated nanotubes are more effective in inhibiting the activity of tyrosine hydroxylase.⁶⁵⁰ In contrast to fullerenes and carbon nanotubes, 2D graphene-based NPs can exhibit strong inhibition of enzyme activity.⁶⁵¹ Strong π - π stacking interactions between graphene and aromatic residues induce denaturation or deformation of active sites of enzymes,⁶⁵² indicating the critical role of the NP shape or curvature in determining its interactions with proteins. Graphene nanosheets were also found to intercalate into the interstrand gap of actin tetramer and cause the breakdown of the tetramer, eventually retarding cellular migration.⁶⁵³ Similar mechanisms were validated for the NP-induced endothelial leakiness. The dimer stability of extracellular domains formed by E-cadherin proteins of two adjacent cells can be reduced by the intercalation of graphene nanosheets.⁶⁵⁴

Besides soluble proteins, membrane proteins are essential for cellular functioning and are potential targets of NP-cell interactions. MD simulations showed that graphene QDs interacted with three representative K^+ channels.⁶⁵⁵ Graphene QDs could form clusters over the extracellular entrance of ion channels, blocking the pore and decreasing the ion current. Electrophysiology measurements demonstrated that the ion channel activity of gramicidin A embedded in lipid bilayers could be reduced by anionic GNPs without disrupting membrane integrity. MD simulations revealed that anionic NPs do not directly interact with the embedded channel but perturb the local properties of the lipid bilayer.⁴² Similar mechanisms were also revealed for graphene-induced activation of $\alpha_v\beta_8$ integrins by extracting surrounding lipids.⁶³³

4.5. ML Approaches to Analyze and Enhance Molecular Simulations

Even though molecular simulation methods have made numerous contributions to better understand the molecular mechanism related to nanotoxicity, they still face the challenges of dealing with massive amounts of data and large complex simulated systems. These challenges greatly limited molecular simulations' length and time scale in nanotoxicology. Fortunately, the above-described ML approaches are well suited and expected to resolve these challenges. Therefore, in this section, we will introduce the applications of ML approaches to analyze and enhance molecular simulations.

Typically, the workflow of ML approaches to molecular simulations is similar to that applied in other fields and mainly includes the generation of molecular simulation data, development of ML models, and prediction of molecular simulation properties. The applications of ML approaches to molecular simulations mainly include analyzing MD trajectories, fitting potential energy surfaces and free energy surfaces, optimization of coarse-graining, force fields, sampling, and thermodynamics, among others.^{656–660} For instance, the DNN model trained on QM calculations could learn the accurate potential for organic molecules and finally acquire DFT accuracy at force field computational cost.⁵³⁴ Similarly, deep-learning-based potentials were also used to speed up the molecular simulation of chemical reactions.⁶⁶¹ A recent study proposed an ML-based dynamic-importance sampling for adaptive multiscale simulations.⁶⁶² The ML-based sampling made it possible to perform macro length- and time-scale simulations at the effective precision of the microscale. Recently, it was proven that ML approaches could directly map the electronic structure of a molecule to CG pseudoatom configurations.⁶⁶³ Furthermore, a series of ML-based packages, such as TorchMD,⁶⁶⁴ DeePMD-kit,⁶⁶⁵ and TensorMol,⁶⁶⁶ has also been developed for speeding up molecular simulations. However, the combination of ML and molecular simulations is still relatively limited in nanotoxicology. We hope that these successful examples from other

fields would facilitate the application of ML approaches to molecular simulations in nanotoxicology.

4.6. Summary of Applying Molecular Simulations in Nanotoxicology

An increasingly large amount of nanotoxicity data is being generated, but it is not properly curated. The mechanisms by which most NMs interact with biological systems have commonalities such as cell or protein binding, cell entry, ROS generation, and induced cytotoxicity. Molecular simulation provides a significant promise to bridge the gap between nanotoxicity data lakes and a mechanistic understanding of nanotoxicology. Unlike experimental methods, molecular simulations are more useful for designing and constructing arrays of NMs with defined physicochemical properties, such as core material, size, shape, surface charge, and coatings. Atomic-level information about these NMs can be easily used to generate useful nanodescriptors. High-throughput simulations using these NMs with well-defined properties can reveal mechanisms underlying NM interactions with specific biomolecules and interfaces. In addition, the effects of specific NM properties can be elucidated. Molecular simulations of NM interactions with specific molecules can also provide information complementary to experiments. These data are very useful for training interpretable AI models and unraveling the relationships between nanostructure and nanotoxicology.

Challenges exist in applying molecular simulations to nanotoxicology studies. Biological responses to NMs are usually cascading events, and NMs in biological systems interact simultaneously or sequentially with multiple molecules. However, such complexity has yet to be attained by molecular simulations. Simulations of NM interactions with single molecules or interfaces need to be improved, and further simulations of NM interactions with multiple related molecules are needed. Current simulations' length and time scales are much smaller than those in actual experiments. This discrepancy hinders direct quantitative comparison between simulation results and experimental measurements. Developing a multiscale simulation method can help resolve this problem by extending the time and length scales to be more appropriate for simulating real world NM interactions with biomolecules. NM-induced generation of excess ROS also needs significantly more exploration. In future studies, chemical mechanisms should be merged with physical mechanisms to comprehensively understand the basis for nanotoxicology.

5. CONCLUDING REMARKS AND FUTURE PERSPECTIVES

Large nanotoxicity data sets have been accumulated through 20+ years of research. To develop safe nanotechnology, we must utilize AI and molecular simulation to transform these valuable nanotoxicology data into useful, predictive QNTR models and mechanistic information. However, there are still substantial challenges ahead.

A major challenge is generating larger, rich, comprehensive, high-quality, and problem-specific data. Roadblocks to progress have been due to several reasons: lack of standardized NM synthesis methods, characterization, biological testing, data generation, reporting, collection, and processing protocols. Experimentalists and modelers must collaborate to improve data quantity and quality and to make consideration of data required for models the first rather than last thought in experimental planning. Model-friendly databases designed under the FAIR

principles are critical for improving data availability for widespread computational use in nanotoxicology. Accurate characterization and annotation of 3D nanostructures are indispensable and critical technical challenges in the short to medium term.

Encoding nanostructures, physicochemical properties, and environmental metadata in nanospecific mathematical descriptors or features is another important challenge to be addressed. Current nanodescriptors are not up to the task. These more advanced nanodescriptors should have, as a minimum, the following characteristics: sufficient information about the structural features and the physicochemical properties; they should be explainable and guide the optimization of bespoke NMs; they must be easily calculated and not require a large amount of computing time and resources; descriptors should be repeatable and their generation methods or codes made publicly accessible.

Another ongoing challenge in ML is increasing model transparency, allowing users to understand how a model learns and what information is behind its decisions. Rather than using feature importance methods to explain black-box models, employing inherently interpretable models is preferable.⁴²⁶ Finally, the value and robustness of ML models and molecular simulation results should be demonstrated more completely by using them to design new, useful NMs and by more extensive experimental testing of model predictions. Models that are never used or tested have little value. ML codes and nanotoxicity data must also be more readily accessible in publicly available repositories such as GitHub, and they must be rigorously described and annotated. The value of data is enhanced when suitable nanospecific ontologies are developed and used.

A primary aim of nanotoxicology is to identify or design NMs with specific desirable properties, such as biocompatibility or higher cellular uptake, and low or no adverse biological effects. It has been shown that this aim can at least be partially achieved using ML, molecular simulation, and virtual screening. Conspicuously, few real success stories have been using these approaches so far. Recent advances in inverse design strategy can change this landscape, generating entirely new molecules with specific properties through continuous optimization. We anticipate substantial growth in inverse design using deep learning to augment or replace the traditional rational design and virtual screening methods and speed up the success rate of bespoke NM design.

Modeling nanosafety in ecosystems is currently one of the most challenging and complex steps in computational nanotoxicology. The future directions should include but not be limited to the following points: modeling processes affecting the environmental fate of NMs such as oxidation, aggregation, and degradation; modeling bioaccumulation of NMs in the food chain; substantially more accurate and quantitative *in vivo* prediction of nanotoxicity by incorporating *in vitro* data; estimating human exposure risk to NMs; and including a more realistic biological system in molecular simulation and ML models. Additionally, modeling perturbation responses is also a great challenge. This is especially useful to account for the distribution of structures in realistic nanotoxicity environments. Recently, ML methods used to predict surface interactions in catalysis may be adapted and generalized to biological interactions to sample the distribution of ligand environments and properties of NMs more efficiently.^{667–669}

Molecular simulation has become an indispensable tool in nanotoxicology studies. While QM-based methods can describe

chemical reactions relevant to NM-induced toxicity, MM-based methods can deal with much larger systems and elucidate physical mechanisms underlying NM interactions with biomolecules. Compared to experiments, molecular simulation has significant advantages: it is possible to construct arrays of NMs with defined nanostructures and properties and generate nanodescriptors with atomic accuracy; parallel simulations can be performed on multiple NMs to elucidate mechanisms underlying physical and chemical processes relevant to nanotoxicity; important mechanism data complementary to data acquired by experiments can be generated; a synergistic combination of AI and molecular simulation techniques can help fill the gap in nanotoxicity data.

Despite yielding important mechanism insight, the quality of data generated by molecular simulation is hampered by the following shortcomings:

1. The accuracy of MD simulations is determined by the force fields used. For NMs, they are empirical and conceptually overly simplistic descriptions of intra- and intermolecular interactions. Although experimental data can rectify some parameters, the quality of experiments needs improvement.
2. Limited computing power means that current simulation length and time scales are much smaller than those in experiments. This discrepancy hinders direct quantitative comparison between simulation results and experimental measurements. Coarse graining is one possible way to extend time and length scales to be appropriate for simulating NM interactions with biomolecules. However, atomistic details are inevitably sacrificed, reducing the simulation precision. In this regard, efforts have been made to develop multiscale simulation methods to extend simulation scales with improved accuracy. Quantum computing may break this nexus in the medium to long term, and in the interim, deep learning methods should allow larger and longer simulations to be achieved.
3. Most simulations treat NM interactions as a single molecule or an interface. However, cell signaling events are based on a network of protein–protein interactions, and NMs in biological systems are expected to interact simultaneously or sequentially with multiple proteins. Therefore, the simulation of NM interactions with single proteins needs to be expanded, and further simulations of NM interactions with multiple related proteins are required.

Molecular simulation methods excel at toxicity mechanism analysis and data generation but typically need more efficiency and scale. AI approaches sometimes lack interpretability but could provide efficiency and scale with the data generated from experiments and simulations. Therefore, the combination of AI and molecular simulation will be decisive in the enhanced parametrization of force fields, improved simulation data analysis, effective nanodescriptors generation, and more transparent and explainable models.⁶⁵⁶

Despite the continued challenges, there are also many successful examples of accelerating the transformation of nanotoxicity data into critical information. In a recent study, the MD simulation and QM calculations were used to uncover the essential atomistic details that chiral NPs site-selectively cleaved capsid in tobacco mosaic virus under sunlight.⁶⁷⁰ Similarly, the MD simulation identified a novel CNT–receptor binding mode mediated by multiple aromatic–aromatic

interactions.⁶⁷¹ Recently, a CNN deep learning model was successfully trained and employed to assist single-vessel analysis of NP permeability in tumor vasculatures.⁶⁷² Using NP library and computational methods (i.e., molecular docking and ML), two novel broad-spectrum adjuvants were identified and experimentally validated.⁶⁷³ Similarly, the NP library and machine learning allowed us to identify seven new GNPs with desired bioactivities, which were experimentally synthesized and confirmed.⁸⁶ These successful examples provided some initial guiding principles for applying AI and molecular simulation in elucidating the molecular mechanism of nanotoxicity, predicting the adverse effects of new NMs, and designing novel NMs with desired properties.

Transforming big data into critical information is a common problem in any data-driven fields, such as drug discovery,¹³ medical diagnosis,¹⁶ agricultural science,⁶⁷⁴ public health⁶⁷⁵ and environmental science.⁶⁷⁶ The difference is the type of data we are dealing with. For example, in the field of drug discovery, we are faced with a large amount of data on the pharmacokinetics of drug molecules; machine learning and other computational methods are expected to learn from these big data and help us discover novel molecules with better therapeutic effects. Therefore, the experience accumulated in a certain field can be used for reference in other fields. We hope that our proposal will galvanize broader discussion from different fields in the future. We strongly encourage researchers interested in this topic to contact us and join this discussion to accelerate the transformation of big data into critical information.

In summary, AI and molecular simulation show considerable potential for closing the gap between the nanotoxicity data and the critical information. However, the applications of these two methods in nanotoxicology are still in their infancy and face significant obstacles, both experimentally and computationally. By adopting AI and molecular simulation methods, substantial benefits will accrue to various stakeholders across the breadth of nanotoxicology research. We are confident that AI and molecular simulation will greatly impact nanotoxicology research processes and help accelerate the path to useful, safe, and sustainable nanotechnology.

AUTHOR INFORMATION

Corresponding Author

Bing Yan — *Institute of Environmental Research at the Greater Bay Area, Key Laboratory for Water Quality and Conservation of the Pearl River Delta, Ministry of Education, Guangzhou University, Guangzhou 510006, China; orcid.org/0000-0002-7970-6764; Email: drbingyan@gzhu.edu.cn*

Authors

Xiliang Yan — *Institute of Environmental Research at the Greater Bay Area, Key Laboratory for Water Quality and Conservation of the Pearl River Delta, Ministry of Education, Guangzhou University, Guangzhou 510006, China; orcid.org/0000-0003-4173-6228*

Tongtao Yue — *Key Laboratory of Marine Environment and Ecology, Ministry of Education, Institute of Coastal Environmental Pollution Control, Ocean University of China, Qingdao 266100, China; orcid.org/0000-0002-8329-167X*

David A. Winkler — *Monash Institute of Pharmaceutical Sciences, Monash University, Parkville, Victoria 3052, Australia; School of Pharmacy, University of Nottingham, Nottingham NG7 2QL, U.K.; Department of Biochemistry*

and Chemistry, La Trobe Institute for Molecular Science, La Trobe University, Melbourne, Victoria 3086, Australia;
orcid.org/0000-0002-7301-6076

Yongguang Yin – State Key Laboratory of Environmental Chemistry and Ecotoxicology, Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, Beijing 100085, China; orcid.org/0000-0002-7287-8598

Hao Zhu – Department of Chemistry and Biochemistry, Rowan University, Glassboro, New Jersey 08028, United States;
orcid.org/0000-0002-3559-6129

Guibin Jiang – State Key Laboratory of Environmental Chemistry and Ecotoxicology, Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, Beijing 100085, China; orcid.org/0000-0002-6335-3917

Complete contact information is available at:
<https://pubs.acs.org/10.1021/acs.chemrev.3c00070>

Notes

The authors declare no competing financial interest.

Biographies

Xiliang Yan received his Ph.D. from Shandong University in computational nanotoxicology in 2020. He is currently working as a lecturer at the Institute of Environmental Research at Greater Bay, Guangzhou University. His current research interests are centered on understanding nanobio interactions through combining big data, machine learning, and molecular simulations.

Tongtao Yue received his Ph.D. in chemical engineering and technology from the Beijing University of Chemical Technology in 2012. From 2012 to 2020, he worked as an associate professor at the China University of Petroleum (East China). In 2020, he became a full professor at the College of Environmental Science and Engineering of Ocean University of China. Prof. Yue's research interests include understanding interactions between various nanoparticles and biological systems at the molecular level and designing new types of nanomaterials for biomedical applications.

David A. Winkler received his Ph.D. degree from Monash University in 1980. He was a research fellow at the Victorian College of Pharmacy (now Monash Institute of Pharmaceutical Sciences, MIPS) and then Research Scientist at the Defence Science and Technology Organization in Adelaide. In 1983, he joined CSIRO as a Senior Research Scientist, remaining there until 2017 as a Senior Principal Research Scientist. He then joined La Trobe University as Professor of Biochemistry and Chemistry and is additionally an adjunct Professor of Medicinal Chemistry at MIPS and visiting Professor in Pharmacy at the University of Nottingham. He has served on the National Committee for Chemistry of the Australian Academy of Science, on IUPAC nomenclature committees, as President of both the Federation of Asian Chemical Societies and Asian Federation for Medicinal Chemistry, and as Chairman of the Board of the Royal Australian Chemical Institute. He is the recipient of the CSIRO Medal for Business Excellence, the RACI Adrien Albert Award, the Herman Skolnik award from the ACS, and Distinguished Fellowship from Royal Academy of Engineering.

Yongguang Yin studied chemistry at Chongqing University and completed his Ph.D. degree in environmental sciences in 2008 at Research Center for Eco-Environmental Sciences (RCEES), Chinese Academy of Sciences (CAS). He was appointed as Professor for Environmental Sciences at RCEES, CAS, in 2017. His research focuses on analytical chemistry, environmental transformation, toxicity, and remediation of nanomaterials/toxic metals, including Hg, Ag, and Cd.

Hao Zhu currently is a Professor in the Department of Chemistry and Biochemistry at Rowan University of USA. He received his Ph.D. in Computational Chemistry from Case Western Reserve University in 2002. Dr. Zhu has authored or coauthored over 100 peer-reviewed publications and book chapters in the applications of machine learning and big data modeling to chemical toxicity assessments, computer-aided drug discovery, and rational nanomaterial design.

Guibin Jiang is a Professor at RCEES, CAS. Prof. Jiang serves as an Associate Editor of *Environmental Science & Technology* and the Academician of Chinese Academy of Sciences. Prof. Jiang's research mainly focuses on environmental analytical chemistry and toxicology of persistent toxic substances (PTS), including metals, emerging organic pollutants, and nanomaterials. Prof. Jiang has contributed more than 1100 papers in peer-reviewed scientific journals with more than 800 lectures including plenary and keynote lectures at international and national meetings.

Bing Yan is Cheung Kong Scholar Professor at Guangzhou University. He received his Ph.D. degree from Columbia University in 1990. He conducted postdoctoral research at the University of Cambridge, U.K., and the University of Texas Medical School in Houston from 1990 to 1993. From 1993 to 2005, he worked at Novartis Biomedical Research Institute and Bristol-Myers Squibb. He has been a full Professor (member) at the Department of Chemical Biology and Therapeutics, St. Jude Children's Research Hospital, in Memphis, TN, from 2007 to 2012, and Cheung Kong Scholar Professor at Shandong University and then Guangzhou University, China, since 2005. He now serves as Co-Editor-in-Chief for *Ecotoxicology and Environmental Safety* and Associate Editor for *NanoImpact*. He was a book series editor for *Critical Reviews in Combinatorial Chemistry* published by the Francis & Taylor Group. He is interested in understanding the health effects of environmental pollution, including nanotoxicology.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (22036002, 22106025), the introduced innovative R&D team project under the "The Pearl River Talent Recruitment Program" of Guangdong province (2019ZT08L387), the National Basic Research Program of China (2022YFC3701301), the Basic and Applied Basic Research Foundation of Guangzhou, China (202201010541), and the Natural Science Foundation of Shandong Province (ZR2021YQ23).

ABBREVIATIONS

| | |
|-----------|---|
| AI | Artificial Intelligence |
| NMs | Nanomaterials |
| NPs | Nanoparticles |
| ML | Machine Learning |
| DFT | Density Functional Theory |
| MD | Molecular Dynamics |
| DPD | Dissipative Particle Dynamics |
| CG | Coarse-Grained |
| MWCNTs | Multi-Walled Carbon Nanotubes |
| HTS | High-Throughput Screening |
| LDH | Lactate Dehydrogenase |
| OECD | Organization for Economic Cooperation and Development |
| SDF | Structure-Data Files |
| PDB | Protein Data Bank |
| caNanoLab | Cancer Nanotechnology Laboratory |
| FAIR | Findability, Accessibility, Interoperability, Reuse |
| CAS | Chemical Abstracts Service |

| | |
|---------|--|
| RNN | Recurrent Neural Networks |
| GAN | Generative Adversarial Networks |
| CNN | Convolutional Neural Networks |
| GNN | Graph Neural Networks |
| RF | Random Forest |
| MLR | Multiple Linear Regression |
| SVM | Support Vector Machine |
| XGBoost | eXtreme Gradient Boosting |
| kNN | k-Nearest Neighbors |
| LASSO | Least Absolute Shrinkage and Selection Operator |
| MLREM | Multiple Linear Regression with Expectation Maximization |
| TEM | Transmission Electron Microscopy |
| AFM | Atomic Force Microscopy |
| SEM | Scanning Electron Microscopy |
| DLS | Dynamic Light Scattering |
| VASP | Vienna Ab initio Simulation Package |
| HOMO | Highest Occupied Molecular Orbital |
| LUMO | Lowest Unoccupied Molecular Orbital |
| SMILES | Simplified Molecular Input Line Entry System |
| GNPs | Gold Nanoparticles |
| MONPs | Metal Oxide Nanoparticles |
| CORAL | CORrelations And Logic |
| NInChI | International Chemical Identifier for NPs |
| SiRMS | Simplex Representation of Molecular Structure |
| LDM | Liquid Drop Model |
| MLB | Metal–Ligand Binding |
| SNPs | Silver Nanoparticles |
| SWCNTs | Single-Wall Carbon Nanotubes |
| ROS | Reactive Oxygen Species |
| AD | Applicability Domain |
| DNN | Deep Neural Networks |
| LIME | Local Interpretable Model-agnostic Explanations |
| SHAP | SHapley Additive exPlanations |
| CAM | Class Activation Mapping |
| AOP | Adverse Outcome Pathway |
| MIE | Molecular Initiating Event |
| MOF | Metal–Organic Framework |
| BP | Black Phosphorus |
| QM | Quantum Mechanics |
| MM | Molecular Mechanics |

REFERENCES

- (1) Kim, B. Y.S.; Rutka, J. T.; Chan, W. C.W. Nanomedicine. *N. Engl. J. Med.* **2010**, *363*, 2434–2443.
- (2) Hussein, A. K. Applications of Nanotechnology in Renewable Energies - A Comprehensive Overview and Understanding. *Renew. Sustain. Energy Rev.* **2015**, *42*, 460–476.
- (3) Nasrollahzadeh, M.; Sajadi, S. M.; Sajjadi, M.; Issaabadi, Z. Applications of Nanotechnology in Daily Life. *Interface Sci. Technol.* **2019**, *28*, 113–143.
- (4) Wu, J.; Liu, W.; Xue, C.; Zhou, S.; Lan, F.; Bi, L.; Xu, H.; Yang, X.; Zeng, F. D. Toxicity and Penetration of TiO₂ Nanoparticles in Hairless Mice and Porcine Skin after Subchronic Dermal Exposure. *Toxicol. Lett.* **2009**, *191*, 1–8.
- (5) Tsuda, A.; Donaghey, T. C.; Konduru, N. V.; Pyrgiotakis, G.; Van Winkle, L. S.; Zhang, Z.; Edwards, P.; Bustamante, J. M.; Brain, J. D.; Demokritou, P. Age-Dependent Translocation of Gold Nanoparticles across the Air-Blood Barrier. *ACS Nano* **2019**, *13*, 10095–10102.
- (6) Souza, M. R.; Mazaro-Costa, R.; Rocha, T. L. Can Nanomaterials Induce Reproductive Toxicity in Male Mammals? A Historical and Critical Review. *Sci. Total Environ.* **2021**, *769*, 144354.
- (7) Guo, Z.; Zhang, P.; Chakraborty, S.; Chetwynd, A. J.; Monikh, F. A.; Stark, C.; Ali-Boucetta, H.; Wilson, S.; Lynch, I.; Valsami-Jones, E. Biotransformation Modulates the Penetration of Metallic Nanomateri-

- als across an Artificial Blood-Brain Barrier Model. *Proc. Natl. Acad. Sci. U. S. A.* **2021**, *118*, e2105245118.
- (8) Krug, H. F.; Wick, P. Nanotoxicology: An Interdisciplinary Challenge. *Angew. Chem., Int. Ed.* **2011**, *50*, 1260–1278.
- (9) Oberdörster, G.; Oberdörster, E.; Oberdörster, J. Nanotoxicology: An Emerging Discipline Evolving from Studies of Ultrafine Particles. *Environ. Health Perspect.* **2005**, *113*, 823–839.
- (10) Colvin, V. L. The Potential Environmental Impact of Engineered Nanomaterials. *Nat. Biotechnol.* **2003**, *21*, 1166–1170.
- (11) Service, R. F. Nanomaterials Show Signs of Toxicity. *Science* **2003**, *300*, 243.
- (12) Ekins, S.; Puhl, A. C.; Zorn, K. M.; Lane, T. R.; Russo, D. P.; Klein, J. J.; Hickey, A. J.; Clark, A. M. Exploiting Machine Learning for End-to-End Drug Discovery and Development. *Nat. Mater.* **2019**, *18*, 435–441.
- (13) Vamathevan, J.; Clark, D.; Czdrowski, P.; Dunham, I.; Ferran, E.; Lee, G.; Li, B.; Madabhushi, A.; Shah, P.; Spitzer, M.; et al. Applications of Machine Learning in Drug Discovery and Development. *Nat. Rev. Drug Discovery* **2019**, *18*, 463–477.
- (14) Butler, K. T. T.; Daniel, W.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine Learning for Molecular and Materials Science. *Nature* **2018**, *559*, 547–555.
- (15) Raccuglia, P.; Elbert, K. C.; Adler, P. D. F.; Falk, C.; Wenny, M. B.; Mollo, A.; Zeller, M.; Friedler, S. A.; Schrier, J.; Norquist, A. J. Machine-Learning-Assisted Materials Discovery Using Failed Experiments. *Nature* **2016**, *533*, 73–76.
- (16) Esteva, A.; Robicquet, A.; Ramsundar, B.; Kuleshov, V.; DePristo, M.; Chou, K.; Cui, C.; Corrado, G.; Thrun, S.; Dean, J. A Guide to Deep Learning in Healthcare. *Nat. Med.* **2019**, *25*, 24–29.
- (17) Myszczyńska, M. A.; Ojames, P. N.; Lacoste, A. M. B.; Neil, D.; Saffari, A.; Mead, R.; Hautbergue, G. M.; Holbrook, J. D.; Ferraiuolo, L. Applications of Machine Learning to Diagnosis and Treatment of Neurodegenerative Diseases. *Nat. Rev. Neurol.* **2020**, *16*, 440–456.
- (18) Tropsha, A.; Mills, K. C.; Hickey, A. J. Reproducibility, Sharing and Progress in Nanomaterial Databases. *Nat. Nanotechnol.* **2017**, *12*, 1111–1114.
- (19) Winkler, D. A. Role of Artificial Intelligence and Machine Learning in Nanosafety. *Small* **2020**, *16*, 2001883.
- (20) Basei, G.; Hristozov, D.; Lamon, L.; Zabeo, A.; Jeliakova, N.; Tsiliki, G.; Marcomini, A.; Torsello, A. Making Use of Available and Emerging Data to Predict the Hazards of Engineered Nanomaterials by Means of in Silico Tools: A Critical Review. *NanoImpact* **2019**, *13*, 76–99.
- (21) Krug, H. F. Nanosafety Research-Are We on the Right Track? *Angew. Chem., Int. Ed.* **2014**, *53*, 12304–12319.
- (22) Foss Hansen, S.; Larsen, B. H.; Olsen, S. I.; Baun, A. Categorization Framework to Aid Hazard Identification of Nanomaterials. *Nanotoxicology* **2007**, *1*, 243–250.
- (23) Jeliakova, N.; Apostolova, M. D.; Andreoli, C.; Barone, F.; Barrick, A.; Battistelli, C.; Bossa, C.; Botea-petcu, A.; Châtel, A.; Angelis, I. De; et al. Towards FAIR Nanosafety Data. *Nat. Nanotechnol.* **2021**, *16*, 644–654.
- (24) Yan, X.; Sedykh, A.; Wang, W.; Yan, B.; Zhu, H. Construction of a Web-Based Nanomaterial Database by Big Data Curation and Modeling Friendly Nanostructure Annotations. *Nat. Commun.* **2020**, *11*, 2519.
- (25) Karakus, C. O.; Winkler, D. A. Overcoming Roadblocks in Computational Roadmaps to the Future for Safe Nanotechnology. *Nano Futur.* **2021**, *5*, 022002.
- (26) Fadeel, B.; Farcal, L.; Hardy, B.; Vázquez-campos, S.; Hristozov, D.; Marcomini, A.; Lynch, I.; Valsami-Jones, E.; Alenius, H.; Savolainen, K. Advanced Tools for the Safety Assessment of Nanomaterials. *Nat. Nanotechnol.* **2018**, *13*, 537–543.
- (27) Lynch, I.; Afantitis, A.; Exner, T.; Himly, M.; Lobaskin, V.; Doganis, P.; Maier, D.; Sanabria, N.; Papadiamantis, A. G.; Rybinska-fryca, A.; et al. Can an Inchi for Nano Address the Need for a Simplified Representation of Complex Nanomaterials across Experimental and Nanoinformatics Studies? *Nanomaterials* **2020**, *10*, 2493.

- (28) Serov, N.; Vinogradov, V. Artificial Intelligence to Bring Nanomedicine to Life. *Adv. Drug Delivery Rev.* **2022**, *184*, 114194.
- (29) He, Y.; Liu, G.; Li, C.; Yan, X. Reaching the Full Potential of Machine Learning in Mitigating Environmental Impacts of Functional Materials. *Rev. Environ. Contam. Toxicol.* **2022**, *260*, 21.
- (30) Shen, W. X.; Zeng, X.; Zhu, F.; Wang, Y. li; Qin, C.; Tan, Y.; Jiang, Y. Y.; Chen, Y. Z. Out-of-the-Box Deep Learning Prediction of Pharmaceutical Properties by Broadly Learned Knowledge-Based Molecular Representations. *Nat. Mach. Intell.* **2021**, *3*, 334–343.
- (31) Chen, D.; Gao, K.; Nguyen, D. D.; Chen, X.; Jiang, Y.; Wei, G. W.; Pan, F. Algebraic Graph-Assisted Bidirectional Transformers for Molecular Property Prediction. *Nat. Commun.* **2021**, *12*, 3521.
- (32) Yan, J.; Yan, X.; Hu, S.; Zhu, H.; Yan, B. Comprehensive Interrogation on Acetylcholinesterase Inhibition by Ionic Liquids Using Machine Learning and Molecular Modeling. *Environ. Sci. Technol.* **2021**, *55*, 14720–14731.
- (33) Yan, X.; Yue, T.; Zhu, H.; Yan, B. Bridging the Gap Between Nanotoxicological Data and the Critical Structure-Activity Relationships. In *Advances in Toxicology and Risk Assessment of Nanomaterials and Emerging Contaminants*; Guo, L. H., Mortimer, M., Eds.; Springer: Singapore, 2022; pp 161–183.
- (34) Wyrzykowska, E.; Mikolajczyk, A.; Lynch, I.; Jeliaskova, N.; Kochev, N.; Sarimveis, H.; Doganis, P.; Karatzas, P.; Afantitis, A.; Melagraki, G.; et al. Representing and Describing Nanomaterials in Predictive Nanoinformatics. *Nat. Nanotechnol.* **2022**, *17*, 924–932.
- (35) Yan, X.; Sedykh, A.; Wang, W.; Zhao, X.; Yan, B.; Zhu, H. In Silico Profiling Nanoparticles: Predictive Nanomodeling Using Universal Nanodescriptors and Various Machine Learning Approaches. *Nanoscale* **2019**, *11*, 8352–8362.
- (36) Yu, H.; Luo, D.; Dai, L.; Cheng, F. In Silico Nanosafety Assessment Tools and Their Ecosystem-Level Integration Prospect. *Nanoscale* **2021**, *13*, 8722–8739.
- (37) Furxhi, I.; Murphy, F.; Mullins, M.; Arvanitis, A. Practices and Trends of Machine Learning Application in Nanotoxicology. *Nanomaterials* **2020**, *10*, 116.
- (38) Artrith, N.; Butler, K. T.; Coudert, F.; Han, S.; Isayev, O.; Jain, A.; Walsh, A. Best Practices in Machine Learning for Chemistry. *Nat. Chem.* **2021**, *13*, 505–508.
- (39) Hollingsworth, S. A.; Dror, R. O. Molecular Dynamics Simulation for All. *Neuron* **2018**, *99*, 1129–1143.
- (40) Karplus, M.; McCammon, J. A. Molecular Dynamics Simulations of Biomolecules. *Nat. Struct. Biol.* **2002**, *9*, 646–652.
- (41) Casalini, T.; Limongelli, V.; Schmutz, M.; Som, C.; Jordan, O.; Wick, P.; Borchard, G.; Perale, G. Molecular Modeling for Nanomaterial-Biology Interactions: Opportunities, Challenges, and Perspectives. *Front. Bioeng. Biotechnol.* **2019**, *7*, 268.
- (42) Foreman-Ortiz, I. U.; Liang, D.; Laudadio, E. D.; Calderin, J. D.; Wu, M.; Keshri, P.; Zhang, X.; Schwartz, M. P.; Hamers, R. J.; Rotello, V. M.; et al. Anionic Nanoparticle-Induced Perturbation to Phospholipid Membranes Affects Ion Channel Function. *Proc. Natl. Acad. Sci. U. S. A.* **2020**, *117*, 27854–27861.
- (43) Kmiecik, S.; Gront, D.; Kolinski, M.; Wieteska, L.; Dawid, A. E.; Kolinski, A. Coarse-Grained Protein Models and Their Applications. *Chem. Rev.* **2016**, *116*, 7898–7936.
- (44) Marrink, S. J.; Corradi, V.; Souza, P. C. T.; Ingólfsson, H. I.; Tieleman, D. P.; Sansom, M. S. P. Computational Modeling of Realistic Cell Membranes. *Chem. Rev.* **2019**, *119*, 6184–6226.
- (45) Feng, R.; Yu, F.; Xu, J.; Hu, X. Knowledge Gaps in Immune Response and Immunotherapy Involving Nanomaterials: Databases and Artificial Intelligence for Material Design. *Biomaterials* **2021**, *266*, 120469.
- (46) Zhu, H. Big Data and Artificial Intelligence Modeling for Drug Discovery. *Annu. Rev. Pharmacol. Toxicol.* **2020**, *60*, 573–589.
- (47) Yang, X.; Wang, Y.; Byrne, R.; Schneider, G.; Yang, S. Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery. *Chem. Rev.* **2019**, *119*, 10520–10594.
- (48) Barnard, A. S.; Motevalli, B.; Parker, A. J.; Fischer, J. M.; Feigl, C. A.; Opletal, G. Nanoinformatics, and the Big Challenges for the Science of Small Things. *Nanoscale* **2019**, *11*, 19190–19201.
- (49) Liang, W.; Tadesse, G. A.; Ho, D.; Li, F. F.; Zaharia, M.; Zhang, C.; Zou, J. Advances, Challenges and Opportunities in Creating Data for Trustworthy AI. *Nat. Mach. Intell.* **2022**, *4*, 669–677.
- (50) Plata, D. L.; Jankovic, N. Z. Achieving Sustainable Nanomaterial Design. *Nat. Nanotechnol.* **2021**, *16*, 612–614.
- (51) Faria, M.; Björnholm, M.; Thurecht, K. J.; Kent, S. J.; Parton, R. G.; Kavallaris, M.; Johnston, A. P. R.; Gooding, J. J.; Corrie, S. R.; Boyd, B. J.; et al. Minimum Information Reporting in Bio-Nano Experimental Literature. *Nat. Nanotechnol.* **2018**, *13*, 777–785.
- (52) Bai, X.; Liu, F.; Liu, Y.; Li, C.; Wang, S.; Zhou, H.; Wang, W.; Zhu, H.; Winkler, D. A.; Yan, B. Toward a Systematic Exploration of Nano-Bio Interactions. *Toxicol. Appl. Pharmacol.* **2017**, *323*, 66–73.
- (53) Domingues, C.; Santos, A.; Alvarez-Lorenzo, C.; Concheiro, A.; Jarak, I.; Veiga, F.; Barbosa, I.; Dourado, M.; Figueiras, A. Where Is Nano Today and Where Is It Headed? A Review of Nanomedicine and the Dilemma of Nanotoxicology. *ACS Nano* **2022**, *16*, 9994–10041.
- (54) Toropova, A. P.; Toropov, A. A.; Benfenati, E.; Leszczynska, D.; Leszczynski, J. QSAR Modeling of Measured Binding Affinity for Fullerene-Based HIV-1 PR Inhibitors by CORAL. *J. Math. Chem.* **2010**, *48*, 959–987.
- (55) Toropov, A. A.; Toropova, A. P.; Benfenati, E.; Leszczynska, D.; Leszczynski, J. SMILES-Based Optimal Descriptors: QSAR Analysis of Fullerene-Based HIV-1 PR Inhibitors by Means of Balance of Correlations. *J. Comput. Chem.* **2010**, *31*, 381–392.
- (56) Durdagi, S.; Mavromoustakos, T.; Chronakis, N.; Papadopoulos, M. G. Computational Design of Novel Fullerene Analogues as Potential HIV-1 PR Inhibitors: Analysis of the Binding Interactions between Fullerene Inhibitors and HIV-1 PR Residues Using 3D QSAR, Molecular Docking and Molecular Dynamics Simulations. *Bioorg. Med. Chem.* **2008**, *16*, 9957–9974.
- (57) Singh, K. P. P.; Gupta, S. Nano-QSAR Modeling for Predicting Biological Activity of Diverse Nanomaterials. *RSC Adv.* **2014**, *4*, 13215–13230.
- (58) Burns, T. D.; Pai, K. N.; Subraveti, S. G.; Collins, S. P.; Krykunov, M.; Rajendran, A.; Woo, T. K. Prediction of MOF Performance in Vacuum Swing Adsorption Systems for Postcombustion CO₂ Capture Based on Integrated Molecular Simulations, Process Optimizations, and Machine Learning Models. *Environ. Sci. Technol.* **2020**, *54*, 4536–4544.
- (59) Chen, Y.; Huang, Y.; Cheng, T.; Goddard, W. A. Identifying Active Sites for CO₂ Reduction on Dealloyed Gold Surfaces by Combining Machine Learning with Multiscale Simulations. *J. Am. Chem. Soc.* **2019**, *141*, 11651–11657.
- (60) Kelkar, A. S.; Dallin, B. C.; Lehn, R. C. Van. Predicting Hydrophobicity by Learning Spatiotemporal Features of Interfacial Water Structure: Combining Molecular Dynamics Simulations with Convolutional Neural Networks. *J. Phys. Chem. B* **2020**, *124*, 9103–9114.
- (61) Webb, M. A.; Jackson, N. E.; Gil, P. S.; Pablo, J. J. De. Targeted Sequence Design within the Coarse-Grained Polymer Genome. *Sci. Adv.* **2020**, *6*, No. eabc6216.
- (62) Sun, Y.; DeJaco, R. F.; Li, Z.; Tang, D.; Glante, S.; Sholl, D. S.; Colina, C. M.; Snurr, R. Q.; Thommes, M.; Hartmann, M.; et al. Fingerprinting Diverse Nanoporous Materials for Optimal Hydrogen Storage Conditions Using Meta-Learning. *Sci. Adv.* **2021**, *7*, No. eabg3983.
- (63) Ma, S.; Huang, S.; Liu, Z. Dynamic Coordination of Cations and Catalytic Selectivity on Zinc-Chromium Oxide Alloys during Syngas Conversion. *Nat. Catal.* **2019**, *2*, 671–677.
- (64) Ma, S.; Liu, Z. Machine Learning for Atomic Simulation and Activity Prediction in Heterogeneous Catalysis: Current Status and Future. *ACS Catal.* **2020**, *10*, 13213–13226.
- (65) Burger, B.; Maffettone, P. M.; Gusev, V. V.; Aitchison, C. M.; Bai, Y.; Wang, X.; Li, X.; Alston, B. M.; Li, B.; Clowes, R.; et al. A Mobile Robotic Chemist. *Nature* **2020**, *583*, 237–241.
- (66) Coley, C. W.; Thomas, D. A.; Lummiss, J. A. M.; Jaworski, J. N.; Breen, C. P.; Schultz, V.; Hart, T.; Fishman, J. S.; Rogers, L.; Gao, H.; et al. A Robotic Platform for Flow Synthesis of Organic Compounds Informed by AI Planning. *Science* **2019**, *365*, No. eaax1566.

- (67) Glavatskikh, M.; Leguy, J.; Hunault, G.; Cauchy, T.; Da Mota, B. Dataset's Chemical Diversity Limits the Generalizability of Machine Learning Predictions. *J. Cheminform.* **2019**, *11*, 69.
- (68) Volk, A. A.; Epps, R. W.; Abolhasani, M. Accelerated Development of Colloidal Nanomaterials Enabled by Modular Microfluidic Reactors: Toward Autonomous Robotic Experimentation. *Adv. Mater.* **2021**, *33*, 2004495.
- (69) Zhao, X.; Bian, F.; Sun, L.; Cai, L.; Li, L.; Zhao, Y. Microfluidic Generation of Nanomaterials for Biomedical Applications. *Small* **2020**, *16*, 1901943.
- (70) Weissleder, R.; Kelly, K.; Sun, E. Y.; Shtatland, T.; Josephson, L. Cell-Specific Targeting of Nanoparticles by Multivalent Attachment of Small Molecules. *Nat. Biotechnol.* **2005**, *23*, 1418–1423.
- (71) Shaw, S. Y.; Westly, E. C.; Pittet, M. J.; Subramanian, A.; Schreiber, S. L.; Weissleder, R. Perturbational Profiling of Nanomaterial Biologic Activity. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 7387–7392.
- (72) Walkey, C. D.; Olsen, J. B.; Song, F.; Liu, R.; Guo, H.; Olsen, D. W. H.; Cohen, Y.; Emili, A.; Chan, W. C. W. Protein Corona Fingerprinting Predicts the Cellular Interaction of Gold and Silver Nanoparticles. *ACS Nano* **2014**, *8*, 2439–2455.
- (73) Anderson, D. G.; Lynn, D. M.; Langer, R. Semi-automated Synthesis and Screening of a Large Library of Degradable Cationic Polymers for Gene Delivery. *Angew. Chem., Int. Ed.* **2003**, *115*, 3261–3266.
- (74) Lynn, D. M.; Anderson, D. G.; Putnam, D.; Langer, R. Accelerated Discovery of Synthetic Transfection Vectors: Parallel Synthesis and Screening of a Degradable Polymer Library. *J. Am. Chem. Soc.* **2001**, *123*, 8155–8156.
- (75) Anderson, D. G.; Akinc, A.; Hossain, N.; Langer, R. Structure/Property Studies of Polymeric Gene Delivery Using a Library of Poly(β -Amino Esters). *Mol. Ther.* **2005**, *11*, 426–434.
- (76) Akinc, A.; Lynn, D. M.; Anderson, D. G.; Langer, R. Parallel Synthesis and Biophysical Characterization of a Degradable Polymer Library for Gene Delivery. *J. Am. Chem. Soc.* **2003**, *125*, 5316–5323.
- (77) Fourches, D.; Pu, D.; Tassa, C.; Weissleder, R.; Shaw, S. Y. Y.; Mumper, R. J. J.; Tropsha, A. Quantitative Nanostructure - Activity Relationship Modeling. *ACS Nano* **2010**, *4*, 5703–5712.
- (78) Epa, V. C.; Burden, F. R.; Tassa, C.; Weissleder, R.; Shaw, S.; Winkler, D. A. Modeling Biological Activities of Nanoparticles. *Nano Lett.* **2012**, *12*, 5808–5812.
- (79) Zhou, H.; Mu, Q.; Gao, N.; Liu, A.; Xing, Y.; Gao, S.; Zhang, Q.; Qu, G.; Chen, Y.; Liu, G.; et al. A Nano-Combinatorial Library Strategy for the Discovery of Nanotubes with Reduced Protein-Binding, Cytotoxicity, and Immune Response. *Nano Lett.* **2008**, *8*, 859–865.
- (80) Niu, G.; Zhang, L.; Ruditskiy, A.; Wang, L.; Xia, Y. A Droplet-Reactor System Capable of Automation for the Continuous and Scalable Production of Noble-Metal Nanocrystals. *Nano Lett.* **2018**, *18*, 3879–3884.
- (81) Gao, N.; Zhang, Q.; Mu, Q.; Bai, Y.; Li, L.; Zhou, H.; Butch, E. R.; Powell, T. B.; Snyder, S. E.; Jiang, G.; et al. Steering Carbon Nanotubes to Scavenger Receptor Recognition by Nanotube Surface Chemistry Modification Partially Alleviates NF κ B Activation and Reduces Its Immunotoxicity. *ACS Nano* **2011**, *5*, 4581–4591.
- (82) Zhou, H.; Jiao, P.; Yang, L.; Li, X.; Yan, B. Enhancing Cell Recognition by Scrutinizing Cell Surfaces with a Nanoparticle Array. *J. Am. Chem. Soc.* **2011**, *133*, 680–682.
- (83) Zhang, B.; Xing, Y.; Li, Z.; Zhou, H.; Mu, Q.; Yan, B. Functionalized Carbon Nanotubes Specifically Bind to α -Chymotrypsin's Catalytic Site and Regulate Its Enzymatic Function. *Nano Lett.* **2009**, *9*, 2280–2284.
- (84) Wu, L.; Zhang, Y.; Zhang, C.; Cui, X.; Zhai, S.; Liu, Y.; Li, C.; Zhu, H.; Qu, G.; Jiang, G.; et al. Tuning Cell Autophagy by Diversifying Carbon Nanotube Surface Chemistry. *ACS Nano* **2014**, *8*, 2087–2099.
- (85) Zhang, Y.; Wang, Y.; Liu, A.; Xu, S. L.; Zhao, B.; Zhang, Y.; Zou, H.; Wang, W.; Zhu, H.; Yan, B. Modulation of Carbon Nanotubes' Perturbation to the Metabolic Activity of CYP3A4 in the Liver. *Adv. Funct. Mater.* **2016**, *26*, 841–850.
- (86) Wang, W.; Sedykh, A.; Sun, H.; Zhao, L.; Russo, D. P. P.; Zhou, H.; Yan, B.; Zhu, H. Predicting Nano-Bio Interactions by Integrating Nanoparticle Libraries and Quantitative Nanostructure Activity Relationship Modeling. *ACS Nano* **2017**, *11*, 12641–12649.
- (87) Bai, X.; Wang, S.; Yan, X.; Zhou, H.; Zhan, J.; Liu, S.; Sharma, V. K.; Jiang, G.; Zhu, H.; Yan, B. Regulation of Cell Uptake and Cytotoxicity by Nanoparticle Core under the Controlled Shape, Size, and Surface Chemistries. *ACS Nano* **2020**, *14*, 289–302.
- (88) Li, S.; Zhai, S.; Liu, Y.; Zhou, H.; Wu, J.; Jiao, Q.; Zhang, B.; Zhu, H.; Yan, B. Experimental Modulation and Computational Model of Nano-Hydrophobicity. *Biomaterials* **2015**, *52*, 312–317.
- (89) Sun, H.; Liu, Y.; Bai, X.; Zhou, X.; Zhou, H.; Liu, S.; Yan, B. Induction of Oxidative Stress and Sensitization of Cancer Cells to Paclitaxel by Gold Nanoparticles with Different Charge Densities and Hydrophobicities. *J. Mater. Chem. B* **2018**, *6*, 1633–1639.
- (90) Su, G.; Zhou, H.; Mu, Q.; Zhang, Y.; Li, L.; Jiao, P.; Jiang, G.; Yan, B. Effective Surface Charge Density Determines the Electrostatic Attraction between Nanoparticles and Cells. *J. Phys. Chem. C* **2012**, *116*, 4993–4998.
- (91) Liu, Y.; Winkler, D. A.; Epa, V. C.; Zhang, B.; Yan, B. Probing Enzyme-Nanoparticle Interactions Using Combinatorial Gold Nanoparticle Libraries. *Nano Res.* **2015**, *8*, 1293–1308.
- (92) Chew, A. K.; Pedersen, J. A.; Van Lehn, R. C. Predicting the Physicochemical Properties and Biological Activities of Monolayer-Protected Gold Nanoparticles Using Simulation-Derived Descriptors. *ACS Nano* **2022**, *16*, 6282–6292.
- (93) Li, S.; Wang, S.; Yan, B.; Yue, T. Surface Properties of Nanoparticles Dictate Their Toxicity by Regulating Adsorption of Humic Acid Molecules. *ACS Sustain. Chem. Eng.* **2021**, *9*, 13705–13716.
- (94) Liu, F.; Li, S.; Feng, H.; Li, L.; Yue, T.; Yan, B. Modulation of Cell Uptake and Cytotoxicity by Nanoparticles with Various Physicochemical Properties after Humic Acid Adsorption. *Environ. Sci. Nano* **2021**, *8*, 3746–3761.
- (95) Elvira, K. S.; I Solvas, X. C.; Wootton, R. C. R.; Demello, A. J. The Past, Present and Potential for Microfluidic Reactor Technology in Chemical Synthesis. *Nat. Chem.* **2013**, *5*, 905–915.
- (96) Chan, E. M.; Xu, C.; Mao, A. W.; Han, G.; Owen, J. S.; Cohen, B. E.; Milliron, D. J. Reproducible, High-Throughput Synthesis of Colloidal Nanocrystals for Optimization in Multidimensional Parameter Space. *Nano Lett.* **2010**, *10*, 1874–1885.
- (97) Tao, H.; Wu, T.; Aldeghi, M.; Wu, T. C.; Aspuru-Guzik, A.; Kumacheva, E. Nanoparticle Synthesis Assisted by Machine Learning. *Nat. Rev. Mater.* **2021**, *6*, 701–716.
- (98) Epps, R. W.; Bowen, M. S.; Volk, A. A.; Abdel-Latif, K.; Han, S.; Reyes, K. G.; Amassian, A.; Abolhasani, M. Artificial Chemist: An Autonomous Quantum Dot Synthesis Bot. *Adv. Mater.* **2020**, *32*, 2001626.
- (99) Han, Y.; Tang, B.; Wang, L.; Bao, H.; Lu, Y.; Guan, C.; Zhang, L.; Le, M.; Liu, Z.; Wu, M. Machine-Learning-Driven Synthesis of Carbon Dots with Enhanced Quantum Yields. *ACS Nano* **2020**, *14*, 14761–14768.
- (100) Krishnadasan, S.; Brown, R. J. C.; DeMello, A. J.; DeMello, J. C. Intelligent Routes to the Controlled Synthesis of Nanoparticles. *Lab Chip* **2007**, *7*, 1434–1441.
- (101) Abdel-Latif, K.; Epps, R. W.; Bateni, F.; Han, S.; Reyes, K. G.; Abolhasani, M. Self-Driven Multistep Quantum Dot Synthesis Enabled by Autonomous Robotic Experimentation in Flow. *Adv. Intell. Syst.* **2021**, *3*, 2000245.
- (102) Xie, Y.; Zhang, C.; Hu, X.; Zhang, C.; Kelley, S. P.; Atwood, J. L.; Lin, J. Machine Learning Assisted Synthesis of Metal-Organic Nanocapsules. *J. Am. Chem. Soc.* **2020**, *142*, 1475–1481.
- (103) Bezing, L.; Maceiczky, R. M.; Lignos, I.; Kovalenko, M. V.; Demello, A. J. Pick a Color MARIA: Adaptive Sampling Enables the Rapid Identification of Complex Perovskite Nanocrystal Compositions with Defined Emission Characteristics. *ACS Appl. Mater. Interfaces* **2018**, *10*, 18869–18878.
- (104) Li, J.; Li, J.; Liu, R.; Tu, Y.; Li, Y.; Cheng, J.; He, T.; Zhu, X. Autonomous Discovery of Optically Active Chiral Inorganic Perovskite Nanocrystals through an Intelligent Cloud Lab. *Nat. Commun.* **2020**, *11*, 2046.

- (105) Li, J.; Chen, T.; Lim, K.; Chen, L.; Khan, S. A.; Xie, J.; Wang, X. Deep Learning Accelerated Gold Nanocluster Synthesis. *Adv. Intell. Syst.* **2019**, *1*, 1900029.
- (106) Salley, D.; Keenan, G.; Grizou, J.; Sharma, A.; Martin, S.; Cronin, L. A Nanomaterials Discovery Robot for the Darwinian Evolution of Shape Programmable Gold Nanoparticles. *Nat. Commun.* **2020**, *11*, 2771.
- (107) Mekki-Berrada, F.; Ren, Z.; Huang, T.; Wong, W. K.; Zheng, F.; Xie, J.; Tian, I. P. S.; Jayavelu, S.; Mahfoud, Z.; Bash, D.; et al. Two-Step Machine Learning Enables Optimized Nanoparticle Synthesis. *npj Comput. Mater.* **2021**, *7*, 55.
- (108) Nikolaev, P.; Hooper, D.; Webber, F.; Rao, R.; Decker, K.; Krein, M.; Poleski, J.; Barto, R.; Maruyama, B. Autonomy in Materials Research: A Case Study in Carbon Nanotube Growth. *npj Comput. Mater.* **2016**, *2*, 16031.
- (109) Comandella, D.; Rauscher, H.; Gottardo, S.; Rio-echavarria, I. M. Quality of Physicochemical Data on Nanomaterials: An Assessment of Data Completeness and Variability. *Nanoscale* **2020**, *12*, 4695–4708.
- (110) Marchese Robinson, R. L.; Lynch, I.; Peijnenburg, W.; Rumble, J.; Klaessig, F.; Marquardt, C.; Rauscher, H.; Puzyn, T.; Purian, R.; Åberg, C.; et al. How Should the Completeness and Quality of Curated Nanomaterial Data Be Evaluated? *Nanoscale* **2016**, *8*, 9919–9943.
- (111) Huang, R.; Xia, M.; Sakamuru, S.; Zhao, J.; Shahane, S. A.; Attene-Ramos, M.; Zhao, T.; Austin, C. P.; Simeonov, A. Modelling the Tox21 10 K Chemical Profiles for in Vivo Toxicity Prediction and Mechanism Characterization. *Nat. Commun.* **2016**, *7*, 10425.
- (112) Trinh, T. X.; Ha, M. K.; Choi, J. S.; Byun, H. G.; Yoon, T. H. Curation of Datasets, Assessment of Their Quality and Completeness, and NanoSAR Classification Model Development for Metallic Nanoparticles. *Environ. Sci. Nano* **2018**, *5*, 1902–1910.
- (113) Ha, M. K.; Trinh, T. X.; Choi, J. S.; Maulina, D.; Byun, H. G.; Yoon, T. H. Toxicity Classification of Oxide Nanomaterials: Effects of Data Gap Filling and PChem Score-Based Screening Approaches. *Sci. Rep.* **2018**, *8*, 3141.
- (114) Fadeel, B.; Fornara, A.; Toprak, M. S.; Bhattacharya, K. Biochemical and Biophysical Research Communications Keeping It Real: The Importance of Material Characterization in Nanotoxicology. *Biochem. Biophys. Res. Commun.* **2015**, *468*, 498–503.
- (115) Frankel, F. C.; Whitesides, G. M. *No Small Matter: Science on the Nanoscale*; Harvard University Press: Cambridge, MA, 2009.
- (116) Caputo, F.; Clogston, J.; Calzolari, L.; Rösslein, M.; Prina-Mello, A. Measuring Particle Size Distribution of Nanoparticle Enabled Medicinal Products, the Joint View of EUNCL and NCI-NCL. A Step by Step Approach Combining Orthogonal Measurements with Increasing Complexity. *J. Controlled Release* **2019**, *299*, 31–43.
- (117) Kilkenny, C.; Browne, W. J.; Cuthill, I. C.; Emerson, M.; Altman, D. G. Improving Bioscience Research Reporting: The ARRIVE Guidelines for Reporting Animal Research. *J. Pharmacol. Pharmacother.* **2010**, *1*, 94–99.
- (118) George, S.; Xia, T.; Rallo, R.; Zhao, Y.; Ji, Z.; Lin, S.; Wang, X.; Zhang, H.; Al, G. E. T. Use of a High-Throughput Screening Approach Coupled with In Vivo Zebra Fi Sh Embryo Screening To Develop Hazard Ranking for Engineered Nanomaterials. *ACS Nano* **2011**, *5*, 1805–1817.
- (119) Nel, A.; Xia, T.; Meng, H.; Wang, X.; Lin, S.; Ji, Z.; Zhang, H. Nanomaterial Toxicity Testing in the 21st Century: Use of a Predictive Toxicological Approach and High-Throughput Screening. *Acc. Chem. Res.* **2013**, *46*, 607–621.
- (120) Sayers, E. W.; Bolton, E. E.; Brister, J. R.; Canese, K.; Chan, J.; Comeau, D. C.; Connor, R.; Funk, K.; Kelly, C.; Kim, S.; et al. Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2022**, *50*, D20–D26.
- (121) Sayers, E. W.; Barrett, T.; Benson, D. A.; Bryant, S. H.; Canese, K.; Chetvernin, V.; Church, D. M.; Dicuccio, M.; Edgar, R.; Federhen, S.; et al. Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2009**, *37*, D5–D15.
- (122) Yamankurt, G.; Berns, E. J.; Xue, A.; Mrksich, M.; Mirkin, C. A.; Lee, A.; Bagheri, N. Exploration of the Nanomedicine-Design Space with High-Throughput Screening and Machine Learning. *Nat. Biomed. Eng.* **2019**, *3*, 318–327.
- (123) Lin, S.; Zhao, Y.; Xia, T.; Meng, H.; Ji, Z.; Liu, R.; George, S.; Xiong, S.; Wang, X.; Zhang, H.; et al. High Content Screening in Zebrafish Speeds up Hazard Ranking of Transition Metal Oxide Nanoparticles. *ACS Nano* **2011**, *5*, 7284–7295.
- (124) Jia, H. R.; Zhu, Y. X.; Duan, Q. Y.; Chen, Z.; Wu, F. G. Nanomaterials Meet Zebrafish: Toxicity Evaluation and Drug Delivery Applications. *J. Controlled Release* **2019**, *311–312*, 301–318.
- (125) Zhang, H.; Ji, Z.; Xia, T.; Meng, H.; Low-Kam, C.; Liu, R.; Pokhrel, S.; Lin, S.; Wang, X.; Liao, Y. P.; et al. Use of Metal Oxide Nanoparticle Band Gap to Develop a Predictive Paradigm for Oxidative Stress and Acute Pulmonary Inflammation. *ACS Nano* **2012**, *6*, 4349–4368.
- (126) Jung, S. K.; Qu, X.; Aleman-Meza, B.; Wang, T.; Riepe, C.; Liu, Z.; Li, Q.; Zhong, W. Multi-Endpoint, High-Throughput Study of Nanomaterial Toxicity in Caenorhabditis Elegans. *Environ. Sci. Technol.* **2015**, *49*, 2477–2485.
- (127) Karatzas, P.; Melagraki, G.; Ellis, L. A. J. A.; Lynch, I.; Varsou, D. D.; Afantitis, A.; Tsoumanis, A.; Doganis, P.; Sarimveis, H. Development of Deep Learning Models for Predicting the Effects of Exposure to Engineered Nanomaterials on Daphnia Magna. *Small* **2020**, *16*, 2001080.
- (128) George, S.; Xia, T.; Rallo, R.; Zhao, Y.; Ji, Z.; Lin, S. S.; Wang, X.; Zhang, H.; France, B.; Schoenfeld, D.; et al. Use of a High-Throughput Screening Approach Coupled with in Vivo Zebrafish Embryo Screening to Develop Hazard Ranking for Engineered Nanomaterials. *ACS Nano* **2011**, *5*, 1805–1817.
- (129) Bai, C.; Tang, M. Toxicological Study of Metal and Metal Oxide Nanoparticles in Zebrafish. *J. Appl. Toxicol.* **2020**, *40*, 37–63.
- (130) Labouta, H. I. I.; Asgarian, N.; Rinker, K.; Cramb, D. T. T. Meta-Analysis of Nanoparticle Cytotoxicity via Data-Mining the Literature. *ACS Nano* **2019**, *13*, 1583–1594.
- (131) Worle-Knirsch, J. M.; Pulskamp, K.; Krug, H. F. Oops They Did It Again! Carbon Nanotubes Hoax Scientists in Viability Assays. *Nano Lett.* **2006**, *6*, 1261–1268.
- (132) Belyanskaya, L.; Manser, P.; Spohn, P.; Bruinink, A.; Wick, P. The Reliability and Limits of the MTT Reduction Assay for Carbon Nanotubes - Cell Interaction. *Carbon* **2007**, *45*, 2643–2648.
- (133) Bilal, M.; Oh, E.; Liu, R.; Breger, J. C.; Medintz, I. L.; Cohen, Y. Toxicity Models: Bayesian Network Resource for Meta-Analysis: Cellular Toxicity of Quantum Dots. *Small* **2019**, *15*, 1970181.
- (134) Oh, E.; Liu, R.; Nel, A.; Gemill, K. B.; Bilal, M.; Cohen, Y.; Medintz, I. L. Meta-Analysis of Cellular Toxicity for Cadmium-Containing Quantum Dots. *Nat. Nanotechnol.* **2016**, *11*, 479–486.
- (135) Halamoda-Kenzaoui, B.; Holzwarth, U.; Roebben, G.; Bogni, A.; Bremer-Hoffmann, S. Mapping of the Available Standards against the Regulatory Needs for Nanomedicines. *Wiley Interdiscip. Rev. Nanomedicine Nanobiotechnology* **2019**, *11*, e1531.
- (136) Schuh, J. A. C. L.; Funk, K. A. Compilation of International Standards and Regulatory Guidance Documents for Evaluation of Biomaterials, Medical Devices, and 3-D Printed and Regenerative Medicine Products. *Toxicol. Pathol.* **2019**, *47*, 344–357.
- (137) Rasmussen, K.; Rauscher, H.; Kearns, P.; González, M.; Riego Sintes, J. Developing OECD Test Guidelines for Regulatory Testing of Nanomaterials to Ensure Mutual Acceptance of Test Data. *Regul. Toxicol. Pharmacol.* **2019**, *104*, 74–83.
- (138) Willemink, M. J.; Koszek, W. A.; Hardell, C.; Wu, J.; Fleischmann, D.; Harvey, H.; Folio, L. R.; Summers, R. M.; Rubin, D. L.; Lungren, M. P. Preparing Medical Imaging Data for Machine Learning. *Radiology* **2020**, *295*, 4–15.
- (139) Kim, E.; Huang, K.; Saunders, A.; McCallum, A.; Ceder, G.; Olivetti, E. Materials Synthesis Insights from Scientific Literature via Text Extraction and Machine Learning. *Chem. Mater.* **2017**, *29*, 9436–9444.
- (140) Tshitoyan, V.; Dagdelen, J.; Weston, L.; Dunn, A.; Rong, Z.; Kononova, O.; Persson, K. A.; Ceder, G.; Jain, A. Unsupervised Word Embeddings Capture Latent Knowledge from Materials Science Literature. *Nature* **2019**, *571*, 95–98.

- (141) Lever, J.; Zhao, E. Y.; Grewal, J.; Jones, M. R.; Jones, S. J. M. CancerMine: A Literature-Mined Resource for Drivers, Oncogenes and Tumor Suppressors in Cancer. *Nat. Methods* **2019**, *16*, 505–507.
- (142) Ban, Z.; Yuan, P.; Yu, F.; Peng, T.; Zhou, Q.; Hu, X. Machine Learning Predicts the Functional Composition of the Protein Corona and the Cellular Recognition of Nanoparticles. *Proc. Natl. Acad. Sci. U. S. A.* **2020**, *117*, 10492–10499.
- (143) Xu, N.; Kang, J.; Ye, Y.; Zhang, Q.; Ke, M.; Wang, Y.; Zhang, Z.; Lu, T.; Peijnenburg, W. J. G. M.; Penuelas, J.; et al. Machine Learning Predicts Ecological Risks of Nanoparticles to Soil Microbial Communities. *Environ. Pollut.* **2022**, *307*, 119528.
- (144) Gul, G.; Yildirim, R.; Ileri-Ercan, N. Cytotoxicity Analysis of Nanoparticles by Association Rule Mining. *Environ. Sci. Nano* **2021**, *8*, 937–949.
- (145) Ma, Y.; Wang, J.; Wu, J.; Tong, C.; Zhang, T. Meta-Analysis of Cellular Toxicity for Graphene via Data-Mining the Literature and Machine Learning. *Sci. Total Environ.* **2021**, *793*, 148532.
- (146) Dang, F.; Wang, Q.; Yan, X.; Zhang, Y.; Yan, J.; Zhong, H.; Zhou, D.; Luo, Y.; Zhu, Y. G.; Xing, B.; et al. Threats to Terrestrial Plants from Emerging Nanoplastics. *ACS Nano* **2022**, *16*, 17157–17167.
- (147) Yu, F.; Wei, C.; Deng, P.; Peng, T.; Hu, X. Deep Exploration of Random Forest Model Boosts the Interpretability of Machine Learning Studies of Complicated Immune Responses and Lung Burden of Nanoparticles. *Sci. Adv.* **2021**, *7*, eabf4130.
- (148) Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *arXiv*, 1810.04805; 2018. DOI: 10.48550/arXiv.1810.04805.
- (149) Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models Are Few-Shot Learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
- (150) Zhu, Q.; Zhang, F.; Huang, Y.; Xiao, H.; Zhao, L.; Zhang, X.; Song, T.; Tang, X.; Li, X.; He, G.; et al. An All-Round AI-Chemist with a Scientific Mind. *Natl. Sci. Rev.* **2022**, *9*, nwac190.
- (151) Gupta, T.; Zaki, M.; Krishnan, N. M. A. MatSciBERT: A Materials Domain Language Model for Text Mining and Information Extraction. *npj Comput. Mater.* **2022**, *8*, 102.
- (152) Zhao, J.; Huang, S.; Cole, J. M. OpticalBERT and OpticalTableSQA: Text- and Table-Based Language Models for the Optical-Materials Domain. *J. Chem. Inf. Model.* **2023**, *63*, 1961–1981.
- (153) Trewartha, A.; Walker, N.; Huo, H.; Persson, K. A. Quantifying the Advantage of Domain-Specific Pre-Training on Named Entity Recognition Tasks in Materials Science. *Patterns* **2022**, *3*, 100488.
- (154) Olah, M.; Mracec, M.; Ostopovici, L.; Rad, R.; Bora, A.; Hadaruga, N. WOMBAT: World of Molecular Bioactivity. In *Cheminformatics in Drug Discovery*; Oprea, T. I., Ed.; John Wiley & Sons: 2004; pp 221–239.
- (155) Bicevskis, J.; Bicevska, Z.; Nikiforova, A.; Oditis, I.; Bicevskis, E. J.; Bicevska, Z.; Oditis, I. An Approach to Data Quality Evaluation. *Fifth Int. Conf. Soc. Networks Anal. Manag. Secur.* **2018**, 196–201.
- (156) Fourches, D.; Muratov, E.; Tropsha, A. Curation of Chemo-genomics Data. *Nat. Chem. Biol.* **2015**, *11*, 535.
- (157) Hendren, C. O.; Powers, C. M.; Hoover, M. D.; Harper, S. L. The Nanomaterial Data Curation Initiative: A Collaborative Approach to Assessing, Evaluating, and Advancing the State of the Field. *Beilstein J. Nanotechnol.* **2015**, *6*, 1752–1762.
- (158) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
- (159) Ciallella, H. L.; Zhu, H. Advancing Computational Toxicology in the Big Data Era by Artificial Intelligence: Data-Driven and Mechanism-Driven Modeling for Chemical Toxicity. *Chem. Res. Toxicol.* **2019**, *32*, 536–547.
- (160) Westbrook, J. D.; Soskind, R.; Hudson, B. P.; Burley, S. K. Impact of the Protein Data Bank on Antineoplastic Approvals. *Drug Discovery Today* **2020**, *25*, 837–850.
- (161) Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Židek, A.; Ballard, A. J.; Cowie, A.; Romera-paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; et al. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596*, 583–589.
- (162) Gaheen, S.; Hinkal, G. W.; Morris, S. A.; Lijowski, M.; Heiskanen, M.; Klemm, J. D. CaNanoLab: Data Sharing to Expedite the Use of Nanotechnology in Biomedicine. *Comput. Sci. Discovery* **2013**, *6*, 014010.
- (163) Pomar-Portillo, V.; Park, B.; Crossley, A.; Vázquez-Campos, S. Nanosafety Research in Europe - Towards a Focus on Nano-Enabled Products. *NanoImpact* **2021**, *22*, 100323.
- (164) Jeliazkova, N.; Chomenidis, C.; Doganis, P.; Fadeel, B.; Grafström, R.; Hardy, B.; Hastings, J.; Hegi, M.; Jeliazkov, V.; Kochev, N.; et al. The ENanoMapper Database for Nanomaterial Safety Information. *Beilstein J. Nanotechnol.* **2015**, *6*, 1609–1634.
- (165) Stone, V.; Gottardo, S.; Bleeker, E. A. J.; Braakhuis, H.; Dekkers, S.; Fernandes, T.; Haase, A.; Hunt, N.; Hristozov, D.; Jantunen, P.; et al. A Framework for Grouping and Read-across of Nanomaterials-Supporting Innovation and Risk Assessment. *Nano Today* **2020**, *35*, 100941.
- (166) Tsiliki, G.; Munteanu, C. R.; Seoane, J. A.; Fernandez-Lozano, C.; Sarimveis, H.; Willighagen, E. L. RRegrs: An R Package for Computer-Aided Model Selection with Multiple Regression Models. *J. Cheminform.* **2015**, *7*, 46.
- (167) Helma, C.; Rautenberg, M.; Gebele, D. Nano-Lazar: Read across Predictions for Nanoparticle Toxicities with Calculated and Measured Properties. *Front. Pharmacol.* **2017**, *8*, 377.
- (168) Chomenidis, C.; Drakakis, G.; Tsiliki, G.; Anagnostopoulou, E.; Valsamis, A.; Doganis, P.; Sopasakis, P.; Sarimveis, H. Jaqpot Quattro: A Novel Computational Web Platform for Modeling and Analysis in Nanoinformatics. *J. Chem. Inf. Model.* **2017**, *57*, 2161–2172.
- (169) Joossens, E.; Macko, P.; Palosaari, T.; Gerlo, K.; Ojea-jiménez, I.; Gilliland, D.; Novak, J.; Torrent, S. F.; Gineste, J.; Römer, I.; et al. A High Throughput Imaging Database of Toxicological Effects of Nanomaterials Tested on HepaRG Cells. *Sci. data* **2019**, *6*, 46.
- (170) Boiko, D. A.; Pentsak, E. O.; Cherepano, V. A.; Ananiko, V. P. Electron Microscopy Dataset for the Recognition of Nanoscale Ordering Effects and Location of Nanoparticles. *Sci. data* **2020**, *7*, 101.
- (171) Levin, B. D. A.; Padgett, E.; Chen, C.; Scott, M. C.; Xu, R.; Theis, W.; Jiang, Y.; Yang, Y.; Ophus, C.; Zhang, H.; et al. Nanomaterial Datasets to Advance Tomography in Scanning Transmission Electron Microscopy. *Sci. data* **2016**, *3*, 160041.
- (172) Boyes, W. K.; Bea, B.; Chan, G.; Thornton, B. L. M.; Harten, P.; Mortensen, H. M. An EPA Database on the Effects of Engineered Nanomaterials- NaKnowBase. *Sci. data* **2022**, *9*, 12.
- (173) Karcher, S.; Willighagen, E. L.; Rumble, J.; Ehrhart, F.; Evelo, C. T.; Fritts, M.; Gaheen, S.; Harper, S. L.; Hoover, M. D.; Jeliazkova, N.; et al. Integration among Databases and Data Sets to Support Productive Nanotechnology: Challenges and Recommendations. *NanoImpact* **2018**, *9*, 85–101.
- (174) Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E.; Bouwman, J.; Brookes, A. J.; Clark, T.; Crosas, M.; Dillo, I.; Dumon, O.; Edmunds, S.; Evelo, C. T.; Finkers, R.; Gonzalez-Beltran, A.; Gray, A. J. G.; Groth, P.; Goble, C.; Grethe, J. S.; Heringa, J.; 't Hoen, P. A. C.; Hooft, R.; Kuhn, T.; Kok, R.; Kok, J.; Lusher, S. J.; Martone, M. E.; Mons, A.; Packer, A. L.; Persson, B.; Rocca-Serra, P.; Roos, M.; van Schaik, R.; Sansone, S.-A.; Schultes, E.; Sengstad, T.; Slater, T.; Strawn, G.; Swertz, M. A.; Thompson, M.; van der Lei, J.; van Mulligen, E.; Velterop, J.; Waagmeester, A.; Wittenburg, P.; Wolstencroft, K.; Zhao, J.; Mons, B. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Sci. data* **2016**, *3*, 160018.
- (175) Hastings, J.; Jeliazkova, N.; Owen, G.; Tsiliki, G.; Munteanu, C. R.; Steinbeck, C.; Willighagen, E. eNanoMapper: Harnessing Ontologies to Enable Data Integration for Nanomaterial Risk Assessment. *J. Biomed. Semantics* **2015**, *6*, 10.
- (176) Toropov, A. A.; Leszczynska, D.; Leszczynski, J. Predicting Water Solubility and Octanol Water Partition Coefficient for Carbon Nanotubes Based on the Chiral Vector. *Comput. Biol. Chem.* **2007**, *31*, 127–128.

- (177) Puzyn, T.; Rasulev, B.; Gajewicz, A.; Hu, X.; Dasari, T. P.; Michalkova, A.; Hwang, H. M.; Toropov, A.; Leszczynska, D.; Leszczynski, J. Using Nano-QSAR to Predict the Cytotoxicity of Metal Oxide Nanoparticles. *Nat. Nanotechnol.* **2011**, *6*, 175–178.
- (178) Zhu, Y.; Huang, R.; Zhu, R.; Xu, W.; Zhu, R.; Cheng, L. DeepScreen: An Accurate, Rapid, and Anti-Interference Screening Approach for Nanoformulated Medication by Deep Learning. *Adv. Sci.* **2018**, *5*, 1800909.
- (179) Russo, D. P.; Yan, X.; Shende, S.; Huang, H.; Yan, B.; Zhu, H. Virtual Molecular Projections and Convolutional Neural Networks for the End-to-End Modeling of Nanoparticle Activities and Properties. *Anal. Chem.* **2020**, *92*, 13971–13979.
- (180) Yan, X.; Zhang, J.; Russo, D. P.; Zhu, H.; Yan, B. Prediction of Nano-Bio Interactions through Convolutional Neural Network Analysis of Nanostructure Images. *ACS Sustain. Chem. Eng.* **2020**, *8*, 19096–19104.
- (181) Kurzweil, R.; Richter, R.; Kurzweil, R.; Schneider, M. L. *The Age of Intelligent Machines*; MIT Press: Cambridge, MA, 1990.
- (182) Haenlein, M.; Kaplan, A. A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence. *Calif. Manage. Rev.* **2019**, *61*, 5–14.
- (183) Weizenbaum, J. ELIZA-A Computer Program for the Study of Natural Language Communication Between Man and Machine. *Commun. ACM* **1966**, *9*, 36–45.
- (184) Barker, V. E.; O'Connor, D. E.; Bachant, J.; Soloway, E. Expert Systems for Configuration at Digital: XCON and Beyond. *Commun. ACM* **1989**, *32*, 298–318.
- (185) LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-Based Learning Applied to Document Recognition. *Proc. IEEE* **1998**, *86*, 2278–2323.
- (186) Stokes, J. M.; Yang, K.; Swanson, K.; Jin, W.; Cubillos-Ruiz, A.; Donghia, N. M.; MacNair, C. R.; French, S.; Carfrae, L. A.; Bloom-Ackerman, Z.; et al. A Deep Learning Approach to Antibiotic Discovery. *Cell* **2020**, *180*, 688–702.
- (187) Zhavoronkov, A.; Ivanenkov, Y. A.; Aliper, A.; Veselov, M. S.; Aladinskiy, V. A.; Aladinskaya, A. V.; Terentiev, V. A.; Polykovskiy, D. A.; Kuznetsov, M. D.; Asadulaev, A.; et al. Deep Learning Enables Rapid Identification of Potent DDR1 Kinase Inhibitors. *Nat. Biotechnol.* **2019**, *37*, 1038–1040.
- (188) Zhong, M.; Tran, K.; Min, Y.; Wang, C.; Wang, Z.; Dinh, C.; De Luna, P.; Yu, Z.; Rasouli, A. S.; Brodersen, P.; et al. Accelerated Discovery of CO₂ Electrocatalysts Using Active Machine Learning. *Nature* **2020**, *581*, 178–183.
- (189) Mei, X.; Lee, H.; Diao, K.; Huang, M.; Lin, B.; Liu, C.; Xie, Z.; Ma, Y.; Robson, P. M.; Chung, M.; et al. Artificial Intelligence - Enabled Rapid Diagnosis of Patients with COVID-19. *Nat. Med.* **2020**, *26*, 1224–1228.
- (190) McKinney, S. M.; Sieniek, M.; Godbole, V.; Godwin, J.; Antropova, N.; Ashrafi, H.; Back, T.; Chesus, M.; Corrado, G. S.; Darzi, A.; Etemadi, M.; Garcia-Vicente, F.; Gilbert, F. J.; Halling-Brown, M.; Hassabis, D.; Jansen, S.; Karthikesalingam, A.; Kelly, C. J.; King, D.; Ledam, J. R.; Melnick, D.; Mostofi, H.; Peng, L.; Reicher, J. J.; Romera-Paredes, B.; Sidebottom, R.; Suleyman, M.; Tse, D.; Young, K. C.; De Fauw, J.; Shetty, S. International Evaluation of an AI System for Breast Cancer Screening. *Nat. Biotechnol.* **2020**, *37*, 89–94.
- (191) Thoppilan, R.; De Freitas, D.; Hall, J.; Shazeer, N.; Kulshreshtha, A.; Cheng, H.-T.; Jin, A.; Bos, T.; Baker, L.; Du, Y. LaMDA: Language Models for Dialog Applications. *arXiv*, 2201.08239; 2022. DOI: 10.48550/arXiv.2201.08239.
- (192) Puzyn, T.; Leszczynska, D.; Leszczynski, J. Toward the Development of “Nano-QSARs”: Advances and Challenges. *Small* **2009**, *5*, 2494–2509.
- (193) Jarvis, R. A. A Perspective on Range Finding for Computer Vision. *IEEE Trans. Pattern Anal. Mach. Intell.* **1983**, *PAMI-S*, 122–139.
- (194) Shu-Hsien Liao. Expert System Methodologies and Applications—a Decade Review from 1995 to 2004. *Expert Syst. Appl.* **2005**, *28*, 93–103.
- (195) Nadkarni, P. M.; Ohno-Machado, L.; Chapman, W. W. Natural Language Processing: An Introduction. *J. Am. Med. Informatics Assoc.* **2011**, *18*, 544–551.
- (196) Chapman, T. Lab Automation and Robotics: Automation on the Move. *Nature* **2003**, *421*, 661–663.
- (197) Jordan, M. I.; Mitchell, T. M. Machine Learning: Trends, Perspectives, and Prospects. *Science* **2015**, *349*, 255–260.
- (198) LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444.
- (199) Hu, S.; Liu, G.; Zhang, J.; Yan, J.; Zhou, H.; Yan, X. Linking Electron Ionization Mass Spectra of Organic Chemicals to Toxicity Endpoints through Machine Learning and Experimentation. *J. Hazard. Mater.* **2022**, *431*, 128558.
- (200) Peterson, L. E. K-Nearest Neighbor. *Scholarpedia* **2009**, *4*, 1883.
- (201) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32.
- (202) Karatzoglou, A.; Meyer, D.; Hornik, K. Support Vector Machines in R. *J. Stat. Softw.* **2006**, *15*, 1–28.
- (203) Scarselli, F.; Gori, M.; Tsoi, A. C.; Hagenbuchner, M.; Monfardini, G. The Graph Neural Network Model. *IEEE Trans. Neural Networks* **2009**, *20*, 61–80.
- (204) Lawrence, S.; Giles, C. L.; Tsoi, A. C.; Back, A. D.; Lawrence, S.; Giles, C. L.; Tsoi, A. C.; Back, A. D. Face Recognition: A Convolutional Neural-Network Approach. *IEEE Trans. Neural Networks* **1997**, *8*, 98–113.
- (205) Schuster, M.; Paliwal, K. K. Bidirectional Recurrent Neural Networks. *IEEE Trans. Signal Process.* **1997**, *45*, 2673–2681.
- (206) Roth, V. The Generalized LASSO. *IEEE Trans. Neural Networks* **2004**, *15*, 16–28.
- (207) Hansch, C.; Maloney, P.; Fujita, T.; Muir, R. Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients. *Nature* **1962**, *194*, 178–180.
- (208) Toropov, A. A.; Leszczynska, D.; Leszczynski, J. QSPR Study on Solubility of Fullerene C₆₀ in Organic Solvents Using Optimal Descriptors Calculated with SMILES. *Chem. Phys. Lett.* **2007**, *441*, 119–122.
- (209) Martin, D.; Maran, U.; Sild, S.; Karelson, M. QSPR Modeling of Solubility of Polyaromatic Hydrocarbons and Fullerene in 1-Octanol and n-Heptane. *J. Phys. Chem. B* **2007**, *111*, 9853–9857.
- (210) Toropov, A. A.; Leszczynski, J. A New Approach to the Characterization of Nanomaterials: Predicting Young's Modulus by Correlation Weighting of Nanomaterials Codes. *Chem. Phys. Lett.* **2006**, *433*, 125–129.
- (211) Jones, D. E.; Ghandehari, H.; Facelli, J. C. A Review of the Applications of Data Mining and Machine Learning for the Prediction of Biomedical Properties of Nanoparticles. *Comput. Methods Programs Biomed.* **2016**, *132*, 93–103.
- (212) Bengio, Y.; Courville, A.; Vincent, P. Representation Learning: A Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1798–1828.
- (213) Guyon, I.; Elisseeff, A. An Introduction to Variable and Feature Selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.
- (214) Shahlai, M. Descriptor Selection Methods in Quantitative Structure-Activity. *Chem. Rev.* **2013**, *113*, 8093–8103.
- (215) Danishuddin; Khan, A. U. Descriptors and Their Selection Methods in QSAR Analysis: Paradigm for Drug Design. *Drug Discovery Today* **2016**, *21*, 1291–1302.
- (216) Chandrashekar, G.; Sahin, F. A Survey on Feature Selection Methods. *Comput. Electr. Eng.* **2014**, *40*, 16–28.
- (217) Algamal, Z. Y.; Lee, M. H. A Novel Molecular Descriptor Selection Method in QSAR Classification Model Based on Weighted Penalized Logistic Regression. *J. Chemom.* **2017**, *31*, e2915.
- (218) Beltran, J. A.; Aguilera-Mendoza, L.; Brizuela, C. A. Optimal Selection of Molecular Descriptors for Antimicrobial Peptides Classification: An Evolutionary Feature Weighting Approach. *BMC Genomics* **2018**, *19*, 672.
- (219) Cano, G.; Garcia-Rodriguez, J.; Garcia-Garcia, A.; Perez-Sanchez, H.; Benediktsson, J. A.; Thapa, A.; Barr, A. Automatic Selection of Molecular Descriptors Using Random Forest: Application to Drug Discovery. *Expert Syst. Appl.* **2017**, *72*, 151–159.

- (220) Zeng, Z.; Zhang, H.; Zhang, R.; Yin, C. A Novel Feature Selection Method Considering Feature Interaction. *Pattern Recognit.* **2015**, *48*, 2656–2666.
- (221) Saey, Y.; Inza, I.; Larrañaga, P. A Review of Feature Selection Techniques in Bioinformatics. *Bioinformatics* **2007**, *23*, 2507–2517.
- (222) Young, S. S.; Yuan, F.; Zhu, M. Chemical Descriptors Are More Important than Learning Algorithms for Modelling. *Mol. Inform.* **2012**, *31*, 707–710.
- (223) Vilar, S.; Cozza, G.; Moro, S. Medicinal Chemistry and the Molecular Operating Environment (MOE): Application of QSAR and Molecular Docking to Drug Discovery. *Curr. Top. Med. Chem.* **2008**, *8*, 1555–1572.
- (224) Mauri, A.; Consonni, V.; Pavan, M.; Todeschini, R. DRAGON Software: An Easy Approach to Molecular Descriptor Calculations. *Match* **2006**, *56*, 237–248.
- (225) Mondini, S.; Ferretti, A. M.; Puglisi, A.; Ponti, A. Pebbles and PebbleJuggler: Software for Accurate, Unbiased, and Fast Measurement and Analysis of Nanoparticle Morphology from Transmission Electron Microscopy (TEM) Micrographs. *Nanoscale* **2012**, *4*, 5356–5372.
- (226) Abramoff, M. D.; Magalhães, P. J.; Ram, S. J. Image Processing with ImageJ. *Biophotonics Int.* **2004**, *11*, 36–42.
- (227) Varsou, D. D.; Afantitis, A.; Tsoumanis, A.; Papadiamantis, A.; Valsami-Jones, E.; Lynch, I.; Melagraki, G. Zeta-Potential Read-Across Model Utilizing Nanodescriptors Extracted via the NanoXtract Image Analysis Tool Available on the Enalos Nanoinformatics Cloud Platform. *Small* **2020**, *16*, 1906588.
- (228) Lee, B.; Yoon, S.; Lee, J. H. J. W. J. S.; Kim, Y.; Chang, J.; Yun, J.; Ro, J. C.; Lee, J. H. J. W. J. S.; Lee, J. H. J. W. J. S. Statistical Characterization of the Morphologies of Nanoparticles through Machine Learning Based Electron Microscopy Image Analysis. *ACS Nano* **2020**, *14*, 17125–17133.
- (229) Wang, X.; Li, J.; Ha, H. D. D.; Dahl, J. C. C.; Ondry, J. C. C.; Moreno-Hernandez, I.; Head-Gordon, T.; Alivisatos, A. P. P. AutoDetect-MNP: An Unsupervised Machine Learning Algorithm for Automated Analysis of Transmission Electron Microscope Images of Metal Nanoparticles. *JACS Au* **2021**, *1*, 316–327.
- (230) Bigdeli, A.; Hormozi-Nezhad, M. R.; Jalali-Heravi, M.; Abedini, M. R.; Sharif-Bakhtiar, F. Towards Defining New Nano-Descriptors: Extracting Morphological Features from Transmission Electron Microscopy Images. *RSC Adv.* **2014**, *4*, 60135–60143.
- (231) Bigdeli, A.; Hormozi-Nezhad, M. R. R.; Parastar, H. Using Nano-QSAR to Determine the Most Responsible Factor(s) in Gold Nanoparticle Exocytosis. *RSC Adv.* **2015**, *5*, 57030–57037.
- (232) Liu, G.; Yan, X.; Wang, S.; Yu, Q.; Jia, J.; Yan, B. Elucidation of the Critical Role of Core Materials in PM2.5-Induced Cytotoxicity by Interrogating Silica- And Carbon-Based Model PM2.5 Particle Libraries. *Environ. Sci. Technol.* **2021**, *55*, 6128–6139.
- (233) Cho, W. S.; Duffin, R.; Thielbeer, F.; Bradley, M.; Megson, I. L.; MacNee, W.; Poland, C. A.; Tran, C. L.; Donaldson, K. Zeta Potential and Solubility to Toxic Ions as Mechanisms of Lung Inflammation Caused by Metal/Metal Oxide Nanoparticles. *Toxicol. Sci.* **2012**, *126*, 469–477.
- (234) Liu, R.; Rallo, R.; Weissleder, R.; Tassa, C.; Shaw, S.; Cohen, Y. Nano-SAR Development for Bioactivity of Nanoparticles with Considerations of Decision Boundaries. *Small* **2013**, *9*, 1842–1852.
- (235) Liu, R.; Rallo, R.; George, S.; Ji, Z.; Nair, S.; Nel, A. E. A. E.; Cohen, Y. Classification NanoSAR Development for Cytotoxicity of Metal Oxide Nanoparticles. *Small* **2011**, *7*, 1118–1126.
- (236) Lazarovits, J.; Sindhwani, S.; Tavares, A. J.; Zhang, Y.; Song, F.; Audet, J.; Krieger, J. R.; Syed, A. M.; Stordy, B.; Chan, W. C. W. Supervised Learning and Mass Spectrometry Predicts the in Vivo Fate of Nanomaterials. *ACS Nano* **2019**, *13*, 8023–8034.
- (237) VandeVondele, J.; Hutter, J. Gaussian Basis Sets for Accurate Calculations on Molecular Systems in Gas and Condensed Phases. *J. Chem. Phys.* **2007**, *127*, 114105.
- (238) HAFNER, J. Ab-Initio Simulations of Materials Using VASP: Density-Functional Theory and Beyond. *J. Comput. Chem.* **2008**, *29*, 2044–2078.
- (239) Liu, R.; Zhang, H. Y.; Ji, Z. X.; Rallo, R.; Xia, T.; Chang, C. H.; Nel, A.; Cohen, Y. Development of Structure-Activity Relationship for Metal Oxide Nanoparticles. *Nanoscale* **2013**, *5*, 5644–5653.
- (240) Gajewicz, A.; Schaeublin, N.; Rasulev, B.; Hussain, S.; Leszczynska, D.; Puzyn, T.; Leszczynski, J. Towards Understanding Mechanisms Governing Cytotoxicity of Metal Oxides Nanoparticles: Hints from Nano-QSAR Studies. *Nanotoxicology* **2015**, *9*, 313–325.
- (241) Sifonte, E. P. P.; Castro-Smirnov, F. A. A.; Soutelo, A. A. A.; Jimenez, A. A. S.; Diez, H. R. G.; Martínez, F. G. Quantum Mechanics Descriptors in a Nano-QSAR Model to Predict Metal Oxide Nanoparticles Toxicity in Human Keratinous Cells. *J. Nanoparticle Res.* **2021**, *23*, 161.
- (242) Mikolajczyk, A.; Sizochenko, N.; Mulkiwicz, E.; Malankowska, A.; Rasulev, B.; Puzyn, T. A Chemoinformatics Approach for the Characterization of Hybrid Nanomaterials: Safer and Efficient Design Perspective. *Nanoscale* **2019**, *11*, 11808–11818.
- (243) Huang, Y.; Li, X.; Xu, S.; Zheng, H.; Zhang, L.; Chen, J.; Hong, H.; Kusko, R.; Li, R. Quantitative Structure - Activity Relationship Models for Predicting Inflammatory Potential of Metal Oxide Nanoparticles. *Environ. Health Perspect.* **2020**, *128*, 067010.
- (244) Jagiello, K.; Chomicz, B.; Avramopoulos, A. Size-Dependent Electronic Properties of Nanomaterials: How This Novel Class of Nanodescriptors Supposed to Be Calculated? *Struct. Chem.* **2017**, *28*, 635–643.
- (245) Sauer, J. Ab Initio Calculations for Molecule-Surface Interactions with Chemical Accuracy. *Acc. Chem. Res.* **2019**, *52*, 3502–3510.
- (246) Puzyn, T.; Suzuki, N.; Haranczyk, M.; Rak, J. Calculation of Quantum-Mechanical Descriptors for QSPR at the DFT Level: Is It Necessary? *J. Chem. Inf. Model.* **2008**, *48*, 1174–1180.
- (247) Le, T. C.; Yan, B.; Winkler, D. A. Robust Prediction of Personalized Cell Recognition from a Cancer Population by a Dual Targeting Nanoparticle Library. *Adv. Funct. Mater.* **2015**, *25*, 6927–6935.
- (248) Wang, W.; Feng, S.; Ye, Z.; Gao, H.; Lin, J.; Ouyang, D. Prediction of Lipid Nanoparticles for mRNA Vaccines by the Machine Learning Algorithm. *Acta Pharm. Sin. B* **2022**, *12*, 2950–2962.
- (249) Li, J.; Yue, L.; Zhao, Q.; Cao, X.; Tang, W.; Chen, F.; Wang, C.; Wang, Z. Prediction Models on Biomass and Yield of Rice Affected by Metal (Oxide) Nanoparticles Using Nano-Specific Descriptors. *NanoImpact* **2022**, *28*, 100429.
- (250) Alves, V. M.; Hwang, D.; Muratov, E.; Sokolsky-papkov, M.; Varlamova, E.; Vinod, N.; Lim, C.; Andrade, C. H.; Tropsha, A.; Kabanov, A. Cheminformatics-Driven Discovery of Polymeric Micelle Formulations for Poorly Soluble Drugs. *Sci. Adv.* **2019**, *5*, No. eaav9784.
- (251) Toropov, A. A.; Toropova, A. P.; Benfenati, E.; Leszczynska, D.; Leszczynski, J. SMILES-Based Optimal Descriptors: QSAR Analysis of Fullerene-Based HIV-1 PR Inhibitors by Means of Balance of Correlations. *J. Comput. Chem.* **2009**, *31*, 381–392.
- (252) Fourches, D.; Pu, D.; Li, L.; Zhou, H.; Mu, Q.; Su, G.; Yan, B.; Tropsha, A. Computer-Aided Design of Carbon Nanotubes with the Desired Bioactivity and Safety Profiles. *Nanotoxicology* **2016**, *10*, 374–383.
- (253) Sizochenko, N.; Rasulev, B.; Gajewicz, A.; Kuz'Min, V.; Puzyn, T.; Leszczynski, J. From Basic Physics to Mechanisms of Toxicity: The “Liquid Drop” Approach Applied to Develop Predictive Classification Models for Toxicity of Metal Oxide Nanoparticles. *Nanoscale* **2014**, *6*, 13986–13993.
- (254) Toropov, A. A.; Toropova, A. P. Quasi-SMILES and Nano-QFAR: United Model for Mutagenicity of Fullerene and MWCNT under Different Conditions. *Chemosphere* **2015**, *139*, 18–22.
- (255) Toropov, A. A.; Toropova, A. P.; Farmacologice, R.; Negri, M.; Negri, V. M.; Carlo, M. Quasi-SMILES as a Basis for the Development of Models for the Toxicity of ZnO Nanoparticles. *Sci. Total Environ.* **2021**, *772*, 145532.
- (256) Leone, C.; Bertuzzi, E. E.; Toropova, A. P.; Toropov, A. A.; Benfenati, E. CORAL: Predictive Models for Cytotoxicity of Functionalized Nanoszeolites Based on Quasi-SMILES. *Chemosphere* **2018**, *210*, 52–56.

- (257) Toropov, A. A.; Toropova, A. P. The Correlation Contradictions Index (CCI): Building up Reliable Models of Mutagenic Potential of Silver Nanoparticles under Different Conditions Using Quasi-SMILES. *Sci. Total Environ.* **2019**, *681*, 102–109.
- (258) Trinh, T. X.; Choi, J. S.; Jeon, H.; Byun, H. G.; Yoon, T. H.; Kim, J. Quasi-SMILES-Based Nano-Quantitative Structure-Activity Relationship Model to Predict the Cytotoxicity of Multiwalled Carbon Nanotubes to Human Lung Cells. *Chem. Res. Toxicol.* **2018**, *31*, 183–190.
- (259) Kuz'min, V. E.; Artemenko, A. G.; Polischuk, P. G.; Muratov, E. N.; Hromov, A. I.; Liahovskiy, A. V.; Andronati, S. A.; Makan, S. Y. Hierarchic System of QSAR Models (1D - 4D) on the Base of Simplex Representation of Molecular Structure. *J. Mol. Model.* **2005**, *11*, 457–467.
- (260) Muratov, E. N.; Varlamova, E. V.; Artemenko, A. G.; Polishchuk, P. G. Existing and Developing Approaches for QSAR Analysis of Mixtures. *Mol. Inform.* **2012**, *31*, 202–221.
- (261) Kar, S.; Gajewicz, A.; Puzyn, T.; Roy, K.; Leszczynski, J. Periodic Table-Based Descriptors to Encode Cytotoxicity Profile of Metal Oxide Nanoparticles: A Mechanistic QSTR Approach. *Ecotoxicol. Environ. Saf.* **2014**, *107*, 162–169.
- (262) De, P.; Kar, S.; Roy, K.; Leszczynski, J. Second Generation Periodic Table-Based Descriptors to Encode Toxicity of Metal Oxide Nanoparticles to Multiple Species: QSTR Modeling for Exploration of Toxicity Mechanisms. *Environ. Sci. Nano* **2018**, *5*, 2742–2760.
- (263) Wigner, E.; Seitz, F. On the Constitution of Metallic Sodium. *Phys. Rev.* **1933**, *43*, 804.
- (264) Tamm, K.; Sikk, L.; Burk, J.; Rallo, R.; Pokhrel, S.; Mädler, L.; Scott-Fordsmand, J. J.; Burk, P.; Tamm, T. Parametrization of Nanoparticles: Development of Full-Particle Nanodescriptors. *Nanoscale* **2016**, *8*, 16243–16250.
- (265) Manshian, B. B.; Pokhrel, S.; Himmelreich, U.; Tamm, K.; Sikk, L.; Fernández, A.; Rallo, R.; Tamm, T.; Mädler, L.; Soenen, S. J. In Silico Design of Optimal Dissolution Kinetics of Fe-Doped ZnO Nanoparticles Results in Cancer-Specific Toxicity in a Preclinical Rodent Model. *Adv. Healthc. Mater.* **2017**, *6*, 1601379.
- (266) Burk, J.; Sikk, L.; Burk, P.; Manshian, B. B.; Soenen, S. J.; Scott-Fordsmand, J. J.; Tamm, T.; Tamm, K. Fe-Doped ZnO Nanoparticle Toxicity: Assessment by a New Generation of Nanodescriptors. *Nanoscale* **2018**, *10*, 21985–21993.
- (267) Liu, G.; Yan, X.; Sedykh, A.; Pan, X.; Zhao, X.; Yan, B.; Zhu, H. Analysis of Model PM2.5-Induced Inflammation and Cytotoxicity by the Combination of a Virtual Carbon Nanoparticle Library and Computational Modeling. *Ecotoxicol. Environ. Saf.* **2020**, *191*, 110216.
- (268) Gebhardt, J.; Kiesel, M.; Riniker, S.; Hansen, N. Combining Molecular Dynamics and Machine Learning to Predict Self-Solvation Free Energies and Limiting Activity Coefficients. *J. Chem. Inf. Model.* **2020**, *60*, 5319–5330.
- (269) Esposito, C.; Wang, S.; Lange, U. E. W.; Oellien, F.; Riniker, S. Combining Machine Learning and Molecular Dynamics to Predict P-Glycoprotein Substrates. *J. Chem. Inf. Model.* **2020**, *60*, 4730–4749.
- (270) Kyaw Zin, P. P.; Borrel, A.; Fourches, D. Benchmarking 2D/3D/MD-QSAR Models for Imatinib Derivatives: How Far Can We Predict? *J. Chem. Inf. Model.* **2020**, *60*, 3342–3360.
- (271) Isayev, O.; Oses, C.; Toher, C.; Gossett, E.; Curtarolo, S.; Tropsha, A. Universal Fragment Descriptors for Predicting Properties of Inorganic Crystals. *Nat. Commun.* **2017**, *8*, 15679.
- (272) AlQuraishi, M. End-to-End Differentiable Learning of Protein Structure. *Cell Syst.* **2019**, *8*, 292–301.
- (273) Hu, Z.; Tang, A.; Singh, J.; Bhattacharya, S.; Butte, A. J. A Robust and Interpretable End-to-End Deep Learning Model for Cytometry Data. *Proc. Natl. Acad. Sci. U. S. A.* **2020**, *117*, 21373–21380.
- (274) Ardila, D.; Kiraly, A. P.; Bharadwaj, S.; Choi, B.; Reicher, J. J.; Peng, L.; Tse, D.; Etemadi, M.; Ye, W.; Corrado, G.; et al. End-to-End Lung Cancer Screening with Three-Dimensional Deep Learning on Low-Dose Chest Computed Tomography. *Nat. Med.* **2019**, *25*, 954–961.
- (275) Moen, E.; Bannon, D.; Kudo, T.; Graf, W.; Covert, M.; Van Valen, D. Deep Learning for Cellular Image Analysis. *Nat. Methods* **2019**, *16*, 1233–1246.
- (276) Smith, K.; Piccinini, F.; Balassa, T.; Koos, K.; Danka, T.; Azizpour, H.; Horvath, P. Phenotypic Image Analysis Software Tools for Exploring and Understanding Big Image Data from Cell-Based Assays. *Cell Syst.* **2018**, *6*, 636–653.
- (277) Ouyang, W.; Aristov, A.; Lelek, M.; Hao, X.; Zimmer, C. Deep Learning Massively Accelerates Super-Resolution Localization Microscopy. *Nat. Biotechnol.* **2018**, *36*, 460–468.
- (278) Rivenson, Y.; Wang, H.; Wei, Z.; Haan, K. De; Zhang, Y.; Wu, Y.; Günaydin, H.; Zuckerman, J. E.; Chong, T.; Sisk, A. E.; et al. Virtual Histological Staining of Unlabelled Tissue-Autofluorescence Images via Deep Learning. *Nat. Biomed. Eng.* **2019**, *3*, 466–477.
- (279) Tang, Y.; Zhang, J.; He, D.; Miao, W.; Liu, W.; Li, Y.; Lu, G.; Wu, F.; Wang, S. GANDA: A Deep Generative Adversarial Network Conditionally Generates Intratumoral Nanoparticles Distribution Pixels-to-Pixels. *J. Controlled Release* **2021**, *336*, 336–343.
- (280) Hollon, T. C.; Pandian, B.; Adapa, A. R.; Urias, E.; Save, A. V.; Khalsa, S. S. S.; Eichberg, D. G.; Amico, R. S. D.; Farooq, Z. U.; Lewis, S.; et al. Near Real-Time Intraoperative Brain Tumor Diagnosis Using Stimulated Raman Histology and Deep Neural Networks. *Nat. Med.* **2020**, *26*, 52–58.
- (281) Esteva, A.; Kuprel, B.; Novoa, R. A.; Ko, J.; Swetter, S. M.; Blau, H. M.; Thrun, S. Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks. *Nature* **2017**, *542*, 115–118.
- (282) Montavon, G.; Samek, W.; Müller, K.-R. Methods for Interpreting and Understanding Deep Neural Networks. *Digit. Signal Process.* **2018**, *73*, 1–15.
- (283) Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-Cam: Visual Explanations from Deep Networks via Gradient-Based Localization. *Proceedings of the IEEE international conference on computer vision*; 2017; pp 618–626.
- (284) Xiong, Z.; Wang, D.; Liu, X.; Zhong, F.; Wan, X.; Li, X.; Li, Z.; Luo, X.; Chen, K.; Jiang, H.; et al. Pushing the Boundaries of Molecular Representation for Drug Discovery with the Graph Attention Mechanism. *J. Med. Chem.* **2020**, *63*, 8749–8760.
- (285) Cha, M.; Emre, E. S. T.; Xiao, X.; Kim, J. Y.; Bogdan, P.; VanEpps, J. S.; Violi, A.; Kotov, N. A. Unifying Structural Descriptors for Biological and Bioinspired Nanoscale Complexes. *Nat. Comput. Sci.* **2022**, *2*, 243–252.
- (286) Sedykh, A.; Zhu, H.; Tang, H.; Zhang, L.; Richard, A.; Rusyn, I.; Tropsha, A. Use of in Vitro HTS-Derived Concentration-Response Data as Biological Descriptors Improves the Accuracy of QSAR Models of in Vivo Toxicity. *Environ. Health Perspect.* **2011**, *119*, 364–370.
- (287) Wang, W.; Kim, M. T.; Sedykh, A.; Zhu, H. Developing Enhanced Blood - Brain Barrier Permeability Models: Integrating External Bio-Assay Data in QSAR Modeling. *Pharm. Res.* **2015**, *32*, 3055–3065.
- (288) Gajewicz, A.; Cronin, M. T. D.; Rasulev, B. Novel Approach for Efficient Predictions Properties of Large Pool of Nanomaterials Based on Limited Set of Species: Nano-Read-Across. *Nanotechnology* **2015**, *26*, 015701.
- (289) Gajewicz, A. What If the Number of Nanotoxicity Data Is Too Small for Developing Predictive Nano-QSAR Models? An Alternative Read-across Based Approach for Filling Data Gaps. *Nanoscale* **2017**, *9*, 8435–8448.
- (290) Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A. V.; Gulin, A. Catboost: Unbiased Boosting with Categorical Features. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 6638–6648.
- (291) Zhang, X.; Yan, C.; Gao, C.; Malin, B. A.; Chen, Y. Predicting Missing Values in Medical Data Via XGBoost Regression. *J. Healthc. informatics Res.* **2020**, *4*, 383–394.
- (292) Lin, W.; Tsai, C.; Hu, Y.; Jhang, J. Clustering-Based Undersampling in Class-Imbalanced Data. *Inf. Sci. (Nijl.)* **2017**, *409*, 17–26.
- (293) Galar, M.; Fern, A.; Barrenechea, E.; Bustince, H. A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and

- Hybrid-Based Approaches. *IEEE Trans. Syst. Man, Cybern. Part C (Applications Rev.)* **2012**, *42*, 463–484.
- (294) Bolón-Canedo, V.; Sánchez-Marño, N.; Alonso-Betanzos, A. A Review of Feature Selection Methods on Synthetic Data. *Knowl. Inf. Syst.* **2013**, *34*, 483–519.
- (295) Xu, Z.; Zhang, H.; Wang, Y.; Chang, X.; Liang, Y. L1/2 Regularization. *Sci. China, Ser. F Inf. Sci.* **2010**, *53*, 1159–1169.
- (296) Abdi, H.; Williams, L. J. Principal Component Analysis. *Wiley Interdiscip. Rev. Comput. Stat.* **2010**, *2*, 433–459.
- (297) Belkina, A. C.; Ciccolella, C. O.; Anno, R.; Halpert, R.; Snyder-cappione, J. E. Automated Optimized Parameters for T-Distributed Stochastic Neighbor Embedding Improve Visualization and Analysis of Large Datasets. *Nat. Commun.* **2019**, *10*, 5415.
- (298) McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv*, 1802.03426; 2018; DOI: 10.48550/arXiv.1802.03426.
- (299) Odhiambo Omuya, E.; Onyango Okeyo, G.; Waema Kimwele, M. Feature Selection for Classification Using Principal Component Analysis and Information Gain. *Expert Syst. Appl.* **2021**, *174*, 114765.
- (300) Li, J.; Gao, X.; Wang, Y.; Xia, T.; Zhao, Y.; Meng, H. Review Precision Design of Engineered Nanomaterials to Guide Immune Systems for Disease Treatment. *Matter* **2022**, *5*, 1162–1191.
- (301) Zhu, M.; Nie, G.; Meng, H.; Xia, T.; Nel, A.; Zhao, Y. Physicochemical Properties Determine Nanomaterial Cellular Uptake, Transport, and Fate. *Acc. Chem. Res.* **2013**, *46*, 622–631.
- (302) McDowell, M. T.; Ryu, L.; Lee, S. W.; Wang, C.; Nix, W. D.; Cui, Y. Studying the Kinetics of Crystalline Silicon Nanoparticle Lithiation with in Situ Transmission Electron Microscopy. *Adv. Mater.* **2012**, *24*, 6034–6041.
- (303) Kasuya, A.; Sivamohan, R.; Barnakov, Y. A.; Dmitruk, I. M.; Nirasawa, T.; Romanyuk, V. R.; Kumar, V.; Mamykin, S. V.; Tohji, K.; Jeyadevan, B.; et al. Ultra-Stable Nanoparticles of CdSe Revealed from Mass Spectrometry. *Nat. Mater.* **2004**, *3*, 99–102.
- (304) Li, T.; Senesi, A. J.; Lee, B. Small Angle X-Ray Scattering for Nanoparticle Research. *Chem. Rev.* **2016**, *116*, 11128–11180.
- (305) Patra, T. K.; Zhang, F.; Daniel, S. S.; Chan, H.; Cherukara, M. J.; Terrones, M.; Das, S.; Narayanan, B.; Sankaranarayanan, S. K. R. S. Defect Dynamics in 2-D MoS₂ Probed by Using Machine Learning, Atomistic Simulations, and High-Resolution Microscopy. *ACS Nano* **2018**, *12*, 8006–8016.
- (306) Luo, Q.; Holm, E. A.; Wang, C. A Transfer Learning Approach for Improved Classification of Carbon Nanomaterials from TEM Images. *Nanoscale Adv.* **2021**, *3*, 206–213.
- (307) Carrera, D.; Manganini, F.; Boracchi, G.; Lanzarone, E. Defect Detection in SEM Images of Nanofibrous Materials. *IEEE Trans. Ind. Informatics* **2017**, *13*, 551–561.
- (308) Maksov, A.; Dyck, O.; Wang, K.; Xiao, K.; Geohegan, D. B.; Sumpter, B. G.; Vasudevan, R. K.; Jesse, S.; Kalinin, S. V.; Ziatdinov, M. Deep Learning Analysis of Defect and Phase Evolution during Electron Beam-Induced Transformations in WS₂. *npj Comput. Mater.* **2019**, *5*, 12.
- (309) Yao, L.; Ou, Z.; Luo, B.; Xu, C.; Chen, Q. Machine Learning to Reveal Nanoparticle Dynamics from Liquid-Phase TEM Videos. *ACS Cent. Sci.* **2020**, *6*, 1421–1430.
- (310) Manshian, B. B.; Moyano, D. F.; Corthout, N.; Munck, S.; Himmelreich, U.; Rotello, V. M.; Soenen, S. J. High-Content Imaging and Gene Expression Analysis to Study Cell-Nanomaterial Interactions: The Effect of Surface Hydrophobicity. *Biomaterials* **2014**, *35*, 9941–9950.
- (311) Lai, R. W. S.; Kang, H. M.; Zhou, G. J.; Yung, M. M. N.; He, Y. L.; Ng, A. M. C.; Li, X. Y.; Djurišić, A. B.; Lee, J. S.; Leung, K. M. Y. Hydrophobic Surface Coating Can Reduce Toxicity of Zinc Oxide Nanoparticles to the Marine Copepod *Tigriopus Japonicus*. *Environ. Sci. Technol.* **2021**, *55*, 6917–6925.
- (312) Wang, W.; Yan, X.; Zhao, L.; Russo, D. P.; Wang, S.; Liu, Y.; Sedykh, A.; Zhao, X.; Yan, B.; Zhu, H. Universal Nanohydrophobicity Predictions Using Virtual Nanoparticle Library. *J. Cheminform.* **2019**, *11*, 4–8.
- (313) Swirow, M.; Mikolajczyk, A.; Jagiello, K.; Jänes, J.; Tamm, K.; Puzyn, T. Predicting Electrophoretic Mobility of TiO₂, ZnO, and CeO₂ Nanoparticles in Natural Waters: The Importance of Environment Descriptors in Nanoinformatics Models. *Sci. Total Environ.* **2022**, *840*, 156572.
- (314) Mikolajczyk, A.; Gajewicz, A.; Rasulev, B.; Schaeublin, N.; Maurer-gardner, E.; Hussain, S.; Leszczynski, J.; Puzyn, T. Zeta Potential for Metal Oxide Nanoparticles: A Predictive Model Developed by a Nano-Quantitative Structure - Property Relationship Approach. *Chem. Mater.* **2015**, *27*, 2400–2407.
- (315) Xia, X. R.; Monteiro-Riviere, N. A.; Riviere, J. E. An Index for Characterization of Nanomaterials in Biological Systems. *Nat. Nanotechnol.* **2010**, *5*, 671–675.
- (316) Wang, X.; Liu, L.; Zhang, W.; Ma, X. Prediction of Plant Uptake and Translocation of Engineered Metallic Nanoparticles by Machine Learning. *Environ. Sci. Technol.* **2021**, *55*, 7491–7500.
- (317) Cedervall, T.; Lynch, I.; Lindman, S.; Berggård, T.; Thulin, E.; Nilsson, H.; Dawson, K. A.; Linse, S. Understanding the Nanoparticle-Protein Corona Using Methods to Quantify Exchange Rates and Affinities of Proteins for Nanoparticles. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 2050–2055.
- (318) Wheeler, K. E.; Chetwynd, A. J.; Fahy, K. M.; Hong, B. S.; Tochihiuti, J. A.; Foster, L. A.; Lynch, I. Environmental Dimensions of the Protein Corona. *Nat. Nanotechnol.* **2021**, *16*, 617–629.
- (319) Chetwynd, A. J.; Lynch, I. The Rise of the Nanomaterial Metabolite Corona, and Emergence of the Complete Corona. *Environ. Sci. Nano* **2020**, *7*, 1041–1060.
- (320) Raesch, S. S.; Tenzer, S.; Storck, W.; Rurainski, A.; Selzer, D.; Ruge, C. A.; Perez-Gil, J.; Schaefer, U. F.; Lehr, C. M. Proteomic and Lipidomic Analysis of Nanoparticle Corona upon Contact with Lung Surfactant Reveals Differences in Protein, but Not Lipid Composition. *ACS Nano* **2015**, *9*, 11872–11885.
- (321) Liu, X.; Zhao, Y.; Dou, J.; Hou, Q.; Cheng, J.; Jiang, X. Bioeffects of Inhaled Nanoplastics on Neurons and Alteration of Animal Behaviors through Deposition in the Brain. *Nano Lett.* **2022**, *22*, 1091–1099.
- (322) Xu, L.; Xu, M.; Wang, R.; Yin, Y.; Lynch, I.; Liu, S. The Crucial Role of Environmental Coronas in Determining the Biological Effects of Engineered Nanomaterials. *Small* **2020**, *16*, 2003691.
- (323) Junaid, M.; Wang, J. Interaction of Nanoplastics with Extracellular Polymeric Substances (EPS) in the Aquatic Environment: A Special Reference to Eco-Corona Formation and Associated Impacts. *Water Res.* **2021**, *201*, 117319.
- (324) Ke, P. C.; Lin, S.; Parak, W. J.; Davis, T. P.; Caruso, F. A Decade of the Protein Corona. *ACS Nano* **2017**, *11*, 11773–11776.
- (325) Mahmoudi, M.; Kalhor, H. R.; Laurent, S.; Lynch, I. Protein Fibrillation and Nanoparticle Interactions: Opportunities and Challenges. *Nanoscale* **2013**, *5*, 2570–2588.
- (326) Cai, R.; Chen, C. The Crown and the Scepter: Roles of the Protein Corona in Nanomedicine. *Adv. Mater.* **2019**, *31*, 1805740.
- (327) Zhang, J.; Guo, W.; Li, Q.; Wang, Z.; Liu, S. The Effects and the Potential Mechanism of Environmental Transformation of Metal Nanoparticles on Their Toxicity in Organisms. *Environ. Sci. Nano* **2018**, *5*, 2482–2499.
- (328) Oh, J. Y.; Kim, H. S.; Palanikumar, L.; Go, E. M.; Jana, B.; Park, S. A.; Kim, H. Y.; Kim, K.; Seo, J. K.; Kwak, S. K.; et al. Cloaking Nanoparticles with Protein Corona Shield for Targeted Drug Delivery. *Nat. Commun.* **2018**, *9*, 4548.
- (329) Zhang, X.; Liu, Y.; Gopalakrishnan, S.; Castellanos-Garcia, L.; Li, G.; Malassiné, M.; Uddin, I.; Huang, R.; Luther, D. C.; Vachet, R. W.; et al. Intracellular Activation of Bioorthogonal Nanozymes through Endosomal Proteolysis of the Protein Corona. *ACS Nano* **2020**, *14*, 4767–4773.
- (330) Treuel, L.; Brandholt, S.; Maffre, P.; Wiegele, S.; Shang, L.; Nienhaus, G. U. Impact of Protein Modification on the Protein Corona on Nanoparticles and Nanoparticle-Cell Interactions. *ACS Nano* **2014**, *8*, 503–513.
- (331) Findlay, M. R.; Freitas, D. N.; Mobed-Miremadi, M.; Wheeler, K. E. Machine Learning Provides Predictive Analysis into Silver

Nanoparticle Protein Corona Formation from Physicochemical Properties. *Environ. Sci. Nano* **2018**, *5*, 64–71.

(332) Duan, Y.; Coreas, R.; Liu, Y.; Bitounis, D.; Zhang, Z.; Parviz, D.; Strano, M.; Demokritou, P.; Zhong, W. Prediction of Protein Corona on Nanomaterials by Machine Learning Using Novel Descriptors. *NanoImpact* **2020**, *17*, 100207.

(333) Liu, R.; Jiang, W.; Walkey, C. D.; Chan, W. C. W. W.; Cohen, Y. Prediction of Nanoparticles-Cell Association Based on Corona Proteins and Physicochemical Properties. *Nanoscale* **2015**, *7*, 9664–9675.

(334) Papa, E.; Doucet, J. P.; Sangion, A.; Doucet-Panaye, A. Investigation of the Influence of Protein Corona Composition on Gold Nanoparticle Bioactivity Using Machine Learning Approaches. *SAR QSAR Environ. Res.* **2016**, *27*, 521–538.

(335) Afantitis, A.; Melagraki, G.; Tsoumanis, A.; Valsami-Jones, E.; Lynch, I. A Nanoinformatics Decision Support Tool for the Virtual Screening of Gold Nanoparticle Cellular Association Using Protein Corona Fingerprints. *Nanotoxicology* **2018**, *12*, 1148–1165.

(336) Doğangün, M.; Hang, M. N.; Troiano, J. M.; McGeachy, A. C.; Melby, E. S.; Pedersen, J. A.; Hamers, R. J.; Geiger, F. M. Alteration of Membrane Compositional Asymmetry by LiCoO₂ Nanosheets. *ACS Nano* **2015**, *9*, 8755–8765.

(337) Konduru, N. V.; Damiani, F.; Stoilova-McPhie, S.; Tresback, J. S.; Pyrgiotakis, G.; Donaghey, T. C.; Demokritou, P.; Brain, J. D.; Molina, R. M. Nanoparticle Wettability Influences Nanoparticle-Phospholipid Interactions. *Langmuir* **2018**, *34*, 6454–6461.

(338) Behyan, S.; Borozenko, O.; Khan, A.; Faral, M.; Badia, A.; Dewolf, C. Nanoparticle-Induced Structural Changes in Lung Surfactant Membranes: An X-Ray Scattering Study. *Environ. Sci. Nano* **2018**, *5*, 1218–1230.

(339) Panikkanvalappil, S. R.; Mahmoud, M. A.; MacKey, M. A.; El-Sayed, M. A. Surface-Enhanced Raman Spectroscopy for Real-Time Monitoring of Reactive Oxygen Species-Induced DNA Damage and Its Prevention by Platinum Nanoparticles. *ACS Nano* **2013**, *7*, 7524–7533.

(340) Liu, Y. L.; Perillo, E. P.; Ang, P.; Kim, M.; Nguyen, D. T.; Blocher, K.; Chen, Y. A.; Liu, C.; Hassan, A. M.; Vu, H. T.; et al. Three-Dimensional Two-Color Dual-Particle Tracking Microscope for Monitoring DNA Conformational Changes and Nanoparticle Landings on Live Cells. *ACS Nano* **2020**, *14*, 7927–7939.

(341) Wang, J.; Li, P.; Yu, Y.; Fu, Y.; Jiang, H.; Lu, M.; Sun, Z.; Jiang, S.; Lu, L.; Wu, M. X. Pulmonary Surfactant-Biomimetic Nanoparticles Potentiate Heterosubtypic Influenza Immunity. *Science* **2020**, *367*, No. eaau0810.

(342) Li, H.; Tao, X.; Song, E.; Song, Y. Iron Oxide Nanoparticles Oxidize Transformed RAW 264.7 Macrophages into Foam Cells: Impact of Pulmonary Surfactant Component Dipalmitoylphosphatidylcholine. *Chemosphere* **2022**, *300*, 134617.

(343) Li, J.; Yang, H.; Sha, S.; Li, J.; Zhou, Z.; Cao, Y. Evaluation of in Vitro Toxicity of Silica Nanoparticles (NPs) to Lung Cells: Influence of Cell Types and Pulmonary Surfactant Component DPPC. *Ecotoxicol. Environ. Saf.* **2019**, *186*, 109770.

(344) Leo, B. F.; Chen, S.; Kyo, Y.; Herpoldt, K. L.; Terrill, N. J.; Dunlop, I. E.; McPhail, D. S.; Shaffer, M. S.; Schwander, S.; Gow, A.; et al. The Stability of Silver Nanoparticles in a Model of Pulmonary Surfactant. *Environ. Sci. Technol.* **2013**, *47*, 11232–11240.

(345) Kelich, P.; Jeong, S.; Navarro, N.; Adams, J.; Sun, X.; Zhao, H.; Landry, M. P.; Vuković, L. Discovery of DNA-Carbon Nanotube Sensors for Serotonin with Machine Learning and Near-Infrared Fluorescence Spectroscopy. *ACS Nano* **2022**, *16*, 736–745.

(346) Yang, Y.; Zheng, M.; Jagota, A. Learning to Predict Single-Wall Carbon Nanotube-Recognition DNA Sequences. *npj Comput. Mater.* **2019**, *5*, 3.

(347) Verma, A.; Stellacci, F. Effect of Surface Properties on Nanoparticle-Cell Interactions. *Small* **2010**, *6*, 12–21.

(348) Jiang, W.; Kim, B. Y. S.; Rutka, J. T.; Chan, W. C. W. Nanoparticle-Mediated Cellular Response Is Size-Dependent. *Nat. Nanotechnol.* **2008**, *3*, 145–150.

(349) Kaksonen, M.; Roux, A. Mechanisms of Clathrin-Mediated Endocytosis. *Nat. Rev. Mol. Cell Biol.* **2018**, *19*, 313–326.

(350) Brunet, S.; Sardon, T.; Zimmerman, T.; Wittmann, T.; Pepperkok, R.; Karsenti, E.; Vernos, I. Characterization of the TPX2 Domains Involved in Microtubule Nucleation and Spindle Assembly in Xenopus Nucleation around Chromatin and Functions in a Network of Other Molecules, Some of Which Also Are Regulated By. *Mol. Biol. Cell* **2004**, *15*, 5318–5328.

(351) Cao, H.; Chen, J.; Awoniyi, M.; Henley, J. R.; McNiven, M. A. Dynamin 2 Mediates Fluid-Phase Micropinocytosis in Epithelial Cells. *J. Cell Sci.* **2007**, *120*, 4167–4177.

(352) Gordon, S. Phagocytosis: An Immunobiologic Process. *Immunity* **2016**, *44*, 463–475.

(353) Drescher, D.; Giesen, C.; Traub, H.; Panne, U.; Kneipp, J.; Jakubowski, N. Quantitative Imaging of Gold and Silver Nanoparticles in Single Eukaryotic Cells by Laser Ablation ICP-MS. *Anal. Chem.* **2012**, *84*, 9684–9688.

(354) Schwertfeger, D. M.; Velicogna, J. R.; Jesmer, A. H.; Saatcioglu, S.; McShane, H.; Scroggins, R. P.; Princz, J. I. Extracting Metallic Nanoparticles from Soils for Quantitative Analysis: Method Development Using Engineered Silver Nanoparticles and SP-ICP-MS. *Anal. Chem.* **2017**, *89*, 2505–2513.

(355) Lu, G.; Cihfield, C. L.; Gattu, S.; Veltri, L. M.; Holland, L. A. Capillary Electrophoresis Separations of Glycans. *Chem. Rev.* **2018**, *118*, 7867–7885.

(356) Gygi, S. P.; Corthals, G. L.; Zhang, Y.; Rochon, Y.; Aebersold, R. Evaluation of Two-Dimensional Gel Electrophoresis-Based Proteome Analysis Technology. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97*, 9390–9395.

(357) Bigdeli, A.; Palchetti, S.; Pozzi, D.; Hormozi-Nezhad, M. R.; Baldelli Bombelli, F.; Caracciolo, G.; Mahmoudi, M. Exploring Cellular Interactions of Liposomes Using Protein Corona Fingerprints and Physicochemical Properties. *ACS Nano* **2016**, *10*, 3723–3737.

(358) Ghorbanzadeh, M.; Fatemi, M. H. M. H.; Karimpour, M. Modeling the Cellular Uptake of Magnetofluorescent Nanoparticles in Pancreatic Cancer Cells: A Quantitative Structure Activity Relationship Study. *Ind. Eng. Chem. Res.* **2012**, *51*, 10712–10718.

(359) Toropov, A. A.; Toropova, A. P.; Puzyn, T.; Benfenati, E.; Gini, G.; Leszczynska, D.; Leszczynski, J. QSAR as a Random Event: Modeling of Nanoparticles Uptake in PaCa2 Cancer Cells. *Chemosphere* **2013**, *92*, 31–37.

(360) Kar, S.; Gajewicz, A.; Puzyn, T.; Roy, K. Nano-Quantitative Structure-Activity Relationship Modeling Using Easily Computable and Interpretable Descriptors for Uptake of Magnetofluorescent Engineered Nanoparticles in Pancreatic Cancer Cells. *Toxicol. Vitro* **2014**, *28*, 600–606.

(361) Chau, Y. T.; Yap, C. W. Quantitative Nanostructure-Activity Relationship Modelling of Nanoparticles. *RSC Adv.* **2012**, *2*, 8489–8496.

(362) Melagraki, G.; Afantitis, A. Enalos InSilicoNano Platform: An Online Decision Support Tool for the Design and Virtual Screening of Nanoparticles. *RSC Adv.* **2014**, *4*, 50713–50725.

(363) Shi, H.; Pan, Y.; Yang, F.; Cao, J.; Tan, X.; Yuan, B.; Jiang, J. Oxide Nanoparticles to PaCa2. *Molecules* **2021**, *26*, 2188.

(364) Luan, F.; Tang, L.; Zhang, L.; Zhang, S.; Monteagudo, M. C.; Cordeiro, M. N. D. S. A Further Development of the QNAR Model to Predict the Cellular Uptake of Nanoparticles by Pancreatic Cancer Cells. *Food Chem. Toxicol.* **2018**, *112*, 571–580.

(365) Qi, R.; Pan, Y.; Cao, J.; Jia, Z.; Jiang, J. The Cytotoxicity of Nanomaterials: Modeling Multiple Human Cells Uptake of Functionalized Magneto-Fluorescent Nanoparticles via Nano-QSAR. *Chemosphere* **2020**, *249*, 126175.

(366) Ojha, P. K.; Kar, S.; Roy, K.; Leszczynski, J. Toward Comprehension of Multiple Human Cells Uptake of Engineered Nano Metal Oxides: Quantitative Inter Cell Line Uptake Specificity (QICLUS) Modeling. *Nanotoxicology* **2019**, *13*, 14–34.

(367) Papa, E.; Doucet, J. P.; Doucet-Panaye, A. Computational Approaches for the Prediction of the Selective Uptake of Magneto-fluorescent Nanoparticles into Human Cells. *RSC Adv.* **2016**, *6*, 68806–68818.

- (368) Trinh, T. X. X.; Ha, M. K. K.; Choi, J. S. S.; Byun, H. G. G.; Yoon, T. H. H. Curation of Datasets, Assessment of Their Quality and Completeness, and NanoSAR Classification Model Development for Metallic Nanoparticles. *Environ. Sci. Nano* **2018**, *5*, 1902–1910.
- (369) Stoliński, F.; Rybińska-Fryca, A.; Gromelski, M.; Mikolajczyk, A.; Puzyn, T. NanoMixHamster: A Web-Based Tool for Predicting Cytotoxicity of TiO₂-Based Multicomponent Nanomaterials toward Chinese Hamster Ovary (CHO-K1) Cells. *Nanotoxicology* **2022**, *16*, 276–289.
- (370) Ahmadi, S. Mathematical Modeling of Cytotoxicity of Metal Oxide Nanoparticles Using the Index of Ideality Correlation Criteria. *Chemosphere* **2020**, *242*, 125192.
- (371) Papa, E.; Doucet, J. P.; Doucet-Panaye, A. Linear and Non-Linear Modelling of the Cytotoxicity of TiO₂ and ZnO Nanoparticles by Empirical Descriptors. *SAR QSAR Environ. Res.* **2015**, *26*, 647–665.
- (372) Le, T. C.; Yin, H.; Chen, R.; Chen, Y.; Zhao, L.; Casey, P. S.; Chen, C.; Winkler, D. A. An Experimental and Computational Approach to the Development of ZnO Nanoparticles That Are Safe by Design. *Small* **2016**, *12*, 3568–3577.
- (373) Manganelli, S.; Leone, C.; Toropov, A. A.; Toropova, A. P.; Benfenati, E. QSAR Model for Predicting Cell Viability of Human Embryonic Kidney Cells Exposed to SiO₂ Nanoparticles. *Chemosphere* **2016**, *144*, 995–1001.
- (374) Huang, Y.; Li, X.; Cao, J.; Wei, X.; Li, Y.; Wang, Z.; Cai, X.; Li, R.; Chen, J. Use of Dissociation Degree in Lysosomes to Predict Metal Oxide Nanoparticle Toxicity in Immune Cells: Machine Learning Boosts Nano-Safety Assessment. *Environ. Int.* **2022**, *164*, 107258.
- (375) Nel, A.; Xia, T.; Mädler, L.; Li, N. Toxic Potential of Materials at the Nanolevel. *Science* **2006**, *311*, 622–627.
- (376) Burello, E.; Worth, A. P. A Theoretical Framework for Predicting the Oxidative Stress Potential of Oxide Nanoparticles. *Nanotoxicology* **2011**, *5*, 228–235.
- (377) Golbamaki, A.; Golbamaki, N.; Sizochenko, N.; Rasulev, B.; Leszczynski, J.; Benfenati, E. Genotoxicity Induced by Metal Oxide Nanoparticles: A Weight of Evidence Study and Effect of Particle Surface and Electronic Properties. *Nanotoxicology* **2018**, *12*, 1113–1129.
- (378) Holden, P. A.; Gardea-Torresdey, J. L.; Klaessig, F.; Turco, R. F.; Mortimer, M.; Hund-Rinke, K.; Cohen Hubal, E. A.; Avery, D.; Barceló, D.; Behra, R.; et al. Considerations of Environmentally Relevant Test Conditions for Improved Evaluation of Ecological Hazards of Engineered Nanomaterials. *Environ. Sci. Technol.* **2016**, *50*, 6124–6145.
- (379) Bradford, S. A.; Shen, C.; Kim, H.; Letcher, R. J.; Rinklebe, J.; Ok, Y. S.; Ma, L. Environmental Applications and Risks of Nanomaterials: An Introduction to CREST Publications during 2018–2021. *Crit. Rev. Environ. Sci. Technol.* **2022**, *52*, 3753–3762.
- (380) Batley, G. E.; Kirby, J. K.; McLaughlin, M. J. Fate and Risks of Nanomaterials in Aquatic and Terrestrial Environments. *Acc. Chem. Res.* **2013**, *46*, 854–862.
- (381) Wang, D.; Zhao, L.; Ma, H.; Zhang, H.; Guo, L. H. Quantitative Analysis of Reactive Oxygen Species Photogenerated on Metal Oxide Nanoparticles and Their Bacteria Toxicity: The Role of Superoxide Radicals. *Environ. Sci. Technol.* **2017**, *51*, 10137–10145.
- (382) Feng, Z. V.; Gunsolus, I. L.; Qiu, T. A.; Hurley, K. R.; Nyberg, L. H.; Frew, H.; Johnson, K. P.; Vartanian, A. M.; Jacob, L. M.; Lohse, S. E.; et al. Impacts of Gold Nanoparticle Charge and Ligand Type on Surface Binding and Toxicity to Gram-Negative and Gram-Positive Bacteria. *Chem. Sci.* **2015**, *6*, 5186–5196.
- (383) Slavin, Y. N.; Asnis, J.; Häfeli, U. O.; Bach, H. Metal Nanoparticles: Understanding the Mechanisms behind Antibacterial Activity. *J. Nanobiotechnology* **2017**, *15*, 65.
- (384) Ameen, F.; Alsamhary, K.; Alabdullatif, J. A.; ALNadhari, S. A Review on Metal-Based Nanoparticles and Their Toxicity to Beneficial Soil Bacteria and Fungi. *Ecotoxicol. Environ. Saf.* **2021**, *213*, 112027.
- (385) Gajewicz, A. Development of Valuable Predictive Read-across Models Based on “Real-Life” (Sparse) Nanotoxicity Data. *Environ. Sci. Nano* **2017**, *4*, 1389–1403.
- (386) Fjodorova, N.; Novic, M.; Gajewicz, A.; Rasulev, B. The Way to Cover Prediction for Cytotoxicity for All Existing Nano-Sized Metal Oxides by Using Neural Network Method. *Nanotoxicology* **2017**, *11*, 475–483.
- (387) Mu, Y.; Wu, F.; Zhao, Q.; Ji, R.; Qie, Y.; Zhou, Y.; Hu, Y.; Pang, C.; Hristozov, D.; Giesy, J. P.; et al. Predicting Toxic Potencies of Metal Oxide Nanoparticles by Means of Nano-QSARs. *Nanotoxicology* **2016**, *10*, 1207–1214.
- (388) Zhai, X.; Chen, M.; Lu, W. Predicting the Toxicities of Metal Oxide Nanoparticles Based on Support Vector Regression with a Residual Bootstrapping Method. *Toxicol. Mech. Methods* **2018**, *28*, 440–449.
- (389) Sizochenko, N.; Mikolajczyk, A.; Jagiello, K.; Puzyn, T.; Leszczynski, J.; Rasulev, B. How the Toxicity of Nanomaterials towards Different Species Could Be Simultaneously Evaluated: A Novel Multi-Nano-Read-across Approach. *Nanoscale* **2018**, *10*, 582–591.
- (390) Sizochenko, N.; Gajewicz, A.; Leszczynski, J.; Puzyn, T. Causation or Only Correlation? Application of Causal Inference Graphs for Evaluating Causality in Nano-QSAR Models. *Nanoscale* **2016**, *8*, 7203–7208.
- (391) Toropov, A. A.; Toropova, A. P.; Benfenati, E.; Gini, G.; Puzyn, T.; Leszczynski, D.; Leszczynski, J. Novel Application of the CORAL Software to Model Cytotoxicity of Metal Oxide Nanoparticles to Bacteria Escherichia Coli. *Chemosphere* **2012**, *89*, 1098–1102.
- (392) Speck-Planche, A.; Kleandrova, V. V.; Luan, F.; Ds Cordeiro, M. N. Computational Modeling in Nanomedicine: Prediction of Multiple Antibacterial Profiles of Nanoparticles Using a Quantitative Structure-Activity Relationship Perturbation Model. *Nanomedicine* **2015**, *10*, 193–204.
- (393) Mirzaei, M.; Furxhi, I.; Murphy, F.; Mullins, M. A Machine Learning Tool to Predict the Antibacterial Capacity of Nanoparticles. *Nanomaterials* **2021**, *11*, 1774.
- (394) Zhang, J.; Yu, F.; Hu, X.; Gao, Y.; Qu, Q. Multifeature Superposition Analysis of the Effects of Microplastics on Microbial Communities in Realistic Environments. *Environ. Int.* **2022**, *162*, 107172.
- (395) Pan, Y.; Li, T.; Cheng, J.; Telesca, D.; Zink, J. I.; Jiang, J. Nano-QSAR Modeling for Predicting the Cytotoxicity of Metal Oxide Nanoparticles Using Novel Descriptors. *RSC Adv.* **2016**, *6*, 25766–25775.
- (396) Gajewicz, A.; Jagiello, K.; Cronin, M. T. D.; Leszczynski, J.; Puzyn, T. Addressing a Bottle Neck for Regulation of Nanomaterials: Quantitative Read-across (Nano-QRA) Algorithm for Cases When Only Limited Data Is Available. *Environ. Sci. Nano* **2017**, *4*, 346–358.
- (397) Pathakoti, K.; Huang, M. J.; Watts, J. D.; He, X.; Hwang, H. M. Using Experimental Data of Escherichia Coli to Develop a QSAR Model for Predicting the Photo-Induced Cytotoxicity of Metal Oxide Nanoparticles. *J. Photochem. Photobiol. B Biol.* **2014**, *130*, 234–240.
- (398) Harper, B.; Thomas, D.; Chikagoudar, S.; Baker, N.; Tang, K.; Heredia-Langner, A.; Lins, R.; Harper, S. Comparative Hazard Analysis and Toxicological Modeling of Diverse Nanomaterials Using the Embryonic Zebrafish (EZ) Metric of Toxicity. *J. Nanoparticle Res.* **2015**, *17*, 250.
- (399) Karcher, S. C.; Harper, B. J.; Harper, S. L.; Hendren, C. O.; Wiesner, M. R.; Lowry, G. V. Visualization Tool for Correlating Nanomaterial Properties and Biological Responses in Zebrafish. *Environ. Sci. Nano* **2016**, *3*, 1280–1292.
- (400) Liu, X.; Tang, K.; Harper, S.; Harper, B.; Steevens, J. A.; Xu, R. Predictive Modeling of Nanomaterial Exposure Effects in Biological Systems. *Int. J. Nanomedicine* **2013**, *8*, 31–43.
- (401) Liu, R.; Lin, S. S.; Rallo, R.; Zhao, Y.; Damoiseaux, R.; Xia, T.; Lin, S. S.; Nel, A.; Cohen, Y. Automated Phenotype Recognition for Zebrafish Embryo Based in Vivo High Throughput Toxicity Screening of Engineered Nano-Materials. *PLoS One* **2012**, *7*, No. e35014.
- (402) Gousiadou, C.; Marchese Robinson, R. L.; Kotzabasaki, M.; Doganis, P.; Wilkins, T. A.; Jia, X.; Sarimveis, H.; Harper, S. L. Machine Learning Predictions of Concentration-Specific Aggregate Hazard Scores of Inorganic Nanomaterials in Embryonic Zebrafish. *Nanotoxicology* **2021**, *15*, 446–476.

- (403) Candan, F.; Markushin, Y.; Ozbay, G. Uptake and Presence Evaluation of Nanoparticles in Cicer Arietinum L. by Infrared Spectroscopy and Machine Learning Techniques. *Plants* **2022**, *11*, 1569.
- (404) Ban, Z.; Zhou, Q.; Sun, A.; Mu, L.; Hu, X. Screening Priority Factors Determining and Predicting the Reproductive Toxicity of Various Nanoparticles. *Environ. Sci. Technol.* **2018**, *52*, 9666–9676.
- (405) Gernand, J. M.; Casman, E. A. A Meta-Analysis of Carbon Nanotube Pulmonary Toxicity Studies-How Physical Dimensions and Impurities Affect the Toxicity of Carbon Nanotubes. *Risk Anal.* **2014**, *34*, 583–597.
- (406) Bahl, A.; Hellack, B.; Balas, M.; Dinischiotu, A.; Wiemann, M.; Brinkmann, J.; Luch, A.; Renard, B. Y.; Haase, A. Recursive Feature Elimination in Random Forest Classification Supports Nanomaterial Grouping. *NanoImpact* **2019**, *15*, 100179.
- (407) Sizochenko, N.; Syzochenko, M.; Fjodorova, N.; Rasulev, B.; Leszczynski, J. Evaluating Genotoxicity of Metal Oxide Nanoparticles: Application of Advanced Supervised and Unsupervised Machine Learning Techniques. *Ecotoxicol. Environ. Saf.* **2019**, *185*, 109733.
- (408) Chen, Z.; Zheng, P.; Han, S.; Zhang, J.; Li, Z.; Zhou, S.; Jia, G. Tissue-Specific Oxidative Stress and Element Distribution after Oral Exposure to Titanium Dioxide Nanoparticles in Rats. *Nanoscale* **2020**, *12*, 20033–20046.
- (409) Peng, T.; Wei, C.; Yu, F.; Xu, J.; Zhou, Q.; Shi, T.; Hu, X. Predicting Nanotoxicity by an Integrated Machine Learning and Metabolomics Approach. *Environ. Pollut.* **2020**, *267*, 115434.
- (410) Tang, Y.; Ma, X.; Wang, S. Interpretable XGBoost-SHAP Model Predicts the Nanoparticles Delivery and Reveals Its Interaction with Tumor Genomic Profiles Affiliations. *bioRxiv*; 2022; DOI: 10.1101/2022.06.06.494964.
- (411) Kingston, B. R.; Syed, A. M.; Ngai, J.; Sindhwani, S.; Chan, W. C. W. Assessing Micrometastases as a Target for Nanoparticles Using 3D Microscopy and Machine Learning. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 14937–14946.
- (412) Lin, Z.; Chou, W. C.; Cheng, Y. H.; He, C.; Monteiro-Riviere, N. A.; Riviere, J. E. Predicting Nanoparticle Delivery to Tumors Using Machine Learning and Artificial Intelligence Approaches. *Int. J. Nanomedicine* **2022**, *17*, 1365–1379.
- (413) Sahigara, F.; Mansouri, K.; Ballabio, D.; Mauri, A.; Consonni, V.; Todeschini, R. Comparison of Different Approaches to Define the Applicability Domain of QSAR Models. *Molecules* **2012**, *17*, 4791–4810.
- (414) Netzeva, T. I.; Worth, A. P.; Aldenberg, T.; Benigni, R.; Cronin, M. T. D.; Gramatica, P.; Jaworska, J. S.; Kahn, S.; Klopman, G.; Marchant, C. A.; et al. Current Status of Methods for Defining the Applicability Domain of (Quantitative) Structure-Activity Relationships. *ATLA Altern. to Lab. Anim.* **2005**, *33*, 155–173.
- (415) Jaworska, J.; Nikolova-Jeliazkova, N.; Aldenberg, T. QSAR Applicability Domain Estimation by Projection of the Training Set in Descriptor Space: A Review. *ATLA Altern. to Lab. Anim.* **2005**, *33*, 445–459.
- (416) Shin, H. K.; Kim, S.; Yoon, S. Use of Size-Dependent Electron Configuration Fingerprint to Develop General Prediction Models for Nanomaterials. *NanoImpact* **2021**, *21*, 100298.
- (417) Zhao, Q.; Yang, K.; Li, W.; Xing, B. Concentration-Dependent Polyparameter Linear Free Energy Relationships to Predict Organic Compound Sorption on Carbon Nanotubes. *Sci. Rep.* **2014**, *4*, 3888.
- (418) Sizochenko, N.; Mikolajczyk, A.; Syzochenko, M.; Puzyn, T.; Leszczynski, J. Zeta Potentials (ζ) of Metal Oxide Nanoparticles: A Meta-Analysis of Experimental Data and a Predictive Neural Networks Modeling. *NanoImpact* **2021**, *22*, 100317.
- (419) Kumar, A.; Kumar, P. Cytotoxicity of Quantum Dots: Use of QuasiSMILES in Development of Reliable Models with Index of Ideality of Correlation and the Consensus Modelling. *J. Hazard. Mater.* **2021**, *402*, 123777.
- (420) Trinh, T. X.; Seo, M.; Yoon, T. H.; Kim, J. Developing Random Forest Based QSAR Models for Predicting the Mixture Toxicity of TiO₂ Based Nano-Mixtures to Daphnia Magna. *NanoImpact* **2022**, *25*, 100383.
- (421) Gajewicz, A. How to Judge Whether QSAR/Read-across Predictions Can Be Trusted: A Novel Approach for Establishing a Model's Applicability Domain. *Environ. Sci. Nano* **2018**, *5*, 408–421.
- (422) Roy, K.; Kar, S.; Ambure, P. On a Simple Approach for Determining Applicability Domain of QSAR Models. *Chemom. Intell. Lab. Syst.* **2015**, *145*, 22–29.
- (423) Esterhuizen, J. A.; Goldsmith, B. R.; Linic, S. Interpretable Machine Learning for Knowledge Generation in Heterogeneous Catalysis. *Nat. Catal.* **2022**, *5*, 175–184.
- (424) Murdoch, W. J.; Singh, C.; Kumbier, K.; Abbasi-Asl, R.; Yu, B. Definitions, Methods, and Applications in Interpretable Machine Learning. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 22071–22080.
- (425) Lundberg, S. M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J. M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67.
- (426) Rudin, C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215.
- (427) Grojean, C.; Paul, A.; Qian, Z.; Strümke, I. Lessons on Interpretable Machine Learning from Particle Physics. *Nat. Rev. Phys.* **2022**, *4*, 284–286.
- (428) Jagiello, K.; Grzonkowska, M.; Swirog, M.; Ahmed, L.; Rasulev, B.; Avramopoulos, A.; Papadopoulos, M. G.; Leszczynski, J.; Puzyn, T. Advantages and Limitations of Classic and 3D QSAR Approaches in Nano-QSAR Studies Based on Biological Activity of Fullerene Derivatives. *J. Nanoparticle Res.* **2016**, *18*, 256.
- (429) González-Durruthy, M.; Alberici, L. C.; Curti, C.; Naal, Z.; Atique-Sawazaki, D. T.; Vázquez-Naya, J. M.; González-Díaz, H.; Munteanu, C. R. Experimental-Computational Study of Carbon Nanotube Effects on Mitochondrial Respiration: In Silico Nano-QSPR Machine Learning Models Based on New Raman Spectra Transform with Markov-Shannon Entropy Invariants. *J. Chem. Inf. Model.* **2017**, *57*, 1029–1044.
- (430) Kar, S.; Gajewicz, A.; Roy, K.; Leszczynski, J.; Puzyn, T. Extrapolating between Toxicity Endpoints of Metal Oxide Nanoparticles: Predicting Toxicity to Escherichia Coli and Human Keratinocyte Cell Line (HaCaT) with Nano-QTTR. *Ecotoxicol. Environ. Saf.* **2016**, *126*, 238–244.
- (431) Wang, Y.; Chen, J.; Tang, W.; Xia, D.; Liang, Y.; Li, X. Modeling Adsorption of Organic Pollutants onto Single-Walled Carbon Nanotubes with Theoretical Molecular Descriptors Using MLR and SVM Algorithms. *Chemosphere* **2019**, *214*, 79–84.
- (432) Krzywinski, M.; Altman, N. Points of Significance: Multiple Linear Regression. *Nat. Methods* **2015**, *12*, 1103–1104.
- (433) Mikolajczyk, A.; Gajewicz, A.; Mulkiewicz, E.; Rasulev, B.; Marchelek, M.; Diak, M.; Hirano, S.; Zaleska-Medynska, A.; Puzyn, T. Nano-QSAR Modeling for Ecosafe Design of Heterogeneous TiO₂-Based Nano-Photocatalysts. *Environ. Sci. Nano* **2018**, *5*, 1150–1160.
- (434) Marvin, H. J. P.; Bouzembrak, Y.; Janssen, E. M.; van der Zande, M.; Murphy, F.; Sheehan, B.; Mullins, M.; Bouwmeester, H. Application of Bayesian Networks for Hazard Ranking of Nanomaterials to Support Human Health Risk Assessment. *Nanotoxicology* **2017**, *11*, 123–133.
- (435) Simeone, F. C.; Costa, A. L. Quantifying Uncertainty in Dose-Response Screenings of Nanoparticles: A Bayesian Data Analysis. *Nanotoxicology* **2022**, *16*, 135–151.
- (436) Jeong, J.; Song, T.; Chatterjee, N.; Choi, I.; Cha, Y. K.; Choi, J. Developing Adverse Outcome Pathways on Silver Nanoparticle-Induced Reproductive Toxicity via Oxidative Stress in the Nematode Caenorhabditis Elegans Using a Bayesian Network Model. *Nanotoxicology* **2018**, *12*, 1182–1197.
- (437) Chen, G.; Peijnenburg, W. J. G. M.; Kovalishyn, V.; Vijver, M. G. Development of Nanostructure-Activity Relationships Assisting the Nanomaterial Hazard Categorization for Risk Assessment and Regulatory Decision-Making. *RSC Adv.* **2016**, *6*, S2227–S2235.
- (438) Jones, D. E.; Ghandehari, H.; Facelli, J. C. Predicting Cytotoxicity of PAMAM Dendrimers Using Molecular Descriptors. *Beilstein J. Nanotechnol.* **2015**, *6*, 1886–1896.

- (439) Sizochenko, N.; Leszczynska, D.; Leszczynski, J. Modeling of Interactions between the Zebrafish Hatching Enzyme ZHE1 and a Series of Metal Oxide Nanoparticles: Nano-QSAR and Causal Analysis of Inactivation Mechanisms. *Nanomaterials* **2017**, *7*, 330.
- (440) Oksel, C.; Winkler, D. A.; Ma, C. Y.; Wilkins, T.; Wang, X. Z. Accurate and Interpretable NanoSAR Models from Genetic Programming-Based Decision Tree Construction Approaches. *Nanotoxicology* **2016**, *10*, 1001–1012.
- (441) Marchwiany, M. E.; Birowska, M.; Popielski, M.; Majewski, J. A.; Jastrzebska, A. M. Surface-Related Features Responsible for Cytotoxic Behavior of Mxenes Layered Materials Predicted with Machine Learning Approach. *Materials* **2020**, *13*, 3083.
- (442) Chen, Z.; Han, S.; Zhang, J.; Zheng, P.; Liu, X.; Zhang, Y.; Jia, G. Exploring Urine Biomarkers of Early Health Effects for Occupational Exposure to Titanium Dioxide Nanoparticles Using Metabolomics. *Nanoscale* **2021**, *13*, 4122–4132.
- (443) Subramanian, N. A.; Palaniappan, A. NanoTox: Development of a Parsimonious in Silico Model for Toxicity Assessment of Metal-Oxide Nanoparticles Using Physicochemical Features. *ACS Omega* **2021**, *6*, 11729–11739.
- (444) Yuan, B.; Wang, P.; Sang, L.; Gong, J.; Pan, Y.; Hu, Y. QNAR Modeling of Cytotoxicity of Mixing Nano-TiO₂ and Heavy Metals. *Ecotox. Environ. Saf.* **2021**, *208*, 111634.
- (445) Altmann, A.; Toloşi, L.; Sander, O.; Lengauer, T. Permutation Importance: A Corrected Feature Importance Measure. *Bioinformatics* **2010**, *26*, 1340–1347.
- (446) Greenwell, B. M. Pdp: An R Package for Constructing Partial Dependence Plots. *R J.* **2017**, *9*, 421–436.
- (447) Palatnik de Sousa, I.; Maria Bernardes Rebuzzi Vellasco, M.; Costa da Silva, E. Local Interpretable Model-Agnostic Explanations for Classification of Lymph Node Metastases. *Sensors* **2019**, *19*, 2969.
- (448) Mangalathu, S.; Hwang, S. H.; Jeon, J. S. Failure Mode and Effects Analysis of RC Members Based on Machine-Learning-Based SHapley Additive ExPlanations (SHAP) Approach. *Eng. Struct.* **2020**, *219*, 110927.
- (449) Yu, H.; Zhao, Z.; Cheng, F. Predicting and Investigating Cytotoxicity of Nanoparticles by Translucent Machine Learning. *Chemosphere* **2021**, *276*, 130164.
- (450) Guo, Y.; Ma, W.; Li, J.; Liu, W.; Qi, P.; Ye, Y.; Guo, B.; Zhang, J.; Qu, C. Effects of Microplastics on Growth, Phenanthrene Stress, and Lipid Accumulation in a Diatom, *Phaeodactylum Tricornutum*. *Environ. Pollut.* **2020**, *257*, 113628.
- (451) Urista, D. V.; Carrué, D. B.; Otero, I.; Arrasate, S.; Quevedo-Tumaili, V. F.; Gestal, M.; González-Díaz, H.; Munteanu, C. R. Prediction of Antimalarial Drug-Decorated Nanoparticle Delivery Systems with Random Forest Models. *Biology* **2020**, *9*, 198.
- (452) Kim, M. T.; Huang, R.; Sedykh, A.; Wang, W.; Xia, M.; Zhu, H. Mechanism Profiling of Hepatotoxicity Caused by Oxidative Stress Using Antioxidant Response Element Reporter Gene Assay Models and Big Data. *Environ. Health Perspect.* **2016**, *124*, 634–641.
- (453) Ankley, G. T.; Bennett, R. S.; Erickson, R. J.; Hoff, D. J.; Hornung, M. W.; Johnson, R. D.; Mount, D. R.; Nichols, J. W.; Russom, C. L.; Schmieder, P. K.; et al. Adverse Outcome Pathways: A Conceptual Framework to Support Ecotoxicology Research and Risk Assessment. *Environ. Toxicol. Chem.* **2010**, *29*, 730–741.
- (454) Ciallella, H. L.; Russo, D. P.; Aleksunes, L. M.; Grimm, F. A.; Zhu, H. Revealing Adverse Outcome Pathways from Public High-Throughput Screening Data to Evaluate New Toxicants by a Knowledge-Based Deep Neural Network Approach. *Environ. Sci. Technol.* **2021**, *55*, 10875–10887.
- (455) Halappanavar, S.; Van Den Brule, S.; Nymark, P.; Gaté, L.; Seidel, C.; Valentino, S.; Zhernovkov, V.; Høgh Danielsen, P.; De Vizcaya, A.; Wolff, H.; et al. Adverse Outcome Pathways as a Tool for the Design of Testing Strategies to Support the Safety Assessment of Emerging Advanced Materials at the Nanoscale. *Part. Fibre Toxicol.* **2020**, *17*, 16.
- (456) Jagiello, K.; Halappanavar, S.; Rybińska-Fryca, A.; Williams, A.; Vogel, U.; Puzyn, T. Transcriptomics-Based and AOP-Informed Structure-Activity Relationships to Predict Pulmonary Pathology Induced by Multiwalled Carbon Nanotubes. *Small* **2021**, *17*, 2003465.
- (457) Bal-Price, A.; Lein, P. J.; Keil, K. P.; Sethi, S.; Shafer, T.; Barenys, M.; Fritsche, E.; Sachana, M.; Meek, M. E. B. Developing and Applying the Adverse Outcome Pathway Concept for Understanding and Predicting Neurotoxicity. *Neurotoxicology* **2017**, *59*, 240–255.
- (458) Groh, K. J.; Carvalho, R. N.; Chipman, J. K.; Denslow, N. D.; Halder, M.; Murphy, C. A.; Roelofs, D.; Rolaki, A.; Schirmer, K.; Watanabe, K. H. Development and Application of the Adverse Outcome Pathway Framework for Understanding and Predicting Chronic Toxicity: II. A Focus on Growth Impairment in Fish. *Chemosphere* **2015**, *120*, 778–792.
- (459) Jia, X.; Wen, X.; Russo, D. P.; Aleksunes, L. M.; Zhu, H. Mechanism-Driven Modeling of Chemical Hepatotoxicity Using Structural Alerts and an in Vitro Screening Assay. *J. Hazard. Mater.* **2022**, *436*, 129193.
- (460) Clippinger, A. J.; Allen, D.; Behrsing, H.; Bérubé, K. A.; Bolger, M. B.; Casey, W.; DeLorme, M.; Gaça, M.; Gehen, S. C.; Glover, K.; et al. Pathway-Based Predictive Approaches for Non-Animal Assessment of Acute Inhalation Toxicity. *Toxicol. Vitro* **2018**, *52*, 131–145.
- (461) MacKay, C.; Davies, M.; Summerfield, V.; Maxwell, G. From Pathways to People: Applying the Adverse Outcome Pathway (AOP) for Skin Sensitization to Risk Assessment. *ALTEX* **2013**, *30*, 473–486.
- (462) Halappanavar, S.; Nymark, P.; Krug, H. F.; Clift, M. J. D.; Rothen-Rutishauser, B.; Vogel, U. Non-Animal Strategies for Toxicity Assessment of Nanoscale Materials: Role of Adverse Outcome Pathways in the Selection of Endpoints. *Small* **2021**, *17*, 2007628.
- (463) Labib, S.; Williams, A.; Yauk, C. L.; Nikota, J. K.; Wallin, H.; Vogel, U.; Halappanavar, S. Nano-Risk Science: Application of Toxicogenomics in an Adverse Outcome Pathway Framework for Risk Assessment of Multi-Walled Carbon Nanotubes. *Part. Fibre Toxicol.* **2015**, *13*, 15.
- (464) Halappanavar, S.; Ede, J. D.; Shatkin, J. A.; Krug, H. F. A Systematic Process for Identifying Key Events for Advancing the Development of Nanomaterial Relevant Adverse Outcome Pathways. *NanoImpact* **2019**, *15*, 100178.
- (465) Horvath, D. A Virtual Screening Approach Applied to the Search for Trypanothione Reductase Inhibitors. *J. Med. Chem.* **1997**, *40*, 2412–2423.
- (466) Sadybekov, A. A.; Sadybekov, A. V.; Liu, Y.; Iliopoulos-Tsoutsouvas, C.; Huang, X. P.; Pickett, J.; Houser, B.; Patel, N.; Tran, N. K.; Tong, F.; et al. Synthon-Based Ligand Discovery in Virtual Libraries of over 11 Billion Compounds. *Nature* **2022**, *601*, 452–459.
- (467) Gorgulla, C.; Boeszoermenyi, A.; Wang, Z. F.; Fischer, P. D.; Coote, P. W.; Padmanabha Das, K. M.; Malets, Y. S.; Radchenko, D. S.; Moroz, Y. S.; Scott, D. A.; et al. An Open-Source Drug Discovery Platform Enables Ultra-Large Virtual Screens. *Nature* **2020**, *580*, 663–668.
- (468) Rosales, A. R.; Wahlers, J.; Limé, E.; Meadows, R. E.; Leslie, K. W.; Savin, R.; Bell, F.; Hansen, E.; Helquist, P.; Munday, R. H.; et al. Rapid Virtual Screening of Enantioselective Catalysts Using CatVS. *Nat. Catal.* **2019**, *2*, 41–45.
- (469) Pyzer-Knapp, E. O.; Suh, C.; Gómez-Bombarelli, R.; Aguilera-Iparraguirre, J.; Aspuru-Guzik, A. What Is High-Throughput Virtual Screening? A Perspective from Organic Materials Discovery. *Annu. Rev. Mater. Res.* **2015**, *45*, 195–216.
- (470) Shoichet, B. K. Virtual Screening of Chemical Libraries. *Nature* **2004**, *432*, 862–865.
- (471) Schneider, G. Virtual Screening: An Endless Staircase? *Nat. Rev. Drug Discovery* **2010**, *9*, 273–276.
- (472) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and Scoring in Virtual Screening for Drug Discovery: Methods and Applications. *Nat. Rev. Drug Discovery* **2004**, *3*, 935–949.
- (473) Gentile, F.; Yaacoub, J. C.; Gleave, J.; Fernandez, M.; Ton, A. T.; Ban, F.; Stern, A.; Cherkasov, A. Artificial Intelligence-Enabled Virtual Screening of Ultra-Large Chemical Libraries with Deep Docking. *Nat. Protoc.* **2022**, *17*, 672–697.

- (474) Graff, D. E.; Shakhnovich, E. I.; Coley, C. W. Accelerating High-Throughput Virtual Screening through Molecular Pool-Based Active Learning. *Chem. Sci.* **2021**, *12*, 7866–7881.
- (475) Li, H.; Sze, K. H.; Lu, G.; Ballester, P. J. Machine-Learning Scoring Functions for Structure-Based Virtual Screening. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2021**, *11*, e1478.
- (476) Ahmed, L.; Rasulev, B.; Kar, S.; Krupa, P.; Mozolewska, M. A.; Leszczynski, J. Inhibitors or Toxins? Large Library Target-Specific Screening of Fullerene-Based Nanoparticles for Drug Design Purpose. *Nanoscale* **2017**, *9*, 10263–10276.
- (477) So, S.; Mun, J.; Rho, J. Simultaneous Inverse Design of Materials and Structures via Deep Learning: Demonstration of Dipole Resonance Engineering Using Core-Shell Nanoparticles. *ACS Appl. Mater. Interfaces* **2019**, *11*, 24264–24268.
- (478) Mao, Y.; He, Q.; Zhao, X. Designing Complex Architected Materials with Generative Adversarial Networks. *Sci. Adv.* **2020**, *6*, No. eaaz4169.
- (479) Freeze, J. G.; Kelly, H. R.; Batista, V. S. Search for Catalysts by Inverse Design: Artificial Intelligence, Mountain Climbers, and Alchemists. *Chem. Rev.* **2019**, *119*, 6595–6612.
- (480) Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse Molecular Design Using Machine Learning: Generative Models for Matter Engineering. *Science* **2018**, *361*, 360–365.
- (481) Zunger, A. Inverse Design in Search of Materials with Target Functionalities. *Nat. Rev. Chem.* **2018**, *2*, 0121.
- (482) Kotsias, P. C.; Arús-Pous, J.; Chen, H.; Engkvist, O.; Tyrchan, C.; Bjerrum, E. J. Direct Steering of de Novo Molecular Generation with Descriptor Conditional Recurrent Neural Networks. *Nat. Mach. Intell.* **2020**, *2*, 254–265.
- (483) Ren, Z.; Tian, S. I. P.; Noh, J.; Oviedo, F.; Xing, G.; Li, J.; Liang, Q.; Zhu, R.; Aberle, A. G.; Sun, S.; et al. An Invertible Crystallographic Representation for General Inverse Design of Inorganic Crystals with Targeted Properties. *Matter* **2022**, *5*, 314–335.
- (484) Yao, Z.; Sánchez-Lengeling, B.; Bobbitt, N. S.; Bucior, B. J.; Kumar, S. G. H.; Collins, S. P.; Burns, T.; Woo, T. K.; Farha, O. K.; Snurr, R. Q.; et al. Inverse Design of Nanoporous Crystalline Reticular Materials with Deep Generative Models. *Nat. Mach. Intell.* **2021**, *3*, 76–86.
- (485) Popova, M.; Isayev, O.; Tropsha, A. Deep Reinforcement Learning for de Novo Drug Design. *Sci. Adv.* **2018**, *4*, eaap788.
- (486) Li, S.; Barnard, A. S. Inverse Design of Nanoparticles Using Multi-Target Machine Learning. *Adv. Theory Simulations* **2022**, *5*, 2100414.
- (487) Peurifoy, J.; Shen, Y.; Jing, L.; Yang, Y.; Cano-Renteria, F.; DeLacy, B. G.; Joannopoulos, J. D.; Tegmark, M.; Soljačić, M. Nanophotonic Particle Simulation and Inverse Design Using Artificial Neural Networks. *Sci. Adv.* **2018**, *4*, eaar420.
- (488) Jia, Y.; Hou, X.; Wang, Z.; Hu, X. Machine Learning Boosts the Design and Discovery of Nanomaterials. *ACS Sustain. Chem. Eng.* **2021**, *9*, 6130–6147.
- (489) Wang, M.; Wang, T.; Cai, P.; Chen, X. Nanomaterials Discovery and Design through Machine Learning. *Small Methods* **2019**, *3*, 1900025.
- (490) Dong, X.; Yu, Z.; Cao, W.; Shi, Y.; Ma, Q. A Survey on Ensemble Learning. *Front. Comput. Sci.* **2020**, *14*, 241–258.
- (491) Sagi, O.; Rokach, L. Ensemble Learning: A Survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discovery* **2018**, *8*, e1249.
- (492) Mai, H.; Le, T. C.; Hisatomi, T.; Chen, D.; Domen, K.; Winkler, D. A.; Caruso, R. A. Use of Metamodels for Rapid Discovery of Narrow Bandgap Oxide Photocatalysts. *iScience* **2021**, *24*, 103068.
- (493) Wang, X.; Wang, L.; Wang, S.; Ren, Y.; Chen, W.; Li, X.; Han, P.; Song, T. QuantumTox: Utilizing Quantum Chemistry with Ensemble Learning for Molecular Toxicity Prediction. *Comput. Biol. Med.* **2023**, *157*, 106744.
- (494) Feng, H.; Zhang, L.; Li, S.; Liu, L.; Yang, T.; Yang, P.; Zhao, J.; Arkin, I. T.; Liu, H. Predicting the Reproductive Toxicity of Chemicals Using Ensemble Learning Methods and Molecular Fingerprints. *Toxicol. Lett.* **2021**, *340*, 4–14.
- (495) Zhang, L.; Ai, H.; Chen, W.; Yin, Z.; Hu, H.; Zhu, J.; Zhao, J.; Zhao, Q.; Liu, H. CarcinoPred-EL: Novel Models for Predicting the Carcinogenicity of Chemicals Using Molecular Fingerprints and Ensemble Learning Methods. *Sci. Rep.* **2017**, *7*, 2118.
- (496) Liu, L.; Zhang, L.; Feng, H.; Li, S.; Liu, M.; Zhao, J.; Liu, H. Prediction of the Blood-Brain Barrier (BBB) Permeability of Chemicals Based on Machine-Learning and Ensemble Methods. *Chem. Res. Toxicol.* **2021**, *34*, 1456–1467.
- (497) Heitler, W.; London, F. Wechselwirkung Neutraler Atome Und Homöopolare Bindung Nach Der Quantenmechanik. *Zeitschrift für Phys.* **1927**, *44*, 455–472.
- (498) Pariser, R.; Parr, R. G. A Semi-Empirical Theory of the Electronic Spectra and Electronic Structure of Complex Unsaturated Molecules. *J. Chem. Phys.* **1953**, *21*, 466–471.
- (499) Frierson, M. R.; Imam, M. R.; Zalkow, V. B.; Allinger, N. L. The MM2 Force Field for Silanes and Polysilanes. *J. Org. Chem.* **1988**, *53*, 5248–5258.
- (500) Shao, Y.; Gan, Z.; Epifanovsky, E.; Gilbert, A. T. B.; Wormit, M.; Kussmann, J.; Lange, A. W.; Behn, A.; Deng, J.; Feng, X.; et al. Advances in Molecular Quantum Chemistry Contained in the Q-Chem 4 Program Package. *Mol. Phys.* **2015**, *113*, 184–215.
- (501) Kresse, G.; Hafner, J. Ab Initio Molecular Dynamics for Liquid Metals. *Phys. Rev. B* **1993**, *47*, 558–561.
- (502) Gonze, X.; Beuken, J. M.; Caracas, R.; Detraux, F.; Fuchs, M.; Rignanese, G. M.; Sindic, L.; Verstraete, M.; Zerah, G.; Jollet, F.; et al. First-Principles Computation of Material Properties: The ABINIT Software Project. *Comput. Mater. Sci.* **2002**, *25*, 478–492.
- (503) Cohen, A. J.; Mori-Sánchez, P.; Yang, W. Insights into Current Limitations of Density Functional Theory. *Science* **2008**, *321*, 792–794.
- (504) Korpelin, V.; Kiljunen, T.; Melander, M. M.; Caro, M. A.; Kristoffersen, H. H.; Mammen, N.; Apaja, V.; Honkala, K. Addressing Dynamics at Catalytic Heterogeneous Interfaces with DFT-MD: Anomalous Temperature Distributions from Commonly Used Thermostats. *J. Phys. Chem. Lett.* **2022**, *13*, 2644–2652.
- (505) Makkar, P.; Ghosh, N. N. A Review on the Use of DFT for the Prediction of the Properties of Nanomaterials. *RSC Adv.* **2021**, *11*, 27897–27924.
- (506) Shen, X.; Wang, Z.; Gao, X.; Zhao, Y. Density Functional Theory-Based Method to Predict the Activities of Nanomaterials as Peroxidase Mimics. *ACS Catal.* **2020**, *10*, 12657–12665.
- (507) Li, L.; Ozden, A.; Guo, S.; Garcia de Arquer, F. P.; Wang, C.; Zhang, M.; Zhang, J.; Jiang, H.; Wang, W.; Dong, H.; et al. Stable, Active CO₂ Reduction to Formate via Redox-Modulated Stabilization of Active Sites. *Nat. Commun.* **2021**, *12*, 5223.
- (508) Shen, X.; Wang, Z.; Gao, X. J.; Gao, X. Reaction Mechanisms and Kinetics of Nanozymes: Insights from Theory and Computation. *Adv. Mater.* **2023**, *2211151*, 2211151.
- (509) Xu, D.; Chang, Y.; Liu, Y.; Qin, W.; Yan, H. Mechanistic Features of Asymmetric Vinylidene Ortho-Quinone Methide Construction and Subsequent Transformations. *ACS Catal.* **2023**, *13*, 2957–2967.
- (510) Ouyang, M.; Papanikolaou, K. G.; Boubnov, A.; Hoffman, A. S.; Giannakakis, G.; Bare, S. R.; Stamatakis, M.; Flytzani-Stephanopoulos, M.; Sykes, E. C. H. Directing Reaction Pathways via in Situ Control of Active Site Geometries in PdAu Single-Atom Alloy Catalysts. *Nat. Commun.* **2021**, *12*, 1549.
- (511) Qi, Y.; Zhang, T.; Jing, C.; Liu, S.; Zhang, C.; Alvarez, P. J. J.; Chen, W. Nanocrystal Facet Modulation to Enhance Transferrin Binding and Cellular Delivery. *Nat. Commun.* **2020**, *11*, 1262.
- (512) Tian, L.; Guan, W.; Ji, Y.; He, X.; Chen, W.; Alvarez, P. J. J.; Zhang, T. Microbial Methylation Potential of Mercury Sulfide Particles Dictated by Surface Structure. *Nat. Geosci.* **2021**, *14*, 409–416.
- (513) Askerka, M.; Li, Z.; Lempen, M.; Liu, Y.; Johnston, A.; Saidaminov, M. I.; Zajacz, Z.; Sargent, E. H. Learning-in-Templates Enables Accelerated Discovery and Synthesis of New Stable Double Perovskites. *J. Am. Chem. Soc.* **2019**, *141*, 3682–3690.
- (514) Zhuo, Y.; Mansouri Tehrani, A.; Oliynyk, A. O.; Duke, A. C.; Brgoch, J. Identifying an Efficient, Thermally Robust Inorganic Phosphor Host via Machine Learning. *Nat. Commun.* **2018**, *9*, 4377.

- (515) Zhang, C.; Yu, Y.; Shi, S.; Liang, M.; Yang, D.; Sui, N.; Yu, W. W.; Wang, L.; Zhu, Z. Machine Learning Guided Discovery of Superoxide Dismutase Nanozymes for Androgenetic Alopecia. *Nano Lett.* **2022**, *22*, 8592–8600.
- (516) Mansouri Tehrani, A.; Oliynyk, A. O.; Parry, M.; Rizvi, Z.; Couper, S.; Lin, F.; Miyagi, L.; Sparks, T. D.; Brgoch, J. Machine Learning Directed Search for Ultraincompressible, Superhard Materials. *J. Am. Chem. Soc.* **2018**, *140*, 9844–9853.
- (517) Ding, H. M.; Ma, Y. Q. Theoretical and Computational Investigations of Nanoparticle-Biomembrane Interactions in Cellular Delivery. *Small* **2015**, *11*, 1055–1071.
- (518) Zhang, X.; Ma, G.; Wei, W. Simulation of Nanoparticles Interacting with a Cell Membrane: Probing the Structural Basis and Potential Biomedical Application. *NPG Asia Mater.* **2021**, *13*, 52.
- (519) Yang, Y. L.; Shao, Q.; Zhang, J.; Yang, L.; Gao, Y. Q. Enhanced Sampling in Molecular Dynamics. *J. Chem. Phys.* **2019**, *151*, 070902.
- (520) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber Biomolecular Simulation Programs. *J. Comput. Chem.* **2005**, *26*, 1668–1688.
- (521) Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. GROMACS: Fast, Flexible, and Free. *J. Comput. Chem.* **2005**, *26*, 1701–1718.
- (522) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. Scalable Molecular Dynamics with NAMD. *J. Comput. Chem.* **2005**, *26*, 1781–1802.
- (523) Franco-Ulloa, S.; Riccardi, L.; Rimembrana, F.; Pini, M.; De Vivo, M. NanoModeler: A Webserver for Molecular Simulations and Engineering of Nanoparticles. *J. Chem. Theory Comput.* **2019**, *15*, 2022–2032.
- (524) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graph.* **1996**, *14*, 33–38.
- (525) Wu, E. L.; Cheng, X.; Jo, S.; Rui, H.; Song, K. C.; Dávila-Contreras, E. M.; Qi, Y.; Lee, J.; Monje-Galvan, V.; Venable, R. M.; et al. CHARMM-GUI Membrane Builder toward Realistic Biological Membrane Simulations. *J. Comput. Chem.* **2014**, *35*, 1997–2004.
- (526) Li, L.; Li, S.; Xu, Y.; Ren, L.; Yang, L.; Liu, X.; Dai, Y.; Zhao, J.; Yue, T. Distinguishing the Nanoplastic-Cell Membrane Interface by Polymer Type and Aging Properties: Translocation, Transformation and Perturbation. *Environ. Sci. Nano* **2023**, *10*, 440–453.
- (527) Unal, M. A.; Bayrakdar, F.; Nazir, H.; Besbinar, O.; Gurcan, C.; Lozano, N.; Arellano, L. M.; Yalcin, S.; Panatli, O.; Celik, D.; et al. Graphene Oxide Nanosheets Interact and Interfere with SARS-CoV-2 Surface Proteins and Cell Receptors to Inhibit Infectivity. *Small* **2021**, *17*, 2101483.
- (528) Gieldoń, A.; Witt, M. M.; Gajewicz, A.; Puzyn, T. Rapid Insight into C60 Influence on Biological Functions of Proteins. *Struct. Chem.* **2017**, *28*, 1775–1788.
- (529) Kang, S. G.; Zhou, G.; Yang, P.; Liu, Y.; Sun, B.; Huynh, T.; Meng, H.; Zhao, L.; Xing, G.; Chen, C.; et al. Molecular Mechanism of Pancreatic Tumor Metastasis Inhibition by Gd@C 82(OH)22 and Its Implication for de Novo Design of Nanomedicine. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 15431–15436.
- (530) Guterres, H.; Im, W. Improving Protein-Ligand Docking Results with High-Throughput Molecular Dynamics Simulations. *J. Chem. Inf. Model.* **2020**, *60*, 2189–2198.
- (531) Barducci, A.; Bussi, G.; Parrinello, M. Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. *Phys. Rev. Lett.* **2008**, *100*, 020603.
- (532) Barnard, A. S.; Zapol, P.; Curtiss, L. A. Modeling the Morphology and Phase Stability of TiO₂ Nanocrystals in Water. *J. Chem. Theory Comput.* **2005**, *1*, 107–116.
- (533) Jinnouchi, R.; Asahi, R. Predicting Catalytic Activity of Nanoparticles by a DFT-Aided Machine-Learning Algorithm. *J. Phys. Chem. Lett.* **2017**, *8*, 4279–4283.
- (534) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: An Extensible Neural Network Potential with DFT Accuracy at Force Field Computational Cost. *Chem. Sci.* **2017**, *8*, 3192–3203.
- (535) Manke, A.; Wang, L.; Rojanasakul, Y. Mechanisms of Nanoparticle-Induced Oxidative Stress and Toxicity. *Biomed Res. Int.* **2013**, *2013*, 942916.
- (536) Canaparo, R.; Foglietta, F.; Limongi, T.; Serpe, L. Biomedical Applications of Reactive Oxygen Species Generation by Metal Nanoparticles. *Materials* **2021**, *14*, 53.
- (537) Fubini, B.; Hubbard, A. Reactive Oxygen Species (ROS) and Reactive Nitrogen Species (RNS) Generation by Silica in Inflammation and Fibrosis. *Free Radic. Biol. Med.* **2003**, *34*, 1507–1516.
- (538) Oberdörster, G.; Maynard, A.; Donaldson, K.; Castranova, V.; Fitzpatrick, J.; Ausman, K.; Carter, J.; Karn, B.; Kreyling, W.; Lai, D.; et al. Principles for Characterizing the Potential Human Health Effects from Exposure to Nanomaterials: Elements of a Screening Strategy. *Part. Fibre Toxicol.* **2005**, *2*, 8.
- (539) Knaapen, A. M.; Borm, P. J. A.; Albrecht, C.; Schins, R. P. F. Inhaled Particles and Lung Cancer. Part A: Mechanisms. *Int. J. Cancer* **2004**, *109*, 799–809.
- (540) Li, Y.; Shu, H.; Niu, X.; Wang, J. Electronic and Optical Properties of Edge-Functionalized Graphene Quantum Dots and the Underlying Mechanism. *J. Phys. Chem. C* **2015**, *119*, 24950–24957.
- (541) Lu, X.; Gao, S.; Lin, H.; Yu, L.; Han, Y.; Zhu, P.; Bao, W.; Yao, H.; Chen, Y.; Shi, J. Bioinspired Copper Single-Atom Catalysts for Tumor Parallel Catalytic Therapy. *Adv. Mater.* **2020**, *32*, 2002246.
- (542) Levard, C.; Hotze, E. M.; Lowry, G. V.; Brown, G. E. Environmental Transformations of Silver Nanoparticles: Impact on Stability and Toxicity. *Environ. Sci. Technol.* **2012**, *46*, 6900–6914.
- (543) Reidy, B.; Haase, A.; Luch, A.; Dawson, K. A.; Lynch, I. Mechanisms of Silver Nanoparticle Release, Transformation and Toxicity: A Critical Review of Current Knowledge and Recommendations for Future Studies and Applications. *Materials* **2013**, *6*, 2295–2350.
- (544) Boukhvalov, D. W. DFT Modeling of the Covalent Functionalization of Graphene: From Ideal to Realistic Models. *RSC Adv.* **2013**, *3*, 7150–7159.
- (545) Lv, J.; Zhang, S.; Luo, L.; Han, W.; Zhang, J.; Yang, K.; Christie, P. Dissolution and Microstructural Transformation of ZnO Nanoparticles under the Influence of Phosphate. *Environ. Sci. Technol.* **2012**, *46*, 7215–7221.
- (546) Laranjeira, J. A. S.; Fabris, G. S. L.; Ferrer, M. M.; Albuquerque, A. R.; Sambrano, J. R. Morphological Transformation Network of Nanoparticles via DFT Simulations. *Cryst. Growth Des.* **2020**, *20*, 4600–4611.
- (547) Krishnadas, K. R.; Baksi, A.; Ghosh, A.; Natarajan, G.; Pradeep, T. Structure-Conserving Spontaneous Transformations between Nanoparticles. *Nat. Commun.* **2016**, *7*, 13447.
- (548) Wang, J.-F.; Chou, K.-C. Molecular Modeling of Cytochrome P450 and Drug Metabolism. *Curr. Drug Metab.* **2010**, *11*, 342–346.
- (549) Gibson, G. G.; Skett, P. *Introduction to Drug Metabolism*; Springer: 2013.
- (550) Fura, A.; Shu, Y. Z.; Zhu, M.; Hanson, R. L.; Roongta, V.; Humphreys, W. G. Discovering Drugs through Biological Transformation: Role of Pharmacologically Active Metabolites in Drug Discovery. *J. Med. Chem.* **2004**, *47*, 4339–4351.
- (551) Wang, Z.; Xia, T.; Liu, S. Mechanisms of Nanosilver-Induced Toxicological Effects: More Attention Should Be Paid to Its Sublethal Effects. *Nanoscale* **2015**, *7*, 7470–7481.
- (552) Djurišić, A. B.; Leung, Y. H.; Ng, A. M. C.; Xu, X. Y.; Lee, P. K. H.; Degger, N.; Wu, R. S. S. Toxicity of Metal Oxide Nanoparticles: Mechanisms, Characterization, and Avoiding Experimental Artefacts. *Small* **2015**, *11*, 26–44.
- (553) Adam, N.; Schmitt, C.; Galceran, J.; Companys, E.; Vakurov, A.; Wallace, R.; Knapen, D.; Blust, R. The Chronic Toxicity of ZnO Nanoparticles and ZnCl₂ to *Daphnia Magna* and the Use of Different Methods to Assess Nanoparticle Aggregation and Dissolution. *Nanotoxicology* **2013**, *8*, 709–717.
- (554) Odzak, N.; Kistler, D.; Behra, R.; Sigg, L. Dissolution of Metal and Metal Oxide Nanoparticles in Aqueous Media. *Environ. Pollut.* **2014**, *191*, 132–138.

- (555) Peng, C.; Shen, C.; Zheng, S.; Yang, W.; Hu, H.; Liu, J.; Shi, J. Transformation of CuO Nanoparticles in the Aquatic Environment: Influence of PH, Electrolytes and Natural Organic Matter. *Nanomaterials* **2017**, *7*, 326.
- (556) Kagan, V. E.; Konduru, N. V.; Feng, W.; Allen, B. L.; Conroy, J.; Volkov, Y.; Vlasova, I. I.; Belikova, N. A.; Yanamala, N.; Kapralov, A.; et al. Carbon Nanotubes Degraded by Neutrophil Myeloperoxidase Induce Less Pulmonary Inflammation. *Nat. Nanotechnol.* **2010**, *5*, 354–359.
- (557) Azuara-Tuexi, G.; Méndez-Cabañas, J. A.; Muñoz-Sandoval, E.; Guirado-López, R. A. Myeloperoxidase-Induced Degradation of N-Doped Carbon Nanotubes: Revealing Possible Atomistic Mechanisms Underlying Hypochlorite-Driven Damage of Nanotube Walls. *Carbon* **2021**, *175*, 387–402.
- (558) Zhang, T.; Wan, Y.; Xie, H.; Mu, Y.; Du, P.; Wang, D.; Wu, X.; Ji, H.; Wan, L. Degradation Chemistry and Stabilization of Exfoliated Few-Layer Black Phosphorus in Water. *J. Am. Chem. Soc.* **2018**, *140*, 7561–7567.
- (559) Zhou, Q.; Chen, Q.; Tong, Y.; Wang, J. Light-Induced Ambient Degradation of Few-Layer Black Phosphorus: Mechanism and Protection. *Angew. Chem., Int. Ed.* **2016**, *55*, 11437–11441.
- (560) Kim, S. T.; Saha, K.; Kim, C.; Rotello, V. M. The Role of Surface Functionality in Determining Nanoparticle Cytotoxicity. *Acc. Chem. Res.* **2013**, *46*, 681–691.
- (561) Quevedo, A. C.; Lynch, I.; Valsami-Jones, E. Silver Nanoparticle Induced Toxicity and Cell Death Mechanisms in Embryonic Zebrafish Cells. *Nanoscale* **2021**, *13*, 6142–6161.
- (562) Guarnieri, D.; Sabella, S.; Muscetti, O.; Belli, V.; Malvindi, M. A.; Fusco, S.; De Luca, E.; Pompa, P. P.; Netti, P. A. Transport across the Cell-Membrane Dictates Nanoparticle Fate and Toxicity: A New Paradigm in Nanotoxicology. *Nanoscale* **2014**, *6*, 10264–10273.
- (563) Behzadi, S.; Serpooshan, V.; Tao, W.; Hamaly, M. A.; Alkawareek, M. Y.; Dreaden, E. C.; Brown, D.; Alkilany, A. M.; Farokhzad, O. C.; Mahmoudi, M. Cellular Uptake of Nanoparticles: Journey inside the Cell. *Chem. Soc. Rev.* **2017**, *46*, 4218–4244.
- (564) Yang, K.; Ma, Y. Q. Computer Simulation of the Translocation of Nanoparticles with Different Shapes across a Lipid Bilayer. *Nat. Nanotechnol.* **2010**, *5*, 579–583.
- (565) Zhang, S.; Gao, H.; Bao, G. Physical Principles of Nanoparticle Cellular Endocytosis. *ACS Nano* **2015**, *9*, 8655–8671.
- (566) Hui, Y.; Yi, X.; Wibowo, D.; Yang, G.; Middelberg, A. P. J.; Gao, H.; Zhao, C. X. Nanoparticle Elasticity Regulates Phagocytosis and Cancer Cell Uptake. *Sci. Adv.* **2020**, *6*, No. eaaz4316.
- (567) Leroueil, P. R.; Hong, S.; Mecke, A.; Baker, J. R.; Orr, B. G.; Holl, M. M. B. Nanoparticle Interaction with Biological Membranes: Does Nanotechnology Present a Janus Face? *Acc. Chem. Res.* **2007**, *40*, 335–342.
- (568) Yi, X.; Shi, X.; Gao, H. Cellular Uptake of Elastic Nanoparticles. *Phys. Rev. Lett.* **2011**, *107*, 098101.
- (569) Mao, J.; Chen, P.; Liang, J.; Guo, R.; Yan, L. T. Receptor-Mediated Endocytosis of Two-Dimensional Nanomaterials Undergoes Flat Vesiculation and Occurs by Revolution and Self-Rotation. *ACS Nano* **2016**, *10*, 1493–1502.
- (570) Van Lehn, R. C.; Alexander-Katz, A. Penetration of Lipid Bilayers by Nanoparticles with Environmentally-Responsive Surfaces: Simulations and Theory. *Soft Matter* **2011**, *7*, 11392–11404.
- (571) Hong, B.; Qiu, F.; Zhang, H.; Yang, Y. Budding Dynamics of Individual Domains in Multicomponent Membranes Simulated by N-Varied Dissipative Particle Dynamics. *J. Phys. Chem. B* **2007**, *111*, 5837–5849.
- (572) Wong-Ekkabut, J.; Baoukina, S.; Triampo, W.; Tang, I. M.; Tieleman, D. P.; Monticelli, L. Computer Simulation Study of Fullerene Translocation through Lipid Membranes. *Nat. Nanotechnol.* **2008**, *3*, 363–368.
- (573) Pogodin, S.; Baulin, V. A. Can a Carbon Nanotube Pierce through a Phospholipid Bilayer? *ACS Nano* **2010**, *4*, 5293–5300.
- (574) Skandani, A. A.; Zeineldin, R.; Al-Haik, M. Effect of Chirality and Length on the Penetrability of Single-Walled Carbon Nanotubes into Lipid Bilayer Cell Membranes. *Langmuir* **2012**, *28*, 7872–7879.
- (575) Guo, Y.; Werner, M.; Seemann, R.; Baulin, V. A.; Fleury, J. B. Tension-Induced Translocation of an Ultrashort Carbon Nanotube through a Phospholipid Bilayer. *ACS Nano* **2018**, *12*, 12042–12049.
- (576) Lelimosin, M.; Sansom, M. S. P. Membrane Perturbation by Carbon Nanotube Insertion: Pathways to Internalization. *Small* **2013**, *9*, 3639–3646.
- (577) Shi, X.; Von Dem Bussche, A.; Hurt, R. H.; Kane, A. B.; Gao, H. Cell Entry of One-Dimensional Nanomaterials Occurs by Tip Recognition and Rotation. *Nat. Nanotechnol.* **2011**, *6*, 714–719.
- (578) Titov, A. V.; Kra, P.; Pearson, R. Sandwiched Graphene-Membrane Superstructures. *ACS Nano* **2010**, *4*, 229–234.
- (579) Chen, P.; Yue, H.; Zhai, X.; Huang, Z.; Ma, G. H.; Wei, W.; Yan, L. T. Transport of a Graphene Nanosheet Sandwiched inside Cell Membranes. *Sci. Adv.* **2019**, *5*, No. eaaw3192.
- (580) Zhu, X.; Li, N.; Huang, C.; Li, Z.; Fan, J. Membrane Perturbation and Lipid Flip-Flop Mediated by Graphene Nanosheet. *J. Phys. Chem. B* **2020**, *124*, 10632–10640.
- (581) Yue, H.; Wei, W.; Yue, Z.; Wang, B.; Luo, N.; Gao, Y.; Ma, D.; Ma, G.; Su, Z. The Role of the Lateral Dimension of Graphene Oxide in the Regulation of Cellular Responses. *Biomaterials* **2012**, *33*, 4013–4021.
- (582) Luo, Z.; Li, S.; Xu, Y.; Ren, H.; Zhang, X.; Hu, G.; Huang, F.; Yue, T. Extracting Pulmonary Surfactants to Form Inverse Micelles on Suspended Graphene Nanosheets. *Environ. Sci. Nano* **2018**, *5*, 130–140.
- (583) Yin, X.; Zhang, S.; Wen, L.; Su, J.; Huang, J.; Duan, G.; Yang, Z. Nonmonotonic Relationship between the Oxidation State of Graphene-Based Materials and Its Cell Membrane Damage Effects. *ACS Appl. Mater. Interfaces* **2022**, *14*, 30306–30314.
- (584) Tu, Y.; Lv, M.; Xiu, P.; Huynh, T.; Zhang, M.; Castelli, M.; Liu, Z.; Huang, Q.; Fan, C.; Fang, H.; et al. Destructive Extraction of Phospholipids from Escherichia Coli Membranes by Graphene Nanosheets. *Nat. Nanotechnol.* **2013**, *8*, 594–601.
- (585) Li, Y.; Yuan, H.; Von Dem Bussche, A.; Creighton, M.; Hurt, R. H.; Kane, A. B.; Gao, H. Graphene Microsheets Enter Cells through Spontaneous Membrane Penetration at Edge Asperities and Corner Sites. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 12295–12300.
- (586) Nel, A. E.; Mädler, L.; Velegol, D.; Xia, T.; Hoek, E. M. V.; Somasundaran, P.; Klaessig, F.; Castranova, V.; Thompson, M. Understanding Biophysicochemical Interactions at the Nano-Bio Interface. *Nat. Mater.* **2009**, *8*, 543–557.
- (587) Chen, P.; Yan, L. T. Physical Principles of Graphene Cellular Interactions: Computational and Theoretical Accounts. *J. Mater. Chem. B* **2017**, *5*, 4290–4306.
- (588) Lin, X.; Li, Y.; Gu, N. Nanoparticle's Size Effect on Its Translocation across a Lipid Bilayer: A Molecular Dynamics Simulation. *J. Comput. Theor. Nanosci.* **2010**, *7*, 269–276.
- (589) Gupta, R.; Rai, B. Effect of Size and Surface Charge of Gold Nanoparticles on Their Skin Permeability: A Molecular Dynamics Study. *Sci. Rep.* **2017**, *7*, 45292.
- (590) Van Lehn, R. C.; Atukorale, P. U.; Carney, R. P.; Yang, Y. S.; Stellacci, F.; Irvine, D. J.; Alexander-Katz, A. Effect of Particle Diameter and Surface Composition on the Spontaneous Fusion of Monolayer-Protected Gold Nanoparticles with Lipid Bilayers. *Nano Lett.* **2013**, *13*, 4060–4067.
- (591) Lin, J.; Miao, L.; Zhong, G.; Lin, C. H.; Dargazangy, R.; Alexander-Katz, A. Understanding the Synergistic Effect of Physicochemical Properties of Nanoparticles and Their Cellular Entry Pathways. *Commun. Biol.* **2020**, *3*, 205.
- (592) Agudo-Canalejo, J.; Lipowsky, R. Critical Particle Sizes for the Engulfment of Nanoparticles by Membranes and Vesicles with Bilayer Asymmetry. *ACS Nano* **2015**, *9*, 3704–3720.
- (593) Gao, H.; Shi, W.; Freund, L. B. Mechanics of Receptor-Mediated Endocytosis. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 9469–9474.
- (594) Vácha, R.; Martínez-Veracoechea, F. J.; Frenkel, D. Receptor-Mediated Endocytosis of Nanoparticles of Various Shapes. *Nano Lett.* **2011**, *11*, 5391–5395.

- (595) Li, Y.; Yue, T.; Yang, K.; Zhang, X. Molecular Modeling of the Relationship between Nanoparticle Shape Anisotropy and Endocytosis Kinetics. *Biomaterials* **2012**, *33*, 4965–4973.
- (596) Li, Y.; Kröger, M.; Liu, W. K. Shape Effect in Cellular Uptake of PEGylated Nanoparticles: Comparison between Sphere, Rod, Cube and Disk. *Nanoscale* **2015**, *7*, 16631–16646.
- (597) Huang, C.; Zhang, Y.; Yuan, H.; Gao, H.; Zhang, S. Role of Nanoparticle Geometry in Endocytosis: Laying down to Stand Up. *Nano Lett.* **2013**, *13*, 4546–4550.
- (598) Verma, A.; Uzun, O.; Hu, Y. Y.; Hu, Y. Y.; Han, H. S.; Watson, N.; Chen, S.; Irvine, D. J.; Stellacci, F. Surface-Structure-Regulated Cell-Membrane Penetration by Monolayer-Protected Nanoparticles. *Nat. Mater.* **2008**, *7*, 588–595.
- (599) Li, Y.; Li, X.; Li, Z.; Gao, H. Surface-Structure-Regulated Penetration of Nanoparticles across a Cell Membrane. *Nanoscale* **2012**, *4*, 3768–3775.
- (600) Lin, J.; Zhang, H.; Chen, Z.; Zheng, Y. Penetration of Lipid Membranes by Gold Nanoparticles: Insights into Cellular Uptake, Cytotoxicity, and Their Relationship. *ACS Nano* **2010**, *4*, 5421–5429.
- (601) Wernert, G. T.; Winkler, D. A.; Holan, G.; Nicoletti, G. Synthesis, Biological Activity, and QSAR Studies of Antimicrobial Agents Containing Biguanide Isosteres. *Aust. J. Chem.* **2004**, *57*, 77–85.
- (602) Lochbaum, C. A.; Chew, A. K.; Zhang, X.; Rotello, V.; Van Lehn, R. C.; Pedersen, J. A. Lipophilicity of Cationic Ligands Promotes Irreversible Adsorption of Nanoparticles to Lipid Bilayers. *ACS Nano* **2021**, *15*, 6562–6572.
- (603) Ding, H.-m.; Ma, Y.-q. Role of Physicochemical Properties of Coating Ligands in Receptor-Mediated Endocytosis of Nanoparticles. *Biomaterials* **2012**, *33*, 5798–5802.
- (604) Li, Y.; Kröger, M.; Liu, W. K. Endocytosis of PEGylated Nanoparticles Accompanied by Structural and Free Energy Changes of the Grafted Polyethylene Glycol. *Biomaterials* **2014**, *35*, 8467–8478.
- (605) Wang, X. X.; Wang, X. X.; Bai, X.; Yan, L.; Liu, T.; Wang, M.; Song, Y.; Hu, G.; Gu, Z.; Miao, Q.; et al. Nanoparticle Ligand Exchange and Its Effects at the Nanoparticle-Cell Membrane Interface. *Nano Lett.* **2019**, *19*, 8–18.
- (606) Kariuki, R.; Penman, R.; Bryant, S. J.; Orrell-Trigg, R.; Meftahi, N.; Crawford, R. J.; McConville, C. F.; Bryant, G.; Voitchovsky, K.; Conn, C. E.; et al. Behavior of Citrate-Capped Ultrasmall Gold Nanoparticles on a Supported Lipid Bilayer Interface at Atomic Resolution. *ACS Nano* **2022**, *16*, 17179–17196.
- (607) Gottwein, E.; Bodem, J.; Müller, B.; Schmechel, A.; Zentgraf, H.; Kräusslich, H.-G. The Mason-Pfizer Monkey Virus PPPY and PSAP Motifs Both Contribute to Virus Release. *J. Virol.* **2003**, *77*, 9474–9485.
- (608) Reynwar, B. J.; Illya, G.; Harmandaris, V. A.; Müller, M. M.; Kremer, K.; Deserno, M. Aggregation and Vesiculation of Membrane Proteins by Curvature-Mediated Interactions. *Nature* **2007**, *447*, 461–464.
- (609) Yue, T.; Zhang, X. Cooperative Effect in Receptor-Mediated Endocytosis of Multiple Nanoparticles. *ACS Nano* **2012**, *6*, 3196–3205.
- (610) Bahrami, A. H.; Lipowsky, R.; Weikl, T. R. The Role of Membrane Curvature for the Wrapping of Nanoparticles. *Soft Matter* **2016**, *12*, 581–587.
- (611) Šarić, A.; Cacciuto, A. Mechanism of Membrane Tube Formation Induced by Adhesive Nanocomponents. *Phys. Rev. Lett.* **2012**, *109*, 188101.
- (612) Jaskiewicz, K.; Larsen, A.; Schaeffel, D.; Koynov, K.; Lieberwirth, I.; Fytas, G.; Landfester, K.; Kroeger, A. Incorporation of Nanoparticles into Polymersomes: Size and Concentration Effects. *ACS Nano* **2012**, *6*, 7254–7262.
- (613) Yue, T.; Wang, X.; Huang, F.; Zhang, X. An Unusual Pathway for the Membrane Wrapping of Rodlike Nanoparticles and the Orientation- and Membrane Wrapping-Dependent Nanoparticle Interaction. *Nanoscale* **2013**, *5*, 9888–9896.
- (614) Xiong, K.; Zhao, J.; Yang, D.; Cheng, Q.; Wang, J.; Ji, H. Cooperative Wrapping of Nanoparticles of Various Sizes and Shapes by Lipid Membranes. *Soft Matter* **2017**, *13*, 4644–4652.
- (615) Yan, Z.; Wu, Z.; Li, S.; Zhang, X.; Yi, X.; Yue, T. Curvature-Mediated Cooperative Wrapping of Multiple Nanoparticles at the Same and Opposite Membrane Sides. *Nanoscale* **2019**, *11*, 19751–19762.
- (616) Wei, Y.; Tang, T.; Pang, H. B. Cellular Internalization of Bystander Nanomaterial Induced by TAT-Nanoparticles and Regulated by Extracellular Cysteine. *Nat. Commun.* **2019**, *10*, 3646.
- (617) Wei, Y.; Chen, H.; Li, Y. X.; He, K.; Yang, K.; Pang, H. B. Synergistic Entry of Individual Nanoparticles into Mammalian Cells Driven by Free Energy Decline and Regulated by Their Sizes. *ACS Nano* **2022**, *16*, 5885–5897.
- (618) Lavagna, E.; Bochicchio, D.; De Marco, A. L.; Güven, Z. P.; Stellacci, F.; Rossi, G. Ion-Bridges and Lipids Drive Aggregation of Same-Charge Nanoparticles on Lipid Membranes. *Nanoscale* **2022**, *14*, 6912–6921.
- (619) Zhang, H.; Ji, Q.; Huang, C.; Zhang, S.; Yuan, B.; Yang, K.; Ma, Y. Q. Cooperative Transmembrane Penetration of Nanoparticles. *Sci. Rep.* **2015**, *5*, 10525.
- (620) Chen, X.; Tieleman, D. P.; Liang, Q. Modulating Interactions between Ligand-Coated Nanoparticles and Phase-Separated Lipid Bilayers by Varying the Ligand Density and the Surface Charge. *Nanoscale* **2018**, *10*, 2481–2491.
- (621) Canepa, E.; Bochicchio, D.; Gasbarri, M.; Odino, D.; Canale, C.; Ferrando, R.; Canepa, F.; Stellacci, F.; Rossi, G.; Dante, S.; et al. Cholesterol Hinders the Passive Uptake of Amphiphilic Nanoparticles into Fluid Lipid Membranes. *J. Phys. Chem. Lett.* **2021**, *12*, 8583–8590.
- (622) Ou, L.; Chen, H.; Yuan, B.; Yang, K. Membrane-Specific Binding of 4 nm Lipid Nanoparticles Mediated by an Entropy-Driven Interaction Mechanism. *ACS Nano* **2022**, *16*, 18090–18100.
- (623) Baumgart, T.; Das, S.; Webb, W. W.; Jenkins, J. T. Membrane Elasticity in Giant Vesicles with Fluid Phase Coexistence. *Biophys. J.* **2005**, *89*, 1067–1080.
- (624) Sheavly, J. K.; Pedersen, J. A.; Van Lehn, R. C. Curvature-Driven Adsorption of Cationic Nanoparticles to Phase Boundaries in Multicomponent Lipid Bilayers. *Nanoscale* **2019**, *11*, 2767–2778.
- (625) McMahon, H. T.; Gallop, J. L. Membrane Curvature and Mechanisms of Dynamic Cell Membrane Remodelling. *Nature* **2005**, *438*, 590–596.
- (626) Li, Y.; Zhang, M.; Niu, X.; Yue, T. Selective Membrane Wrapping on Differently Sized Nanoparticles Regulated by Clathrin Assembly: A Computational Model. *Colloids Surf., B* **2022**, *214*, 112467.
- (627) Li, Y.; Niu, X.; Li, L.; Zhang, X.; Yang, K.; Yue, T. Size, Geometry and Mobility of Protein Assemblage Regulate the Kinetics of Membrane Wrapping on Nanoparticles. *J. Mol. Liq.* **2021**, *333*, 115990.
- (628) Chen, H.; Zhang, W.; Zhu, G.; Xie, J.; Chen, X. Rethinking Cancer Nanotheranostics. *Nat. Rev. Mater.* **2017**, *2*, 17024.
- (629) Yue, T.; Zhou, H.; Sun, H.; Li, S.; Zhang, X.; Cao, D.; Yi, X.; Yan, B. Why Are Nanoparticles Trapped at Cell Junctions When the Cell Density Is High? *Nanoscale* **2019**, *11*, 6602–6609.
- (630) Yu, M.; Xu, L.; Tian, F.; Su, Q.; Zheng, N.; Yang, Y.; Wang, J.; Wang, A.; Zhu, C.; Guo, S.; et al. Rapid Transport of Deformation-Tuned Nanoparticles across Biological Hydrogels and Cellular Barriers. *Nat. Commun.* **2018**, *9*, 2607.
- (631) Yu, M.; Song, W.; Tian, F.; Dai, Z.; Zhu, Q.; Ahmad, E.; Guo, S.; Zhu, C.; Zhong, H.; Yuan, Y.; et al. Temperature- and Rigidity-Mediated Rapid Transport of Lipid Nanovesicles in Hydrogels. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 5362–5369.
- (632) Xu, Z.; Dai, X.; Bu, X.; Yang, Y.; Zhang, X. X.; Man, X.; Zhang, X. X.; Doi, M.; Yan, L. T. Enhanced Heterogeneous Diffusion of Nanoparticles in Semiflexible Networks. *ACS Nano* **2021**, *15*, 4608–4616.
- (633) Chen, S. H.; Perez-Aguilar, J. M.; Zhou, R. Graphene-Extracted Membrane Lipids Facilitate the Activation of Integrin $\text{Av}\beta 8$. *Nanoscale* **2020**, *12*, 7939–7949.
- (634) Baimanov, D.; Wang, J.; Zhang, J.; Liu, K.; Cong, Y.; Shi, X.; Zhang, X.; Li, Y.; Li, X.; Qiao, R.; et al. In Situ Analysis of Nanoparticle Soft Corona and Dynamic Evolution. *Nat. Commun.* **2022**, *13*, 5389.
- (635) Ge, C.; Du, J.; Zhao, L.; Wang, L.; Liu, Y.; Li, D.; Yang, Y.; Zhou, R.; Zhao, Y.; Chai, Z.; et al. Binding of Blood Proteins to Carbon

Nanotubes Reduces Cytotoxicity. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 16968–16973.

(636) Wang, L.; Li, J.; Pan, J.; Jiang, X.; Ji, Y.; Li, Y.; Qu, Y.; Zhao, Y.; Wu, X.; Chen, C. Revealing the Binding Structure of the Protein Corona on Gold Nanorods Using Synchrotron Radiation-Based Techniques: Understanding the Reduced Damage in Cell Membranes. *J. Am. Chem. Soc.* **2013**, *135*, 17359–17368.

(637) Wang, X. X.; Lei, R.; Li, L.; Fei, X.; Ju, R.; Sun, X.; Cao, H.; Zhang, Q.; Chen, C.; Wang, X. X. Rearrangement of Protein Structures on a Gold Nanoparticle Surface Is Regulated by Ligand Adsorption Modes. *Nanoscale* **2021**, *13*, 20425–20436.

(638) Hu, Q.; Bai, X.; Hu, G.; Zuo, Y. Y. Unveiling the Molecular Structure of Pulmonary Surfactant Corona on Nanoparticles. *ACS Nano* **2017**, *11*, 6832–6842.

(639) Wang, X.; Wang, X.; Wang, M.; Zhang, D.; Yang, Q.; Liu, T.; Lei, R.; Zhu, S.; Zhao, Y.; Chen, C. Probing Adsorption Behaviors of BSA onto Chiral Surfaces of Nanoparticles. *Small* **2018**, *14*, 1703982.

(640) Lu, X.; Xu, P.; Ding, H. M.; Yu, Y. S.; Huo, D.; Ma, Y. Q. Tailoring the Component of Protein Corona via Simple Chemistry. *Nat. Commun.* **2019**, *10*, 4520.

(641) Ding, H.-m.; Ma, Y.-q. Computer Simulation of the Role of Protein Corona in Cellular Delivery of Nanoparticles. *Biomaterials* **2014**, *35*, 8703–8710.

(642) Duan, G.; Kang, S. G.; Tian, X.; Garate, J. A.; Zhao, L.; Ge, C.; Zhou, R. Protein Corona Mitigates the Cytotoxicity of Graphene Oxide by Reducing Its Physical Interaction with Cell Membrane. *Nanoscale* **2015**, *7*, 15214–15224.

(643) Lara, S.; Alnasser, F.; Polo, E.; Garry, D.; Lo Giudice, M. C.; Hristov, D. R.; Rocks, L.; Salvati, A.; Yan, Y.; Dawson, K. A. Identification of Receptor Binding to the Biomolecular Corona of Nanoparticles. *ACS Nano* **2017**, *11*, 1884–1893.

(644) Lee, H. Corona Formation: Effects of Nanoparticle Electrostatics and Protein-Protein Interactions on Corona Formation: Conformation and Hydrodynamics. *Small* **2020**, *16*, 2070054.

(645) Bernardi, R. C.; Melo, M. C. R.; Schulten, K. Enhanced Sampling Techniques in Molecular Dynamics Simulations of Biological Systems. *Biochim. Biophys. Acta - Gen. Subj.* **2015**, *1850*, 872–877.

(646) Noon, W. H.; Kong, Y.; Jianpeng, M. Molecular Dynamics Analysis of a Buckyball-Antibody Complex. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 6466–6470.

(647) An, D.; Su, J.; Weber, J. K.; Gao, X.; Zhou, R.; Li, J. A Peptide-Coated Gold Nanocluster Exhibits Unique Behavior in Protein Activity Inhibition. *J. Am. Chem. Soc.* **2015**, *137*, 8412–8418.

(648) Zuo, G.; Huang, Q.; Wei, G.; Zhou, R.; Fang, H. Plugging into Proteins: Poisoning Protein Function by a Hydrophobic Nanoparticle. *ACS Nano* **2010**, *4*, 7508–7514.

(649) Xue, X.; Yang, J. Y.; He, Y.; Wang, L. R.; Liu, P.; Yu, L. S.; Bi, G. H.; Zhu, M. M.; Liu, Y. Y.; Xiang, R. W.; et al. Aggregated Single-Walled Carbon Nanotubes Attenuate the Behavioural and Neurochemical Effects of Methamphetamine in Mice. *Nat. Nanotechnol.* **2016**, *11*, 613–620.

(650) Yu, Y.; Sun, H.; Gilmore, K.; Hou, T.; Wang, S.; Li, Y. Aggregated Single-Walled Carbon Nanotubes Absorb and Deform Dopamine-Related Proteins Based on Molecular Dynamics Simulations. *ACS Appl. Mater. Interfaces* **2017**, *9*, 32452–32462.

(651) Sun, X.; Feng, Z.; Hou, T.; Li, Y. Mechanism of Graphene Oxide as an Enzyme Inhibitor from Molecular Dynamics Simulations. *ACS Appl. Mater. Interfaces* **2014**, *6*, 7153–7163.

(652) Zuo, G.; Zhou, X.; Huang, Q.; Fang, H.; Zhou, R. Adsorption of Villin Headpiece onto Graphene, Carbon Nanotube, and C60: Effect of Contacting Surface Curvatures on Binding Affinity. *J. Phys. Chem. C* **2011**, *115*, 23323–23328.

(653) Tian, X.; Yang, Z.; Duan, G.; Wu, A.; Gu, Z.; Zhang, L.; Chen, C.; Chai, Z.; Ge, C.; Zhou, R. Graphene Oxide Nanosheets Retard Cellular Migration via Disruption of Actin Cytoskeleton. *Small* **2017**, *13*, 1602133.

(654) Yan, Z.; Li, L.; Li, S.; Xu, Y.; Yue, T. Extracellular Interactions between Graphene Nanosheets and E-Cadherin. *Environ. Sci. Nano* **2021**, *8*, 2152–2164.

(655) Gu, Z.; Baggetta, A. M.; Chong, Y.; Plant, L. D.; Meng, X. Y.; Zhou, R. Multifaceted Regulation of Potassium-Ion Channels by Graphene Quantum Dots. *ACS Appl. Mater. Interfaces* **2021**, *13*, 27784–27795.

(656) Noé, F.; Tkatchenko, A.; Müller, K. R.; Clementi, C. Machine Learning for Molecular Simulation. *Annu. Rev. Phys. Chem.* **2020**, *71*, 361–390.

(657) Friederich, P.; Häse, F.; Proppe, J.; Aspuru-guzik, A. Machine-Learned Potentials for next-Generation Matter Simulations. *Nat. Mater.* **2021**, *20*, 750–761.

(658) Bai, Q.; Liu, S.; Tian, Y.; Xu, T.; Banegas-Luna, A. J.; Perez-Sanchez, H.; Huang, J.; Liu, H.; Yao, X. Application Advances of Deep Learning Methods for de Novo Drug Design and Molecular Dynamics Simulation. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2022**, *12*, No. e1581.

(659) Wang, Y.; Lamim Ribeiro, J. M.; Tiwary, P. Machine Learning Approaches for Analyzing and Enhancing Molecular Dynamics Simulations. *Curr. Opin. Struct. Biol.* **2020**, *61*, 139–145.

(660) Gkeka, P.; Stoltz, G.; Farimani, A. B.; Belkacemi, Z.; Ceriotti, M.; Chodera, J. D.; Dinner, A. R.; Ferguson, A. L.; Maillet, J.; Minoux, H.; et al. Machine Learning Force Fields and Coarse-Grained Variables in Molecular Dynamics: Application to Materials and Biological Systems. *J. Chem. Theory Comput.* **2020**, *16*, 4757–4775.

(661) Feng, T.; Yang, B.; Lu, G. Investigation on the Local Structure and Properties of Molten Li₂CO₃-K₂CO₃ Binary Salts by Machine Learning Potentials. *J. Mol. Liq.* **2022**, *356*, 118979.

(662) Bhatia, H.; Carpenter, T. S.; Ingolfsson, H. I.; Dharuman, G.; Karande, P.; Liu, S.; Oppelstrup, T.; Neale, C.; Lightstone, F. C.; Van Essen, B.; Glosli, J. N.; Bremer, P.-T.; et al. Machine-Learning-Based Dynamic-Importance Sampling for Adaptive Multiscale Simulations. *Nat. Mach. Intell.* **2021**, *3*, 401–409.

(663) Jackson, N. E.; Bowen, A. S.; Antony, L. W.; Webb, M. A.; Vishwanath, V.; de Pablo, J. J. Electronic Structure at Coarse-Grained Resolutions from Supervised Machine Learning. *Sci. Adv.* **2019**, *5*, No. eaav1190.

(664) Doerr, S.; Majewski, M.; Perez, A.; Kramer, A.; Clementi, C.; Noe, F.; Giorgino, T.; De Fabritiis, G. TorchMD: A Deep Learning Framework for Molecular Simulations. *J. Chem. Theory Comput.* **2021**, *17*, 2355–2363.

(665) Wang, H.; Zhang, L.; Han, J.; Weinan, E. DeePMD-Kit: A Deep Learning Package for Many-Body Potential Energy Representation and Molecular Dynamics. *Comput. Phys. Commun.* **2018**, *228*, 178–184.

(666) Yao, K.; Herr, J. E.; Toth, D. W.; Mckintyre, R.; Parkhill, J. The TensorMol-0.1 Model Chemistry: A Neural Network Augmented with Long-Range Physics. *Chem. Sci.* **2018**, *9*, 2261–2269.

(667) Price, C. C.; Singh, A.; Frey, N. C.; Shenoy, V. B. Efficient Catalyst Screening Using Graph Neural Networks to Predict Strain Effects on Adsorption Energy. *Sci. Adv.* **2022**, *8*, No. eabq5944.

(668) Li, Z.; Ma, X.; Xin, H. Feature Engineering of Machine-Learning Chemisorption Models for Catalyst Design. *Catal. Today* **2017**, *280*, 232–238.

(669) Fung, V.; Ganesh, P.; Sumpter, B. G. Machine Learned Features from Density of States for Accurate Adsorption Energy Prediction. *Nat. Commun.* **2021**, *12*, 88.

(670) Gao, R.; Xu, L.; Sun, M.; Xu, M.; Hao, C.; Guo, X.; Colombari, F. M.; Zheng, X.; Král, P.; de Moura, A. F.; et al. Site-Selective Proteolytic Cleavage of Plant Viruses by Photoactive Chiral Nanoparticles. *Nat. Catal.* **2022**, *5*, 694–707.

(671) Yamaguchi, S. I.; Xie, Q.; Ito, F.; Terao, K.; Kato, Y.; Kuroiwa, M.; Omori, S.; Taniura, H.; Kinoshita, K.; Takahashi, T.; et al. Carbon Nanotube Recognition by Human Siglec-14 Provokes Inflammation. *Nat. Nanotechnol.* **2023**, DOI: 10.1038/s41565-023-01363-w.

(672) Zhu, M.; Zhuang, J.; Li, Z.; Liu, Q.; Zhao, R.; Gao, Z.; Midgley, A. C.; Qi, T.; Tian, J.; Zhang, Z.; et al. Machine-Learning-Assisted Single-Vessel Analysis of Nanoparticle Permeability in Tumour Vasculatures. *Nat. Nanotechnol.* **2023**, DOI: 10.1038/s41565-023-01323-4.

(673) Ma, J.; Wang, S.; Zhao, C.; Yan, X.; Ren, Q.; Dong, Z.; Qiu, J.; Liu, Y.; Shan, Q.; Xu, M.; et al. Computer-aided Discovery of Potent

Broad-spectrum Vaccine Adjuvants. *Angew. Chem., Int. Ed.* **2023**, *135*, No. e202301059.

(674) Zhang, P.; Guo, Z.; Ullah, S.; Melagraki, G.; Afantitis, A.; Lynch, I. Nanotechnology and Artificial Intelligence to Enable Sustainable and Precision Agriculture. *Nat. Plants* **2021**, *7*, 864–876.

(675) Mhasawade, V.; Zhao, Y.; Chunara, R. Machine Learning and Algorithmic Fairness in Public and Population Health. *Nat. Mach. Intell.* **2021**, *3*, 659–666.

(676) Zhong, S.; Zhang, K.; Bagheri, M.; Burken, J. G.; Gu, A.; Li, B.; Ma, X.; Marrone, B. L.; Ren, Z. J.; Schrier, J.; et al. Machine Learning: New Ideas and Tools in Environmental Science and Engineering. *Environ. Sci. Technol.* **2021**, *55*, 12741–12754.

Recommended by ACS

Evidence-Based Prediction of Cellular Toxicity for Amorphous Silica Nanoparticles

, Kenji Mizuguchi, *et al.*

MAY 30, 2023
ACS NANO

READ 

Data-Driven Design of Classes of Ruthenium Nanoparticles Using Multitarget Bayesian Inference

Jonathan Y. C. Ting, Amanda S. Barnard, *et al.*

JANUARY 09, 2023
CHEMISTRY OF MATERIALS

READ 

Big Data in a Nano World: A Review on Computational, Data-Driven Design of Nanomaterials Structures, Properties, and Synthesis

Ruo Xi Yang, Kristin A. Persson, *et al.*

NOVEMBER 15, 2022
ACS NANO

READ 

Interfacing Nanomaterials with Biology through Ligand Engineering

Aarohi Gupta, Vincent M. Rotello, *et al.*

JULY 28, 2023
ACCOUNTS OF CHEMICAL RESEARCH

READ 

Get More Suggestions >

Prediction of Nano–Bio Interactions through Convolutional Neural Network Analysis of Nanostructure Images

Xiliang Yan, Jin Zhang, Daniel P. Russo, Hao Zhu,* and Bing Yan*

Cite This: *ACS Sustainable Chem. Eng.* 2020, 8, 19096–19104

Read Online

ACCESS |



Metrics & More



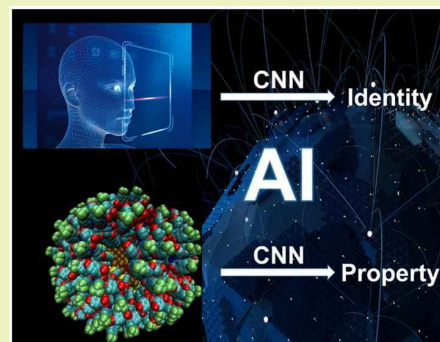
Article Recommendations



Supporting Information

ABSTRACT: Artificial intelligence approaches, such as machine learning and deep learning, may predict nano–bio interactions. However, such a prediction is now hindered by the paucity of suitable nanodescriptors with applicable nanostructure annotation methods. Inspired by face recognition technology, we have developed a novel nanostructure annotation method to automatically convert nanostructures to images for convolutional neural network modeling. In this operation, nanostructure features were directly learned from nanoparticle images without complicated nanodescriptor calculations. The constructed convolutional neural network models were successfully used to predict physicochemical properties (i.e., logP and zeta potential) and biological activities (i.e., cellular uptake and protein adsorption) of 147 unique nanoparticles, including 123 gold nanoparticles, 12 platinum nanoparticles, and 12 palladium nanoparticles. Our nanostructure diversity and wide distribution of experimental values are beneficial for building predictive deep learning models. The deep learning models provide highly accurate predictions with all determination coefficients (R^2) higher than 0.68 for both cross validation and external prediction. In addition, the constructed model is explainable because we can visualize how it learns from the class activation map. This approach enables a much more efficient end-to-end deep learning modality suitable for design of next generation nanomaterials.

KEYWORDS: Artificial intelligence, Face recognition, Nano–Bio interaction, Nanostructure annotation, Nanomaterial design



INTRODUCTION

The ability to predict nano–bio interactions is critical for the design of novel nanomaterials.^{1,2} Artificial intelligence approaches, such as machine learning and deep learning, have been applied to predict a wide variety of nanomaterial properties, such as logP,³ zeta potential,⁴ cellular uptake,⁵ cytotoxicity,⁶ drug delivery,⁷ and even *in vivo* fate.⁸ Unfortunately, accurate predictions are still hindered by the lack of larger sets of high-quality nano–bio interaction data and suitable nanodescriptors. Most previous nanomaterial modeling studies are based on small data sets that normally consisted of fewer than 40 congeneric nanomaterials.^{6–11} The nanodescriptors previously used in nanomaterial modeling were mainly experimental, empirical, or ligand descriptors. Experimental nanodescriptors, such as nanomaterial size,⁹ zeta potential,¹² relaxivities (magnetic properties),¹³ and protein corona fingerprint,⁵ have been widely used in previous machine learning models for predicting nano–bio interactions. For example, eight qualitatively experimental properties (e.g., nanoparticle type and shape) and 13 quantitatively experimental properties (e.g., nanoparticle size and zeta potential) were applied to predict the functional composition of the protein corona and the cellular recognition of nanoparticles.¹⁴ However, these experimental values vary from laboratory to laboratory and are not reliable without universal reference

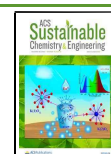
materials and standardized procedures. More importantly, it is not possible to acquire these descriptors for new nanomaterials before chemical synthesis. Empirical descriptors, usually involving the application of quantum chemistry⁶ or molecular simulation,^{15,16} are rather limited by the high demand for computational resources. Ligand descriptors, merely calculated from surface ligands,¹⁷ cannot fully represent the entire nanostructures including nanomaterial type, shape, size, ligand position and density. The major issue of lacking suitable nanodescriptors is due to the annotation of nanostructures is much more complicated than small molecules and even proteins. Therefore, a new approach for nano–bio interaction prediction without nanodescriptor calculations is much needed.

The convolutional neural network deep learning algorithm has been successfully applied to face recognition,¹⁸ self-driving vehicles,¹⁹ and medical diagnosis.²⁰ Compared to the traditional machine learning approaches, the convolutional neural

Received: October 11, 2020

Revised: November 14, 2020

Published: December 15, 2020



network deep learning algorithm can build end-to-end models through extracting molecular features automatically from molecular images,^{21,22} simplifying nanostructure annotations, and completely skipping the step of descriptor calculations. This strategy has been used to predict properties of small molecules,²³ crystals,²⁴ and proteins²⁵ from their corresponding structure images. For example, various deep learning approaches were reported to predict toxicity of small molecules merely from their graphic images²³ or classify the space groups from crystal structure images.²⁴ However, prediction of nano-bio interactions using the “face recognition” approach was only reported recently as a proof of concept by using two-dimensional images of a relatively small data set.²⁶ To demonstrate the applicability of a convolutional neural network with a large nanomaterial data set of three-dimensional images for nano-bio interaction prediction, we carried out convolutional neural network modeling of 147 diverse nanoparticles by directly learning nanoparticle features from three-dimensional nanostructure images in this study.

MATERIALS AND METHODS

Nanoparticle Data Sets. The 147 nanoparticles used to generate nanostructure images were curated from our previous publications.^{27–30} All nanoparticles were synthesized by the nanocombinatorial library approach.³¹ The synthesis and extensive characterization of nanoparticles were under strict quality control, which ensured the reliability of data sets used for deep learning studies. As shown in Table 1, based on different physicochemical properties and biological

Table 1. Overview of Nanoparticle Data Sets Used in This Study

| Data set | Number of nanoparticles | Physicochemical properties or biological activities | refs |
|----------|-------------------------|---|------------|
| 1 | 147 | logP | 27–30 |
| 2 | 119 | Zeta potential | 27, 29, 30 |
| 3 | 77 | Cellular uptake | 29, 30 |
| 4 | 36 | Protein adsorption | 30 |

activities, four data sets were generated to validate the convolutional neural network modeling. The first data set contained 147 nanoparticles tested for their nanohydrophobicity. The nanohydrophobicity was indicated by the experimental logP value that was defined by the ratio of the concentration of nanoparticles in water to their concentration in octanol. In the second data set, 119 nanoparticles were tested for their zeta potential, which was used to evaluate the surface charge of nanoparticles. The experimental value of the zeta potential was measured in a Malvern Nano Z Zetasizer by suspending nanoparticles in Millipore water with pH 7. The last two data sets were used to construct convolutional neural network models for nano-bio interaction prediction. In the third data set, 77 nanoparticles were tested for their cellular uptake potential. Here, adenocarcinomic human alveolar-based epithelial (A549) cells were used to incubate with nanoparticles at 50 $\mu\text{g mL}^{-1}$ for 24 h. The last data set contained 36 nanoparticles tested for their protein adsorption capacity. To this end, 1×10^{16} nm² nanoparticles were added to 1.0 mL of 10% serum in phosphate-buffered saline (PBS) and incubated at 37 °C for 24 h.

Annotation of Nanostructures. Nanostructure annotation or nanostructure digitalization can convert nanostructures into a format that is suitable for convolutional neural network modeling. In the present study, two steps were employed to annotate each nanoparticle: (1) simulation of the nanostructure to generate a three-dimensional virtual nanoparticle and (2) nanostructure image generation based on the resulted virtual nanoparticle. An in-house program, coded in Python 3.5, was used to generate the three-dimensional conformations of nanostructures based on four basic

nanoparticle attributes: (1) type of core material, (2) size of nanoparticle core, (3) chemical structure of surface ligand, and (4) surface ligand density. As shown in Figure S1a, after inputting the above nanostructure parameters, the protein data bank (PDB) files containing the three-dimensional nanostructural information were generated. The construction of a PDB file in the program is briefly described below. First, based on the material type (i.e., gold, platinum, and palladium), a different unit cell was used to stack into a cube with a specific size. Then, extra atoms were deleted from the cube to form a sphere with the nanoparticle core size. Then, the associated surface ligands were randomly placed on the nanoparticle core surface to reach the input ligand density. At last, the resulting virtual nanoparticle with detailed three-dimensional nanostructure information (i.e., atomic types and coordinates, connection between atoms) was saved as a protein data bank (PDB) file. All PDB files are freely available from a nano-bio interaction database (<http://www.pubvinas.com/>).³²

Prior to modeling, the PDB files of all nanoparticles were converted to nanostructure images, which were suitable for convolutional neural network input. Using an in-house program, the PDB files of 147 nanoparticles were automatically imported into visual molecular dynamics (VMD) software one by one and then converted to uniform images. The in-house program was written in a tool command language (Figure S1b) and contained the basic settings of nanostructure images (i.e., atom size, atom color, and image format). The program can be executed by entering the codes in the VMD console or by running the TCL format text that contained all the codes. Here, the nanostructure in each image was represented by the van der Waals (VDW) drawing method. Briefly, the nanoparticle atoms were drawn as various spheres. The sphere size reflected the van der Waals radius of each atom, and different atoms were represented in different colors (Figure S2a). Here, in order to better reflect the three-dimensional nature of nanoparticle structures, 18 (360°/20°) images were generated for each nanoparticle by taking a screenshot after every 20° rotation around the y axis (Figure S2b). As a result, a total of 2646 images were generated for 147 nanoparticles. All images were cropped to the same size in a portable network graphics (PNG) format, and the background is set to black.

Convolutional Neural Network Modeling. The LeNet architecture is mainly used to construct the convolutional neural network model, which is implemented with TensorFlow 1.14.0 and Keras 2.2.5. The convolutional neural network model mainly contains two parts. One is the feature extractor including the first four pairs of convolutional layers and max-pooling layers; the other part is the data predictor with two fully connected layers. To be more specific, there are 32, 64, 128, and 128 filters in the four convolutional layers, and the kernel sizes are all set as (3, 3), with strides as (1, 1). The pool sizes of all max-pooling layers are set as (2, 2), with strides as (2, 2). The results of the last max-pooling layer are flattened and passed to the following two dense layers with 512 units. Before the output layer, a dropout layer with dropout rate of 0.3 is added to reduce overfitting. Finally, an output layer is attached with only one neuron to generate the predicted value. The architecture of the LeNet convolutional neural network is shown in Figure S3. The ReLU (Rectified Linear Unit) is used as the activation function to perform nonlinear transformation. The learning rate is set as 10^{-3} , and MSE (Mean Square Error) is set as the loss function. Training is performed using RMSprop (Root Mean Square prop) optimization with batches of 32 images for 300 epochs when no significant changes of MSE were observed (Figure S4). There was no significant difference between training loss and validation loss, indicating that the model results were not overfitting after data augmentation (i.e., screenshot from multiviews) and dropout regularization.

In order to evaluate the convolutional neural network model, each data set is randomly split into a training set (80% of the whole set) and test set (20% of the whole set). The training set is used to build an initial model, and the test set is used to estimate the model's predictive ability. The model robustness is verified by internal validation using the 5-fold cross validation algorithm, which has been widely used in our previous studies.^{4,33} In short, the training set is

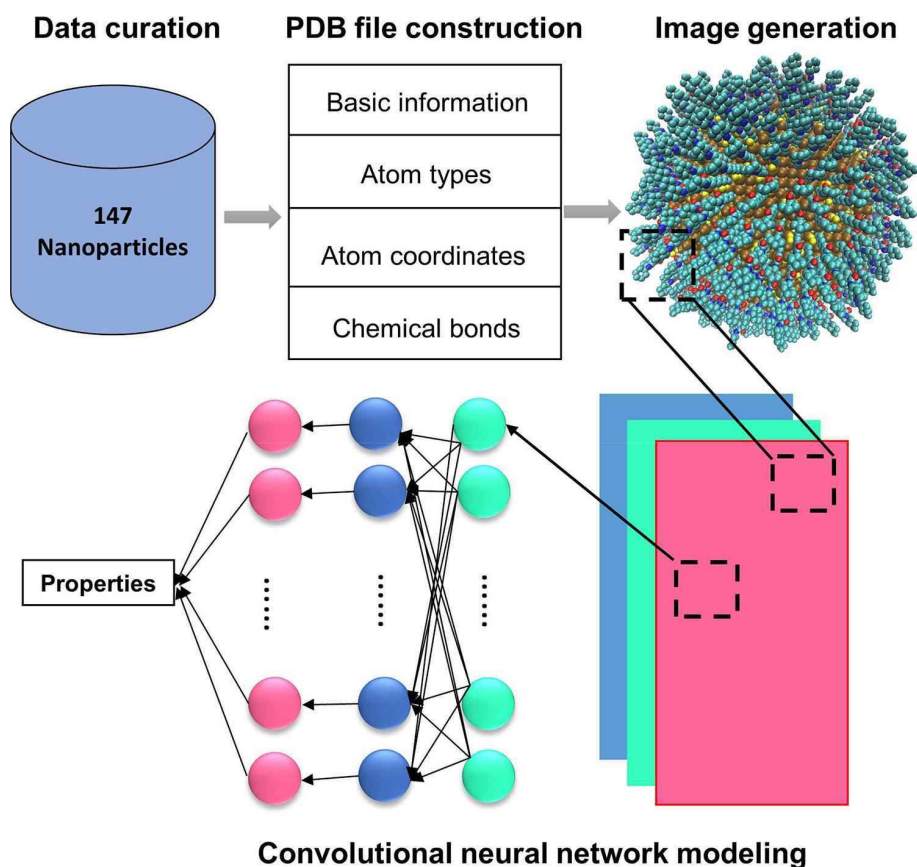


Figure 1. Schematic of the computational workflow. The workflow mainly includes nanomaterial data curation, PDB file construction, image generation, and prediction of nanomaterial properties using a convolutional neural network. In the present study, the convolutional neural network can directly extract features from nanostructure images without complicated nanodescriptor calculations.

further split into five subsets. The model is developed using four of the five subsets, and the remaining subset is used for validation. This procedure is repeated five times until all subsets are used for validation once. In this study, for both 5-fold cross validation and external prediction, all 18 images for this nanomaterial are used if the nanomaterial is selected for training purposes, and only one image is randomly selected if the nanomaterial is used for validation purposes. In order to improve calculation efficiency, original images (600×600 pixels) were preprocessed to 200×200 pixel images using the *Image* module of *Keras* before putting them into the convolutional neural network model. Here, the nearest-neighbor interpolation method that replaced every pixel with the nearest pixel in the output was used to downsample all the original images. The model performance is accessed by the coefficient of determination (R^2) for 5-fold cross validation ($R2_SCV$) and external prediction ($R2_val$).

RESULTS AND DISCUSSION

Workflow, Diverse Nanoparticle Structures, and Properties. Our multistage approach for nanomaterial properties prediction is schematically illustrated in Figure 1. It consists of three integral components: nanomaterial data curation, nanostructure annotation (i.e., PDB file construction and image generation), and convolutional neural network model construction. The data sets were curated as in our previous studies.^{27–30} On the basis of the experimental data, an in-house program was used to construct the PDB file that contained the detailed atom information of a nanoparticle. Then, uniform nanostructure images can be generated from the corresponding PDB files. Different atoms are represented

by balls of different colors and sizes. For clarification purposes, all atomic radii were set as the same in the schematic diagram. Compared to the transmission electron images (TEM) images, these nanostructure images contained more detailed atom information. Inspired by the face recognition technique, the convolutional neural network was then applied to directly extract critical nanostructure image features for nano–bio interaction modeling and predictions. These critical image features (shown as dotted box) usually indicate important nanostructure features (e.g., aromatic and aliphatic rings) that are responsible for the nano–bio interactions.

Figure 2 shows the distribution of the experimental values of material type, nanoparticle size, surface ligand number, physicochemical properties, and biological activities. The data sets contain 147 unique nanoparticles covering three material types (Figure 2a), i.e., 123 gold nanoparticles (GNPs), 12 platinum nanoparticles (PtNPs), and 12 palladium nanoparticles (PdNPs). Primary sizes of most nanoparticles are less than 10 nm (Figure 2b), which are on a size scale commensurate with proteins.³⁴ As a result, these nanomaterials can be used to construct peptide-functioned biomimetic surfaces that are suitable for biomedical applications.^{35–37} Due to the diversity of the nanomaterial type, size, and ligand property, the surface ligand number also exhibits large difference between different nanoparticles (Figure 2c). For example, similar-sized GNP (~5.9 nm) can have over 700 neutral ligands per particle (e.g., GNP106), while ligands with positively charged (e.g., GNP84) or negatively charged (e.g.,

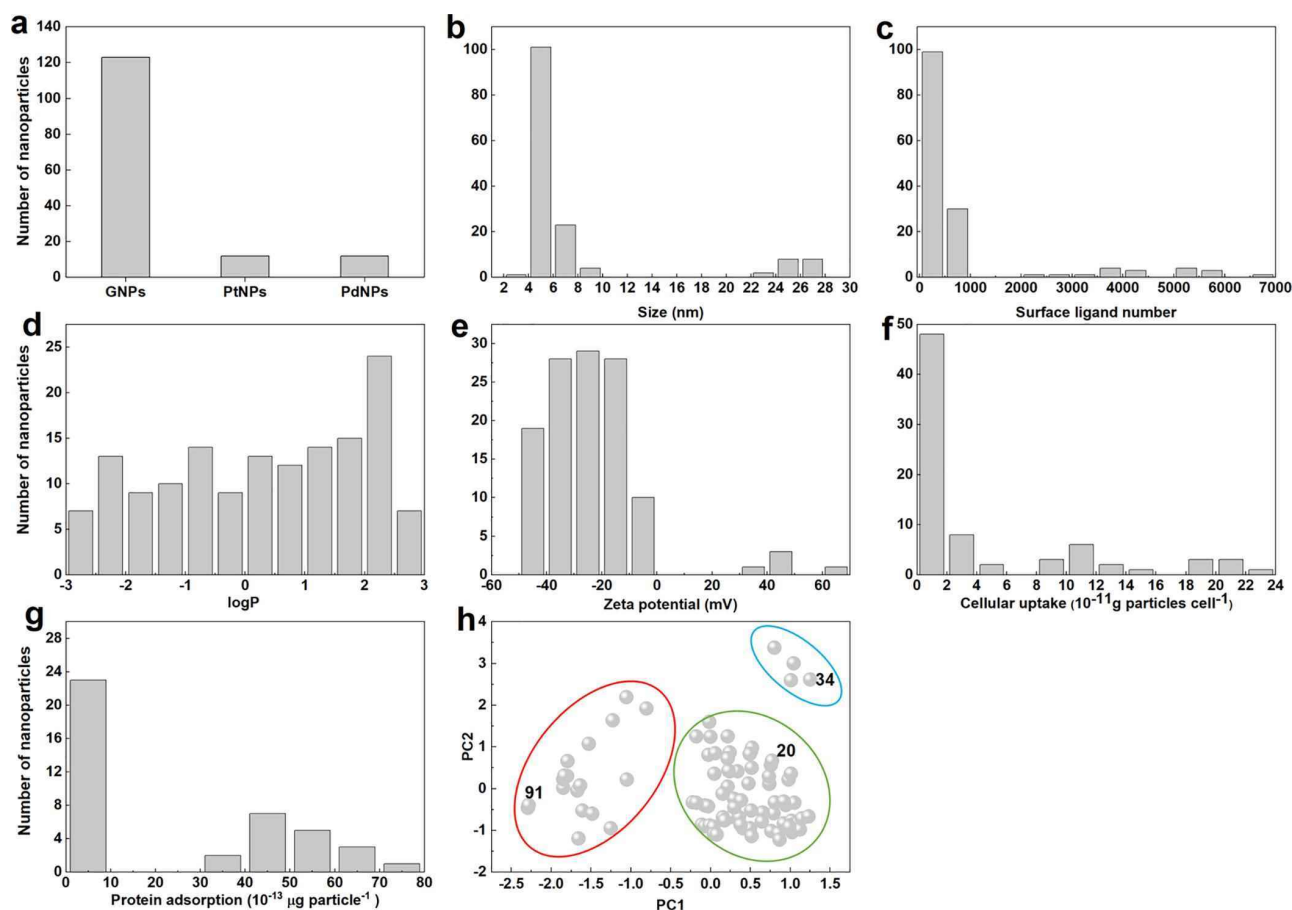


Figure 2. Summary of the nanoparticle data sets. Distributions of experimental values for (a) material types, (b) nanoparticle size, (c) surface ligand number, (d) logP, (e) zeta potential, (f) cellular uptake, (g) protein adsorption, and (h) principal component analysis of 91 unique surface ligands; the top two components account for 45% of the total descriptor variance. These surface ligands are divided into three clusters, reflecting the nanostructure diversity.

GNP92) ligands can only pack up to around 200 ligands per particle. Among these nanoparticles, there are totally 91 unique surface ligands. In addition, the experimental values of physicochemical properties and biological activities also exhibit wide distribution (Figure 2d–g). The logP values of these nanoparticles, which describe the hydrophobicity of relevant nanoparticles, ranged from -2.68 to 2.72 (Figure 2d). Zeta potential indicating the charge at the interface between the nanoparticle surface and its liquid medium ranged from -49.40 to 65.30 mV (Figure 2e). Cellular uptake and protein adsorption, which were highly related to the *in vivo* fate of nanoparticles ranged from 0.02×10^{-11} to 22.81×10^{-11} g (Figure 2f) and from 4.2×10^{-13} to 71×10^{-13} μg (Figure 2g). Furthermore, the principal component analysis (PCA) of the 91 unique surface ligands also reflects the surface chemistry diversity of the nanostructures (Figure 2h). Due to their structural differences, these surface ligands are generally divided into three clusters: surface ligands with short chains (red circle), surface ligands with long chains (green circle), and surface ligands with long chains and special groups (e.g., multiple hydroxyl and fluorine atoms at the end; blue circle). Three representative ligands from the clusters can be seen in Figure S5. Detailed information of 91 surface ligands and all experimental values are shown in Figure S10 and Table S1.

Nanostructure Images Depict Nanoparticle Structures and Structural Differences. The next step of this study is nanostructure annotation, aiming to generate a suitable input format for convolutional neural network modeling. First, a three-dimensional virtual nanostructure was created for each nanoparticle using an in-house program. Then, virtual nanostructures were converted into uniform images containing atomic information (i.e., atom types and coordinates) of nanoparticles. Figure 3a shows six nanostructure images generated from six different nanoparticles covering three material types (i.e., gold, platinum, and palladium). Although these images show us an intuitive impression of the nanostructure diversity and differences (e.g., nanomaterial type, nanomaterial size, and surface ligand), it still cannot fully reflect the actual experimental values due to the complexity of the nanostructures. In the future, it is necessary to develop an advanced representation method that can more accurately reflect actual nanostructures. Although each nanostructure image is a two-dimensional matrix in space, the three-dimensional nanostructures can be better reflected by different images obtained from multiviews (Figure S2b). Overall, to some extent, these color images could depict the full nanoparticle structures and structural differences among these nanoparticles. In order to better explain how a color image is processed and fed to the convolutional neural network

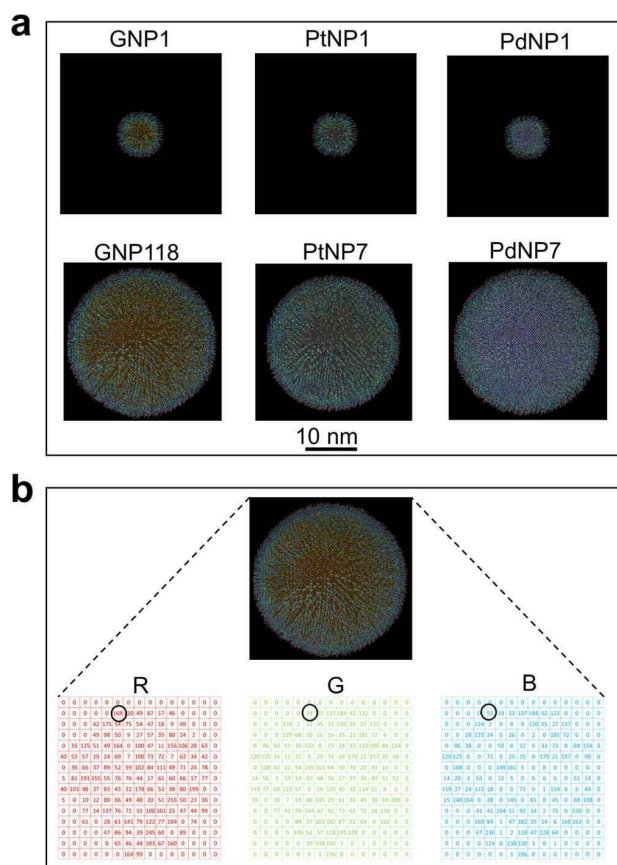


Figure 3. Generated nanostructure images and digital representations of the image. (a) The images were generated from six different nanoparticles, which vary with different cores, sizes, and surface ligands. Different atoms are represented by balls with different colors and sizes. (b) In the computer coding, each color image was split into three channels, i.e., red (R), green (G), and blue (B). Each channel is a two-dimensional matrix where the values represent the discrete pixels with conventional brightness intensities between 0 and 255. The values in the black circle indicate that there are more oxygen atoms in this location.

for feature extraction, the digital representation of one nanostructure image was shown. As for a color imaging system, each color is typically represented by three or four component intensities (integer values) such as red (R), green (G), and blue (B). For example, the RGB value for black was (0, 0, 0). On the basis of the commonly used RGB color space, the nanostructure image (i.e., GNP118) was decomposed into three two-dimensional matrices as shown in Figure 3b. Because each atom was represented as a ball of different size and color (Figure S2a), the values of the RGB channel can reflect the nanostructure features (e.g., atom types) in certain locations. For example, the values in the second row and sixth column of the RGB channels were (168, 19, 53) (scarlet red), indicating more oxygen atoms in this location of GNP118. Then, these matrices were directly fed into the convolutional neural network and operated with the convolutional kernel (a 3×3 matrix) for nanostructure feature extractions.

Convolutional Neural Network Models Predict Nano-Bio Interactions without Complicated Nanodescriptor Calculations. In the present study, two physicochemical properties (i.e., logP and zeta potential) and two nano-bio

interactions (i.e., cellular uptake and protein adsorption) were selected to verify the effectiveness of the deep learning method. The logP and zeta potential indicated the hydrophobicity and surface charge of nanoparticles, respectively. These properties were highly related to cellular uptakes and protein adsorptions. Previous studies²⁹ showed that hydrophobic and positively charged nanoparticles entered cells more readily than hydrophilic or negatively nanoparticles. Physicochemical properties of nanoparticles may also affect their *in vivo* fate. For nanoparticles, the cellular uptake is usually a prerequisite testing for their application in drug delivery³⁸ and bioimaging.³⁹ The chemical and physical properties of nanoparticles are immediately altered as they absorb proteins. Hence, the accurate prediction of these properties and activities may guide future nanomaterial design for biomedical applications.

Here, two convolutional neural network methods, LeNet⁴⁰ and GoogLeNet,⁴¹ were used to construct the deep learning models. The LeNet was one of the earliest convolutional neural networks proposed by LeCun et al. in 1998⁴⁰ and promoted the development of deep learning. Here, the LeNet convolutional neural network model contained 13 layers (Figure S3). Through continuous learning in a layer-by-layer manner, the relationships between the nanostructure images and nano-bio interaction data were built. By continuous refinement, the deep learning models adjusted the parameters from layer to layer to improve the model performance, which was accessed by the coefficient of determination (R^2), root mean squared error (RMSE) for 5-fold cross validation (R^2_{5CV} and $RMSE_{5CV}$), and external prediction (R^2_{val} and $RMSE_{val}$). Figure 4 shows the correlations between the predicted values and the experimental values of various physicochemical properties and biological activities for all nanoparticles, which also includes R^2 and RMSE. Overall, both R^2 and RMSE for 5-fold cross validation (R^2_{5CV} and $RMSE_{5CV}$) and external prediction (R^2_{val} and $RMSE_{val}$) are at the same order of magnitude, indicating that the 5-fold cross-validation process and external prediction yielded similar results. All determination coefficients (both R^2_{5CV} and R^2_{val}) were above 0.68, indicating that all deep learning models successfully predicted the relationships between the nanostructures and target activities. Especially, the cellular uptake and protein adsorption were predicted with better accuracy, demonstrating the strong predictive power of convolutional neural network models for nano-bio interactions. To avoid correlation by chance of the convolutional neural network model, we additionally performed the Y-scrambling numerical experiment. We built 100 random “models” utilizing the same nanostructure image but correlated it with the physicochemical property or biological activity data randomly shuffled every time. Since the values of both $RMSE_{5CV}$ and $RMSE_{val}$ for the convolutional neural network model were at least two times lower than these for the randomly obtained models (Figure S6), we have confirmed that the convolutional neural network model was not obtained by chance.

Although the properties of most nanoparticles can be correctly predicted, a couple of predictive outliers were also noticed. For example, the experimental logP values of GNP22 and GNP30 were -0.78 and -2.08 , while the predicted values were 1.61 and 0.55, respectively. In order to explore why GNP22 and GNP30 were outliers, we checked images of the two nanoparticles (Figure S7). Both nanoparticles contained fluorine (F) atoms (lemon) at the end of the surface ligands.

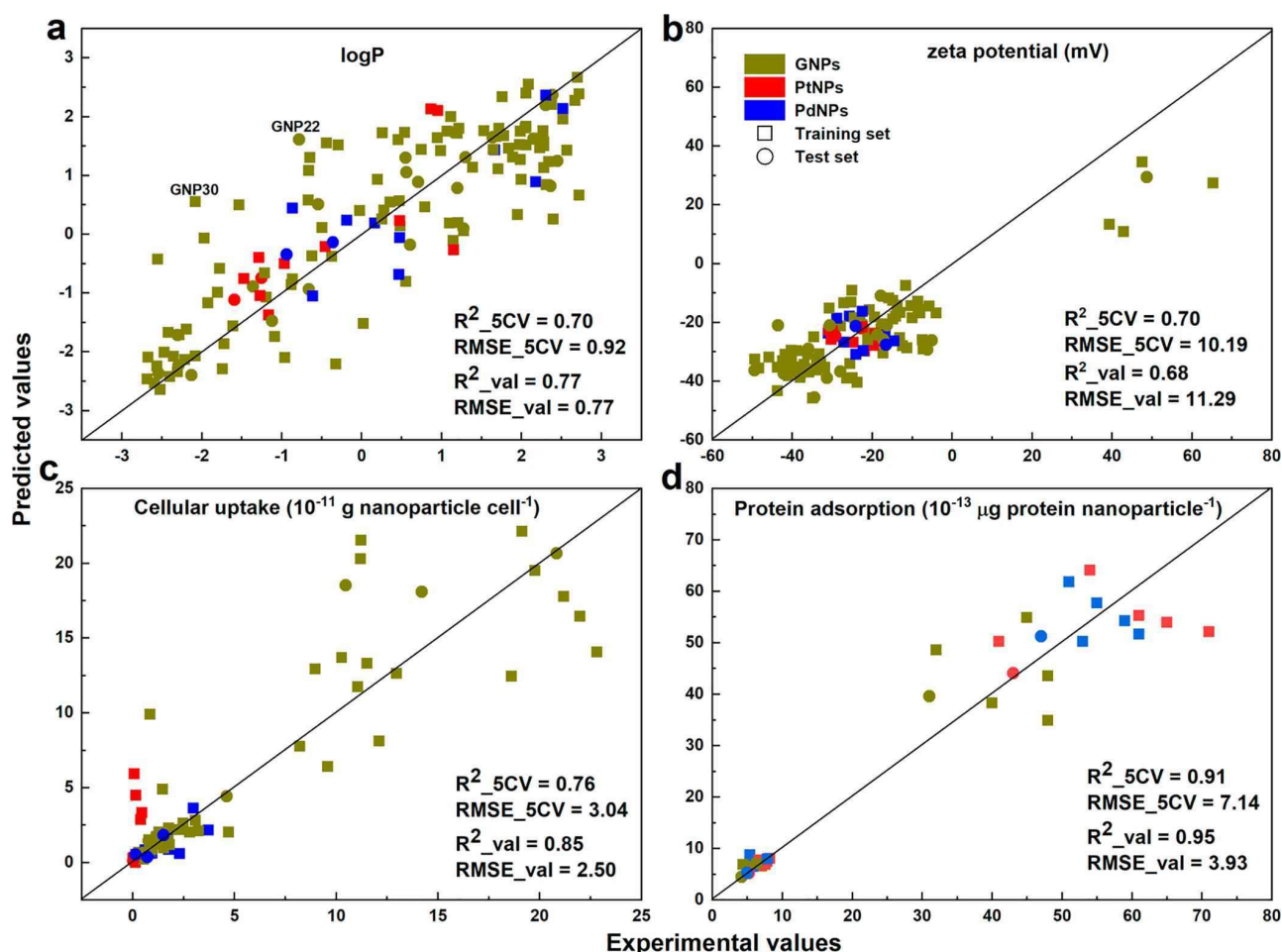


Figure 4. LeNet convolutional neural network modeling results. The correlations between the experimental values and predicted values of (a) $\log P$, (b) zeta potential, (c) cellular uptake, and (d) protein adsorption are shown. The coefficients of determination (R^2), root mean squared error (RMSE) for 5-fold cross validation (R^2_{5CV} , and $RMSE_{5CV}$), and external prediction (R^2_{val} , and $RMSE_{val}$) are also shown.

Nanoparticles in the data sets only contain regular organic ligands and lack fluorinated ligands. It is well known that fluorinated compounds have a different intermiscibility compared to normal organic compounds containing mainly C, H, N, O, and S. As a result, these two nanoparticles are structural outliers compared to other nanoparticles and cannot be handled well in the modeling process. This problem can be corrected in the future by including more fluorine-containing nanoparticles. In fact, even for models with high prediction accuracy, there is a defined domain of applicability. For example, the nanoparticle with surface ligands that is structurally different from the 91 unique ligands in the data set (Figure 2h) is likely to be predicted as an outlier. In addition, the current method cannot distinguish highly similar ligands such as chiral molecules, because the convolutional neural network cannot distinguish such two images due to rotation invariance.⁴² In the future, more advanced three-dimensional nanostructural imaging methods are needed.

In addition, we also constructed convolutional neural network models using GoogLeNet architecture (Inception v1),⁴¹ which was the winner of the ILSVRC (ImageNet Large Scale Visual Recognition Competition) in 2014. The GoogLeNet contained 22 layers and utilized inception modules that allowed the network to choose between multiple

convolutional filter sizes in each block. As shown in Figure S8, with the more advanced GoogLeNet architecture, the prediction results did not show much difference compared to that of LeNet models. The main reason is that the current data set (2646 images) is particularly small compared to a data set suitable for GoogLeNet (such as the ImageNet data with more than 14 millions). Similar results have been observed in previous studies,^{43,44} and as for small data sets, the deep learning models did not show better predictivity than traditional machine learning models.

Convolutional Neural Network Modeling Mechanism Visualized through Class Activation Maps. Although deep learning models are often referred to as “black boxes”, the nanostructure features learned by a convolutional neural network are highly amenable to visualization.^{45,46} A class activation map is a simple technique that allows us to see which regions in the image are relevant to this class.⁴⁷ Deep convolutional neural network modeling effectively acts as an information distillation pipeline, with raw data being transformed to filter out irrelevant information and magnify and refine useful information. As a result, the class activation maps of higher layers carry less information about the specific input but more information about the target. In this study, we created class activation maps from the last convolutional layer

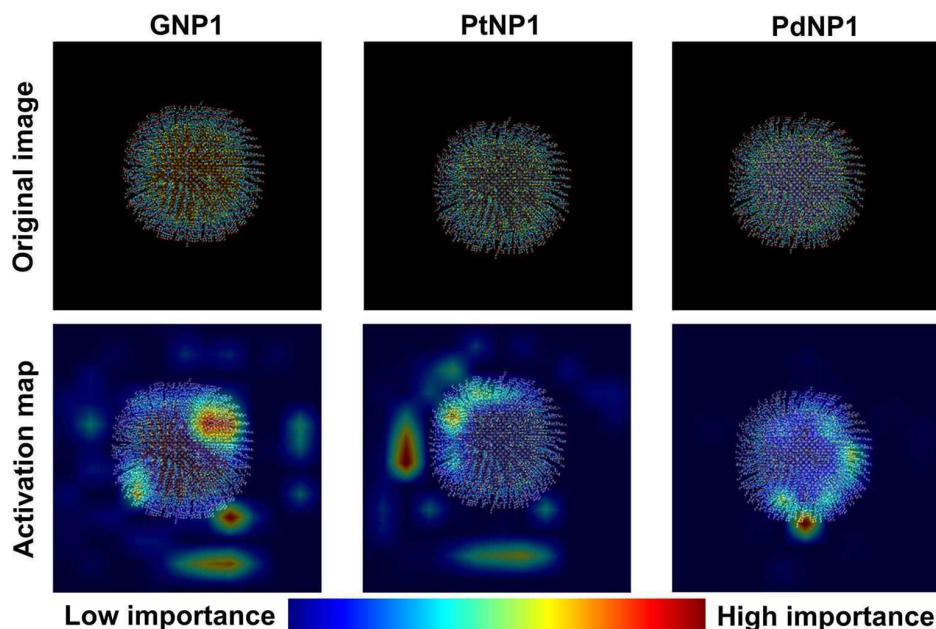


Figure 5. Class activation maps from the last convolutional layer in the convolutional neural network used for logP prediction. A class activation map is a two-dimensional grid of scores associated with a specific output value, computed for every location in any input image, indicating how important each location is with respect to the output value under consideration. The deeper the color of the map is, the more important the location of the nanostructure image is. Original images are shown at the top, and the corresponding class activation maps are shown at the bottom. For clarity, some excess background is cropped.

to visualize the specific parts of an image the convolutional neural network is looking at in order to make a decision. For example, as shown in Figure 5, the class activation maps of three different nanoparticles are generated from the last convolutional layer in the convolutional neural network used for logP prediction. The class activation maps predominantly highlight the regions of the nanostructure outer layers, indicating that surface ligands contributed the most to the convolutional neural network models for logP prediction in this study. In addition, the activation maps of nanoparticles were generated to explore the common trend of the relationship of colors to nanostructures and further to the activities. Figure S9 shows the activation maps of five hydrophobic nanoparticles with high cellular uptake potential. These gold nanoparticles had the same surface ligands but different ligand densities (Figure S9f). It can be seen that the convolutional neural network models mainly extracted the outermost portion features (i.e., aromatic and aliphatic rings) of the surface ligands, indicating that π -bond density plays an important role in hydrophobicity and cellular uptake of these nanoparticles. In fact, in our previous study,⁴⁸ we found that specific interactions between the surface ligands and cell surface receptors are responsible for the cellular uptake potential. However, machine learning and deep learning mainly focused on building relationships between nanostructures and various end points of nano–bio interactions, such as protein binding, cell surface receptor binding, cell uptake, and *in vivo* effects. Other computational modeling methods such as molecular simulation can be applied to better understand the nanobiological interaction mechanism. In addition, it should be noted that some extracted features are beyond the edge of materials. In general, a class activation map is the weighted sum of the original input and the extracted feature after the last convolutional layer in deep learning networks. In this work, the

extracted feature after the last convolution layer is the matrix with a size of 21×21 , which is much smaller than the input nanoparticle images (200×200). For calculations, the extracted features are upsampled to the same size as the input images and then added with the original images with a weight of 0.3 to 1. Due to the upsampling effect, some extracted feature is beyond the edge of these nanoparticles and occurs in the background.

CONCLUDING REMARKS

In summary, this end-to-end convolutional neural network modeling approach predicted various physicochemical properties and biological activities of a large and diverse set of nanoparticles without employing tedious nanostructure annotation and nanodescriptor calculation. This approach not only establishes quantitative nanostructure–activity relationship models and target nanoparticle prediction but also allows elucidation of such a decision-making process. The current modeling strategy based on nanostructure images and convolutional neural network can be a universal tool for the rational material design purpose. The modeling tool will guide future development of novel nanomaterials with desirable biological properties by quickly virtual screening newly designed nanomaterials.

However, it should also be noted that pitfalls still exist in the current nanostructure images and convolutional neural network modeling. First of all, compared to commonly used small molecule data sets (more than a few thousands), the nano–bio interaction data sets used in this study for modeling are still small (less than 100). In order to construct deep learning models that can be generally applicable for nanomaterial design, more high-quality data should be experimentally generated by the combination of nanocombinatorial chemistry and high-throughput screening. In addition, this approach was

only verified for sphere nanoparticles. As for nonspherical nanomaterials, the acquired nanostructure images are quite different when taking screenshots from different perspectives. In the future, more advanced three-dimensional nanostructural imaging methods are needed.

■ ASSOCIATED CONTENT

SI Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acssuschemeng.0c07453>.

Index used in PubVINAS (XLSX)

Illustration of two in-house programs used in the present study (Figure S1), atomic settings of nanostructures and screenshots of nanostructures from multiviews (Figure S2), different blocks of the LeNet convolutional neural network model (Figure S3), validation loss and training loss against epochs (Figure S4), chemical structures of three representative surface ligands (Figure S5), results of the Y-scrambling test (Figure S6), nanostructure images for two outliers (Figure S7), GoogLeNet convolutional neural network modeling results (Figure S8), class activation maps of five hydrophobic nanoparticles (Figure S9), chemical structures of 91 surface ligands (Figure S10), and experimental values of physicochemical properties and biological activities for 147 nanoparticles (Table S1) (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Bing Yan – Institute of Environmental Research at Greater Bay Area, Key Laboratory for Water Quality and Conservation of the Pearl River Delta, Ministry of Education, Guangzhou University, Guangzhou 510006, China; School of Environmental Science and Engineering, Shandong University, Jinan 250100, China; orcid.org/0000-0002-7970-6764; Email: drbingyan@yahoo.com

Hao Zhu – The Rutgers Center for Computational and Integrative Biology, Camden, New Jersey 08102, United States; Department of Chemistry, Rutgers University, Camden, New Jersey 08102, United States; Phone: (856) 225-6781; Email: hao.zhu99@rutgers.edu

Authors

Xiliang Yan – Institute of Environmental Research at Greater Bay Area, Key Laboratory for Water Quality and Conservation of the Pearl River Delta, Ministry of Education, Guangzhou University, Guangzhou 510006, China; The Rutgers Center for Computational and Integrative Biology, Camden, New Jersey 08102, United States

Jin Zhang – School of Food Science, Guizhou Medical University, Guiyang 550025, China; orcid.org/0000-0003-3074-5647

Daniel P. Russo – The Rutgers Center for Computational and Integrative Biology, Camden, New Jersey 08102, United States

Complete contact information is available at: <https://pubs.acs.org/doi/10.1021/acssuschemeng.0c07453>

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

X.Y. and B.Y. were supported by the National Natural Science Foundation of China (22036002 and 91643204), the National Key R&D Program of China (2016YFA0203103), and the introduced innovative R&D team project under the “The Pearl River Talent Recruitment Program” of Guangdong Province (2019ZT08L387). Codes for building convolutional neural network models can be accessed from <https://github.com/yanxiliang1991/CNN-modeling-for-nano-bio-interaction-prediction>. All nanostructure images can be accessed from https://drive.google.com/drive/folders/1FtPKYPz_5TgO7bh2mhYiKjVxW0QjLx3z?usp=sharing.

■ REFERENCES

- (1) Singh, A. V.; Ansari, M. H. D.; Rosenkranz, D.; Maharjan, R. S.; Krieger, F. L.; Gandhi, K.; Kanase, A.; Singh, R.; Laux, P.; Luch, A. Artificial Intelligence and Machine Learning In Computational Nanotoxicology: Unlocking and Empowering Nanomedicine. *Adv. Healthcare Mater.* **2020**, *9*, 1901862.
- (2) Singh, A. V.; Rosenkranz, D.; Ansari, M. H. D.; Singh, R.; Kanase, A.; Singh, S. P.; Johnston, B.; Tentschert, J.; Laux, P.; Luch, A. Artificial Intelligence and Machine Learning Empower Advanced Biomedical Material Design to Toxicity Prediction. *Adv. Intell. Syst.* **2020**, 2000084.
- (3) Wang, W.; Sedykh, A.; Sun, H.; Zhao, L.; Russo, D. P.; Zhou, H.; Yan, B.; Zhu, H. Predicting Nano-Bio Interactions by Integrating Nanoparticle Libraries and Quantitative Nanostructure Activity Relationship Modeling. *ACS Nano* **2017**, *11* (12), 12641–12649.
- (4) Yan, X.; Sedykh, A.; Wang, W.; Zhao, X.; Yan, B.; Zhu, H. *In Silico* Profiling Nanoparticles: Predictive Nanomodeling Using Universal Nanodescriptors and Various Machine Learning Approaches. *Nanoscale* **2019**, *11* (17), 8352–8362.
- (5) Walkey, C. D.; Olsen, J. B.; Song, F.; Liu, R.; Guo, H.; Olsen, D. W. H.; Cohen, Y.; Emili, A.; Chan, W. C. W. Protein Corona Fingerprinting Predicts the Cellular Interaction of Gold and Silver Nanoparticles. *ACS Nano* **2014**, *8* (3), 2439–2455.
- (6) Puzyn, T.; Rasulev, B.; Gajewicz, A.; Hu, X.; Dasari, T. P.; Michalkova, A.; Hwang, H. M.; Toropov, A.; Leszczynska, D.; Leszczynski, J. Using Nano-QSAR to Predict the Cytotoxicity of Metal Oxide Nanoparticles. *Nat. Nanotechnol.* **2011**, *6* (3), 175–178.
- (7) Alves, V. M.; Hwang, D.; Muratov, E.; Sokolsky-Papkov, M.; Varlamova, E.; Vinod, N.; Lim, C.; Andrade, C. H.; Tropsha, A.; Kabanov, A. Cheminformatics-Driven Discovery of Polymeric Micelle Formulations for Poorly Soluble Drugs. *Sci. Adv.* **2019**, *5* (6), No. eaav9784.
- (8) Lazarovits, J.; Sindhvani, S.; Tavares, A. J.; Zhang, Y.; Song, F.; Audet, J.; Krieger, J. R.; Syed, A. M.; Stordy, B.; Chan, W. C. W. Supervised Learning and Mass Spectrometry Predicts the in Vivo Fate of Nanomaterials. *ACS Nano* **2019**, *13* (7), 8023–8034.
- (9) Liu, R.; Rallo, R.; George, S.; Ji, Z.; Nair, S.; Nel, A. E.; Cohen, Y. Classification NanoSAR Development for Cytotoxicity of Metal Oxide Nanoparticles. *Small* **2011**, *7* (8), 1118–1126.
- (10) Liu, R.; Zhang, H. Y.; Ji, Z. X.; Rallo, R.; Xia, T.; Chang, C. H.; Nel, A.; Cohen, Y. Development of Structure–Activity Relationship for Metal Oxide Nanoparticles. *Nanoscale* **2013**, *5* (12), S644–S653.
- (11) Mikolajczyk, A.; Gajewicz, A.; Mulkiewicz, E.; Rasulev, B.; Marchelek, M.; Diak, M.; Hirano, S.; Zaleska-medynska, A.; Puzyn, T. Nano-QSAR Modeling for Ecosafe Design of Heterogeneous TiO₂-Based Nano-Photocatalysts. *Environ. Sci.: Nano* **2018**, *5* (5), 1150–1160.
- (12) Liu, R.; Rallo, R.; Weissleder, R.; Tassa, C.; Shaw, S.; Cohen, Y. Nano-SAR Development for Bioactivity of Nanoparticles with Considerations of Decision Boundaries. *Small* **2013**, *9* (9–10), 1842–1852.
- (13) Fourches, D.; Pu, D.; Tassa, C.; Weissleder, R.; Shaw, S. Y.; Mumper, R. J.; Tropsha, A. Quantitative Nanostructure - Activity Relationship Modeling. *ACS Nano* **2010**, *4* (10), S703–S712.

- (14) Ban, Z.; Yuan, P.; Yu, F.; Peng, T.; Zhou, Q.; Hu, X. Machine Learning Predicts the Functional Composition of the Protein Corona and the Cellular Recognition of Nanoparticles. *Proc. Natl. Acad. Sci. U. S. A.* **2020**, *117* (19), 10492–10499.
- (15) Ahmed, L.; Rasulev, B.; Kar, S.; Krupa, P.; Mozolewska, M. A.; Leszczynski, J. Inhibitors or Toxins? Large Library Target-Specific Screening of Fullerene-Based Nanoparticles for Drug Design Purpose. *Nanoscale* **2017**, *9* (29), 10263–10276.
- (16) Singh, A. V.; Jahnke, T.; Wang, S.; Xiao, Y.; Alapan, Y.; Kharratian, S.; Onbasli, M. C.; Kozielski, K.; David, H.; Richter, G.; Bill, J.; Laux, P.; Luch, A.; Sitti, M. Anisotropic Gold Nanostructures: Optimization via in Silico Modeling for Hyperthermia. *ACS Appl. Nano Mater.* **2018**, *1* (11), 6205–6216.
- (17) Le, T. C.; Yan, B.; Winkler, D. A. Robust Prediction of Personalized Cell Recognition from a Cancer Population by a Dual Targeting Nanoparticle Library. *Adv. Funct. Mater.* **2015**, *25* (44), 6927–6935.
- (18) Deng, J.; Guo, J.; Xue, N.; Zafeiriou, S. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. *Proc. IEEE Comput. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* **2019**, June, 4685–4694.
- (19) Lin, S.-C.; Zhang, Y.; Hsu, C.-H.; Skach, M.; Haque, M. E.; Tang, L.; Mars, J. The Architectural Implications of Autonomous Driving: Constraints and Acceleration. *ACM SIGPLAN Not* **2018**, *53* (2), 751–766.
- (20) Esteva, A.; Robicquet, A.; Ramsundar, B.; Kuleshov, V.; DePristo, M.; Chou, K.; Cui, C.; Corrado, G.; Thrun, S.; Dean, J. A Guide to Deep Learning in Healthcare. *Nat. Med.* **2019**, *25* (1), 24–29.
- (21) Cortes-Ciriano, I.; Bender, A. KekuleScope: Prediction of Cancer Cell Line Sensitivity and Compound Potency Using Convolutional Neural Networks Trained on Compound Images. *J. Cheminf.* **2019**, *11* (1), 1–16.
- (22) Torng, W.; Altman, R. B. Graph Convolutional Neural Networks for Predicting Drug-Target Interactions. *J. Chem. Inf. Model.* **2019**, *59* (10), 4131–4149.
- (23) Asilar, E.; Hemmerich, J.; Ecker, G. F. Image Based Liver Toxicity Prediction. *J. Chem. Inf. Model.* **2020**, *60* (3), 1111–1121.
- (24) Ziletti, A.; Kumar, D.; Scheffler, M.; Ghiringhelli, L. M. Insightful Classification of Crystal Structures Using Deep Learning. *Nat. Commun.* **2018**, *9* (1), 1–10.
- (25) Maddhuri Venkata Subramaniya, S. R.; Terashi, G.; Kihara, D. Protein Secondary Structure Detection in Intermediate-Resolution Cryo-EM Maps Using Deep Learning. *Nat. Methods* **2019**, *16* (9), 911–917.
- (26) Russo, D. P.; Yan, X.; Shende, S.; Huang, H.; Yan, B.; Zhu, H. Virtual Molecular Projections and Convolutional Neural Networks for End-to-End Modeling of Nanoparticle Activities and Properties. *Anal. Chem.* **2020**, *92* (20), 13971–13979.
- (27) Li, S.; Zhai, S.; Liu, Y.; Zhou, H.; Wu, J.; Jiao, Q.; Zhang, B.; Zhu, H.; Yan, B. Experimental Modulation and Computational Model of Nano-Hydrophobicity. *Biomaterials* **2015**, *52* (1), 312–317.
- (28) Wang, W.; Yan, X.; Zhao, L.; Russo, D. P.; Wang, S.; Liu, Y.; Sedykh, A.; Zhao, X.; Yan, B.; Zhu, H. Universal Nanohydrophobicity Predictions Using Virtual Nanoparticle Library. *J. Cheminf.* **2019**, *11* (1), 1–5.
- (29) Sun, H.; Liu, Y.; Bai, X.; Zhou, X.; Zhou, H.; Liu, S.; Yan, B. Induction of Oxidative Stress and Sensitization of Cancer Cells to Paclitaxel by Gold Nanoparticles with Different Charge Densities and Hydrophobicities. *J. Mater. Chem. B* **2018**, *6* (11), 1633–1639.
- (30) Bai, X.; Wang, S.; Yan, X.; Zhou, H.; Zhan, J.; Liu, S.; Sharma, V. K.; Jiang, G.; Zhu, H.; Yan, B. Regulation of Cell Uptake and Cytotoxicity by Nanoparticle Core under the Controlled Shape, Size, and Surface Chemistries. *ACS Nano* **2020**, *14* (1), 289–302.
- (31) Zhou, H.; Mu, Q.; Gao, N.; Liu, A.; Xing, Y.; Gao, S.; Zhang, Q.; Qu, G.; Chen, Y.; Liu, G.; Zhang, B.; Yan, B. A Nano-Combinatorial Library Strategy for the Discovery of Nanotubes with Reduced Protein-Binding, Cytotoxicity, and Immune Response. *Nano Lett.* **2008**, *8* (3), 859–865.
- (32) Yan, X.; Sedykh, A.; Wang, W.; Yan, B.; Zhu, H. Construction of a Web-Based Nanomaterial Database by Big Data Curation and Modeling Friendly Nanostructure Annotations. *Nat. Commun.* **2020**, *11* (1), 1–10.
- (33) Wang, W.; Kim, M. T.; Sedykh, A.; Zhu, H. Developing Enhanced Blood-Brain Barrier Permeability Models: Integrating External Bio-Assay Data in QSAR Modeling. *Pharm. Res.* **2015**, *32* (9), 3055–3065.
- (34) Erickson, H. P. Size and Shape of Protein Molecules at the Nanometer Level Determined by Sedimentation, Gel Filtration, and Electron Microscopy. *Biol. Proced. Online* **2009**, *11* (1), 32–51.
- (35) Jiang, Y.; Huo, S.; Mizuhara, T.; Das, R.; Lee, Y. W.; Hou, S.; Moyano, D. F.; Duncan, B.; Liang, X. J.; Rotello, V. M. The Interplay of Size and Surface Functionality on the Cellular Uptake of Sub-10 nm Gold Nanoparticles. *ACS Nano* **2015**, *9* (10), 9986–9993.
- (36) Huo, S.; Jin, S.; Ma, X.; Xue, X.; Yang, K.; Kumar, A.; Wang, P. C.; Zhang, J.; Hu, Z.; Liang, X. J. Ultrasmall Gold Nanoparticles as Carriers for Nucleus-Based Gene Therapy Due to Size-Dependent Nuclear Entry. *ACS Nano* **2014**, *8* (6), 5852–5862.
- (37) Huang, K.; Ma, H.; Liu, J.; Huo, S.; Kumar, A.; Wei, T.; Zhang, X.; Jin, S.; Gan, Y.; Wang, P. C.; He, S.; Zhang, X.; Liang, X. J. Size-Dependent Localization and Penetration of Ultrasmall Gold Nanoparticles in Cancer Cells, Multicellular Spheroids, and Tumors in Vivo. *ACS Nano* **2012**, *6* (5), 4483–4493.
- (38) Jin, Q.; Deng, Y.; Chen, X.; Ji, J. Rational Design of Cancer Nanomedicine for Simultaneous Stealth Surface and Enhanced Cellular Uptake. *ACS Nano* **2019**, *13* (2), 954–977.
- (39) Scanone, A. C.; Santamarina, S. C.; Heredia, D. A.; Durantini, E. N.; Durantini, A. M. Functionalized Magnetic Nanoparticles with BODIPYs for Bioimaging and Antimicrobial Therapy Applications. *ACS Appl. Bio Mater.* **2020**, *3* (2), 1061–1070.
- (40) LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86* (11), 2278–2324.
- (41) Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper with Convolutions. *Proc. IEEE Comput. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015, 07–12 June, 1–9*.
- (42) Cheng, G.; Zhou, P.; Han, J. Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing* **2016**, *54* (12), 7405–7415.
- (43) Russo, D. P.; Zorn, K. M.; Clark, A. M.; Zhu, H.; Ekins, S. Comparing Multiple Machine Learning Algorithms and Metrics for Estrogen Receptor Binding Prediction. *Mol. Pharmaceutics* **2018**, *15* (10), 4361–4370.
- (44) Rodríguez-Pérez, R.; Miyao, T.; Jasial, S.; Vogt, M.; Bajorath, J. Prediction of Compound Profiling Matrices Using Machine Learning. *ACS Omega* **2018**, *3* (4), 4713–4723.
- (45) Tang, Z.; Chuang, K. V.; DeCarli, C.; Jin, L. W.; Beckett, L.; Keiser, M. J.; Dugger, B. N. Interpretable Classification of Alzheimer's Disease Pathologies with a Convolutional Neural Network Pipeline. *Nat. Commun.* **2019**, *10* (1), 1–14.
- (46) Rudin, C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nat. Mach. Intell.* **2019**, *1* (5), 206–215.
- (47) Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-Cam: Visual Explanations from Deep Networks via Gradient-Based Localization. *Proceedings of the IEEE international conference on computer vision* **2017**, 618–626.
- (48) Zhou, H.; Jiao, P.; Yang, L.; Li, X.; Yan, B. Enhancing Cell Recognition by Scrutinizing Cell Surfaces with a Nanoparticle Array. *J. Am. Chem. Soc.* **2011**, *133* (4), 680–682.

Article

The Glutamatergic System Regulates Feather Pecking Behaviors in Laying Hens Through the Gut–Brain Axis

Xiliang Yan ¹ , Chao Wang ¹, Yaling Li ¹, Yating Lin ¹, Yinbao Wu ¹ and Yan Wang ^{1,2,3,*}
¹ Heyuan Branch, Guangdong Laboratory for Lingnan Modern Agriculture, College of Animal Science, South China Agricultural University, Guangzhou 510642, China; yanxiliang1991@163.com (X.Y.); 13355502761@163.com (C.W.); 18241674143@163.com (Y.L.); 19875216900@163.com (Y.L.); wuyinbao@scau.edu.cn (Y.W.)

² Guangdong Provincial Key Lab of Agro-Animal Genomics and Molecular Breeding, South China Agricultural University, Guangzhou 510642, China

³ National Engineering Research Center for Breeding Swine Industry, College of Animal Science, South China Agricultural University, Guangzhou 510642, China

* Correspondence: ywang@scau.edu.cn

Simple Summary: This study investigates the biological mechanism of harmful FP behavior in laying hens induced by chronic stress. We found that gut microbes like *Romboutsia* may increase the plasma arginine and histidine levels by enhancing their biosynthesis and suppressing catabolic pathways, thereby elevating glutamate levels and *GRIN2A* and *SLC17A6* gene expression in the hippocampus. These neurochemical shifts ultimately regulate the glutamatergic system to influence the FP behavior in laying hens. These findings highlight the gut–brain axis as a critical regulator of FP, offering scientific support to develop targeted strategies to mitigate FP behavior.

Abstract: Feather pecking (FP) is a significant welfare and economic problem in laying hen husbandry. While there is growing evidence that the glutamatergic system plays a crucial role in regulating FP behavior, the biological mechanisms remain unclear, largely due to the limited uptake of peripheral glutamate across the blood–brain barrier (BBB). Here, we applied a multi-omics approach combined with physiology assays to answer this question from the perspective of the gut–brain axis. A total of 108 hens were randomly assigned to two groups (treatment and control) with six replicates each, and the treatment group was subjected to chronic environmental stressors including re-housing, noise, and transport. We found that chronic exposure to environmental stressors induced severe FP, accompanied by reduced production performance and increased anxiety- and depression-related behaviors, compared to controls. In addition, the immune system was potentially disrupted in FP chickens. Notably, gut microbiota diversity and composition were significantly altered, leading to decreased microbial community stability. Non-targeted metabolomic analysis identified a variety of differential metabolites, primarily associated with arginine and histidine biosynthesis. A significant increase in glutamate levels was also observed in the hippocampus of FP chickens. Transcriptome analysis revealed the upregulated expressions of glutamate-related receptors *GRIN2A* and *SLC17A6* in the hippocampus. Correlation analysis indicated that *GRIN2A* and *SLC17A6* are positively associated with arginine levels in the duodenum, while *Romboutsia* in the duodenum is negatively correlated with arginine. These findings suggest that intestinal bacteria, including *Romboutsia*, may influence FP behavior by altering plasma arginine and histidine levels. These changes, in turn, affect glutamate levels and receptor gene expression in the hippocampus, thereby regulating the glutamatergic system. Our research offers insights into novel strategies for mitigating harmful behaviors in poultry farming, with potential benefits for animal performance and welfare.



Academic Editor: Sabine Gebhardt-Henrich

Received: 5 April 2025

Revised: 26 April 2025

Accepted: 28 April 2025

Published: 30 April 2025

Citation: Yan, X.; Wang, C.; Li, Y.; Lin, Y.; Wu, Y.; Wang, Y. The Glutamatergic System Regulates Feather Pecking Behaviors in Laying Hens Through the Gut–Brain Axis. *Animals* **2025**, *15*, 1297. <https://doi.org/10.3390/ani15091297>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: feather pecking; glutamatergic system; gut–brain axis; intestinal microbiota; animal welfare

1. Introduction

For centuries, laying hens have been used to provide a protein source for humanity. In China, the total population of laying hens in 2022 was approximately 946 million (data from <http://www.moa.gov.cn/> (accessed on 11 May 2023)), making the poultry industry one of the most large-scale, intensive, and industrialized agricultural sectors. FP is an abnormal and harmful behavior that often occurs in intensive farming of laying hens [1]. It may occur among birds of any age or any breed and can lead to skin and tissue damage, increasing the risk of infections [2]. In addition, FP increases the heat loss and feed intake of laying hens, thus negatively affecting their egg production efficiency. While pecking behavior can be effectively reduced through various measures, like beak trimming, these methods raise significant welfare concerns due to stress and pain in birds, and a loss of beak sensation and function. Therefore, understanding the biological mechanism of pecking behavior can develop more effective humane mitigating strategies [3].

Although FP has been extensively studied, its underlying biological mechanisms remain incompletely understood [4,5]. Several studies have linked FP to multiple neurotransmitters in the central nervous system [6–9], including serotonin (5-hydroxytryptamine, 5-HT), dopamine (DA), γ -aminobutyric acid (GABA), and glutamate. Glutamate, the most prevalent neurotransmitter in the brain and spinal cord, is involved in at least 90% of excitatory synapses and regulates numerous brain functions, including learning, memory, and cognition. In the hippocampus of chickens displaying intense FP, glutamate levels were found to be 1.51 times higher than in control chickens, and with significantly increased levels of its precursor, histidine [9]. The upregulated expression of genes (*GRIN1*, *GRIN2A*, and *GRIN2B*) encoding the voltage-sensitive ionotropic glutamate receptor (N-Methyl-D-Aspartic acid, NMDA) has also been observed in pecking chickens [10]. However, the precise role of the glutamatergic system in FP remains unclear due to the BBB's limited permeability to peripheral glutamate.

As a virtual endocrine organ, the gut microbiota can influence host behavior by regulating glutamatergic, GABA receptor, and the corresponding gene expression. Studies have shown that gut microbiota from patients with schizophrenia can alter the glutamate–glutamine–GABA cycle and induce schizophrenia-relevant behaviors in mice [11]. Similarly, microbiota transplantation from depressed patients can alter arginine, proline, and histidine metabolism, leading to emotionally impaired phenotypes in mice [12]. Gut microbiota can also synthesize various neurotransmitters and modulate gene expressions in the central nervous system, influencing a range of host behaviors. For example, *Lactobacillus rhamnosus* can produce GABA neurotransmitters and reduce depression and anxiety-like behavior in mice [13]. Our previous research revealed significantly higher glutamate levels in the hippocampus of FP laying hens, along with significant alterations in the cecal bacterial community [9]. Based on these findings, we hypothesized that gut microbiota may influence the central glutamatergic system by altering the metabolism of histidine and other glutamate precursor substances.

2. Materials and Methods

2.1. Experimental Instruments and Biochemical Kits

Details of experimental instruments and biochemical kits can be found in Tables S1 and S2.

2.2. Animals and Housing Conditions

A total of 108 healthy, 27-week-old Hyland Gray laying hens with similar egg production rates and body weights were purchased from a large-scale commercial market in Maoming City, Guangdong Province, China. The chickens were randomly assigned to two experimental groups ($n = 54$ per group), each consisting of six replicates with 9 chickens per replicate (cage). Birds were housed in a closed single level cage (1.5 m \times 0.6 m \times 0.5 m) under conventional commercial farm management conditions. During the experiment, artificial lighting was provided from 6:00 a.m. to 10:00 p.m., with light intensity maintained at 12–14 lx, and the temperature inside the house was kept at 25 ± 4 °C. The treatment group was exposed to chronic and unpredictable stressors, while the control group was raised under standard conditions. A numbered silicone backpack (8 cm \times 6 cm \times 0.5 cm) affixed to each chicken's back served as an individual identifier.

2.3. Stress Treatments

The experimental timeline and stress treatment workflow are depicted in Figure 1. A two-week pre-feeding period (Weeks 27 and 28) allowed all birds to acclimate to the environment. During Week 28, baseline FP behavior was assessed over four days. Beginning at 29 weeks of age, the treatment group was exposed to four weeks of chronic stressors, consisting of re-housing, noise, and transport. Re-housing involved dividing the chickens within each cage into two subgroups (4 or 5 chickens per subgroup) and mixing them with subgroups from different cages. The noise stressor consisted of 100 dB sound pressure at frequencies between 605 and 3112 Hz, delivered in four second bursts with three replicates every 2–10 min. Transport stressor involved moving each cage back and forth for two minutes.

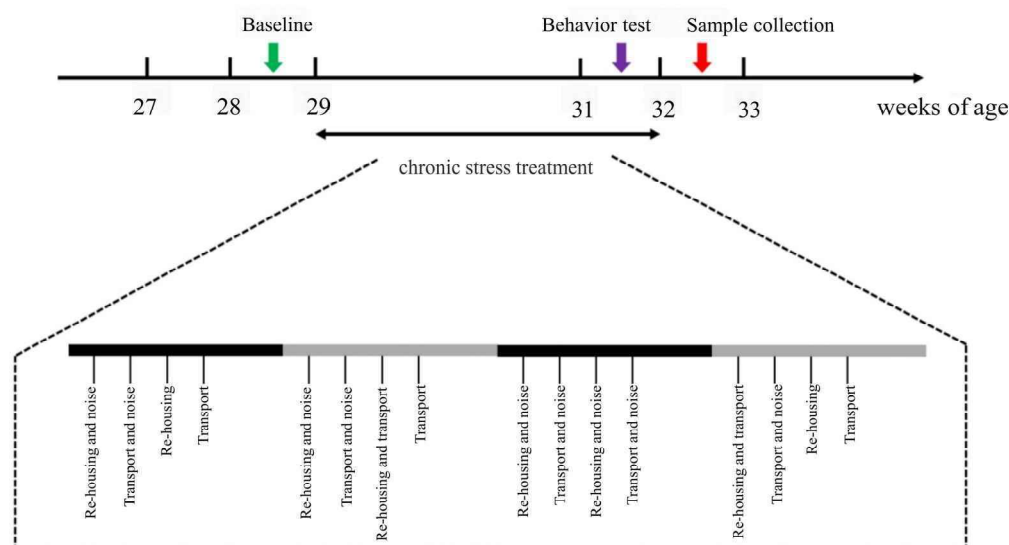


Figure 1. A schematic diagram designed for chronic stress treatment. During Weeks 29–32, the treatment group of laying hens were subjected to chronic and unpredicted stressors including re-housing, noise, and transport. Behavior assessments during Week 28 were assigned as the baseline of FP. The open field (OF) test was performed at Week 31, and blood and tissue samples were collected for further analysis after stress treatments.

2.4. Behavioral Assessments

During the stress treatment period (Weeks 29–32), behavioral observations were recorded five times per week for 20 min (10 min at 9:00 a.m. and 10 min at 3:00 p.m.). FP behavior was defined as continuous pecks directed toward the same body part of the same chicken, and categorized as severe FP or gentle FP. Gentle FP was further subdivided into

exploratory FP and stereotyped FP. Details about the different types of FP can be found in Table S3. For each observation period, the initiator and receiver of FP interactions, the number of pecks, and the type of peck were recorded. A “feather pecker” was defined as a chicken initiating an aggressive peck, a “feather victim” as a chicken receiving an aggressive peck, and a “neutral chicken” as a chicken neither initiating nor receiving aggressive pecks. Two trained observers conducted all behavioral observations.

2.5. Sample Collection

At 32 weeks of age, 24 chickens were selected for biological sample collection: 12 exhibiting the most severe FP behavior and 12 neutral chickens from the control group. Blood samples were collected from the brachial wing veins, placed on ice for one hour, and centrifuged ($860 \times g$, $4\text{ }^{\circ}\text{C}$, 10 min) to prepare plasma. Following plasma collection, the selected chickens were euthanized by cervical dislocation. Ileal, cecal, and duodenal contents were collected. The brain was rapidly dissected, and the hippocampus and amygdala were collected. Plasma, brain tissue, and intestinal contents were immediately transferred to cryotubes and stored at $-80\text{ }^{\circ}\text{C}$ for subsequent analysis.

2.6. Quantification of Immune Response Markers

Enzyme-linked immunosorbent assay (ELISA) kits were used to measure immune response markers, including immunoglobulin A (IgA), immunoglobulin G (IgG), immunoglobulin M (IgM), interleukin 1 (IL-1), interleukin 6 (IL-6), tumor necrosis factor α (TNF- α), corticosterone (CORT), epinephrine (EPI), and norepinephrine (NE). Optical density was measured at 450 nm using Nessler’s reagent spectrophotometry (Thermo, MA, America). As an advanced immunological technique, the ELISA kits were commonly used to measure proteins, antibodies, antigens, and hormones in biological samples, including chicken plasma. Examples include the determination of cytokine levels (CORT and IL-1 β) [14] and immune parameters (IgG and IgM) [15] in chicken serum.

2.7. The 16S rRNA Gene Quantification

The 16S rRNA gene quantification comprised three steps: DNA isolation, library preparation, and sequencing. Microbial DNA was extracted from duodenal, ileal, and cecal samples using the QIAamp PowerFecal DNA Kit (QIAGEN, Hilden, Germany), following the manufacturer’s protocol. DNA quality was assessed by measuring purity and concentration using an ultra-micro spectrophotometer. DNA was diluted to $1\text{ ng}/\mu\text{L}$. Using diluted DNA as a template, the 16S rRNA gene was amplified with barcoded primers (341F: 5′-CCTAYGGGRBGCASCAG-3′; 806R: 5′-GGACTACNNGGGTATCTAAT-3′) flanking the V3-V4 hypervariable regions. PCR products were assessed by 2% agarose gel electrophoresis and pooled in equal proportions based on PCR concentrations. DNA libraries were constructed using the NEBNext® UltraTM II DNA Library Prep Kit (Illumina, San Diego, CA, USA) and quantified by Qubit and Q-PCR. The final library was sequenced on a NovaSeq6000 system (Illumina, San Diego, CA, USA). PCR amplification and 16S rRNA gene sequencing were performed by Beijing Novel Biosciences Co., Ltd. (Beijing, China).

2.8. Eukaryotic Transcriptome Sequencing

Total RNA was purified from the hippocampus and amygdala tissues using the RNeasy Pure Tissue Kit (TIANGEN, Beijing, China). RNA concentration was measured using an ultra-micro spectrophotometer, and samples with a concentration $> 100\text{ ng}/\mu\text{L}$, $\text{OD}_{260/280} \geq 1.5$, and $\text{OD}_{260/230} \geq 1.5$ were retained. rRNA was removed using the Ribo-zero kit (Illumina, San Diego, CA, USA). Libraries were generated using the Illumina TruseqTM RNA sample prep Kit (Illumina, San Diego, CA, USA). PolyA-tailed eukaryotic mRNA was enriched, fragmented by sonication, and used as a template for first-strand

cDNA synthesis with M-MuLV reverse transcriptase. The RNA strand was degraded with RNase H, followed by second-strand cDNA synthesis with DNA polymerase I using dNTPs as substrate. Double-stranded cDNA was ligated to adapter sequences. The cDNA was amplified by PCR and purified with AMPure XP beads (Beckman Coulter, Brea, CA, USA). The final products were sequenced and analyzed by Shanghai BioZero Biotech Co., Ltd. (Shanghai, China).

2.9. Statistical Analysis

SPSS 26 was used for independent samples *t*-test. Statistical significance was defined as $p < 0.05$. Data are presented as mean \pm SE (standard error). GraphPad Prism 8.0 was used for data visualization (FP behavior, production performance, egg quality, immune and stress response markers). The processing details of omics data are described in Methods S1–S3. Correlation analysis and heatmap generation were performed using R (Version 4.1.1). Network visualization and analysis were performed using Gephi 0.10.1 and Cytoscape 3.8.0.

3. Results

3.1. Stress-Induced FP Initiates a Cascade of Adverse Physiological and Behavioral Changes in Laying Hens

As expected, there was no significant difference in the frequency of severe FP or gentle FP behaviors between the two groups of 28-week-old laying hens before stress treatment began. However, the two groups exhibited distinct FP behaviors during Weeks 29–32. As shown in Figure 2A, the frequency of severe FP behaviors in the stress treatment group was significantly higher than in the control group ($p < 0.05$) and showed a steady increase over time. At Week 32, there was a four-fold difference in the number of birds exhibiting severe FP behavior between the two groups, with average threat frequencies per bird of 0.059 and 0.015 pecks/min for the stress treatment and control groups, respectively. Additionally, the stress treatment group displayed more frequent exploratory FP behaviors than the control group, particularly during Weeks 30–31 ($p < 0.05$) (Figure 2B). Interestingly, the control group tended to exhibit more stereotyped FP behaviors than the stress treatment group (Figure 2C). This form of FP typically does not result in significant feather damage and is considered normal investigatory behavior. These results indicate that stressed birds primarily exhibited severe FP behaviors, causing serious injury or even death to the recipients.

Figure 2D–G show the behavioral performance of the two groups in the open field (OF) test (Method S4). The vocalization latency of laying hens in the stress treatment group was significantly higher compared with the control group ($p < 0.05$). However, there was no significant difference in the number of vocalizations. Furthermore, the stress-treated birds took significantly longer to ambulate ($p < 0.05$) and the number of steps was significantly reduced ($p < 0.05$), indicating lower activity, compared with the control group. These results demonstrate that laying hens subjected to chronic and unpredictable stressors exhibit higher levels of fear and depression.

Production performance and egg quality are important economic traits in laying hens (Method S5). These indicators, including daily feed intake, egg weight, egg shape index, eggshell thickness, etc., are summarized in Figures S1 and S2. No significant differences were observed between the two groups in average daily feed intake, average egg weight, or feed conversion ratio. However, the egg-laying rate of the stress treatment group was significantly lower than that of the control group ($p < 0.05$), decreasing by 5.52% and 7.57% at Weeks 30 and 32, respectively. Moreover, the total egg weight of the stress treatment group also significantly decreased by 6.93% at Week 32 ($p < 0.05$). Chronic and unpredictable

stress decreased egg shape index ($p < 0.05$), eggshell thickness ($p < 0.01$), and eggshell strength ($p < 0.05$). Albumen height and Haugh unit did not differ between the groups. Overall, FP behaviors adversely affected the production performance and egg quality of the laying hens.

The effect of FP behaviors on the immune system was further investigated. As shown in Table S4, the levels of proinflammatory cytokines (IL-1, IL-6, and TNF- α) and immunoglobulins (IgA, IgG, and IgM) were significantly reduced in the stress-treated laying hens at the end of the four-week experimental period ($p < 0.05$). These indicators are critical for initiating the inflammatory response and maintaining the immune system. Furthermore, a significant increase in catecholamines (EPI and NE) was observed in the stress treatment group ($p < 0.05$). Extreme levels of individual catecholamines contribute to numerous adverse effects, such as anxiety, fatigue, and depression. These results suggest that long-term stress causes a potential perturbation in the immune system of FP chickens.

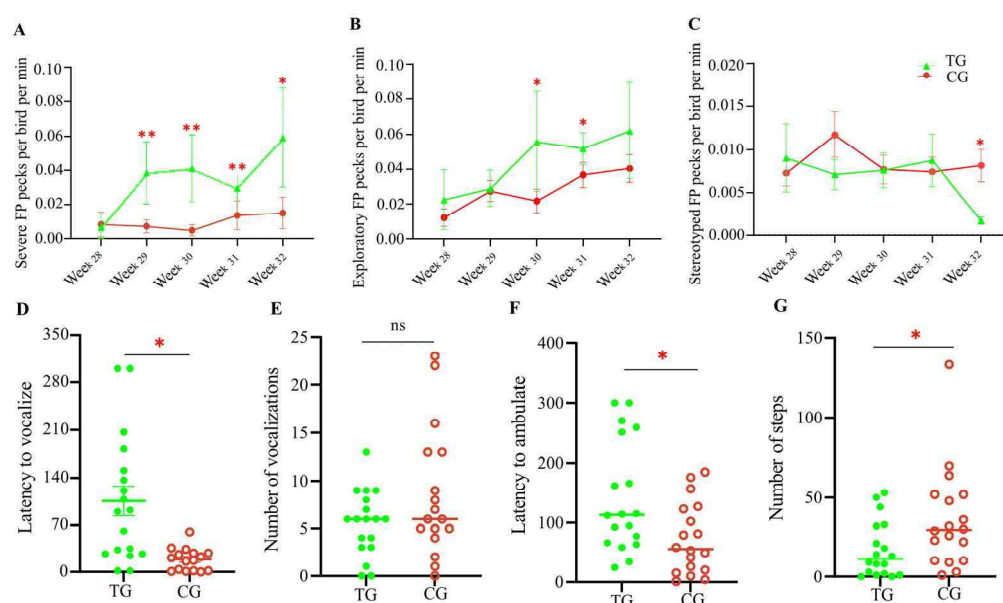


Figure 2. Stress-induced behavior changes in laying hens. The pecking frequency of severe FP (A), exploratory FP (B), and stereotyped FP (C) behaviors between two groups of laying hens during four-week stress treatments are shown. The behaviors of laying hens in OF tests in terms of vocalization (D,E) and ambulating (F,G). TG, treatment group; CG, control group. * $p < 0.05$; ** $p < 0.01$. Each group in the pecking frequency test consisted of six chickens; each group in the OF test consisted of eighteen chickens. ns means no statistical difference.

3.2. Gut Microbiota Diversity and Composition of Laying Hens Are Altered by Stress-Induced FP

The gut microbiota in chickens is closely related to their health status and production performance. To further understand the relationship between gut microbiota and FP in laying hens, 16S rRNA gene sequencing was used to evaluate microbiota changes between the two groups. Figure 3A,B depict the alpha diversity of the microbiota in the cecum, duodenum, and ileum. Chao1 and Shannon indices were used to quantify species richness and diversity, respectively. Species richness in the cecum and ileum of feather peckers was significantly lower than in neutral chickens ($p < 0.05$). Species diversity was also significantly reduced in the cecum and ileum of feather-pecking chickens ($p < 0.05$). The alpha diversity of gut microbiota in the duodenum did not differ significantly between the two groups. To determine whether FP was associated with altered microbiota composition, beta diversity analysis was performed. As shown in Figure 3C–E, the principal coordinate analysis (PCoA) of the microbiome beta diversity further demonstrated that cecal and

ileal microbiota composition differed significantly between chickens with FP and neutral chickens ($p < 0.05$). There was no significant difference in duodenal microbiota beta diversity between the groups. These results indicate that stress-induced FP significantly alters gut microbiota diversity and composition in laying hens.

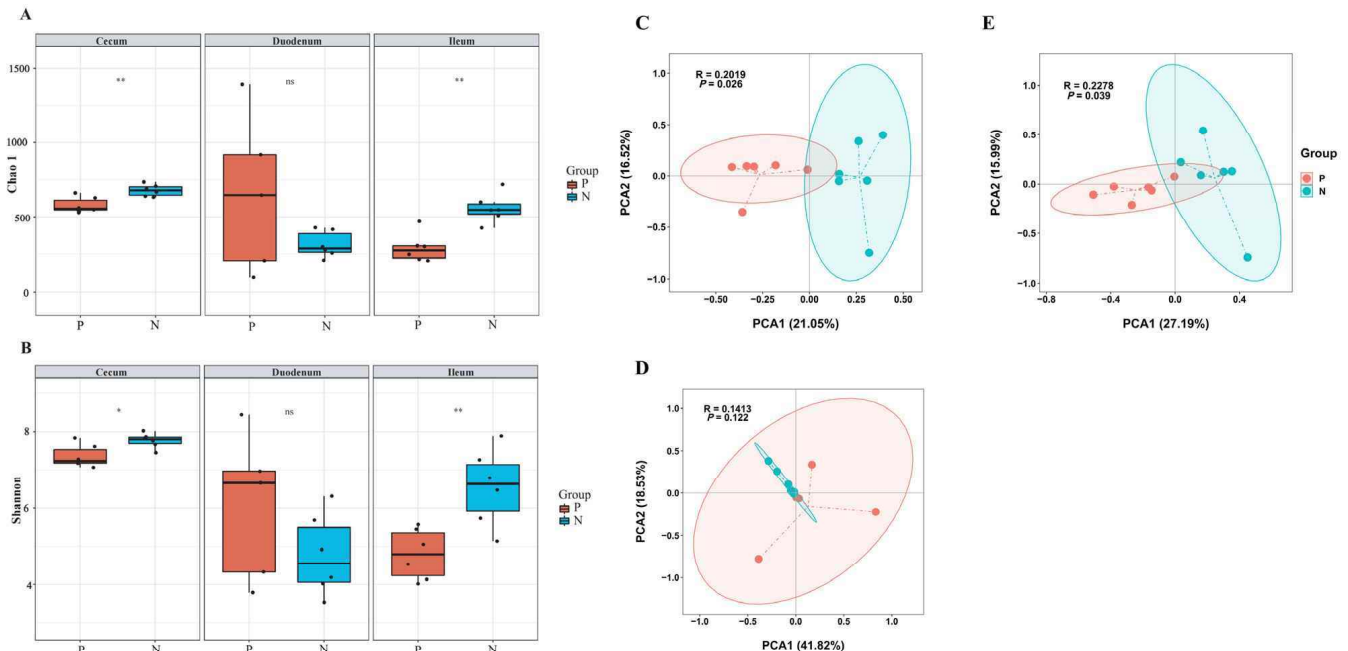


Figure 3. Diversity analysis of gut microbiota. The alpha diversity of gut microbiota is based on the Chao1 index (A) and the Shannon index (B). The beta diversity of gut microbiota in the Cecum (C), Duodenum (D), and Ileum (E) is based on principal coordinate analysis. P, pecker; N, neutral. Each group consisted of six chickens. ns means no statistical difference. * means $p < 0.05$ and ** represents $p < 0.01$.

The relative abundance of gut microbiota at the phylum level is shown in Figure S3. Three major bacterial phyla were identified in the cecum (Figure S3A): *Firmicutes* (~55.7%), *Bacteroidota* (~32.7%), and *Actinobacteriota* (~6.47%). In the duodenum (Figure S3B), the dominant phyla were *Firmicutes* (~73.3%), *Campilobacterota* (~11.9%), and *Proteobacteria* (~5.48%). In the ileum (Figure S3C), the dominant phyla were *Firmicutes* (~82.1%), *Bacteroidota* (~9.24%), and *Actinobacteriota* (~5.13%). Figure S4 shows the relative abundance of gut microbiota at the genus level. In the cecum (Figure S4A), the dominant genera were *Lactobacillus* (~23.3%), *Bacteroides* (~17.9%), and *Olsenella* (~5.71%). In the duodenum (Figure S4B), the dominant genera were *Lactobacillus* (~60.9%), *Helicobacter* (~12.2%), and *Aeriscardovia* (~3.27%). In the ileum (Figure S4C), the dominant genera were *Lactobacillus* (~52.1%), *Romboutsia* (~15.2%), and *Bacteroides* (~5.70%). Linear discriminant analysis effect size (LEfSe) was used to identify differential gut microbiota between the groups. At the genus level, *Lactobacillus* and *Enterococcus* were enriched in the cecum of FP chickens, while the *Olsenella* and *Ruminococcus torques* group were enriched in neutral chickens (Figure 4). In the duodenum, *Sphingomonas* and *Pseudomonas* characterized FP chickens, while *Helicobacter* and *Romboutsia* were identified in neutral chickens. In the ileum, *Enterococcus* and *Staphylococcus* were enriched in FP chickens, while *Bacteroides* and *Olsenella* were enriched in neutral chickens. These results demonstrate that FP is characterized by disturbed gut microbiota.

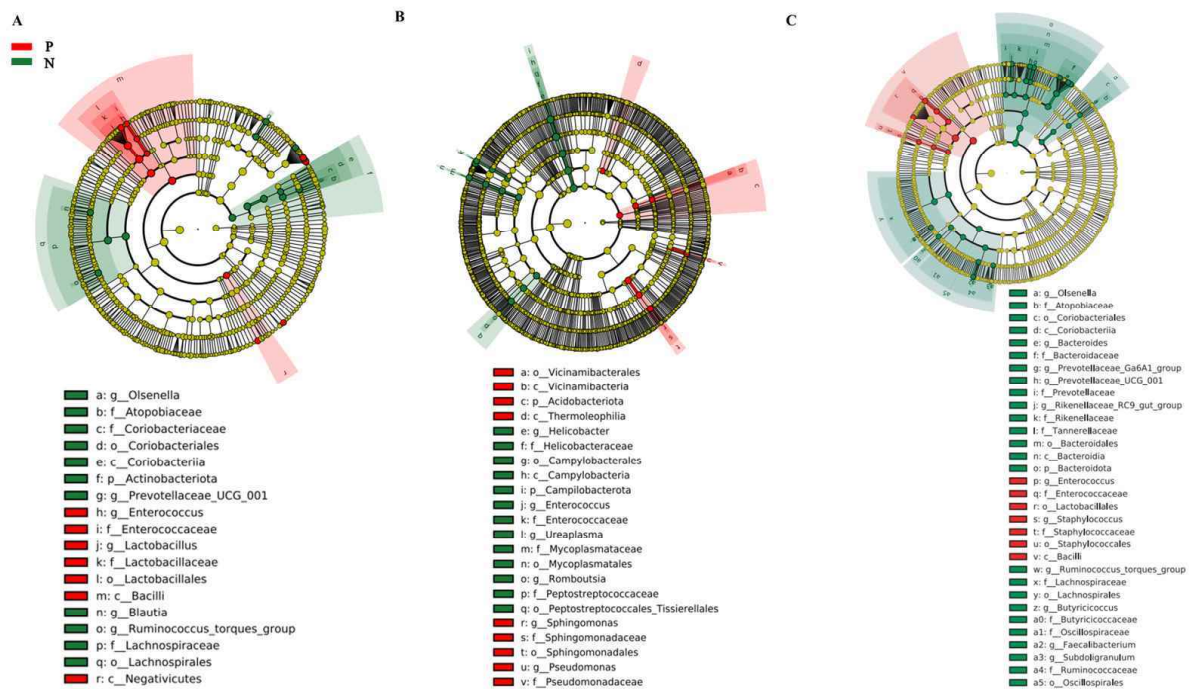


Figure 4. Identification of differential gut microbiota. LEfSe analysis of gut microbiota in the cecum (A), duodenum (B), and ileum (C). Each analysis involved six chickens.

As described above, cecal and duodenal microbiota compositions differed significantly between the two groups. To investigate gut microbiota stability, Spearman's correlation coefficient was used to visualize the co-occurrence network of core bacteria (Figure S5; $|R_{\text{spearman}}| > 0.6, p < 0.05$). The network structure differed distinctly between the groups. In the cecum, the FP group had 689 links and 159 nodes, while the control group had 917 links and 187 nodes. Similarly, in the duodenum, the FP group had fewer links and nodes than the control group (decreases of 259 and 16, respectively). These results indicate decreased gut microbiota stability in FP chickens. PICRUSt2 was used to predict gut microbiota functionality. PICRUSt2 analysis revealed 377, 425, and 402 predicted metabolic pathways from cecal, duodenal, and ileal microbiota gene expression, respectively (Figure S6). Compared with neutral chickens, several gene expressions related to amino acid biosynthesis were upregulated in FP chickens. For example, arginine biosynthesis II and L-histidine biosynthesis were significantly enhanced in the cecum of FP chickens. Arginine biosynthesis I and arginine biosynthesis IV were significantly activated in the duodenum. In the ileum, histidine biosynthesis, arginine biosynthesis III, the superpathway of arginine and polyamines, and other pathways were upregulated.

3.3. A Variety of Distinct Metabolites Are Identified in the FP Group

The gut microbiota plays a significant role in regulating host metabolism, and its disruption can lead to a wide range of metabolic diseases. Therefore, we investigated how FP behaviors perturb microbial metabolism in laying hens. Non-targeted metabolomics, a global and comprehensive analysis method, allowed us to identify a range of metabolic features in these organisms (Method S6 and Figure S7). In total, 724, 572, 599, and 389 metabolite features in positive ion mode and 303, 320, 288, and 219 in negative ion mode were extracted from peripheral blood plasma, cecum, duodenum, and ileum, respectively. Additionally, 758 and 639 metabolites were identified from hippocampus and amygdala samples, respectively. To determine metabolic differences between the two groups, orthogonal partial least squares–discriminant analysis (OPLS-DA) was performed. Metabolites were considered differential if their variable importance in projection (VIP)

was not less than 1.0 and the p -value of the t -test was less than 0.05. Using these criteria, 121 and 28 differential metabolites were detected in hippocampus and amygdala samples, respectively. Of these, 101 and 48 metabolites showed higher levels in FP chickens. Univariate and multivariate analysis also revealed 147 and 81 significantly altered metabolites in the intestine and peripheral blood plasma of FP chickens, respectively. Details of these differential metabolites are provided in Table S5.

The identified differential metabolites were mapped to KEGG metabolic pathways for functional enrichment analysis. As shown in Figure 5, a total of 20 and 10 significantly enriched metabolic pathways were identified in the hippocampus and amygdala, respectively, including “arachidonic acid metabolism”, “tyrosine metabolism”, “steroid hormone biosynthesis”, “arginine biosynthesis”, and “alanine, aspartate and glutamate metabolism”. There were 5, 11, 19, and 22 significantly enriched metabolic pathways in the cecum, duodenum, ileum, and peripheral blood plasma, respectively, related to “arginine biosynthesis”, “arachidonic acid metabolism”, “arginine”, “proline metabolism”, “pyrimidine metabolism”, and others (Figure S8).

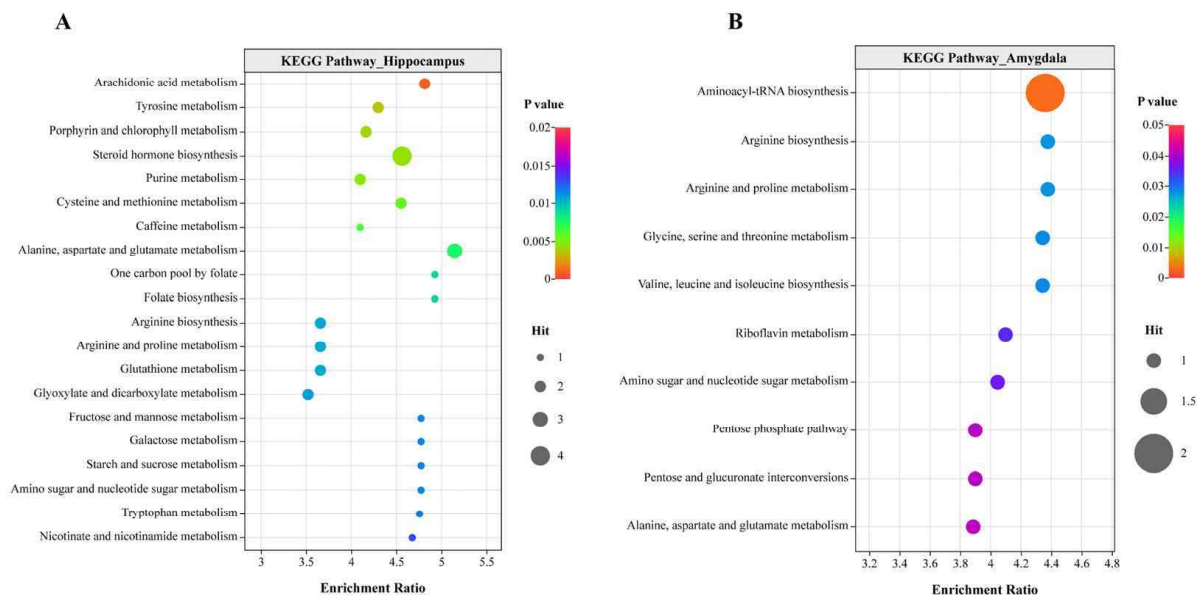


Figure 5. KEGG pathway enrichment analysis. Pathway enrichment of differentially accumulated metabolites in the hippocampus (A) and amygdala (B). The bubble size represents the number of metabolites; the color bar represents the corrected p value. Each analysis involved six chickens. Permission was obtained from Kanehisa Laboratories to use the KEGG pathway database [16].

To further explore the relationship between the peripheral (plasma/intestine) and central (brain) glutamatergic systems, the metabolic features of glutamate and its precursors (histidine, arginine, proline, and ornithine) were analyzed. L-glutamate and its important synthetic precursors (L-histidine and L-ornithine) were significantly increased in the hippocampus of FP chickens ($p < 0.05$; Figure S9). Glutamate levels were not statistically different in the amygdala, but DL-arginine levels were significantly increased ($p < 0.05$; Figure S10). Arginine (L-arginine and DL-arginine) and N-acetylornithine (a glutamate transformation product) levels were significantly increased in the plasma of FP chickens ($p < 0.05$) (Figure S11). In the cecum and duodenum, significant increases in glutamate metabolite (N-acetylglutamic acid and N-acetyl-DL-glutamic acid) and precursor (L-arginine) levels were observed, respectively ($p < 0.05$; Figures S12 and S13). There was no significant difference in the levels of substances related to glutamate synthesis and

metabolism in the ileum of laying hens between the two groups (Figure S14). These results indicate the dysregulation of the glutamatergic system in FP chickens.

3.4. Transcriptome Profiling of Hippocampus and Amygdala Exhibits Distinct Gene Expression in FP Chickens

Transcriptome sequencing was used to explore gene expression differences affected by FP behaviors. Genes with a false discovery rate (FDR) ≤ 0.05 and fold change (FC) ≥ 1.5 were considered significantly differentially expressed. Using these criteria, 666 and 445 significantly differentially expressed genes were identified in the hippocampus and amygdala, respectively. Among them, 475 genes were significantly upregulated, and 636 genes were significantly downregulated (Table S6). Gene Ontology (GO) analysis was performed to understand the biological functions of these differentially expressed genes. GO analysis identifies associations between differentially expressed genes and specific cellular components, molecular functions, or biological processes. Fisher's exact test with FDR correction for multiple testing was used. Figure S15 shows the top 30 enriched GO terms for significantly downregulated and upregulated genes in the hippocampus. These GO terms were mainly associated with biological processes. Downregulated genes were enriched in 22 biological processes, including the regulation of leukocyte-mediated immunity, the negative regulation of immune response, the sensory perception of sound, and the regulation of neurotransmitter levels. Upregulated genes were mainly enriched in biological processes such as feeding behavior, reproductive behavior, synaptic signaling, and the G protein-coupled receptor signaling pathway. As shown in Figure S16, downregulated amygdala genes were significantly enriched in GO, such as ion transport, regulation of natural killer cell-mediated immunity, and signaling receptor activity, while upregulated genes were mainly associated with behavior, the regulation of trans-synaptic signaling, and the G protein-coupled receptor signaling pathway.

KEGG pathway enrichment analysis was performed to infer gene functions mapped to biological pathways in the KEGG database. As shown in Figure 6A, hippocampal differential genes were significantly enriched in 12 KEGG pathways, including "Neuroactive ligand-receptor interaction, (ko04080)", "cAMP signaling pathway (ko04024)", "Cholinergic synapse (ko04725)", and "Calcium signaling pathway (ko04020)". To explore the expression of genes related to the glutamatergic system, the KEGG pathways "Arginine and proline metabolism (ko00330, $p = 0.134$)", "histidine metabolism (ko00340, $p = 0.252$)", "Glutamatergic synapse (ko04724, $p = 0.093$)", and "GABAergic synapse (ko04727, $p = 0.087$)" were selected for further analysis. Four genes (*AGMAT*, *ALGH1A3*, *CARNS1*, and *P4HA3*) enriched in "arginine and proline metabolism" were significantly downregulated (Figure S17A). Two genes (*ALDH1A3* and *CARNS1*) enriched in "histidine metabolism" were also significantly downregulated (Figure S17A). Notably, genes enriched in the glutamatergic synapse, such as *GRIN2A* (encoding glutamate ionotropic receptor NMDA subunit 2A), *SLC17A6* (encoding vesicular glutamate transporter 2), and *KCNJ3* (encoding protein activates inwardly rectifying potassium channels 1), were significantly upregulated ($p < 0.05$). Amygdala differential genes were involved in nine KEGG pathways (Figure 6B), including "Neuroactive ligand-receptor interaction (ko04080)", "Circadian rhythm (ko04710)", "Cholinergic synapse (ko04725)", and "Taste transduction (ko04742)". Two genes related to the glutamatergic system, *ADCY1* (encoding adenylyl cyclase 1) and *GATM* (encoding glycine amidinyltransferase), were significantly downregulated ($p < 0.05$; Figure S17B).

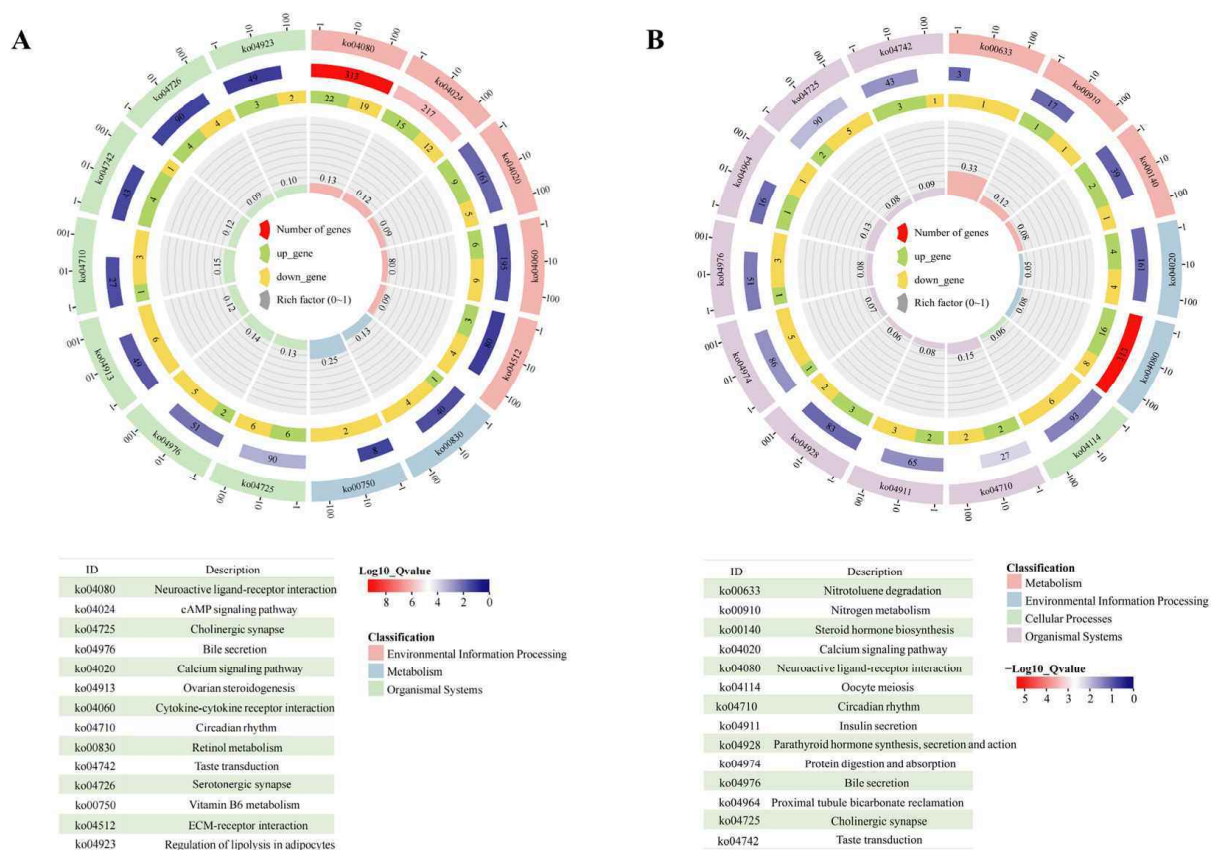


Figure 6. Pathway enrichment of differentially accumulated genes. KEGG cluster plot of differentially accumulated genes in the hippocampus (A) and amygdala (B). Each analysis involved six chickens. Permission has been obtained from Kanehisa Laboratories to use the KEGG pathway database [16].

3.5. Differential Gut Microbiomes Are Associated with Metabolites and Gene Expression Changes

The relationships between differential gut microbiota and metabolites were investigated using Spearman's rank correlation analysis, with a selection criterion of $|\text{Rspearman}| \geq 0.6$ and $p < 0.05$. As visualized in the correlation network (Figures S18–S20), there was a significant negative correlation between *Olsenella* and cortisol, N-acetylglutamate, and N-acetyl-DL-glutamate in the cecum. *Enterococcus* also exhibited a significant negative correlation with N-acetylglutamate and N-acetyl-DL-glutamate. In the duodenum, a significant negative correlation was observed between *Romboutsia* and L-arginine, N-acetyl-L-methionine, N-acetylvaline, L-epinephrine, and N-acetylalanine. In the ileum, a significant negative correlation was found between *Olsenella*, *Prevotellaceae_UCG_001*, *Prevotellaceae_Ga61_group*, and spermidine, 4-guanidinobutyric acid.

To understand how gut-derived metabolites influence hippocampal gene expression, Spearman's rank correlation was used to analyze the relationship between differential gut metabolites and glutamate-related genes. As shown in Figure 7, no significant correlation was observed between *SLC17A6* and differential cecal metabolites, but a significant positive correlation was observed between *GRIN2A* and N-acetyl-DL-glutamate ($p < 0.05$). In the duodenum (Figure S21), both *GRIN2A* and *SLC17A6* exhibited significant positive correlations with arginine and glycine-valine ($p < 0.05$). In the ileum (Figure S22), *SLC17A6* showed significant positive correlations with riboflavin and prostaglandin E1.

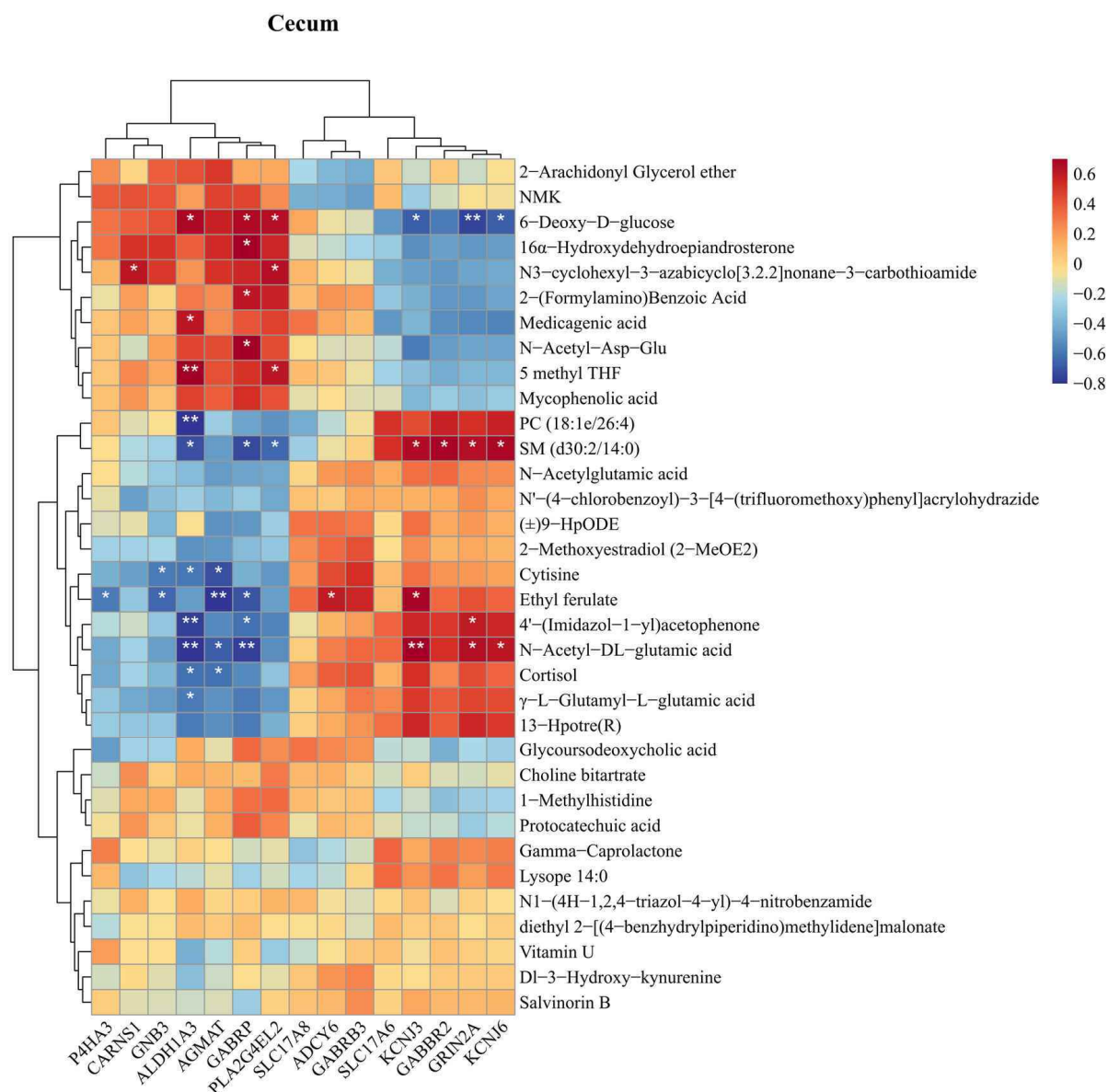


Figure 7. Correlation analysis of intestinal differential metabolites and hippocampal differential genes. Spearman correlation coefficients between the hippocampal differential genes and the differential metabolites in the cecum. * $p < 0.05$, ** $p < 0.01$.

4. Discussion

4.1. The Hippocampal Glutamatergic System Affects FP in Laying Hens

In this work, we constructed an FP model in which the laying hens were subjected to chronic and unpredicted stress. Multi-omics analyses, including non-targeted metabolomics and eukaryotic transcriptome sequencing, demonstrated that the hippocampal glutamatergic system played a significant role in the development of FP. We observed that the levels of glutamate neurotransmitter and its precursors (i.e., histidine and ornithine) were significantly increased in the hippocampus of FP chickens. The expression of *GRIN2A* (encoding ionotropic glutamate receptor NMDA subunit 2A) and *SLC17A6* (encoding vesicular glutamate transporter 2) was also significantly upregulated in the hippocampus of FP chickens (Figure S17). Increasing evidence suggests that FP behavior is associated with depression and fear traits, and the glutamatergic system plays a vital regulatory role in the pathophysiology of depression and fear. The OF test results indicated that FP laying hens exhibited

higher levels of depression and fear, manifested by a significant increase in latency to vocalize and ambulate and a reduction in walking steps (Figure 2).

Glutamate, an important excitatory neurotransmitter in the central nervous system, is widely distributed and highly concentrated in the hippocampus, cerebral cortex, and thalamus [17,18]. It plays a major role in shaping memory and learning and is closely associated with various neurological diseases such as depression and schizophrenia. Our present study found significantly high glutamate levels in the hippocampus of FP chickens, confirming our previous findings [9]. Most previous studies on FP have focused on the potential role of alteration in the central serotonergic and dopaminergic systems. They found that FP genotype and phenotype affected the serotonin metabolism of laying hens; low serotonin level was a possible reason for FP behavior of baby chicks, while high serotonin levels were observed in adult hens [7]. Dopamine metabolism in the central nervous system also differs significantly between low-FP and high-FP laying hens and is age-related [8,19]. In this study, serotonin (or dopamine) levels and related compounds did not differ significantly between pecking and neutral chickens, which may be related to the specific breeds used and indirectly indicates the complexity of the pecking mechanism.

Central neurons express various glutamate receptors, which are categorized into two main types: voltage-sensitive ionotropic glutamate receptors and ligand-sensitive metabotropic glutamate receptors. Glutamate receptors mediate excitatory synaptic transmission and are important for the maintenance of various functions such as learning and memory [20]. In this study, the expression of *GRIN2A* (NMDA receptor subunit 2A) and *SLC17A6* (encoding vesicular glutamate transporter 2) was significantly upregulated in the hippocampus of FP chickens. Previous studies also have observed significantly increased *GRIN2A* levels in the hippocampus of mice with depressive-like behavior [21]. Vesicular glutamate transporter 2 can activate the septohippocampal system, which is critical for learning adverse events and can lead to mood disorders such as aggression, aversion, and depression-related anhedonia [22].

In conclusion, the altered glutamatergic systems, such as the release of glutamate neurotransmitters and the upregulated expression of corresponding receptors, may be an important neuroregulatory mechanism of FP behavior in laying hens.

4.2. Gut Microbiota Regulates the Hippocampal Glutamatergic System Influencing FP Behavior in Laying Hens

The gut microbiota, sometimes referred to as the “second brain”, is intimately involved in various physiological processes such as host nutrient metabolism, and regulates the central nervous system, thus affecting host behaviors [23,24]. In the current study, we found that the composition of gut microbiota was significantly altered in FP chickens (Figure 4). Specifically, the relative abundance of *Sphingomonas* and *Pseudomonas* was significantly increased in the duodenum, and the abundance of bacterial genera such as *Romboutsia* and *Enterococcus* was significantly decreased ($p < 0.05$). The PICRUSt2-based functional prediction of bacterial community demonstrated that the gene expression of arginine and histidine biosynthesis was significantly increased (Figure S6). We found that the level of arginine in the FP chickens was significantly higher than that of neutral chickens, and there was a significant negative correlation between *Romboutsia* and arginine in the duodenum (Figure S19). High levels of *Romboutsia* have been observed in the feces of stress-resilient mice [25]. Arginine supplementation in the diet of birds significantly increased the relative abundance of ileal *Romboutsia*, indicating that the proliferation of gut *Romboutsia* may be determined by arginine metabolism [26]. Several studies have also shown that high or low dietary levels of arginine and methionine are associated with the occurrence of abnormal behaviors such as FP and anal pecking [27]. In the present study, although the histidine biosynthetic function was improved, the histidine level showed no obvious differences

between the two groups of laying hens. This may be due to the fact that a large amount of histidine was consumed when they participated in the energy metabolism and antioxidant activity of intestinal epithelial cells exposed to environmental stressors [28].

Blood circulation is an important pathway for the gut microbiota to regulate the central nervous system through its metabolites. Our results indicated that the arginine level in the peripheral plasma was significantly increased; the histidine level also increased but not significantly ($p = 0.076$), which was probably due to the enhanced permeability of the BBB to histidine under environmental stressors [29]. Arginine and histidine supplementation increased their blood plasma concentrations [30,31]. Thus, in this study, the increased levels of plasma arginine and histidine can be mainly attributed to the enhancement of their biosynthesis in the gut bacteria, and the reduction in arginine metabolism. Previous studies have focused on the potential role of tryptophan and its metabolites and catecholamine hormones (e.g., dopamine, norepinephrine, and epinephrine) in FP behaviors of laying hens, while few studies have addressed the relationship between glutamate and its precursors and FP behaviors. Compared with gentle FP, the corticosterone level in severe FP laying hens was significantly decreased, while the epinephrine increased more significantly in the manual restraint test. A decreased level of tryptophan in peripheral blood, such as conversion into 5-HT [10], or the enhancement of tryptophan-KYN metabolic pathway [32], may be related to FP and aggressive behaviors in laying hens [33]. Similarly to previous results [34], we did not observe significant changes in the levels of tryptophan, 5-HT, and other metabolites in plasma, and analysis from ELISA revealed significantly increased levels of norepinephrine and epinephrine in plasma of FP chickens, indicating the stress state of FP chickens.

Metabolites in peripheral plasma can cross the BBB and ultimately affect animal behavior by regulating the central nervous system. Glutamate in the central nervous system is mainly derived from α -ketoglutaric acid in the citric acid cycle [17], and some studies have found that amino acids such as glutamine, arginine, proline, and histidine can also be converted to glutamate [12,35,36]. In addition, arginine and ornithine in the peripheral blood can cross the BBB into the central nervous system through the cationic amino acid transport protein 1 (CAT 1) [37,38], and histidine in the peripheral blood can cross the BBB through the large neutral amino acid transporter [39]. Our findings revealed a trend toward elevated levels of arginine and proline in the hippocampus; however, this difference did not reach statistical significance ($p = 0.051$ and 0.057 , respectively). This is mainly attributed to the conversion of arginine and proline into glutamate [12], which also explains the significantly increased glutamate level observed in the hippocampus of FP chickens. Growing evidence indicates that the gut microbiota can regulate the central nervous system through the metabolism, and thus influence host behaviors. For example, the chronic treatment of mice with *Lactobacillus rhamnosus* could alleviate depressive and anxiety-like behaviors by producing gamma-aminobutyric acid (GABA) or altering the expression of GABA receptors [13]. In this study, the *GRIN2A* gene in the hippocampus exhibited a significant positive correlation with the levels of N-acetyl-glutamate in the cecum and arginine in the duodenum; there was a significant positive correlation between the *SLC17A6* gene and arginine in the duodenum and between riboflavin and prostaglandin E1 in the ileum (Figure 7, Figures S21 and S22). Elevated levels of arginine and riboflavin were also found in the peripheral blood of FP chickens, suggesting a potential role of these two differential metabolites. Arginine supplementation can activate central NMDA receptors through the arginine–nitric oxide pathway, causing depressive-like behavior in mice [40]. Riboflavin can inhibit the release of glutamate in the nervous system and affect glutamatergic neurotransmission, thereby exerting neuroprotective effects [41].

Therefore, we proposed a possible mechanism by which gut microbiota regulates the glutamatergic system to affect FP in laying hens, as shown in Figure 8. Gut microbiota such as *Romboutsia* may increase the levels of arginine and histidine in plasma by increasing the biosynthesis of arginine and histidine or reducing the metabolism of arginine, thereby increasing the levels of glutamate and expression of *GRIN2A* and *SLC17A6* genes in the hippocampus. The glutamatergic system was then altered in several brain regions associated with mood disorders such as anxiety and depression, eventually leading to FP behaviors. In this study, biological sampling was conducted only on the pecking chickens (feather peckers), following previous research. However, since pecked chickens (feather victims) are likely the most stressed, they should also be included in future analyses to provide a deeper understanding of the mechanisms behind FP.

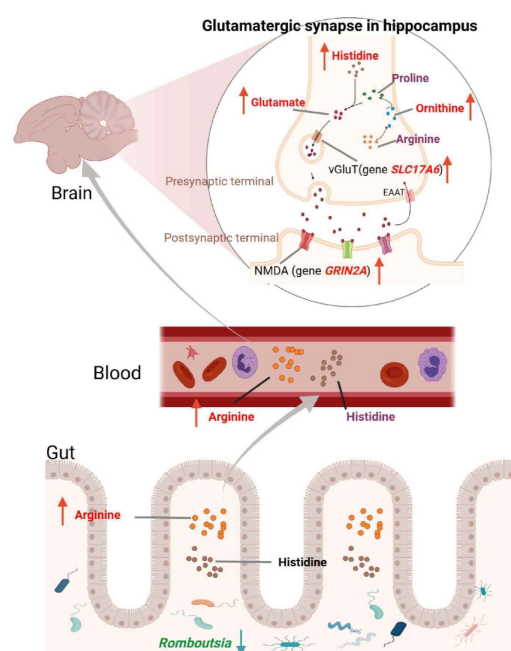


Figure 8. The proposed biological mechanism of FP behaviors. The gut microbiota regulates the glutamatergic system and affects FP in laying hens.

5. Conclusions

This study revealed that FP in laying hens can reduce the production performance and cause potential immune system perturbation, concurrent with heightened fear responses and depressive-like behaviors. This may be related to alterations in gut microbiota composition and function, particularly changes in the abundance of *Romboutsia* and other genera in the intestine. These changes increase intestinal arginine and histidine biosynthesis and reduce arginine metabolism, leading to increased plasma arginine and histidine levels. Since both arginine and histidine can cross the BBB, they can be converted to glutamate in the hippocampus. This process increases *GRIN2A* and *SLC17A6* gene expression, thus regulating the central glutamatergic system and affecting FP behavior. While this study provides valuable insights into the correlations between differential gut microbiomes, metabolites, and gene expression changes, it is crucial to explore potential causal relationships to further understand the underlying mechanisms. Future investigations should be focused on the key microorganisms that affect FP in laying hens by regulating the central glutamatergic system through arginine and histidine. For instance, using germ-free animal models, fecal microbiota transplantation, and stable isotope tracing can provide deeper insights into how specific microbiome alterations influence metabolite production and gene expression.

over time. Overall, our findings provide theoretical support and new biological control strategies for reducing the occurrence of FP, especially severe FP, in the laying hen industry.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/ani15091297/s1>. Method S1: Analysis of non-targeted metabolomics data; Method S2: Analysis of 16S rRNA gene sequencing data; Method S3: Analysis of eukaryotic transcriptome sequencing data; Method S4: The open field test; Method S5: Production performance and egg quality; Method S6: Non-targeted metabolomic analysis; Table S1: The main instruments used in the experiments; Table S2: The main biochemical kits used in the experiments; Table S3: Types and descriptions of FP behaviors; Table S4: Quantification of immune response markers in two groups of laying hens; Table S5: The identified differential metabolite features in different biological organisms of feather peckers; Table S6: Differential genes in the hippocampus and amygdala of feather pecking chickens; Figure S1: Production performance of laying hens in two groups; Figure S2: Egg quality of laying hens in two groups; Figure S3: The relative abundance of gut microbiota at the phylum level; Figure S4: The relative abundance of gut microbiota at the genus level; Figure S5: The co-occurrence network of the core bacteria; Figure S6: Functional prediction of differential gut microbiota; Figure S7: The number of identified metabolic features in biological organisms; Figure S8: KEGG pathway enrichment analysis; Figure S9: The metabolic features of glutamate acid and its precursors in the hippocampus; Figure S10: The metabolic features of glutamate acid and its precursors in the amygdala; Figure S11: The metabolic features of glutamate acid and its precursors in the plasma; Figure S12: The metabolic features of glutamate acid and its precursors in the cecum; Figure S13: The metabolic features of glutamate acid and its precursors in the duodenum; Figure S14: The metabolic features of glutamate acid and its precursors in the ileum; Figure S15: The top 30 enriched GO terms for differential genes in the hippocampus of feather pecking chickens; Figure S16: The top 30 enriched GO terms for differential genes in the amygdala of feather pecking chickens; Figure S17: Analysis of differentially accumulated genes; Figure S18: The correlation network of cecum microbiota and metabolites; Figure S19: The correlation network of duodenum microbiota and metabolites; Figure S20: The correlation network of ileum microbiota and metabolites; Figure S21: Spearman correlation coefficients between the hippocampal differential genes and the differential metabolites in the duodenum; Figure S22: Spearman correlation coefficients between the hippocampal differential genes and the differential metabolites in the ileum.

Author Contributions: Conceptualization, X.Y. and Y.W. (Yan Wang); data curation, C.W.; funding acquisition, X.Y. and Y.W. (Yan Wang); methodology, C.W., Y.L. (Yaling Li) and Y.L. (Yating Lin); supervision, Y.W. (Yan Wang); writing—original draft, X.Y.; writing—review and editing, Y.W. (Yan Wang) and Y.W. (Yinbao Wu). All authors are involved in the discussion and finalization of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (32372931, 31972610), the National Key R&D Program of China (2023YFD1301904), the Construction Project of Modern Agricultural Science and Technology Innovation Alliance in Guangdong Province (2023KJ128, 2024KJ128), the Earmarked Fund for Modern Agro-industry Technology Research System (CARS-40), and the Specific University Discipline Construction Project (2023B10564001).

Institutional Review Board Statement: All experimental procedures were approved by the Experiment Animal Ethics Committee of South China Agricultural University (2024G030).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data used in this work are available via Zenodo (<https://doi.org/10.5281/zenodo.10579290> (accessed on 29 January 2024)). This provides the original data for enzyme-linked immunosorbent assays, production performance, behaviors, transcriptomics, and metabolomics. In addition, 16s rRNA sequencing data that support the findings of this study have been deposited in the National Center for Biotechnology Information (<https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1069351> (accessed on 26 January 2024), accession: PRJNA1069351).

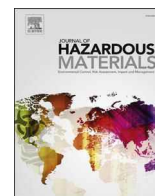
Conflicts of Interest: The authors declare no competing interests.

References

1. Rodenburg, T.B.; Van Krimpen, M.M.; De Jong, I.C.; De Haas, E.N.; Kops, M.S.; Riedstra, B.J.; Nordquist, R.E.; Wagenaar, J.P.; Bestman, M.; Nicol, C.J. The Prevention and Control of Feather Pecking in Laying Hens: Identifying the Underlying Principles. *World's Poult. Sci. J.* **2013**, *69*, 361–374. [\[CrossRef\]](#)
2. Cronin, G.M.; Glatz, P.C. Causes of Feather Pecking and Subsequent Welfare Issues for the Laying Hen: A Review. *Anim. Prod. Sci.* **2021**, *61*, 990–1005. [\[CrossRef\]](#)
3. Rodenburg, T.B.; Van Hierden, Y.M.; Buitenhuis, A.J.; Riedstra, B.; Koene, P.; Korte, S.M.; Van Der Poel, J.J.; Groothuis, T.G.G.; Blokhuis, H.J. Feather Pecking in Laying Hens: New Insights and Directions for Research? *Appl. Anim. Behav. Sci.* **2004**, *86*, 291–298. [\[CrossRef\]](#)
4. Fijn, L.B.; Josef van der Staay, F.; Goerlich-Jansson, V.C.; Arndt, S.S. Importance of Basic Research on the Causes of Feather Pecking in Relation to Welfare. *Animals* **2020**, *10*, 213. [\[CrossRef\]](#)
5. Wysocki, M.; Bessei, W.; Kjaer, J.B.; Bennewitz, J. Genetic and Physiological Factors Influencing Feather Pecking in Chickens. *World's Poult. Sci. J.* **2010**, *66*, 659–672. [\[CrossRef\]](#)
6. Falker-Gieske, C.; Mott, A.; Preuß, S.; Franzenburg, S.; Bessei, W.; Bennewitz, J.; Tetens, J. Analysis of the Brain Transcriptome in Lines of Laying Hens Divergently Selected for Feather Pecking. *BMC Genom.* **2020**, *21*, 595. [\[CrossRef\]](#)
7. de Haas, E.N.; van der Eijk, J.A.J. Where in the Serotonergic System Does It Go Wrong? Unravelling the Route by Which the Serotonergic System Affects Feather Pecking in Chickens. *Neurosci. Biobehav. Rev.* **2018**, *95*, 170–188. [\[CrossRef\]](#)
8. Kops, M.S.; Kjaer, J.B.; Güntürkün, O.; Westphal, K.G.C.; Korte-Bouws, G.A.H.; Olivier, B.; Korte, S.M.; Bolhuis, J.E. Brain Monoamine Levels and Behaviour of Young and Adult Chickens Genetically Selected on Feather Pecking. *Behav. Brain Res.* **2017**, *327*, 11–20. [\[CrossRef\]](#)
9. Wang, C.; Li, Y.; Wang, H.; Li, M.; Rong, J.; Liao, X.; Wu, Y.; Wang, Y. Differences in Peripheral and Central Metabolites and Gut Microbiome of Laying Hens with Different Feather-Pecking Phenotypes. *Front. Microbiol.* **2023**, *14*, 1132866. [\[CrossRef\]](#)
10. Buitenhuis, A.J.; Kjaer, J.B.; Labouriau, R.; Juul-Madsen, H.R. Altered Circulating Levels of Serotonin and Immunological Changes in Laying Hens Divergently Selected for Feather Pecking Behavior. *Poult. Sci.* **2006**, *85*, 1722–1728. [\[CrossRef\]](#)
11. Zhang, Y.; Fan, Q.; Hou, Y.; Zhang, X.; Yin, Z.; Cai, X.; Wei, W.; Wang, J.; He, D.; Wang, G.; et al. Bacteroides Species Differentially Modulate Depression-like Behavior via Gut-Brain Metabolic Signaling. *Brain. Behav. Immun.* **2022**, *102*, 11–22. [\[CrossRef\]](#) [\[PubMed\]](#)
12. Mayneris-Perxachs, J.; Castells-Nobau, A.; Arnoriaga-Rodríguez, M.; Martin, M.; de la Vega-Correa, L.; Zapata, C.; Burokas, A.; Blasco, G.; Coll, C.; Escrichs, A.; et al. Microbiota Alterations in Proline Metabolism Impact Depression. *Cell Metab.* **2022**, *34*, 681–701. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Bravo, J.A.; Forsythe, P.; Chew, M.V.; Escaravage, E.; Savignac, H.M.; Dinan, T.G.; Bienenstock, J.; Cryan, J.F. Ingestion of Lactobacillus Strain Regulates Emotional Behavior and Central GABA Receptor Expression in a Mouse via the Vagus Nerve. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 16050–16055. [\[CrossRef\]](#) [\[PubMed\]](#)
14. Wang, H.; Yang, F.; Song, Z.-W.; Shao, H.-T.; Bai, D.-Y.; Ma, Y.-B.; Kong, T.; Yang, F. The Influence of Immune Stress Induced by *Escherichia coli* Lipopolysaccharide on the Pharmacokinetics of Danofloxacin in Broilers. *Poult. Sci.* **2022**, *101*, 101629. [\[CrossRef\]](#)
15. Akinyemi, F.; Adewole, D. Effects of Brown Seaweed Products on Growth Performance, Plasma Biochemistry, Immune Response, and Antioxidant Capacity of Broiler Chickens Challenged with Heat Stress. *Poult. Sci.* **2022**, *101*, 102215. [\[CrossRef\]](#)
16. Kanehisa, M.; Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **2000**, *28*, 27–30. [\[CrossRef\]](#)
17. Zhou, Y.; Danbolt, N.C. Glutamate as a Neurotransmitter in the Healthy Brain. *J. Neural Transm.* **2014**, *121*, 799–817. [\[CrossRef\]](#)
18. Meldrum, B.S. Glutamate as a Neurotransmitter in the Brain: Review of Physiology and Pathology. *J. Nutr.* **2000**, *130*, 1007S–1015S. [\[CrossRef\]](#)
19. Dennis, R.L.; Cheng, H.W. Effects of Selective Serotonin Antagonism on Central Neurotransmission. *Poult. Sci.* **2012**, *91*, 817–822. [\[CrossRef\]](#)
20. Riedel, G.; Platt, B.; Micheau, J. Glutamate Receptor Function in Learning and Memory. *Behav. Brain Res.* **2003**, *140*, 1–47. [\[CrossRef\]](#)
21. Karisetty, B.C.; Maitra, S.; Wahul, A.B.; Musalamadugu, A.; Khandelwal, N.; Guntupalli, S.; Garikapati, R.; Jhansyrani, T.; Kumar, A.; Chakravarty, S. Differential Effect of Chronic Stress on Mouse Hippocampal Memory and Affective Behavior: Role of Major Ovarian Hormones. *Behav. Brain Res.* **2017**, *318*, 36–44. [\[CrossRef\]](#) [\[PubMed\]](#)
22. Szo, A.; Zichó, K.; Barth, A.M.; Gönczi, R.T.; Schlingloff, D.; Török, B.; Sipos, E.; Major, A.; Bardóczi, Z.; Sos, K.E.; et al. Median Raphe Controls Acquisition of Negative Experience in the Mouse. *Science* **2019**, *366*, eaay8746.
23. Morais, L.H.; Schreiber, H.L.; Mazmanian, S.K. The Gut Microbiota–Brain Axis in Behaviour and Brain Disorders. *Nat. Rev. Microbiol.* **2021**, *19*, 241–255. [\[CrossRef\]](#) [\[PubMed\]](#)

24. Fetissov, S.O. Role of the Gut Microbiota in Host Appetite Control: Bacterial Growth to Animal Feeding Behaviour. *Nat. Rev. Endocrinol.* **2017**, *13*, 11–25. [\[CrossRef\]](#)
25. He, H.; Zhao, Z.; Xiao, C.; Li, L.; Liu, Y.-E.; Fu, J.; Liao, H.; Zhou, T.; Zhang, J. Gut Microbiome Promotes Mice Recovery from Stress-Induced Depression by Rescuing Hippocampal Neurogenesis. *Neurobiol. Dis.* **2024**, *191*, 106396. [\[CrossRef\]](#)
26. Ruan, D.; Fouad, A.M.; Fan, Q.L.; Huo, X.H.; Kuang, Z.X.; Wang, H.; Guo, C.Y.; Deng, Y.F.; Zhang, C.; Zhang, J.H.; et al. Dietary L-Arginine Supplementation Enhances Growth Performance, Intestinal Antioxidative Capacity, Immunity and Modulates Gut Microbiota in Yellow-Feathered Chickens. *Poult. Sci.* **2020**, *99*, 6935–6945. [\[CrossRef\]](#)
27. Van Krimpen, M.M.; Kwakkel, R.P.; Reuvekamp, B.F.J.; Van Der Peet-Schwering, C.M.C.; Den Hartog, L.A.; Verstegen, M.W.A. Impact of Feeding Management on Feather Pecking in Laying Hens. *World's Poult. Sci. J.* **2005**, *61*, 663–686. [\[CrossRef\]](#)
28. Han, Y.; Koshio, S.; Ishikawa, M.; Yokoyama, S. Interactive Effects of Dietary Arginine and Histidine on the Performances of Japanese Flounder *Paralichthys Olivaceus* Juveniles. *Aquaculture* **2013**, *414–415*, 173–182. [\[CrossRef\]](#)
29. Taghadosi, Z.; Zarifkar, A.; Razban, V.; Owjifard, M.; Aligholi, H. Effect of Chronically Electric Foot Shock Stress on Spatial Memory and Hippocampal Blood Brain Barrier Permeability. *Behav. Brain Res.* **2021**, *410*, 113364. [\[CrossRef\]](#)
30. Lackner, J.; Hess, V.; Marx, A.; Hosseini-Ghaffari, M.; Sauerwein, H. Effects of Dietary Supplementation with Histidine and β -Alanine on Blood Plasma Metabolome of Broiler Chickens at Different Ages. *PLoS ONE* **2022**, *17*, e0277476. [\[CrossRef\]](#)
31. Brugaletta, G.; Zampiga, M.; Laghi, L.; Indio, V.; Oliveri, C.; De Cesare, A.; Sirri, F. Feeding Broiler Chickens with Arginine above Recommended Levels: Effects on Growth Performance, Metabolism, and Intestinal Microbiota. *J. Anim. Sci. Biotechnol.* **2023**, *14*, 33. [\[CrossRef\]](#) [\[PubMed\]](#)
32. Birkel, P.; Chow, J.; Forsythe, P.; Gostner, J.M.; Kjaer, J.B.; Kunze, W.A.; McBride, P.; Fuchs, D.; Harlander-Matauschek, A. The Role of Tryptophan-Kynurenine in Feather Pecking in Domestic Chicken Lines. *Front. Vet. Sci.* **2019**, *6*, 209. [\[CrossRef\]](#) [\[PubMed\]](#)
33. Birkel, P.; Franke, L.; Bas Rodenburg, T.; Ellen, E.; Harlander-Matauschek, A. A Role for Plasma Aromatic Amino Acids in Injurious Pecking Behavior in Laying Hens. *Physiol. Behav.* **2017**, *175*, 88–96. [\[CrossRef\]](#) [\[PubMed\]](#)
34. Dennis, R.L. Adrenergic and Noradrenergic Regulation of Poultry Behavior and Production. *Domest. Anim. Endocrinol.* **2016**, *56*, S94–S100. [\[CrossRef\]](#)
35. Andersen, J.V.; Markussen, K.H.; Jakobsen, E.; Schousboe, A.; Waagepetersen, H.S.; Rosenberg, P.A.; Aldana, B.I. Glutamate Metabolism and Recycling at the Excitatory Synapse in Health and Neurodegeneration. *Neuropharmacology* **2021**, *196*, 108719. [\[CrossRef\]](#)
36. Zhu, H.; Wang, N.; Yao, L.; Chen, Q.; Zhang, R.; Qian, J.; Hou, Y.; Guo, W.; Fan, S.; Liu, S.; et al. Moderate UV Exposure Enhances Learning and Memory by Promoting a Novel Glutamate Biosynthetic Pathway in the Brain. *Cell* **2018**, *173*, 1716–1727. [\[CrossRef\]](#)
37. Tachikawa, M.; Hirose, S.; Akanuma, S.-I.; Matsuyama, R.; Hosoya, K.-I. Developmental Changes of L-Arginine Transport at the Blood-Brain Barrier in Rats. *Microvasc. Res.* **2018**, *117*, 16–21. [\[CrossRef\]](#)
38. Henriques, C.; Miller, M.P.; Catanho, M.; De Carvalho, T.M.U.; Krieger, M.A.; Probst, C.M.; De Souza, W.; Degraive, W.; Amara, S.G. Identification and Functional Characterization of a Novel Arginine/Ornithine Transporter, a Member of a Cationic Amino Acid Transporter Subfamily in the Trypanosoma Cruzi Genome. *Parasites Vectors* **2015**, *8*, 346. [\[CrossRef\]](#)
39. Oldendorf, W.H.; Crane, P.D.; Braun, L.D.; Gosschalk, E.A.; Diamond, J.M. PH Dependence of Histidine Affinity for Blood-Brain Barrier Carrier Transport Systems for Neutral and Cationic Amino Acids. *J. Neurochem.* **1988**, *50*, 857–861. [\[CrossRef\]](#)
40. Yildiz, F.; Erden, B.F.; Ulak, G.; Utkan, T.; Gacar, N. Antidepressant-like Effect of 7-Nitroindazole in the Forced Swimming Test in Rats. *Psychopharmacology* **2000**, *149*, 41–44. [\[CrossRef\]](#)
41. Huang, S.-K.; Lu, C.-W.; Lin, T.-Y.; Wang, S.-J. Neuroprotective Role of the B Vitamins in the Modulation of the Central Glutamatergic Neurotransmission. *CNS Neurol. Disord. Drug Targets* **2022**, *21*, 292–301. [\[CrossRef\]](#)

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.



Linking electron ionization mass spectra of organic chemicals to toxicity endpoints through machine learning and experimentation

Song Hu^{a,b}, Guohong Liu^a, Jin Zhang^c, Jiachen Yan^a, Hongyu Zhou^a, Xiliang Yan^{a,*}

^a Institute of Environmental Research at Greater Bay Area, Key Laboratory for Water Quality and Conservation of the Pearl River Delta, Ministry of Education, Guangzhou University, Guangzhou 510006, China

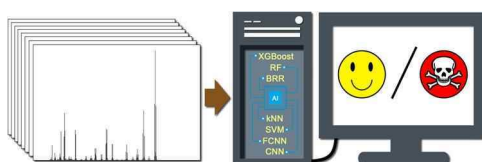
^b School of Environmental Science and Engineering, Shandong University, Qingdao 266237, China

^c School of Public Health, Guizhou Medical University, Guiyang 550025, China

HIGHLIGHTS

- A novel QSAR method was proposed to predict the toxicity of chemicals without knowing chemical structures.
- Toxicity of chemicals were linked to the EI-MS using machine learning and experimentation.
- Critical spectra features were identified to be responsible for chemical-induced toxicity.
- Application of applicability domain allowed for making reasonable toxicity predictions for unknown chemicals.

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: Jorg Rinklebe

Keywords:

Chemical structure identification
Machine learning
Mass spectra
Chemical toxicity prediction
Environmental health and safety

ABSTRACT

Quantitative structure-activity relationship (QSAR) modeling has been widely used to predict the potential harm of chemicals, in which the prediction heavily relies on the accurate annotation of chemical structures. However, it is difficult to determine the accurate structure of an unknown compound in many cases, such as in complex water environments. Here, we solved the above problem by linking electron ionization mass spectra (EI-MS) of organic chemicals to toxicity endpoints through various machine learning methods. The proposed method was verified by predicting 50% growth inhibition of *Tetrahymena pyriformis* (*T. pyriformis*) and liver toxicity. The optimal model performance obtained an $R^2 > 0.7$ or balanced accuracy > 0.72 for both the training set and test set. External experimentation further verified the application potential of our proposed method in the toxicity prediction of unknown chemicals. Feature importance analysis allowed us to identify critical spectral features that were responsible for chemical-induced toxicity. Our approach has the potential for toxicity prediction in such fields that it is difficult to determine accurate chemical structures.

1. Introduction

As of May 8, 2019, the number of chemicals registered in the Chemical Abstracts Service (CAS) Registry reached 150 million and continued to grow exponentially (Linda Wang, 2019). Chemicals, while

improving the quality of human life and bringing convenience to society, also pose potential risks to the environment and human health. Indeed, chemical pollution has been shown to threaten human survival on Earth (Diamond et al., 2015). According to a report from the Lancet Commission on Pollution and Health (Landrigan et al., 2018), an

* Corresponding author.

E-mail address: yanxiliang1991@gzhu.edu.cn (X. Yan).

<https://doi.org/10.1016/j.jhazmat.2022.128558>

Received 6 December 2021; Received in revised form 25 January 2022; Accepted 21 February 2022

Available online 23 February 2022

0304-3894/© 2022 Elsevier B.V. All rights reserved.

estimated 9 million premature deaths in 2015 were related to the contamination of air, water and soil. These statistics highlight the necessity and urgency of environmental health risk assessment for chemicals before commercial production and use (Kavlock et al., 2018). However, chemicals with comprehensive toxicological data only account for a very small portion of tens of thousands of commercial chemicals (Judson et al., 2009). Furthermore, the current toxicity assessment of chemicals mainly relies on time-consuming and laborious in vitro or in vivo experiments. Therefore, there is an urgent need to develop effective computational methods for rapid toxicity assessment of chemicals and to reduce animal studies (Ciallella and Zhu, 2019; Russo et al., 2019).

In previous studies (Muratov et al., 2020; Piir et al., 2018), QSAR modeling has been proven to be an efficient approach for generating quantitative risk estimation for chemicals. As for QSAR models constructed by traditional machine learning methods or deep learning methods, accurate structures of chemicals must be obtained before toxicity prediction (Muratov et al., 2020). For example, traditional machine learning methods (She and Judson, 2019; Zorn et al., 2020), such as the *k*-nearest neighbors (kNN) and random forest (RF), use descriptors calculated from molecular structures as the input. The end-to-end deep learning methods (Asilar et al., 2020; Ciriano and Bender, 2019), such as convolutional neural networks (CNN), take molecular images converted from molecular structures as input. However, in many cases, e.g., in complex water (Escher et al., 2020) and atmospheric environments (Li et al., 2017), it is difficult to determine the accurate structure of an unknown compound. In fact, the identification of molecules is still one of the key issues in many fields of analytical chemistry, such as environmental pollutants analysis (Gago-Ferrero et al., 2015) and natural product detection (Dührkop et al., 2019). Although some advanced analytical methods, such as mass spectrometry, provide technical support for molecule identification, the transformation of spectral data to molecular structure still relies on laborious manual annotation (Feider et al., 2019; Kind et al., 2018) or advanced computational algorithms (Dührkop et al., 2021; Ji et al., 2020; Picache et al., 2020). To this end, we proposed such a hypothesis, that is, whether it is possible to directly use the spectral data to predict the toxicity of molecules without knowing their chemical structures.

Previous studies (Beger and Wilkes, 2001; Tie et al., 2012) used NMR (Nuclear Magnetic Resonance) features to construct QSDAR (Quantitative Spectrometric Data—Activity Relationship) models to predict the binding of organic compounds to the protein, these results provided theoretical support for our hypothesis. However, the feasibility of this hypothesis needs to be further verified due to small data sets (mainly less than 200 compounds) and the arbitrary selection of modeling approaches (mainly linear regression) in previous studies. To further verify our hypothesis, we proposed a spectrum-based machine learning framework to estimate the toxicity of organic chemicals using large data set and various machine learning approaches. Unlike the existing QSAR models, the proposed method can directly extract features from electron ionization mass spectra (EI-MS) and then output the toxicity value of corresponding compounds. The constructed machine learning models were used to predict 50% growth inhibition of *T. pyriformis* and drug-induced liver injury. The high determination coefficient (R^2) and accuracy score indicated that the constructed machine learning models could be used for both regression and classification analysis tasks. Interpretation of the machine learning predictions allowed us to identify possible structure alerts inferred from spectral features, which could be used for designing green chemicals in the future. In addition, external experimentation verified the application of our proposed method in the toxicity prediction of unknown chemicals. This is the first study to link the EI-MS of chemicals to the toxicity endpoints, demonstrating huge application potential for toxicity prediction in such fields that it is difficult to determine the accurate chemical structures. We believe that the proposed method can well complement the existing QSAR modeling that needs to know the chemical structures before toxicity prediction.

2. Materials and methods

2.1. Data sets

In the present study, we trained and validated the machine learning models using two different data sets, i.e., *T. pyriformis* growth inhibition data set (IGC50 set) and drug-induced liver injury data set (DILI set). The IGC50 set measured the concentration of organic chemicals that causes 50% growth inhibition of *T. pyriformis* after 40 h. The DILI set reported the hepatotoxicity of organic compounds and aimed to evaluate the drug-induced liver injury in the drug-development process. The IGC50 set was obtained from the ecotoxicology database (<http://cfpub.epa.gov/ecotox/>), and the DILI set was collected from the literature (Ai et al., 2018). The concentration values in the IGC50 set were log-transformed with the unit of mol/L, and the hepatotoxic effects in the DILI set were labeled as “positive” or “negative”. The IGC50 set and DILI set were respectively used for building regression and classification models, which could meet the needs of different modeling purposes. In addition, the application of the proposed method in two data sets can provide supports for the toxicity prediction of unknown compounds in complex environments or during drug development. Before building machine learning models, we further curated the data to ensure no salts, mixtures, and duplicate chemicals exist in each data set.

2.2. Spectral data acquisition and preprocessing

The spectral data were obtained from the NIST (National Institute of Standard and Technology) Chemistry WebBook (<https://webbook.nist.gov/>). The EI-MS of each compound was stored in a *idx* format file and downloaded according to their CAS number. The *in silico* spectral data predicted from the deep learning model (Wei et al., 2019) were used if no experimental data were available. Finally, we obtained the EI-MS of 2531 compounds, of which 1306 compounds were used in the IGC50 set for regression and 1237 compounds were used in the DILI set for binary classification (i.e., there was an overlap of 12 compounds in the above two data sets). Herein, the EI-MS of all compounds in the IGC50 set were experimental data, and the EI-MS of most compounds in the DILI set were simulated data. For the DILI set, 808 compounds were labeled as positive and 429 compounds were labeled as negative.

Preprocessing of the EI-MS was essential prior to machine learning analysis, and if absent or inadequate, will lead to biased or biologically irrelevant conclusions. An EI-MS is usually composed of the mass-to-charge ratio and its corresponding intensity (Fig. 1a). Firstly, based on the maximum and minimum of mass-to-charge ratios in the IGC50 set or DILI set, the mass-to-charge ratio of each EI-MS was unified to the same interval (Fig. 1b). Next, an interpolating step was performed (Fig. 1c), in which three interpolation methods (i.e., linear interpolation, cubic spline interpolation, and zero-interpolation) were used to fill in empty values of the preprocessed EI-MS. To determine the most suitable interpolation method, we compared the model performances under different interpolation methods. Herein, the extreme gradient boost (XGBoost) model with default parameters was selected to predict the *T. pyriformis* inhibition and liver toxicity. As shown in Table S1, the machine learning models exhibited better performances when using zero-interpolation method. The main reason was that the EI-MS was in the form of relative ion abundances and was a discrete function of mass-to-charge ratio, while the linear interpolation and cubic spline interpolation were prone to generate irrelative noise peaks, so the discrete interpolation method (i.e., zero-interpolation method) was more suitable. Based on the mean (\bar{x} , Eq. 1) and standard deviation (σ , Eq. 1) of the relative intensities of each mass-to-charge ratio in each data set, the relative intensity (x , Eq. 1) of the mass-to-charge ratio of each molecule was then standardized using the zero-mean normalization method (Fig. 1d).

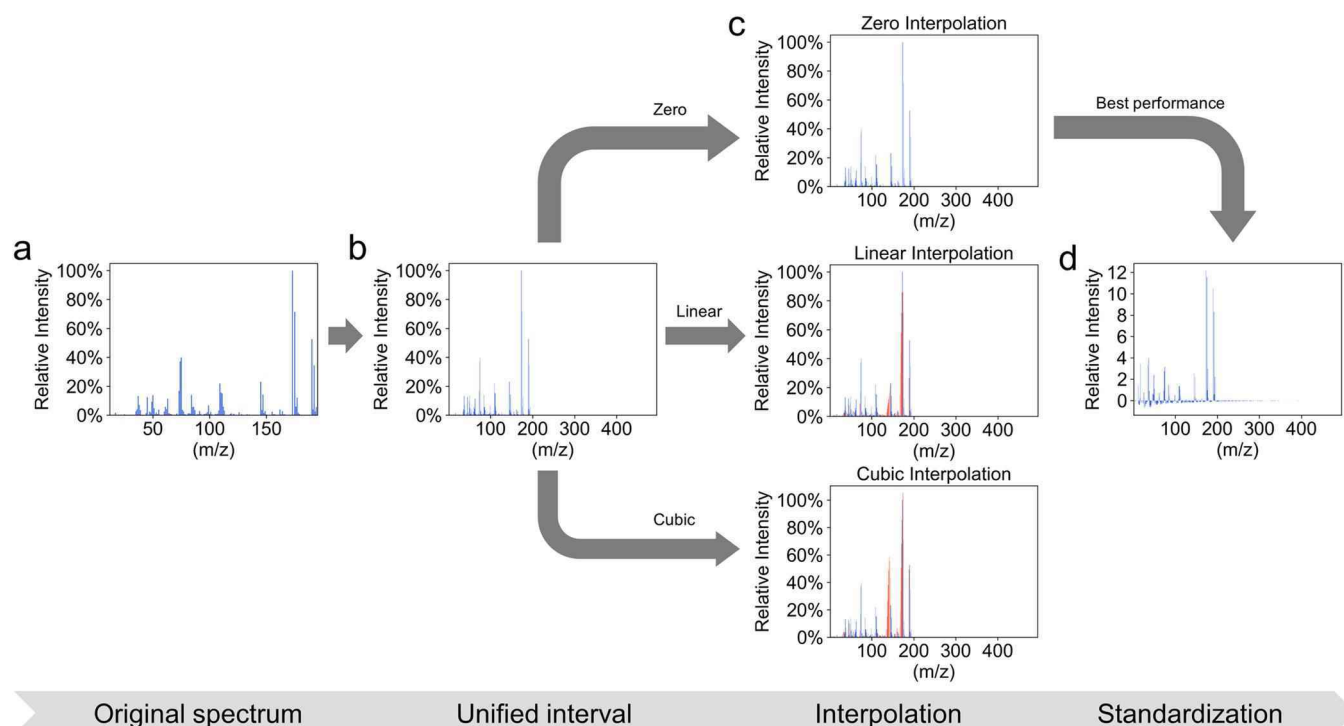


Fig. 1. Preprocessing of the EI-MS. Before feeding into the machine learning models, the original EI-MS (a) was first preprocessed by unified interval (b), interpolation (c), and standardization (d). The interpolation values were marked as red lines. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

$$x' = \frac{x - \bar{x}}{\sigma} \quad (1)$$

Where, x is the original relative intensity of the mass-to-charge ratio of each molecule, \bar{x} and σ are respectively the mean and standard deviation of the relative intensities of the mass-to-charge ratio in the IGC50 set or DILI set.

To better understand the preprocessing of spectral data, an example spectrum (i.e., 2,6-Dichlorobenzoic acid, CAS number: 50-30-6) from IGC50 set was provided. Overall, the EI-MS of each molecule can be regarded as a vector, in which mass-to-charge ratios corresponded to indices and relative intensities were values (Fig. S1). Firstly, the interval of mass-to-charge ratio changed from [12,195] to [1,393] after interval unification. Then, the missing values of relative intensities were filled with zero. The relative intensity was then standardized using the mean and standard deviation of the relative intensities of each mass-to-charge ratio in IGC50 set. For instance, the mean and standard deviation of the relative intensities at $m/z = 44$ were respectively 4.79 and 15.91. So, the relative intensity of 2,6-Dichlorobenzoic acid at $m/z = 44$ was standardized as $-0.215 = \frac{1.37-4.79}{15.91}$ (Eq. 1). After preprocessing, the EI-MS of each molecule was linked to its corresponding toxicity value using various machine learning and deep learning methods.

2.3. Machine learning and deep learning methods

In the present study, we applied five classic machine learning (i.e., Bayesian ridge regression, RF, support vector machines, kNN, and XGBoost) and two deep learning methods (i.e., fully connected neural network and convolutional neural network) to build the prediction models. All classic machine learning models were constructed using the scikit-learn package in Python. Briefly, RF is an ensemble machine learning algorithm consisting of many decision trees and generates predictions by combining outputs from these decision trees (Breiman, 2001). The basic idea of a support vector machine (SVM) is to find an optimal hyperplane for separating different classes and maximize the

margin between these classes (Karatzoglou et al., 2006). The Bayesian ridge regression (BRR) is an approach to linear regression and applies Bayesian inference for statistical analysis (Shi et al., 2016). The kNN method outputs the prediction by averaging the results of its k -nearest neighbors (Ajmani et al., 2006), while the XGBoost method is an implementation of gradient boosted decision trees designed for speed and performance (Chen et al., 2015). To obtain optimal model performance, the grid-search algorithm was used to tune the hyperparameters of machine learning models on the training set. Briefly, the grid-search algorithm iterated through every hyperparameter combination and stored a model for each combination. Then, the model with the best combination of hyperparameters was retained and used for prediction on the test set.

The fully connected neural network (FCNN) and CNN are both subclasses of deep neural networks (DNNs) that contain one input layer, one output layer, and more than one hidden layer. The FCNN consists of a series of fully connected layers that connect every neuron in one layer to every neuron in the other (LeCun et al., 2015). The CNN is a class of end-to-end DNN that can recognize features from raw input such as images (Xu et al., 2014). Typically, CNN contains three types of layers, i.e., convolutional layer, pooling layer, and fully connected layer. The convolutional layer is the fundamental component of the CNN architecture and performs a convolution operation to filter useful information from input data. In this study, the 1-dimensional CNN was applied to learn features from EI-MS of chemicals. All the deep learning models were implemented in PyTorch package of Python.

2.4. Model development and evaluation metrics

Each data set was divided into a training set (80% of the whole set) and a test set (20% of the whole set). The training set was used for constructing an initial model, and the test set was applied to evaluate the model's predictive ability. The performances of regression models were evaluated by the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE). R^2 denotes the strength

of the linear association between real values and predicted values (Eq. 2), RMSE measures the standard deviation of the prediction errors (Eq. 3), MAE is a measure of the average difference between experimentations and predictions (Eq. 4). With the maximum R^2 value of 1, lower RMSE and MAE values indicate better model performance. For the DILI set, the F1 score, accuracy and the area under the receiver operating characteristic curve (AUC-ROC) were employed to evaluate the performance of the classification models. The F1 score is the weighted average of precision and recall (Eq. 5), while accuracy measures the fraction of the number of correctly classified compounds to the total number of compounds. By contrast, AUC-ROC assesses the false positive rate against the true positive rate at various thresholds. Considering the imbalance of DILI set, the balanced accuracy (Eq. 6) was calculated as the indicator for classification models in the present study. The values of F1 score, balanced accuracy, and AUC-ROC range from 0 to 1, with higher values indicating better model performance. All models were validated using 5-fold cross validation and external validation. Eqs. 2–6 are listed as follows:

$$R^2 = \frac{\sum_{i=1}^n (y_i^{obs} - y_i^{pred})^2}{\sum_{i=1}^n (y_i^{obs} - \bar{y}^{obs})^2} \quad (2)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i^{obs} - y_i^{pred})^2}{n}} \quad (3)$$

$$MAE = \frac{\sum_{i=1}^n |y_i^{obs} - y_i^{pred}|}{n} \quad (4)$$

$$F1score = 2 \times \frac{precision \times recall}{precision + recall} = \frac{TP}{TP + 0.5 \times (FP + TN)} \quad (5)$$

$$Balanced \ accuracy = \frac{\frac{TP}{TP+FN} + \frac{TN}{TN+FP}}{2} \quad (6)$$

where, y_i^{pred} is the predicted toxicity value for each molecule, y_i^{obs} is the experimental toxicity value for each molecule, \bar{y}^{obs} is the mean toxicity value over all molecules, and n is the number of molecules; TP and TN are cases correctly predicted as positive and negative, respectively; FP and FN are cases falsely predicted as positive and negative, respectively.

2.5. Applicability domain

As stated by the OECD guiding principles, a QSAR model should have a defined applicability domain (AD). Generally, it is impractical for a QSAR model to reliably predict the target property of every compound in the entire chemical space. The predictions will be affected by a large error if a compound exhibits a huge difference with all compounds in the modeling set. Therefore, the concept of AD was developed and applied to identify the region where the model's prediction can be reliable. In the present study, a proposed cutoff value (Zhang et al., 2008), D_c , was used to determine the AD (Eq. 7):

$$D_c = \bar{y} + Z\sigma \quad (7)$$

Where, \bar{y} represents the average of the Euclidean distances between each compound and its k nearest neighbors in the training set, σ represents the standard deviation of these Euclidean distances, and Z is an empirical parameter to adjust the significance level. Therefore, if the Euclidean distance for an external compound exceeds D_c , the prediction is considered unreliable. As described above (Fig. S1), the EI-MS of each molecule can be regarded as a vector, in which mass-to-charge ratios corresponded to indices and relative intensities were values. So, the similarity of two molecules can be quantitatively evaluated by the

Euclidean distance between two vectors (Eq. 8).

$$d = \sqrt{\sum_{m=1}^m (x_{im} - x_{jm})^2} \quad (8)$$

Where, x_{im} or x_{jm} is the relative intensity of the i_{th} or j_{th} molecule at $m/z = m$.

2.6. Experimental validation

To further verify the effectiveness of the proposed method, an external validation set of compounds was constructed and tested for IGC50. These compounds were not included in the IGC50 set. The compounds for toxicity testing were purchased from J&K Chemical Ltd., Shanghai, China. The IGC50 values were determined by a population growth impairment test (Schultz et al., 1997) and calculated from the dose-response curves. Specifically, the compounds were firstly dissolved in DMSO (dimethyl sulfoxide) and diluted to eight different concentrations. Then, *T. pyriformis* were exposed to the compounds at different concentrations. Simultaneously, two control groups of *T. pyriformis* were incubated with either DMSO or culture medium to provide a basis for interpreting data from other chemical treatments. After 40-hour incubation, the population density of *T. pyriformis* was quantified using an automated cell counter according to the manufacturer's protocol (Ruiyu Biotechnology Co. Ltd., Shanghai, China). Based on the generated growth inhibition rates coupled with the corresponding chemical concentrations, the IGC50 value for each compound was finally calculated using SigmaPlot software (Systat Software Inc., San Jose, USA). The experimental protocol described above is same with that used to generate the IGC50 set (Schultz et al., 1997), which can ensure the reliability of external experimental validation.

3. Results and discussion

3.1. Visualization of chemical structure diversity and toxicity values distribution

Prior to machine learning, we firstly analyzed chemical structure diversity and toxicity value distribution, as the way to evaluate the suitability of the data set for modeling. As described in the part of spectral data acquisition and preprocessing, the original EI-MS of each compound was transformed into a unified matrix. As shown in Fig. 2a & c, the m/z values of the IGC50 set ranged from 1 to 393, while the m/z values of DILI set ranged from 1 to 999. The difference was highly related to the chemical structures in the two data sets. The compounds in IGC50 set were mainly small molecules with molecular weights between 46 and 391 (e.g., 2,6-Diiodo-4-nitrophenol, CAS number: 305-85-1, Fig. S2a), while there were several macromolecule compounds with molecular weights higher than 1000 (e.g., corticotropin, CAS number: 9002-60-2, Fig. S2b) in the DILI set.

Using the preprocessed EI-MS, a principal component analysis (PCA) was then performed to visualize the chemical space. Here, the molecular features generated from the preprocessed EI-MS were dimensionally reduced to several principal components. To better understand the descriptive variables used for PCA and machine learning models, a schematic diagram was added to illustrate the dimensional reduction of spectral features of molecules in the IGC50 set. As shown in Fig. S3a, the original feature matrix contained 1306 rows (i.e., 1306 compounds) and 393 columns (i.e., 393 mass-to-charge ratios). Herein, the mass-to-charge ratios corresponded to feature names and the relative intensities were feature values. After PCA, 393 features were dimensionally reduced to three principal components (Fig. S3b). The top three principal components were used to show the distribution of investigated compounds in a 3D chemical space. When viewed as the chemical space of organic compounds in the two data sets, all compounds were almost structurally different due to various chemical compositions and

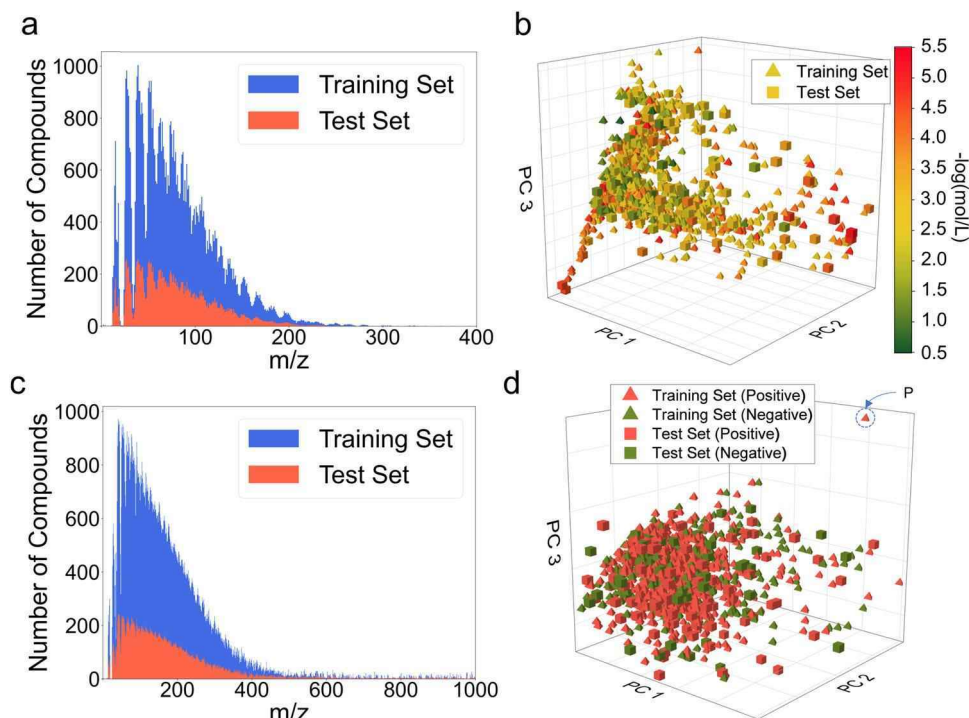


Fig. 2. Visualization of chemical structure diversity and toxicity value distribution. Distribution of m/z values in (a) the IGC50 set, (c) the DILI set, and the principal component analysis results of molecular features generated from EI-MS in (b) the IGC50 set and (d) DILI set. The toxicity values or labels are marked with different colors in the figure. The P data point represents a structural outlier.

occupied most of this chemical space (Fig. 2b & d). The results indicated that most molecules can be effectively distinguished by spectral features extracted from the EI-MS. The diversity of chemical structures and wide distribution of toxicity values were conducive to prediction performance and generalizability of machine learning models. As can be seen in Fig. 2d, there were also several structural outliers. For example, due to the strong peaks at $m/z = 778$ and 779 (Fig. S2c), the P data point (vinorelbine, CAS number: 71486-22-1) was significantly different from others (Fig. 2d). Usually, the structural outliers will lead to larger prediction errors. Therefore, we applied the applicability domain to avoid such an inaccurate extrapolation of toxicity prediction, which will be discussed below. On the other hand, since the NIST database has been assembled over many years and under different conditions, these experimental conditions may result in several fragmentation ions of one molecule that are significantly different with other molecules. This requires experimentalists to work together with the QSAR modellers to ensure the data quality. Only with high-quality data, the machine learning models can make reliable decision. The above analysis can help preclude the performance of the machine learning model to be built.

3.2. Comparison of model performances of different machine learning and deep learning methods

As described above, we planned to verify our hypothesis on two data sets. The IGC50 set was used for building regression models, while the DILI set was applied to construct classification models. In this study, five machine learning methods (i.e., RF, SVM, BRR, kNN, and XGBoost) and two deep learning methods (i.e., FCNN and CNN) were employed to link the EI-MS of chemicals to their corresponding toxicity values or labels. To optimize the model performance, we firstly applied a grid-search algorithm to tune the parameters on the training sets. Due to a large number of adjustable parameters, it is impossible and also unnecessary to evaluate all possible parameter combinations. Therefore, we tuned the parameters within a reasonable range and evaluated the performance of the corresponding models. The main parameters and their

ranges were listed in Table S2.

With the best combination of parameters (Table S3), almost all machine learning and deep learning models exhibited acceptable performance ($R^2 > 0.5$ or balanced accuracy > 0.7 , Table 1 and Table 2) (Golbraikh et al., 2003). In addition, the higher performance of regression models or classification models was further verified by lower RMSE/MAE or higher F1 score/AUC-ROC. Overall, both the 5-fold cross validation process and external prediction obtained similar performance, indicating that the model parameters selected by the 5-fold cross validation were efficient. Especially, regardless of regression models or classification models, both XGBoost and RF algorithms showed better performances than other algorithms. Similar to what we have concluded from our previous studies (Yan et al., 2019, 2020, 2021), the deep learning model did not show better performance than the traditional machine learning model when the volume of data was small. Although deep learning has brought us many surprises, such as AlphaFold (Ronneberger et al., 2021), its success still benefits from big data in the field.

Due to the noticeable difference in model performances between different machine learning algorithms, we finally selected the model with the best performance instead of the consensus model (Yan et al., 2021). Herein, the XGBoost and RF models with the best parameter combinations were used to predict the IGC50 and liver toxicity, respectively. To avoid chance prediction of machine learning models, the additional Y-scrambling permutation tests for finally adopted models (i.e., RF and XGBoost models) were also performed. Briefly, we constructed 100 random machine learning models, where input spectra features remain the same but output toxicity values undergo different permutations. As shown in Fig. S4, both XGBoost and RF models showed better prediction performance than the random models, indicating the prediction results were not obtained by chance.

As shown in Fig. 3, most experimental values of IGC50 were close to the model predictions, and the liver toxicity of most compounds was also correctly labeled. For machine learning models, discrimination of similar compounds (e.g., the homologues and isomers) will show a great impact on the model performance, because the predicted toxicity value

Table 1

The performance of various machine learning models for IGC50 prediction.

| Model | XGBoost | | RF | | BRR | | SVM | | FCNN | | CNN | |
|----------------|---------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| | Train | Test | Train | Test | Train | Test | Train | Test | Train | Test | Train | Test |
| R ² | 0.705 | 0.734 | 0.676 | 0.704 | 0.475 | 0.555 | 0.586 | 0.599 | 0.617 | 0.657 | 0.598 | 0.640 |
| RMSE | 0.488 | 0.440 | 0.511 | 0.464 | 0.651 | 0.569 | 0.578 | 0.541 | 0.556 | 0.500 | 0.569 | 0.512 |
| MAE | 0.384 | 0.353 | 0.410 | 0.375 | 0.512 | 0.439 | 0.445 | 0.401 | 0.431 | 0.376 | 0.447 | 0.405 |

Table 2

The performance of various machine learning models for liver toxicity classification.

| Model | XGBoost | | RF | | kNN | | SVM | | FCNN | | CNN | |
|-------------------|---------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| | Train | Test | Train | Test | Train | Test | Train | Test | Train | Test | Train | Test |
| Balanced accuracy | 0.710 | 0.735 | 0.734 | 0.723 | 0.704 | 0.708 | 0.677 | 0.665 | 0.756 | 0.753 | 0.583 | 0.582 |
| F1 Score | 0.835 | 0.863 | 0.856 | 0.854 | 0.776 | 0.791 | 0.821 | 0.821 | 0.844 | 0.850 | 0.809 | 0.814 |
| AUC-ROC | 0.710 | 0.735 | 0.734 | 0.723 | 0.704 | 0.708 | 0.677 | 0.665 | 0.756 | 0.753 | 0.583 | 0.582 |

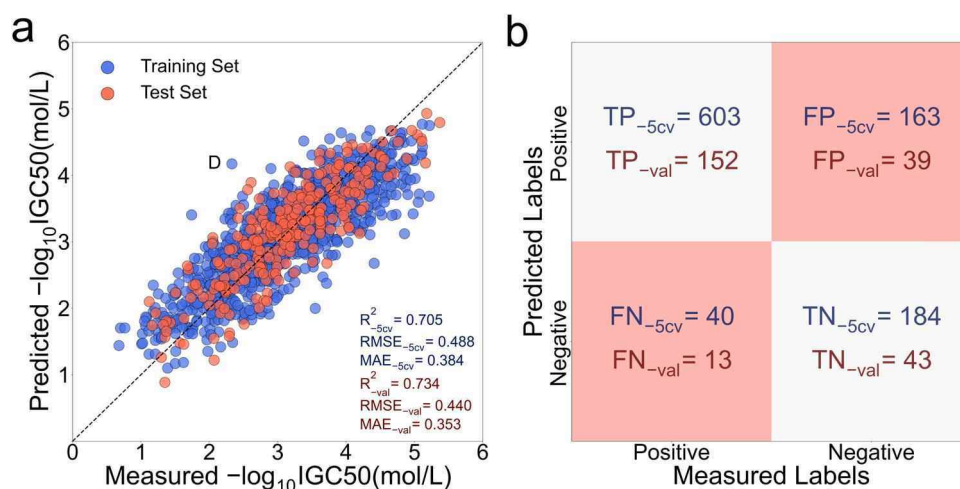


Fig. 3. Model predictions of the optimal machine learning models. (a) Correlations between the experimental values of IGC50 and the predictions from the XGBoost model. The D data point represents a prediction outlier generated from the XGBoost model. The dashed line is $y = x$. (b) Confusion matrix of the classification models for liver toxicity prediction using RF method. 5CV and val indicate the results from 5-fold cross validation and test set validation, respectively.

of one molecule was usually calculated from the experimental toxicity values of other structural analogs in the training set. In the present study, the high prediction accuracy indicated that most similar compounds can be well discriminated. For instance, although the alcohols in IGC50 set differed from each other by only one or more $-\text{CH}_2$, these molecules can still be effectively distinguished by their EI-MS (Fig. 4a). Furthermore, the machine learning model can effectively extract

spectral features from the EI-MS and then accurately predict the toxicity of these compounds (Fig. 4b). The results showed that the constructed machine learning model could learn slight differences in the EI-MS to make accurate judgments about the toxicity of organic compounds. However, some prediction outliers were also noticeable. For example, the experimental IGC50 value of the D data point (Fig. 3a) was 2.33, but the prediction from the XGBoost model was 4.17. Since the D data point

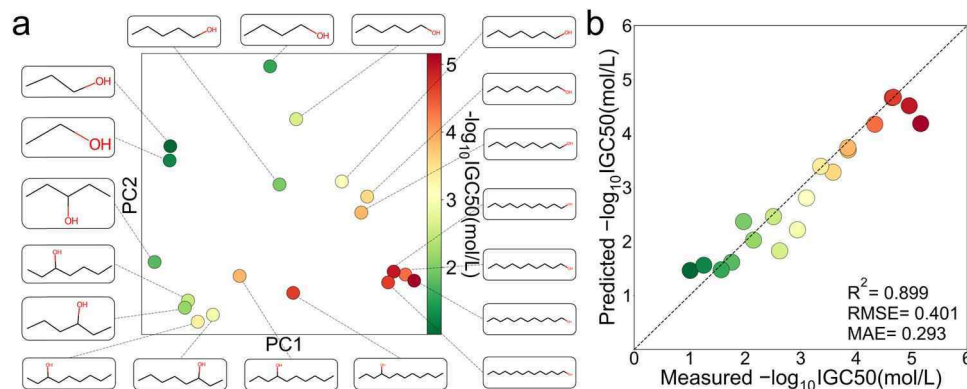


Fig. 4. Discrimination and toxicity prediction of homologues in the IGC50 set. (a) PCA of 19 alcohols in the IGC50 set. (b) Correlations between the experimental toxicity values of 19 alcohols and the predictions from the XGBoost model. The structures of 19 alcohols were also shown in the figure.

belonged to the training set, we checked its structural analog (its IGC50 was 3.62) in the other four folds of the training set. As shown in Fig. S5, the EI-MS of D1 and its nearest structural analog almost overlapped. Obviously, the current features extracted from EI-MS cannot effectively distinguish these structural analogs. Similarly, as we expected that the structural outlier (Fig. 2d) would lead to a large prediction error, it was indeed misclassified by the RF model. These prediction outliers suggest a direction for future improvement of the spectrum-based features.

3.3. Mechanism analysis of chemical-induced toxicity using spectrum-based features

According to the OECD principles, a machine learning model should be explainable. Analysis of the machine learning models allowed us to identify a number of molecular descriptors (i.e., spectrum-based features) responsible for the model performance, which can be used to elucidate the potential mechanism of chemical-induced toxicity. In the present study, the ranking of molecular descriptors was generated from the optimal machine learning models, i.e., XGBoost models for IGC50 and RF models for liver toxicity. Herein, the built-in function, *feature importances*, was used to compute the contributions of different molecular descriptors. The high ranking of a descriptor indicated its critical contribution to the machine learning model.

As shown in Fig. 5, several spectra features were found to be important for IGC50 and liver toxicity. Overall, many more peaks showed important effects on liver toxicity than the peaks showing on IGC50. In addition, the importance of different peaks for liver toxicity did not show much difference (Fig. 5b). Considering the imbalance of DILI set (i.e., the positive samples were more abundant than negative samples), we wondered if too many important features were induced by the imbalanced data. To verify our hypothesis, a balanced data set was randomly selected from the DILI set to build the machine learning model. The data set contained all 429 negative samples and 429 randomly selected positive samples. As shown in Fig. S6a, the optimal RF model obtained acceptable accuracy (i.e., higher than 0.75 for both training set and test set). However, the feature importance did not show much difference with the result from the imbalanced data (Fig. S6b and Fig. 5b). The results indicated that the machine learning model built for liver toxicity was more complex than that for IGC50, and it was difficult to distinguish different molecules in the DILI set by a few spectral features.

To explore the reason why these spectral features played important roles in the machine learning models, the nature of the chemical-induced toxicity was correlated to the structural alerts represented by the corresponding m/z values. As for IGC50, the most important m/z value was 91 (Fig. 5a), which accounted for 11.69% of all variables used in the XGBoost model. In an EI-MS, one single peak may represent several kinds of fragment ions. To identify the possible structure alerts

represented by the peak at $m/z = 91$, we checked the EI-MS of several highly toxic compounds containing the peak at $m/z = 91$ in the IGC50 set. As shown in Fig. S7, the peak at $m/z = 91$ can be induced by benzyl or other fragment ions in the organic compounds. The result suggested that although these compounds could inhibit the survival of *T. piriformis*, the toxic mechanism may be different. For the DILI set, the peak at $m/z = 75$ played the most important role in liver toxicity prediction. Similarly, we also checked the related compounds in the DILI set. It can be seen that the peak was mainly induced by a subgroup of chlorobenzene (Fig. S8). Compared with aliphatic organic compounds, the aromatic organic compounds (e.g., polycyclic aromatic hydrocarbons and polychlorinated biphenyls) were more stable and difficult to decompose, which could cause higher toxicity to the organism and severe pollution to the environment (Gao et al., 2018; Grimm et al., 2015). To be more efficient for green design, we further checked the possible structural alerts represented by other important peaks (e.g., the peak at $m/z = 11$ in IGC50 set and the peak at $m/z = 52$ in DILI set). As shown in Fig. S9 and Fig. S10, several other substructures (e.g., benzyl chloride) were also critical structural alerts that could cause *T. piriformis* growth inhibition or liver toxicity. The above findings provided new insights into understanding the toxicity mechanisms of organic compounds from the perspective of EI-MS. The ranking of more spectral features (Top 50 mass-to-charge ratios) can be seen in Table S4.

3.4. The effect of applicability domain and experimental validation

As described above, the model performance was not the same across different clusters of molecules. The prediction cannot be reliable if a compound is too dissimilar from others in the training set. Introducing the AD to the machine learning model can help us estimate the uncertainty in the prediction of a particular molecule. Compared with the prediction of molecules outside AD, the prediction of molecules within AD was more reliable. In the present study, we explored the effect of varying the threshold values of AD on the prediction results. Fig. 6 showed the model performances when applying different threshold values to predict the compounds of the test set. Overall, the effect of AD on liver toxicity prediction was more significant than that on IGC50 prediction. The difference was mainly caused by the evaluation metrics for classification and regression models. As depicted in Eq. 2 and Eq. 6, the removal of a prediction outlier had a greater impact on the accuracy score than on R^2 . As shown in Fig. 6, the coverage of the test set decreased after introducing AD, but the model performance improved. For the IGC50 set, the R^2 value for the RF model and XGBoost model improved from 0.68 to 0.79 when the coverage of the test set decreased from 90% to 40%. For the DILI set, implementing AD improved the balanced accuracy score for the XGBoost model and RF model from 0.72 to 0.91 when the coverage of the test set decreased from 90% to 30%. As one of the principles of QSAR models for regulatory purposes, the

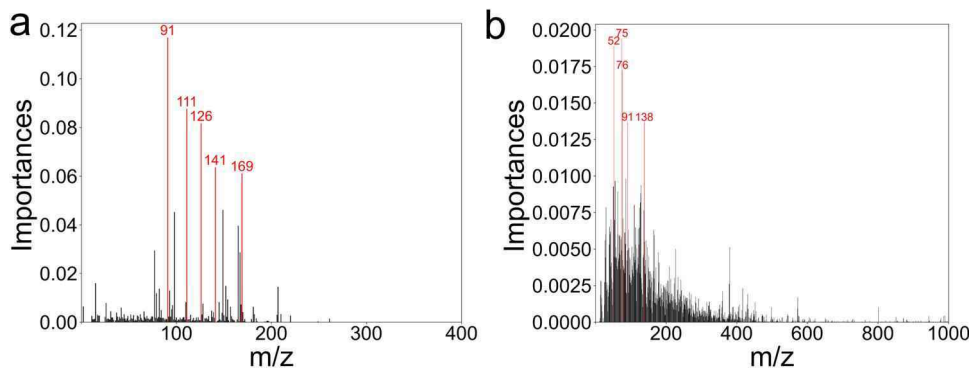


Fig. 5. Importance analysis of spectra features. The ranking of different m/z values related to (a) IGC50 and (b) liver toxicity. The feature importance was calculated from optimal models, i.e., (a) XGBoost and (b) RF models. The top5 ranking m/z values are marked in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

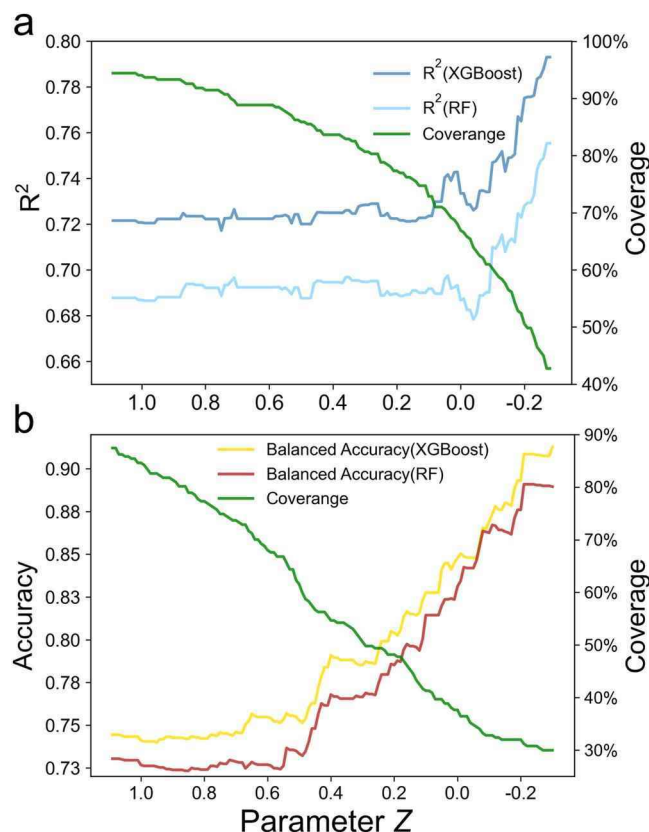


Fig. 6. The effect of AD on model predictions. Predictivity of XGBoost and RF models for (a) IGC50 and (b) liver toxicity using different ADs. The coverage of the test set is also shown in the figure.

application of AD could help regulatory agencies make more reasonable regulatory decisions (Gramatica, 2007).

The proposed method was further verified on an external data set, which contained ten new compounds that were not included in the IGC50 set. These compounds covered multiple types, such as alcohol, ketone and ester. The chemical structures and their corresponding EI-MS were shown in Fig. S11. As described above, the preprocessed spectral data of these new compounds were imported into the machine learning model without considering the specific structures. Herein, the optimal XGBoost model was used to predict the IGC50 of these new compounds. The experimental IGC50 of each compound was calculated from the dose-response curve (Fig. S12). As shown in Table 3, the toxicity of most compounds can be accurately evaluated, indicating that the constructed machine learning model could provide effective support for the toxicity prediction of unknown chemicals, even without knowing their chemical structures. Especially, the predictions for the compounds within AD were very close to the experimental values ($R^2 = 0.852$, $MAE = 0.220$, $RMSE = 0.154$; Table 3 and Fig. S13), while the predictions for the compounds outside AD exhibited much difference with the experimental values (absolute error > 0.8; Table 3 and Fig. S13). These results further indicated that AD should be a prudent consideration when the QSAR

model was used to make a regulatory decision.

4. Pitfalls and perspectives

4.1. Discrimination of similar compounds

Discrimination of similar compounds can reflect the validity of spectral features and the learning ability of machine learning models. In this study, we further explored the discrimination of isomers by spectral features extracted from EI-MS. Here, 211 nonylphenol isomers were used (Zenkevich et al., 2009). Considering the lack of experimental EI-MS and toxicity data, we applied the simulated EI-MS (Wei et al., 2019) to perform unsupervised learning (i.e., PCA). Although most isomers can be effectively distinguished, some isomers still overlapped (Fig. S14). In the future, more advanced machine learning methods or the combination of other spectral features (e.g., NMR and infrared spectra data) are needed for the discrimination of structural analogs.

4.2. The inadequacy of experimental spectral data

When considering the broad chemical space or even the number of known compounds in public databases, the availability of experimental spectral data is quite limited. Previous QSDAR models successfully constructed relationships between the predicted 1D ^{13}C NMR data and biological activity (Beger and Wilkes, 2001), our results also led us believe that the predicted EI-MS data was feasible. With advanced computational software or predictive models, it is expected that *in silico* spectra data can effectively complement the inadequacy of experimental spectra data.

4.3. The potential application of the proposed method

The main purpose of the proposed method is to perform toxicity prediction when it is difficult to determine the chemical structures. In the present study, the proposed method was proved to be effective for toxicity prediction of unknown compounds. In the future, we plan to apply the proposed method for chemical mixtures in complex water environment (Escher et al., 2020) or pharmaceuticals (such as traditional Chinese medicines) (Wong et al., 2016). Considering the lack of toxicity data of mixtures, the transfer learning can be developed on a pre-trained model that was used for 'pure' compound. To do this, more comprehensive spectra data should be harvested from other spectral libraries, such as MassBank (Horai et al., 2010) and GNPS (Wang et al., 2016).

4.4. Systematic classification of unknown compounds

As described above, the proposed method has the potential to be applied for toxicity prediction of chemical mixtures in complex environments, and non-targeted tandem mass spectrometry can detect thousands of molecules in such cases. Recently, some advanced methods, such as CANOPUS (Dührkop et al., 2021), can predict compound classes even for which neither spectral nor structural reference data are available. The systematic classification of unknown compounds can help us identify toxicophores and make reasonable regulation when

Table 3
Toxicity prediction of ten new compounds not included in the IGC50 set.

| NO. ^a | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|--------------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| Pred. ^a | 2.656 | 2.916 | 2.866 | 2.562 | 3.210 | 3.649 | 3.771 | 3.978 | 4.380 | 3.869 |
| AD ^a | in | in | in | in | in | in | in | in | out | out |
| Exp. ^a | 2.657 | 2.936 | 2.812 | 2.638 | 3.328 | 3.496 | 3.455 | 4.467 | 5.259 | 2.167 |
| Abs. ^a | 0.001 | 0.020 | 0.054 | 0.078 | 0.118 | 0.153 | 0.316 | 0.489 | 0.879 | 1.702 |

^a The NO., Pred., AD, Exp., and Abs. were abbreviations for the serial number of compounds, predicted value, applicability domain, experimental value, absolute error, respectively. The unit of all the values was $-\log(\text{mol/L})$.

the toxicity of mixtures was predicted by our proposed method.

5. Concluding remarks

Accurate annotation of chemical structures was a prerequisite of QSAR modeling to predict the chemical-induced toxicity. However, the determination of a chemical structure required extremely advanced methods and complicated steps. Especially, in many cases, such as in complex water environments, it is almost impossible to identify an unknown compound from thousands of compounds. In the present study, we proposed a machine learning-based method to link the EI-MS of compounds to the toxicity endpoint. After a series of mathematical transformations, the processed spectral features were imported into the machine learning models. The constructed models exhibited high predictability of *T. pyriformis* growth inhibition and liver toxicity induced by organic chemicals. Feature important analysis helped us understand the potential toxicity mechanism from the perspective of EI-MS. It would be useful to adopt the above method to predict the chemical toxicity when it is difficult to determine the chemical structures. The focus of our future research will be the extension of the proposed method in predicting the toxicity of mixtures. In addition, the combination with other spectra (e.g., Raman Spectroscopy and Infrared Spectroscopy) will be useful to distinguish similar chemicals, thereby improving the model performance. Overall, the proposed method is a good complement to existing QSAR models that needs to know the chemical structures before toxicity prediction.

CRedit authorship contribution statement

Xiliang Yan: Conceived the idea. **Xiliang Yan, Song Hu:** Designed experiments. **Song Hu, Jin Zhang, Jiachen Yan:** Collected the data. **Song Hu:** Constructed the model. **Guohong Liu, Hongyu Zhou:** Performed the experimental validation. **Xiliang Yan, Song Hu:** Co-wrote the manuscript and all authors discussed and approved the paper.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (22106025, 22036002), the Introduced Innovative R&D Team Project under the “The Pearl River Talent Recruitment Program” of Guangdong Province (2019ZT08L387). We thank Prof. Guangbo Qu for generously providing the *T. pyriformis* strain.

Code availability

The source codes of all machine learning and deep learning models can be found at <https://github.com/YanLabAI/SpectraTox>.

Appendix A. Supporting information

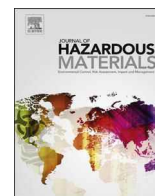
Supplementary data associated with this article can be found in the online version at [doi:10.1016/j.jhazmat.2022.128558](https://doi.org/10.1016/j.jhazmat.2022.128558).

References

- Ai, H., Chen, W., Zhang, L., Huang, L., Yin, Z., Hu, H., Zhao, Q., Zhao, J., Liu, H., 2018. Predicting drug-induced liver injury using ensemble learning methods and molecular fingerprints. *Toxicol. Sci.* 165, 100–107. <https://doi.org/10.1093/toxsci/kfy121>.
- Ajmani, S., Jadhav, K., Kulkarni, S.A., 2006. Three-dimensional QSAR using the k-nearest neighbor method and its interpretation. *J. Chem. Inf. Model.* 46, 24–31.

- Asilar, E., Hemmerich, J., Ecker, G.F., 2020. Image based liver toxicity prediction. *J. Chem. Inf. Model.* 60, 1111–1121. <https://doi.org/10.1021/acs.jcim.9b00713>.
- Beger, R.D., Wilkes, J.G., 2001. Developing ¹³C NMR quantitative spectrometric data-activity relationship (QSDAR) models of steroid binding to the corticosteroid binding globulin. *J. Comput. Aided Mol. Des.* 15, 659–669. <https://doi.org/10.1023/A:1011959120313>.
- Breiman, L., 2001. Random forests. *Mach. Learn.* 45, 5–32. <https://doi.org/10.1023/A:1010933404324>.
- Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., 2015. Xgboost: extreme gradient boosting. R Packag. version 0.4–2 1, 1–4. <https://doi.org/10.1515/9783110671124-021>.
- Ciallella, H.L., Zhu, H., 2019. Advancing computational toxicology in the big data era by artificial intelligence: data-driven and mechanism-driven modeling for chemical toxicity. *Chem. Res. Toxicol.* 32, 536–547. <https://doi.org/10.1021/acs.chemrestox.8b00393>.
- Ciriano, I.C., Bender, A., 2019. KekuleScope: prediction of cancer cell line sensitivity and compound potency using convolutional neural networks trained on compound images. *J. Cheminf.* 1–16. <https://doi.org/10.1186/s13321-019-0364-5>.
- Diamond, M.L., de Wit, C.A., Molander, S., Schreiner, P.C., Backhaus, T., Lohmann, R., Arvidsson, R., Bergman, Å., Hauschild, M., Holoubek, I., Persson, L., Suzuki, N., Vigli, M., Zetzsch, C., 2015. Exploring the planetary boundary for chemical pollution. *Environ. Int.* 78, 8–15. <https://doi.org/10.1016/j.envint.2015.02.001>.
- Dührkop, K., Fleischauer, M., Ludwig, M., Aksenov, A.A., Melnik, A.V., Meusel, M., Dorrestein, P.C., Rousu, J., Böcker, S., 2019. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat. Methods* 16, 299–302. <https://doi.org/10.1038/s41592-019-0344-8>.
- Dührkop, K., Nothias, L.F., Fleischauer, M., Reher, R., Ludwig, M., Hoffmann, M.A., Petras, D., Gerwick, W.H., Rousu, J., Dorrestein, P.C., Böcker, S., 2021. Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nat. Biotechnol.* 39, 462–471. <https://doi.org/10.1038/s41587-020-0740-8>.
- Escher, B.I., Stapleton, H.M., Schymanski, E.L., 2020. Tracking complex mixtures of chemicals in our changing environment. *Science* 367, 388–392.
- Feider, C.L., Krieger, A., Dehoog, R.J., Eberlin, L.S., 2019. Ambient ionization mass spectrometry: recent developments and applications. *Anal. Chem.* 91, 4266–4290. <https://doi.org/10.1021/acs.analchem.9b00807>.
- Gago-Ferrero, P., Schymanski, E.L., Bletsou, A.A., Aalizadeh, R., Hollender, J., Thomaidis, N.S., 2015. Extended suspect and non-target strategies to characterize emerging polar organic contaminants in raw wastewater with LC-HRMS/MS. *Environ. Sci. Technol.* 49, 12333–12341. <https://doi.org/10.1021/acs.est.5b03454>.
- Gao, P., da Silva, E., Hou, L., Denslow, N.D., Xiang, P., Ma, L.Q., 2018. Human exposure to polycyclic aromatic hydrocarbons: metabolomics perspective. *Environ. Int.* 119, 466–477. <https://doi.org/10.1016/j.envint.2018.07.017>.
- Golbraikh, A., Shen, M., Xiao, Z., Xiao, Y., De Lee, K.H., Tropsha, A., 2003. Rational selection of training and test sets for the development of validated QSAR models. *J. Comput. Aided Mol. Des.* 17, 241–253. <https://doi.org/10.1023/A:1025386326946>.
- Gramatica, P., 2007. Principles of QSAR models validation: internal and external. *QSAR Comb. Sci.* 26, 694–701. <https://doi.org/10.1002/qsar.200610151>.
- Grimm, F.A., Hu, D., Kania-Korwel, I., Lehmler, H.J., Ludewig, G., Hornbuckle, K.C., Duffel, M.W., Bergman, Å., Robertson, L.W., 2015. Metabolism and metabolites of polychlorinated biphenyls. *Crit. Rev. Toxicol.* 45, 245–272. <https://doi.org/10.3109/10408444.2014.999365>.
- Horai, H., Arita, M., Kanaya, S., Nihei, Y., Ikeda, T., Suwa, K., Ojima, Y., Tanaka, Kenichi, Tanaka, S., Aoshima, K., Oda, Y., Kakazu, Y., Kusano, M., Tohge, T., Matsuda, F., Sawada, Y., Hirai, M.Y., Nakanishi, H., Ikeda, K., Akimoto, N., Maoka, T., Takahashi, H., Ara, T., Sakurai, N., Suzuki, H., Shibata, D., Neumann, S., Iida, T., Tanaka, Ken, Funatsu, K., Matsuura, F., Soga, T., Taguchi, R., Saito, K., Nishioka, T., 2010. MassBank: a public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.* 45, 703–714. <https://doi.org/10.1002/jms.1777>.
- Ji, H., Deng, H., Lu, H., Zhang, Z., 2020. Predicting a molecular fingerprint from an electron ionization mass spectrum with deep neural networks. *Anal. Chem.* 92, 8649–8653. <https://doi.org/10.1021/acs.analchem.0c01450>.
- Judson, R., Richard, A., Dix, D.J., Houck, K., Martin, M., Kavlock, R., Dellarco, V., Henry, T., Holderman, T., Sayre, P., Tan, S., Carpenter, T., Smith, E., 2009. The toxicity data landscape for environmental chemicals. *Environ. Health Perspect.* 117, 685–695. <https://doi.org/10.1289/ehp.0800168>.
- Karatzoglou, A., Meyer, D., Hornik, K., 2006. Support Vector Mach. *R. J. Stat. Softw.* 15, 1–28.
- Kavlock, R.J., Bahadori, T., Barton-Maclaren, T.S., Gwinn, M.R., Rasenberg, M., Thomas, R.S., 2018. Accelerating the pace of chemical risk assessment. *Chem. Res. Toxicol.* 31, 287–290. <https://doi.org/10.1021/acs.chemrestox.7b00339>.
- Kind, T., Tsugawa, H., Cajka, T., Ma, Y., Lai, Z., Mehta, S.S., Wohlgemuth, G., Barupal, D. K., Showalter, M.R., Arita, M., Fiehn, O., 2018. Identification of small molecules using accurate mass MS/MS search. *Mass Spectrom. Rev.* 37, 513–532. <https://doi.org/10.1002/mas.21535>.
- Landrigan, P.J., Fuller, R., Acosta, N.J.R., Adeyi, O., Arnold, R., Basu, N., (Nil), Baldé, A. B., Bertollini, R., Bose-O'Reilly, S., Boufford, J.L., Breysse, P.N., Chiles, T., Mahidol, C., Coll-Seck, A.M., Cropper, M.L., Fobil, J., Fuster, V., Greenstone, M., Haines, A., Hanrahan, D., Hunter, D., Khare, M., Krupnick, A., Lanphear, B., Lohani, B., Martin, K., Mathiasen, K.V., McTeer, M.A., Murray, C.J.L., Ndahimananjara, J.D., Perera, F., Potočník, J., Preker, A.S., Ramesh, J., Rockström, J., Salinas, C., Samson, L.D., Sandilya, K., Sly, P.D., Smith, K.R., Steiner, A., Stewart, R.B., Suk, W.A., van Schayck, O.C.P., Yadama, G.N., Yumkella, K., Zhong, M., 2018. The Lancet commission on pollution and health. *Lancet* 391, 462–512. [https://doi.org/10.1016/S0140-6736\(17\)32345-0](https://doi.org/10.1016/S0140-6736(17)32345-0).

- LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. *Nature* 521, 436–444.
- Li, C., Martin, R.V., Van Donkelaar, A., Boys, B.L., Hammer, M.S., Xu, J.W., Marais, E.A., Reff, A., Strum, M., Ridley, D.A., Crippa, M., Brauer, M., Zhang, Q., 2017. Trends in chemical composition of global and regional population-weighted fine particulate matter estimated for 25 years. *Environ. Sci. Technol.* 51, 11185–11195. <https://doi.org/10.1021/acs.est.7b02530>.
- Wang, L., 2019. CAS reaches 150 millionth substance. *CEN Glob. Enterp.* 97 <https://doi.org/10.1021/cen-09722-acnews1>.
- Muratov, E.N., Bajorath, J., Sheridan, R.P., Tetko, I.V., Filimonov, D., Poroikov, V., Oprea, T.I., Baskin, I.I., Varnek, A., Roitberg, A., Isayev, O., Curtalolo, S., Fourches, D., Cohen, Y., Aspuru-Guzik, A., Winkler, D.A., Agrafiotis, D., Cherkasov, A., Tropsha, A., 2020. QSAR without borders. *Chem. Soc. Rev.* 49, 3525–3564. <https://doi.org/10.1039/d0cs00098a>.
- Picache, J.A., May, J.C., McLean, J.A., 2020. Chemical class prediction of unknown biomolecules using ion mobility-mass spectrometry and machine learning: supervised inference of feature taxonomy from ensemble randomization. *Anal. Chem.* 92, 10759–10767. <https://doi.org/10.1021/acs.analchem.0c02137>.
- Piir, G., Kahn, I., García-Sosa, A.T., Sild, S., Ahte, P., Maran, U., 2018. Best practices for QSAR model reporting: physical and chemical properties, ecotoxicity, environmental fate, human health, and toxicokinetics endpoints. *Environ. Health Perspect.* 126, 1–20. <https://doi.org/10.1289/EHP3264>.
- Ronneberger, O., Tunyasuvunakool, K., Bates, R., Židek, A., Ballard, A.J., Cowie, A., Romera-paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu, K., 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589. <https://doi.org/10.1038/s41586-021-03819-2>.
- Russo, D.P., Strickland, J., Karmaus, A.L., Wang, W., Shende, S., Hartung, T., Aleksunes, L.M., Zhu, H., 2019. Nonanimal models for acute toxicity evaluations: applying data-driven profiling and read-across. *Environ. Health Perspect.* 127, 3614. <https://doi.org/10.1289/EHP3614>.
- Schultz, T.W., Sinks, G.D., Cronin, M.T.D., 1997. Quinone-induced toxicity to tetrahymena: structure-activity relationships. *Aquat. Toxicol.* 39, 267–278. [https://doi.org/10.1016/S0166-445X\(97\)00031-3](https://doi.org/10.1016/S0166-445X(97)00031-3).
- She, T.Y., Judson, R.S., 2019. Ensemble QSAR modeling to predict multispecies fish toxicity lethal concentrations and points of departure. *Environ. Sci. Technol.* 53, 12793–12802. <https://doi.org/10.1021/acs.est.9b03957>.
- Shi, Q., Abdel-Aty, M., Lee, J., 2016. A bayesian ridge regression analysis of congestion's impact on urban expressway safety. *Accid. Anal. Prev.* 88, 124–137. <https://doi.org/10.1016/j.aap.2015.12.001>.
- Tie, Y., McPhail, B., Hong, H., Pearce, B.A., Schnackenberg, L.K., Ge, W., Buzatu, D.A., Wilkes, J.G., Fuscoe, J.C., Tong, W., Fowler, B.A., Beger, R.D., Demchuk, E., 2012. Modeling chemical interaction profiles: II. Molecular docking, spectral data-activity relationship, and structure-activity relationship models for potent and weak inhibitors of cytochrome P450 CYP3A4 isozyme. *Molecules*. <https://doi.org/10.3390/molecules17033407>.
- Wang, M., Carver, J.J., Phelan, V.V., Sanchez, L.M., Garg, N., Peng, Y., Nguyen, D.D., Watrous, J., Kapono, C.A., Luzzatto-Knaan, T., Porto, C., Bouslimani, A., Melnik, A. V., Meehan, M.J., Liu, W.T., Crisemann, M., Boudreau, P.D., Esquenazi, E., Sandoval-Calderón, M., Kersten, R.D., Pace, L.A., Quinn, R.A., Duncan, K.R., Hsu, C. C., Floros, D.J., Gavilan, R.G., Kleigrew, K., Northen, T., Dutton, R.J., Parrot, D., Carlson, E.E., Aigle, B., Michelsen, C.F., Jelsbak, L., Sohlenkamp, C., Pevzner, P., Edlund, A., McLean, J., Piel, J., Murphy, B.T., Gerwick, L., Liaw, C.C., Yang, Y.L., Humpf, H.U., Maansson, M., Keyzers, R.A., Sims, A.C., Johnson, A.R., Sidebottom, A. M., Sedio, B.E., Klitgaard, A., Larson, C.B., Boya, C.A.P., Torres-Mendoza, D., Gonzalez, D.J., Silva, D.B., Marques, L.M., Demarque, D.P., Pociute, E., O'Neill, E.C., Briand, E., Helfrich, E.J.N., Granatosky, E.A., Glukhov, E., Ryffel, F., Houson, H., Mohimani, H., Kharbush, J.J., Zeng, Y., Vorholt, J.A., Kurita, K.L., Charusanti, P., McPhail, K.L., Nielsen, K.F., Vuong, L., Elfeki, M., Traxler, M.F., Engene, N., Koyama, N., Vining, O.B., Baric, R., Silva, R.R., Mascuch, S.J., Tomasi, S., Jenkins, S., Macherla, V., Hoffman, T., Agarwal, V., Williams, P.G., Dai, J., Neupane, R., Gurr, J., Rodriguez, A.M.C., Lamsa, A., Zhang, C., Dorresteijn, K., Duggan, B.M., Almaliti, J., Allard, P.M., Phapale, P., Nothias, L.F., Alexandrov, T., Litaudon, M., Wolfender, J. L., Kyle, J.E., Metz, T.O., Peryea, T., Nguyen, D.T., VanLeer, D., Shinn, P., Jadhav, A., Müller, R., Waters, K.M., Shi, W., Liu, X., Zhang, L., Knight, R., Jensen, P.R., Palsson, B., Pogliano, K., Linington, R.G., Gutiérrez, M., Lopes, N.P., Gerwick, W.H., Moore, B.S., Dorresteijn, P.C., Bandeira, N., 2016. Sharing and community curation of mass spectrometry data with global natural products social molecular networking. *Nat. Biotechnol.* 34, 828–837. <https://doi.org/10.1038/nbt.3597>.
- Wei, J.N., Belanger, D., Adams, R.P., Sculley, D., 2019. Rapid prediction of electron-ionization mass spectrometry using neural networks. *ACS Cent. Sci.* 5, 700–708. <https://doi.org/10.1021/acscentsci.9b00085>.
- Wong, M.Y.M., So, P.K., Yao, Z.P., 2016. Direct analysis of traditional Chinese medicines by mass spectrometry. *J. Chromatogr. B Anal. Technol. Biomed. Life Sci.* 1026, 2–14. <https://doi.org/10.1016/j.jchromb.2015.11.032>.
- Xu, L., Ren, J.S.J., Liu, C., Jia, J., 2014. Deep convolutional neural network for image deconvolution. *Adv. Neural Inf. Process. Syst.* 2, 1790–1798.
- Yan, J., Yan, X., Hu, S., Zhu, H., Yan, B., 2021. Comprehensive interrogation on acetylcholinesterase inhibition by ionic liquids using machine learning and molecular modeling. *Environ. Sci. Technol.* 55, 14720–14731. <https://doi.org/10.1021/acs.est.1c02960>.
- Yan, X., Sedykh, A., Wang, W., Yan, B., Zhu, H., 2020. Construction of a web-based nanomaterial database by big data curation and modeling friendly nanostructure annotations. *Nat. Commun.* 11, 1–10. <https://doi.org/10.1038/s41467-020-16413-3>.
- Yan, X., Sedykh, A., Wang, W., Zhao, X., Yan, B., Zhu, H., 2019. *In silico* profiling nanoparticles: predictive nanomodeling using universal nanodescriptors and various machine learning approaches. *Nanoscale* 11, 8352–8362. <https://doi.org/10.1039/C9NR00844F>.
- Zenkovich, I.G., Makarov, A.A., Schrader, S., Moeder, M., 2009. A new version of an additive scheme for the prediction of gas chromatographic retention indices of the 211 structural isomers of 4-nonylphenol. *J. Chromatogr. A* 1216, 4097–4106. <https://doi.org/10.1016/j.chroma.2009.03.021>.
- Zhang, L., Zhu, H., Oprea, T.I., Golbraikh, A., Tropsha, A., 2008. QSAR modeling of the blood-brain barrier permeability for diverse organic compounds. *Pharm. Res.* 25, 1902–1914. <https://doi.org/10.1007/s11095-008-9609-0>.
- Zorn, K.M., Foil, D.H., Lane, T.R., Russo, D.P., Hillwalker, W., Feifarek, D.J., Jones, F., Klaren, W.D., Brinkman, A.M., Ekins, S., 2020. Machine learning models for estrogen receptor bioactivity and endocrine disruption prediction. *Environ. Sci. Technol.* 54, 12202–12213. <https://doi.org/10.1021/acs.est.0c03982>.



Implementing comprehensive machine learning models of multispecies toxicity assessment to improve regulation of organic compounds

Ying He^{a,1}, Guohong Liu^{a,c,1}, Song Hu^b, Xiaohong Wang^a, Jianbo Jia^a, Hongyu Zhou^{a,*}, Xiliang Yan^{a,c,**}

^a Institute of Environmental Research at Greater Bay Area, Key Laboratory for Water Quality and Conservation of the Pearl River Delta, Ministry of Education, Guangzhou University, Guangzhou 510006, China

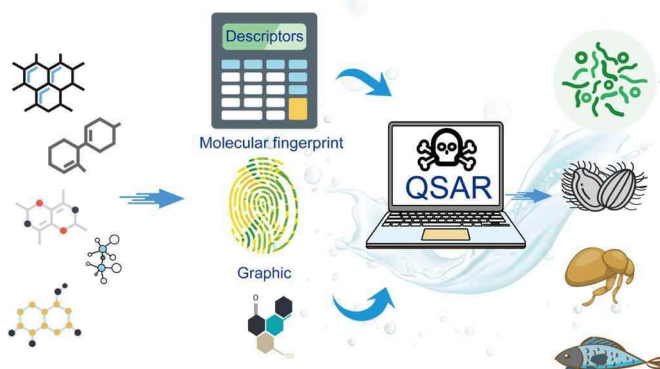
^b School of Environmental Science and Engineering, Shandong University, Qingdao 266237, China

^c School of Agriculture and Biological Sciences, Qiannan Normal University for Nationalities, Duyun 558000, China

HIGHLIGHTS

- Chemical regulation was determined by 21 machine learning models of multi-species toxicity assessment.
- The toxicity of chemicals to aquatic organisms was species sensitivity.
- The aquatic toxicity mechanism was comprehensively analyzed from the critical structure features of chemicals.
- The optimal models were used to identify high-risk chemicals from over 16,000 compounds.

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: Lingxin Chen

Keywords:

Aquatic toxicity
Chemical risk assessment
Machine learning
End-to-end deep learning
Species sensitivity

ABSTRACT

Machine learning has made significant progress in assessing the risk associated with hazardous chemicals. However, most models were constructed by randomly selecting one algorithm and one toxicity endpoint towards single species, which may cause biased regulation of chemicals. In the present study, we implemented comprehensive prediction models involving multiple advanced machine learning and end-to-end deep learning to assess the aquatic toxicity of chemicals. The generated optimal models accurately unravel the quantitative structure-toxicity relationships, with the correlation coefficients of all training sets from 0.59 to 0.81 and of the test sets from 0.56 to 0.83. For each chemical, its ecological risk was determined from the toxicity information towards multiple species. The results also revealed the toxicity mechanism of chemicals was species sensitivity, and the high-level organisms were faced with more serious side effects from hazardous substances. The proposed

* Corresponding author.

** Corresponding author at: Institute of Environmental Research at Greater Bay Area, Key Laboratory for Water Quality and Conservation of the Pearl River Delta, Ministry of Education, Guangzhou University, Guangzhou 510006, China.

E-mail addresses: hyzhou001@gzhu.edu.cn (H. Zhou), yanxiliang1991@gzhu.edu.cn (X. Yan).

¹ Y. He and G. Liu contributed equally to this work.

<https://doi.org/10.1016/j.jhazmat.2023.131942>

Received 15 April 2023; Received in revised form 12 June 2023; Accepted 24 June 2023

Available online 27 June 2023

0304-3894/© 2023 Elsevier B.V. All rights reserved.

approach was finally applied to screen over 16,000 compounds and identify high-risk chemicals. We believe that the current approach can provide a useful tool for predicting the toxicity of diverse organic chemicals and help regulatory authorities make more reasonable decisions.

1. Introduction

Water quality issues have increasingly become one of the primary challenges facing humanity in the 21st century [11,24], and most water pollution accidents are caused by chemical pollution [12,54]. Chemical contaminants in water can lead to a series of adverse effects to eco-environment and human health [25,40]. Aquatic toxicity test is an important step to evaluate the environmental hazard and risk assessment of chemicals in water. Many governments usually use the results of toxicology studies to support regulation of chemicals and other toxic substances [37]. During the past decades, toxicity testing in animals have been routinely conducted to determine the potential toxic effects of chemicals. However, the number of new chemicals is growing exponentially, and the Chemical Abstracts Service Registry has contained more than 150 million chemicals in 2019 [48]. It is impossible to evaluate the toxicity of such a large number of compounds only relying on in vivo procedures. Furthermore, the use of animals for toxicology studies raises ethical issues, the U.S. Environmental Protection Agency has decided to eliminate all mammal testing by 2035 [16]. Therefore, developing alternative methods to animal testing for chemical risk assessment is highly needed by regulatory agencies [18,30].

Previously, quantitative structure-activity relationship (QSAR) modeling has been used to predict the aquatic toxicity of chemicals, and proven to be an efficient method to decrease the number and cost of animal testing [17,2,42]. However, most current QSAR models focus on the toxicity of a small set of chemicals to specific aquatic species [9]. Furthermore, the predictions are usually based on one or two arbitrarily selected machine learning methods, often employing linear regression or read-across [32]. These critical limitations greatly restricted the external predictive ability of the QSAR models when predicting the toxicity of new chemicals not included in the modeling set. In addition, the selection of irrelevant descriptors also leads to the poor performance of QSAR models. Although several chemical toxicity assessment tools, such as ECOSAR (ecological structure-activity relationship) and TEST (toxicity estimation software tool) have been developed to predict the aquatic toxicity, these QSAR modeling tools cannot spontaneously accept user data to improve prediction accuracy.

Therefore, it is necessary to develop novel techniques that could take advantage of public big data for aquatic toxicity prediction. The emergence of some advanced machine learning algorithms, especially deep learning methods, provides the possibility to solve these challenges. In previous studies, deep neural networks (DNN) with multi-task learning exhibited better performance in predicting chemical toxicity [31] and protein-ligand interactions [56]. Some end-to-end deep learning methods, such as convolutional neural networks (CNN) and recurrent neural networks (RNN), can directly predict the molecular properties through images or Simplified Molecular Input Line Entry System (SMILES) format of chemicals without descriptors calculation [27,3,38,8].

In order to identify the most appropriate models for regulatory purpose, we developed a series of QSAR models based on various machine learning and end-to-end deep learning algorithms. These QSAR models were verified on the toxicity datasets of organic chemicals towards four aquatic organisms, i.e., *fathead minnow*, *daphnia magna*, *tetrahymena pyriformis* and *vibrio fischeri*. Although most models can achieve acceptable prediction accuracy, the model performance varies greatly with algorithms and descriptors used. Results from feature importance analysis revealed that the aquatic toxicity of organic chemicals were highly correlated to the oil-water partition coefficient, aqueous solubility and molecular weight. The optimal model was further

selected to screen potentially toxic chemicals from the Inventory of Existing Chemical Substances in China (IECSC). Our results demonstrated that chemicals were prone to exhibit severer toxicity in high-level organisms. The present study provided an important supplement for the ecotoxicological assessment of organic chemicals, allowing stakeholders to make reasonable regulatory decisions.

2. Materials and methods

2.1. Overview of the datasets

As shown in Table S1, we trained and validated the QSAR models for regulatory purpose using four different datasets, i.e., 15-minute *Vibrio fischeri* 50% bioluminescence inhibition concentration dataset (IBC50 set), 40-hour *Tetrahymena pyriformis* 50% growth inhibition concentration dataset (IGC50 set), 48-hour *daphnia magna* 50% lethal concentration dataset (LC50-DM set), 96-hour *fathead minnow* 50% lethal concentration dataset (LC50 set). The IBC50 set contains the concentration of 1188 organic chemicals causing the 50% bioluminescence inhibition of *Vibrio fischeri* after 15 min. Similarly, the IGC50 set measures the concentration of 1792 organic chemicals that causes the 50% growth inhibition of *Tetrahymena pyriformis* after 40 h. The LC50-DM set reports the concentration of 353 organic chemicals inducing 50% of *daphnia magna* to die after 48 h. The LC50 set involves 823 organic chemicals and aims to predict the concentration of chemicals causing 50% of *fathead minnow* to die after 96 h. The first dataset (i.e., IBC50 set) was collected from the literature [57], while the last three datasets (i.e., IGC50 set, LC50-DM set, and LC50 set) were obtained from the ECOTOX aquatic toxicity database (<http://cfpub.epa.gov/ecotox/>). These datasets covered four representative types of model organisms in water environment, which can be used to better understand the toxic effects of chemicals on different aquatic species. All data were generated from acute toxicity tests, i.e., the one-generation toxicity assay. All concentration values are transformed to -Log10 form with the unit of mol/L. Before QSAR modeling, we further curate all chemicals to ensure no salts, mixtures, and duplicates exist in each dataset. The detailed information of chemical structure curation can be seen in Method S1.

2.2. Molecular representations and machine learning methods

Herein, four types of molecular descriptors are generated for traditional machine learning models, including extended connectivity fingerprints (ECFP), functional connectivity fingerprints (FCFP), molecular access system (MACCS) keys, and molecular operating environment (MOE) descriptors (Fig. S1). Additionally, the molecules were also converted to graphs, images and SMILES for end-to-end deep learning. Herein, the ECFP, FCFP, and MACCS descriptors are calculated using the chemoinformatics package RDKit, while the MOE descriptors are generated from the MOE software [44]. The ECFP, FCFP and MACCS descriptors are three representative molecular fingerprints that can be used to describe the structure features of chemicals. The MOE descriptors contain a large amount of information (e.g., structure, shape and electronic properties) that can be used to describe the physico-chemical properties of chemicals from multiple perspectives.

The traditional machine learning methods involved Bayesian ridge regression (BRR), random forest (RF), support vector machine (SVM), and extreme gradient boost (XGBoost). The deep learning included three end-to-end methods, i.e., CNN, graph neural network (GNN), and RNN. Three recent CNN architectures, i.e., Visual Geometry Group Network (VGG) [51], ResNet [55], and DenseNet [39] were used in this study. A

newly proposed architecture called AttentiveFP is used to construct the GNN model [1]. The gate recurrent unit (GRU) of RNN is used to extract molecular features from SMILES representations [4]. All traditional machine learning models were constructed using the Scikit-learn v0.24.1 in Python. All the end-to-end deep learning models were implemented in PyTorch v1.0.2 of Python. Detailed information about molecular representations and machine learning methods can be seen in **Method S2&S3**.

2.3. QSAR model development

Individual regression model was developed using one of the end-to-end deep learning methods (i.e., VGG, ResNet, DenseNet, AttentiveFP, and GRU), and the combination of one type of descriptors (i.e., ECFP, FCFP, MACCS, and MOE descriptors) and one of the traditional machine learning methods (i.e., RF, SVM, BRR, and XGBoost) (**Method S3**). Therefore, 21 individual models were finally generated for each aquatic toxicity endpoint. Each dataset was divided into training set (80% of the whole set) and test set (20% of the whole set). All models use a unified partitioning data result. The training set was employed to construct an initial model, and the test set was used for evaluating the predictive ability of constructed models. The hyperparameters optimization was performed using grid search algorithm, the model with best combination of hyperparameters was retained and used for prediction on test sets. The regression models were evaluated by the coefficient of determination (R^2) (Eq. 1), root mean square error (RMSE) (Eq. 2), and mean absolute error (MAE) (Eq. 3). All models were validated using 5-fold cross validation and external validation.

$$R^2 = \frac{\sum_{i=1}^n (y_i^{obs} - y_i^{pred})^2}{\sum_{i=1}^n (y_i^{obs} - \bar{y}_i^{obs})^2} \quad (1)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i^{obs} - y_i^{pred})^2}{n}} \quad (2)$$

$$MAE = \frac{\sum_{i=1}^n |y_i^{obs} - y_i^{pred}|}{n} \quad (3)$$

Where, y_i^{pred} is the predicted value for each particular molecule; y_i^{obs} is the observed value for each particular molecule; \bar{y}_i^{obs} is the mean value over all molecules; and n is the number of molecules.

2.4. Chemical risk prediction for regulatory purpose

In the present study, we further employed the constructed QSAR models to screen potentially toxic chemicals from the IECSC. The IECSC mainly aims to record chemical substances that have been produced, processed, sold, used in China. The chemical structures with their SMILES are generated from a recent study [49]. Each compound was rigorously checked to ensure the data quality. There remain 16,543 compounds after removing the duplicates and inorganic molecules in training sets. The optimal models with MOE descriptors were finally applied to perform the high-throughput virtual screening. Furthermore, the applicability domain (AD) was used to identify unreliable prediction (**Method S4**).

3. Results and discussion

The core of our present study is to construct reliable machine learning models, which can be used to prioritize chemicals for further risk evaluation and regulatory purpose. To achieve this goal, a well-designed workflow was depicted in Fig. 1. The whole workflow was consisted of four main components, i.e., data generation, descriptor calculation, model construction and chemical risk assessment.

For data-driven modeling, enough high-quality data is a prerequisite to ensure the predictive ability of machine learning models. Herein, all molecules in four datasets are rigorously checked for subsequent machine learning. In the application of machine learning, the most influential factors are effective descriptors and appropriate algorithms. The appropriate descriptor determines the upper limit of the prediction

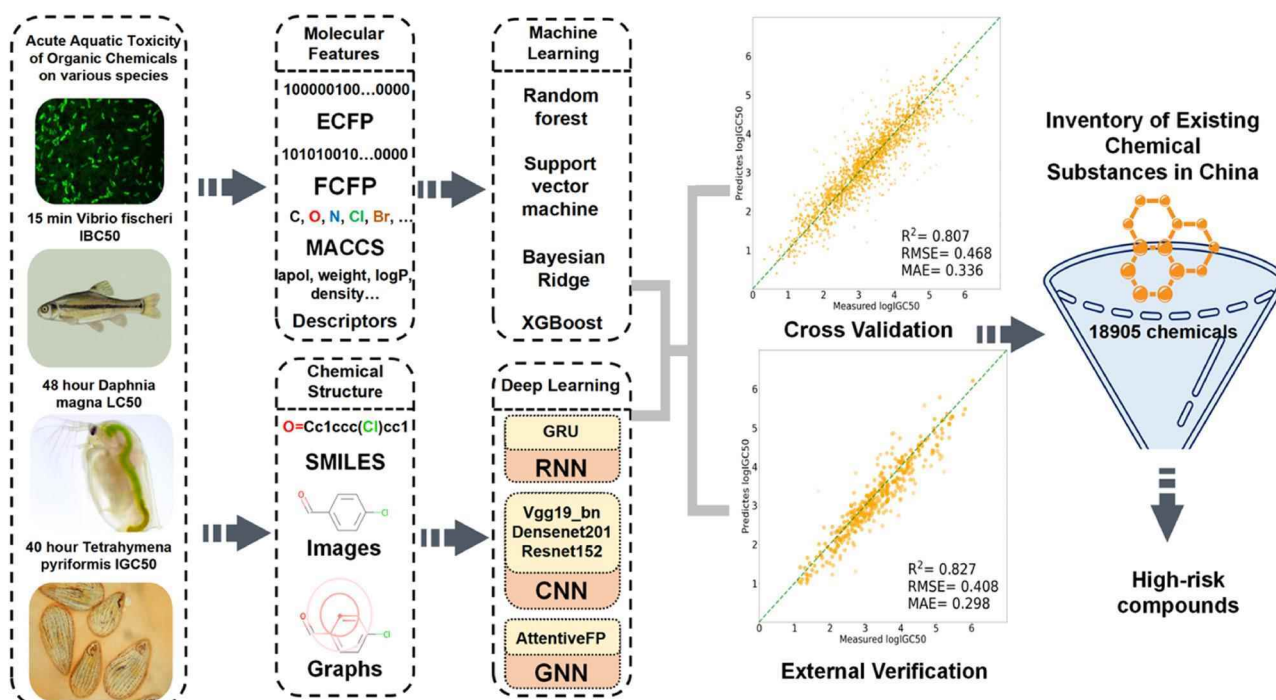


Fig. 1. Illustration of the workflow used in the present study. The whole workflow is consisted of four main steps, i.e., data generation, descriptor calculation, model construction and chemical risk assessment.

accuracy of the machine learning algorithm. Therefore, various molecular descriptors and machine learning algorithms are applied to generate possible optimal models. Furthermore, the feature importance analysis was also employed to interpret the prediction results and elucidate the possible toxicity mechanism. Finally, we use the established aquatic toxicity prediction model to screen the high-risk compounds from IECSC.

3.1. Analysis of chemical structure diversity and aquatic toxicity values

A dataset with chemical structure diversity is an important prerequisite to acquire machine learning models with high predictive accuracy and good generalization ability. To view the chemical space of organic compounds used in the present study, the t-distributed stochastic neighbor embedding (t-SNE) was applied for dimensionality reduction of MOE descriptors. The t-SNE is a non-linear dimensionality reduction technique that maps high-dimensional data to a lower-dimensional space for better visualization of the data distribution [22]. In chemical

space, t-SNE can map the high-dimensional representation of chemical descriptors to two- or three-dimensional space and display them as points on a coordinate system. Thus, t-SNE can help us discover similarities and differences between compounds, identify compound populations and clusters, and perform structure visualization and analysis. There remained 2121 compounds after removing duplicates in four datasets. The chemical space indicated that most molecules were structurally different due to various functional groups (Fig. 2a). Furthermore, the compounds of both training and test sets have a wide distribution in the chemical space, indicating the effectiveness of dataset partitioning. Additionally, the training set with chemical structure diversity was helpful to enhance the robustness of machine learning models, and the current test set was able to verify the models' generalization. According to the functional groups of organic compounds, these chemicals can be classified into 17 main categories involving hydrocarbons, organic halides, organic oxygen compounds and others (Fig. S2). Among them, benzenoids accounted for 44.41%, nearly half of the total compounds. Benzenoids include benzene, toluene,

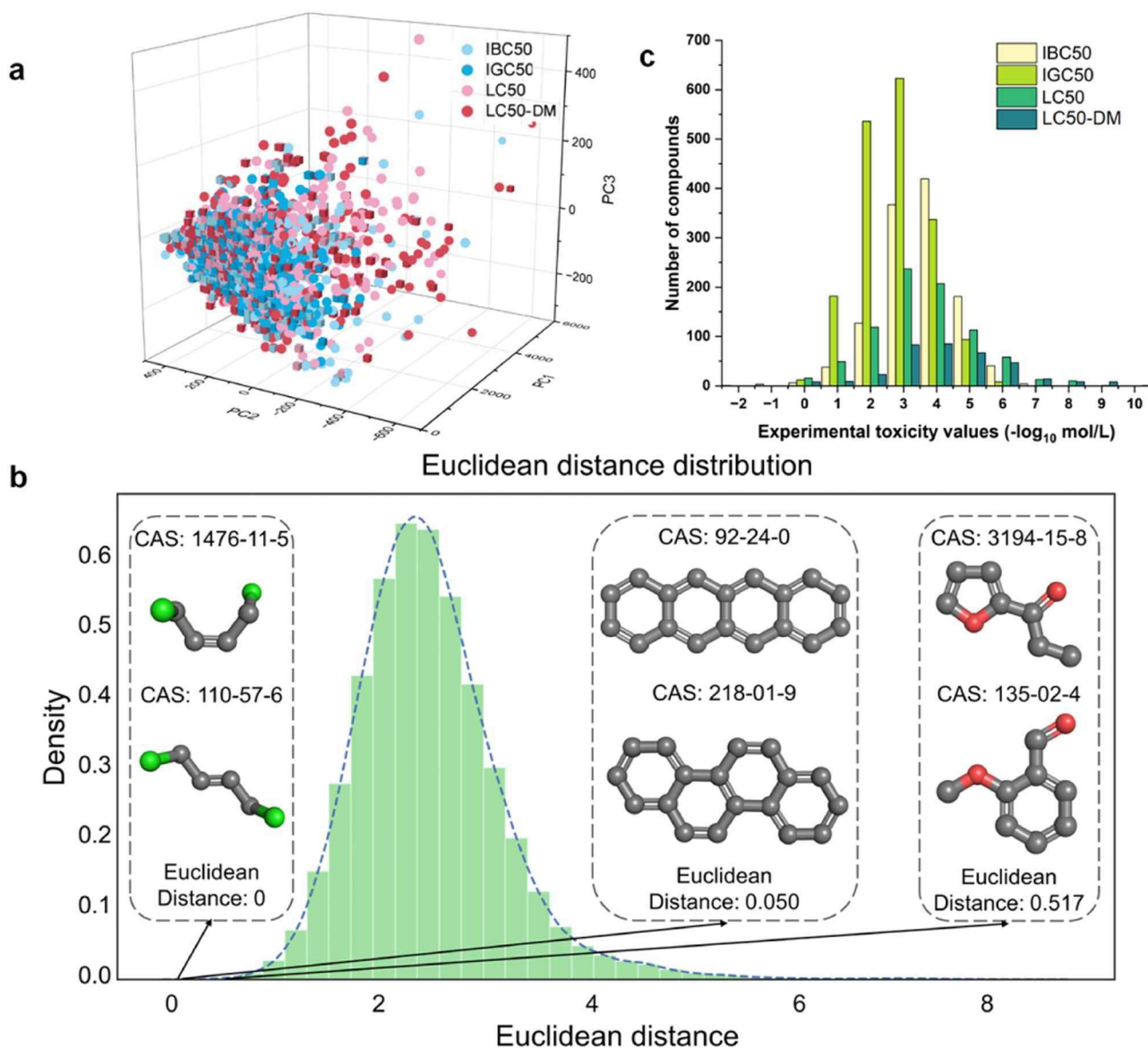


Fig. 2. Visualization of chemical similarity and toxicity value distribution. (a) Distribution of four aquatic organisms the principal component analysis results of molecular features generated from MOE descriptors. The training set is represented by a spherical shape, while the test set is represented by a cubic shape. (b) Distribution of the 2248,260 Euclidean distances calculated from each pair of molecules in the datasets. (c) Histogram of aquatic toxicity values. The diversity of chemical structures and nearly normal distribution of toxicity values were beneficial for machine learning.

ethylbenzene and xylene, which were mainly from vehicle exhaust and waste water and gas discharged beyond the standard [5]. In addition, there are also a considerable number of organic heterocyclic and organic oxygen compounds with the proportion of 11.1% and 11.81%, respectively. These chemicals are widely used in nature, and some of them have been released into the environment, such as drugs, pesticides, herbicides, dyes and plastics [49]. In the datasets, the number of organic acids and derivatives accounted for 8.78%. These chemicals included some emerging pollutants such as fluoride. The above results indicated that most molecules can be effectively separated by the calculated MOE descriptors, and the aquatic toxicity of molecules may be highly related with their functional groups.

In a QSAR model, the basic assumption is that similar chemicals have similar toxicities. To quantitatively evaluate the chemical similarity in four datasets, the MOE descriptors were further used to generate the pairwise Euclidean distances of all molecules. A total of 2248,260 distances were finally calculated among each two of the 2121 molecules (Method S5). As shown in Fig. 2b, these distances ranged from 0 to 8.722 with the mean of 2.520. A shorter distance between two molecules indicated a higher similarity, such as naphthacene (CAS number: 92-24-0) and chrysene (CAS number: 218-01-9) with Euclidean distance of 0.050. These two similar compounds belong to polycyclic aromatic hydrocarbons (PAHs) that are potent atmospheric pollutants. In the datasets, some molecules that belong to different categories, also exhibit higher structure similarity. For instance, the Euclidean distance of 1-Propanone,1-(2-furanyl) (CAS number: 3194-15-8) and 2-Methoxybenzaldehyde (CAS number: 135-02-4) is 0.517. This structure similarity is beneficial for screening targeted molecules from a wider range, i.e., not limited to the same category. However, some organic compounds with different structures are seen as completely identical. For example, cis-1,4-Dichloro-2-butene (CAS number: 1476-11-5) and trans-1,4-Dichloro-2-butene (CAS number: 110-57-6) are cis-trans isomers but their Euclidean distance is calculated as Zero. The result indicated that there were certain defects in distinguishing similar molecules by current descriptors, which provided a direction for the development or adoption of more advanced descriptors such as Dragon descriptors.

The prediction ability of a machine learning model is also related to the nature of toxicity data. The biased or imbalanced data make the machine learning model unable to learn all-round information [23]. Therefore, the histogram was used to show the frequency distribution of the experimental acute toxicity values. As shown in Fig. 2c, the nearly normal distributions of toxicity values demonstrated that the selected four datasets were suitable for building useful machine learning models.

3.2. Comparison of prediction results generated from traditional machine learning and end-to-end deep learning

One of our goals in this study is to effectively integrate the chemical feature information for quantitatively relating to the toxicity values so that the prediction of machine learning models can be improved. Therefore, various types of chemical features including ECFP, FCFP, MACCS, MOE descriptors were generated from molecular structures. At the same time, four traditional machine learning models (i.e., RF, SVM, BRR and XGBoost) were used to build quantitative relationships between the chemical features and corresponding toxicity values. In addition, multiple end-to-end deep learning models (i.e., VGG19_bn, ResNet, DenseNet, AttentiveFP and GRU) were also developed to extract features from molecular images, graphs and their SMILES. At present study, 21 prediction models were finally generated for each dataset. The hyperparameters and optimal models' final parameters were listed in Table S2 and Table S3, respectively. Table S4 showed the prediction results of different machine learning and deep learning models. A criterion of minimum R^2 value ($R^2 > 0.5$) was used to determine the acceptable model parameters [13]. Moreover, the performance of these regression models was further evaluated by RMSE and MAE.

As shown in Table S4, almost all machine learning and deep learning models exhibit acceptable performance with an optimal combination of parameters ($R^2 > 0.5$). Furthermore, the results from 5-fold cross validation and external prediction did not show much difference, indicating that the selected model parameters were useful and the constructed models were not overfitting. As shown in Fig. S3, the average R^2 value of each algorithm in four datasets were calculated. It can be found that the SVM ($R^2 = 0.597$) algorithm showed slightly better performance than other traditional machine learning algorithms. SVM is a powerful tool to solve some problems in data mining with the help of optimization methods [19]. In this study, the deep learning model did not show significant better performance than the traditional machine learning model [20,52]. Similar to our previous results, this is mainly due to the relatively small datasets. Compared with the big data in other fields such as drug discovery and medical diagnosis [7,28], the higher predictive ability of this method was driven by big data. For such small data sets, the advantage of deep learning cannot be reflected.

Compared with molecular fingerprints (i.e., ECFP, FCFP and MACCS), machine learning models with MOE descriptor showed more accurate and more stable prediction results. This can be attributed that MOE descriptors cover more comprehensive feature information including molecular structures and their physicochemical properties, while fingerprints can only represent the molecular fragments. Additionally, the feature selection was performed to explore the effect of number of parameters on model performance. As shown in Table S5, these models did not show much difference with or without feature selection, further indicating the constructed machine learning models were not overfitting. The results also demonstrated that the machine learning model was able to automatically select important variables from the pooling of features. Currently, the machine learning models are mainly driven by data. Therefore, we further explored the effect of datasets on model performance. In IGC50 set, both 5-fold cross-validation and external validation achieved excellent prediction results ($R^2 > 0.6$) no matter what machine learning methods and features are used. In the present study, the IGC50 set is the largest one and thus has a wider diversity in chemical structures. Clearly, enough high-quality data is still the most important factor determining model performance, followed by molecular descriptors and machine learning algorithms. Our results provide theoretical support to design more advanced QSAR models for regulatory purpose.

Given that the model performance varies with molecular features and machine learning algorithms in each dataset, we finally selected the model with best performance for further analysis. Herein, the XGBoost, AttentiveFP, and SVM models with the best parameter combinations were used to predict the IBC50, IGC50, LC50 and LC50-DM. The correlations between measured toxicity values and the predicted results are plotted in Fig. 3a-d. It was found that most experimental toxicity values were close to the model predictions. Overall, the R^2 of all 5-fold cross validations are higher than 0.56, and the R^2 of external validations are higher than 0.6. In IGC50 set, the optimal machine learning models obtained excellent performance with R^2 higher than 0.8 and the loss curve of this model was shown in Fig. S4. As the learning epochs increased, the loss value kept decreasing and eventually levelled off. Furthermore, there was no significant difference between the training loss and validation loss, indicating that the AttentiveFP model was not overfitting. The results demonstrate that the selected machine learning models can accurately evaluate the aquatic toxicity of most organic chemicals from the molecular structures. However, some predicted outliers were also obvious. For example, the experimental value of data point A (methylenedithiocyanate, CAS: 6317-18-6) is 6.57, but the predicted value of XGBoost model is 2.64 (Fig. 3a). Additionally, the experimental value of data point B (1-iodobutane, CAS: 542-69-8) is 2.61, but the predicted value of AttentiveFP model is 5.79 (Fig. 3b). Generally, the predicted values of organic chemicals were generated from toxicity information learned from their structural analogues. This is a common phenomenon known as "activity cliff", where the

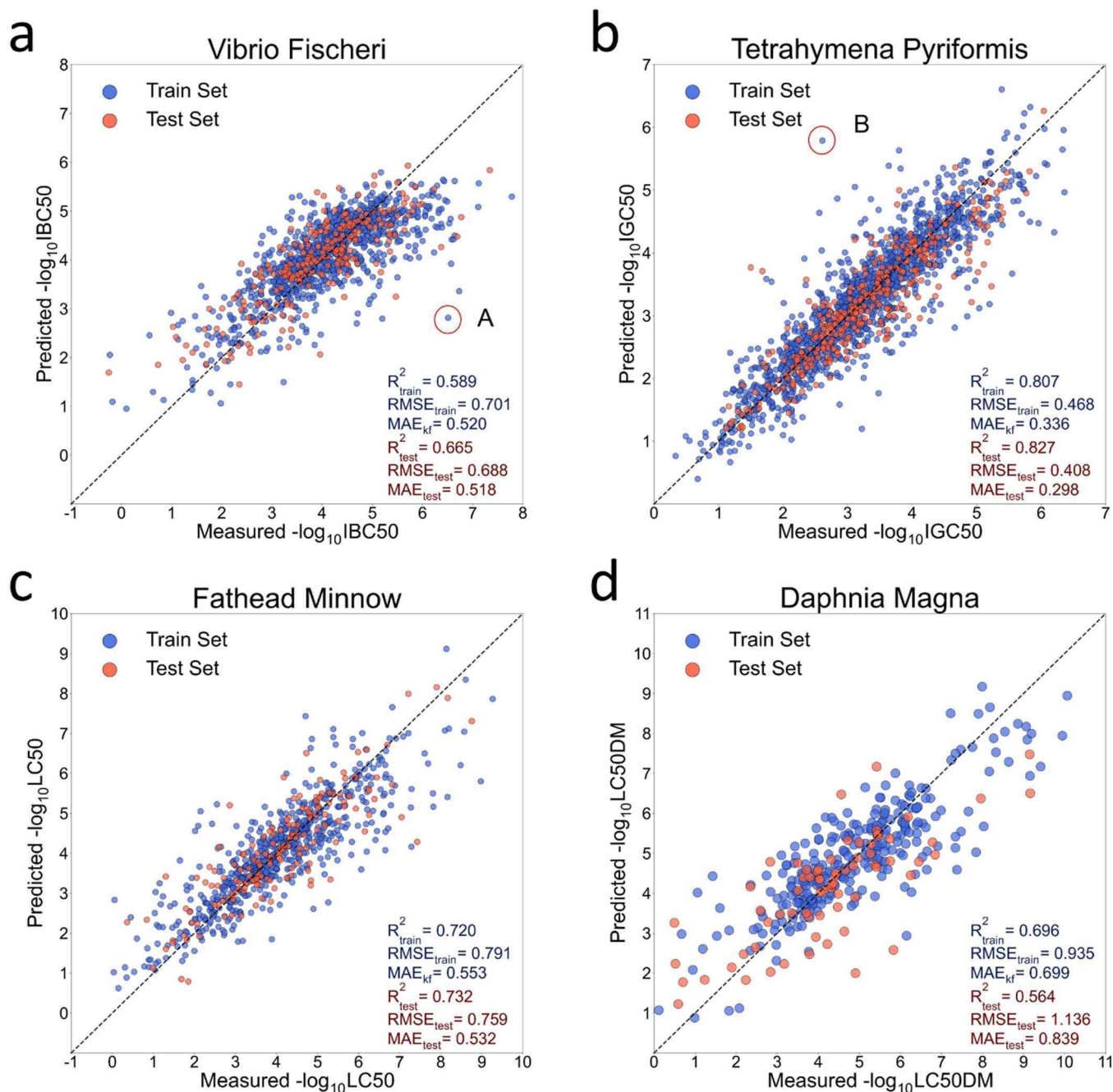


Fig. 3. Plot of the measured and predicted values of chemical toxicity in training and test set. (a) Correlations between the experimental values of IBC50 and the predictions from the XGBoost model, (b) Correlations between the experimental values of IGC50 and the predictions from the AttentiveFP model, (c) Correlations between the experimental values of LC50 and the predictions from the SVM model, and (d) Correlations between the experimental values of LC50-DM and the predictions from the SVM model.

structurally similar molecules exhibited different toxicity [46]. The main reason for this can be attributed that structural descriptors alone are not enough to distinguish similar molecules. Additionally, it may be challenging for traditional feature extraction methods to capture subtle structural variations of similar chemicals [6]. To mitigate activity cliff, it may require the integration of larger and more diverse datasets, as well as the development of advanced models and feature representation methods. However, the constructed machine learning models have limitations and cannot predict the toxicity of all compounds universally.

The optimal model performances from the present machine learning models and TEST were 0.732 VS 0.728 (LC50 set), 0.696 VS 0.739 (LC50-DM set), and 0.807 VS 0.764 (IGC50 set). Although the present

study achieved similar prediction results as the TEST, our models have the following advantages: 1) the present machine learning models adopted more descriptors and advanced algorithms, especially end-to-end deep learning; 2) more toxicity endpoint (i.e., IBC50 set) was involved in the present study; 3) our models can spontaneously accept user data to improve prediction accuracy. These models are specific to predicting the toxicity of certain organic compounds as they are trained on the characteristics of organic compounds. Therefore, the structure differences between organic compounds and non-organic compounds make it challenging to achieve precise predictions using our current model. To improve the applicability of the model, it is necessary to gather toxicity data for a wider range of compounds, including non-

organic compounds, and incorporate data from various aquatic organisms. Expanding the dataset will enable the model to make predictions across a broader spectrum, enhancing its effectiveness and reliability.

3.3. Toxicity mechanism analysis based on model interpretation

When applying machine learning in toxicology, we hope not only to obtain predictions from models but also to understand how the models are making predictions (so-called model interpretation). Interpretation of the machine learning models can help us identify a number of chemical properties responsible for aquatic toxicity, which can be used to elucidate the potential toxicity mechanism. In the present study, feature importance was applied to interpret the machine learning models and infer how much each feature contributes to the final predictions [59]. We selected the RF model of each aquatic toxicity dataset for feature important analysis. Here, the Gini index was used to evaluate the feature importance of RF model [58]. The decrease in impurity caused by each feature is calculated and used to rank the importance of the features. The features that cause the greatest decrease in impurity are considered the most important. Fig. 4 showed the ranking of top 4 important features. Details of these features can be seen in Table S6. It was found that aquatic toxicity of chemicals was highly related with certain physicochemical properties such as octanol-water partition coefficient ($\log P(o/w)$, $h_{\log P}$, and $S_{\log P}$) and solubility in water ($\log S$ and $h_{\log S}$). Compared with hydrophilic compounds, lipophilic compounds are more likely to pass through the cell membrane and thus cause greater harm to the organisms. Lipophilic compounds can affect the exchange and signal transduction inside and outside the cell by changing the permeability, stability and fluid properties of the membrane [14].

For lower aquatic organisms such as *Vibrio fischeri*, they are more sensitive to organic compounds with higher lipophilic in the environment since their relatively simple composition of organisms [41]. Moreover, aquatic organisms are easily exposed to organic chemicals with high water solubility since the high concentration of these chemicals in aquatic ecosystems. Although organic compounds with high hydrophobicity and strong solubility can both cause damage to aquatic organisms, the toxic mechanisms are different. At the same concentration, hydrophobic compounds are more harmful to aquatic organisms; in natural water environment, the high concentrations make hydrophilic compounds more harmful. In addition, molecular weight is also an important indicator determining the aquatic toxicity of organic chemicals. Organic compounds with smaller molecular weight typically have higher rates of absorption and better tissue permeability, and thus more harmful to aquatic organisms [45,47,53]. Previous studies also demonstrated that molecules containing phenol groups and multiple halogen substituents typically induce high toxicity towards organisms [21,36].

Further, we also analyzed the AttentiveFP model to identify the important structure features from the molecular graphs that were consisted of atoms and chemical bonds. Naturally, an organic molecule can be represented as a graph consisting of a set of atoms (nodes) and bonds (edges). Detailed information about an atom environment can be given using molecular graph, and thus improving the model performances in various tasks such as drug design, reaction prediction, and toxicity assessment [10,33,43]. This can help us understand the toxicity mechanism from the atomic level. In the AttentiveFP algorithm, the attention mechanism was introduced to learn the effects of different atoms on the aquatic toxicity of organic chemicals. The feature importance was determined by analyzing the attention weight of all atoms in the organic structure. The weight value calculations were conducted on all atoms of each molecule, as shown in Fig. S5. To comprehensively evaluate the importance of atomic characteristics, we randomly selected three representative compounds with high, medium and low toxicity molecules from four aquatic toxicity datasets. It was found that the C atom is ranked the highest, indicating that the substructures (e.g., CH_3 and CH_2) of carbon chain have great influence on the chemical toxicity. Previous studies have shown that the toxicity of organic compounds increased with the length of carbon chain [29]. For example, higher alcohols (alcohols that have more than 2 carbons) are more toxic than propanol and ethanol [35]. In addition, the N atoms are also marked with a darker color, indicating their greater impact on the chemical toxicity. Studies have shown that the toxicity of aromatic hydrocarbons is significantly increased by replacing hydrogen atoms with amino or nitro groups [34]. Although the DNNs are once regarded as “black-box” approaches, several emerging tools (e.g., class activation map and attention mechanism) are improving the transparency and accountability of deep learning models [50,52]. Herein, the interpretation of attentive FP model allows us explore the effect of each atom of organic molecules on their toxicity and make a deeper analysis of the potential toxicity mechanisms. More importantly, feature importance analysis from machine learning and deep learning models can help us design molecules with low toxicity.

3.4. The effect of applicability domain on model performance

It is believed that the prediction is unreliable if a molecule is quite different with others in the training set. Therefore, it is necessary to introduce AD to estimate the uncertainty in predicting the toxicity of a particular molecule. In this study, we discussed the impact of setting AD threshold on the prediction results. Fig. 5 showed the model performances and the coverage of test set when introducing AD to the machine learning model. It can be found that the coverage of the test set decreased after introducing AD, but the model performance improved in most cases of four aquatic toxicity datasets. However, it was also noted

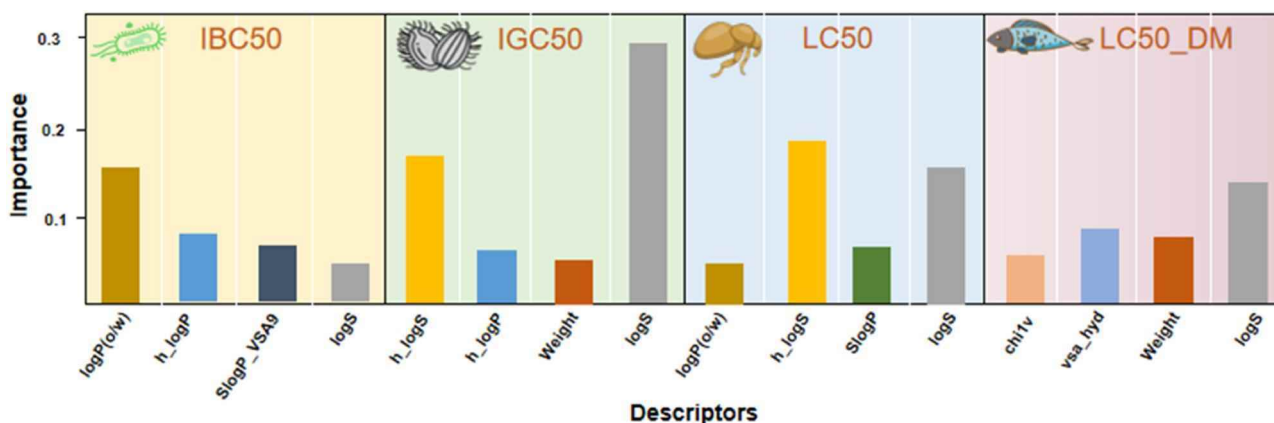


Fig. 4. Contributions of top 4 features to chemical toxicity. The contributions were calculated from feature importance analysis of the selected optimal machine learning models. The greater weight of the descriptor indicates a higher contribution to the chemical toxicity.

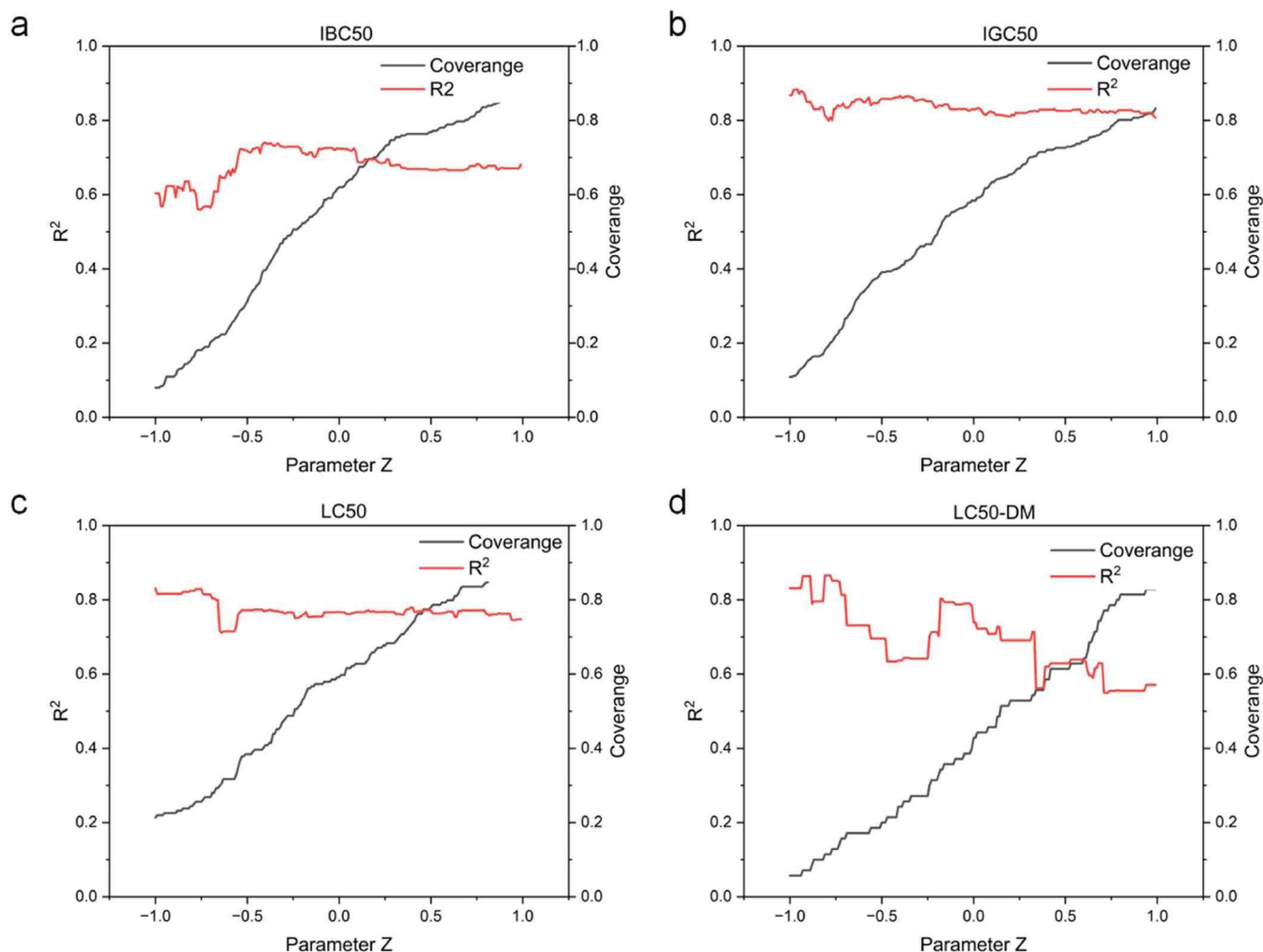


Fig. 5. The effect of AD on model predictions. (a-d) Predictive ability of the optimized models for four aquatic toxicity datasets using different ADs. The coverage of the test set is also shown in the figure.

that the model accuracy sometimes did not significantly improve even though a considerable amount of compounds were removed from the test set. For instance, the R^2 value for LC50 set only increased from 0.732 to 0.774 when the coverage of test set decreased 51% (Fig. S6). The main reason is that the machine learning models have acquired acceptable performances even for the entire test sets, and thus there were not too much significant prediction outliers.

Typically, the coverage of test set was related with the training set and the threshold of AD. A rigorous strict threshold could reduce the coverage of test set, while the structural diversity of training set could improve the coverage of test set. Clearly, the aquatic toxicity model becomes more reliable with the exclusion of several prediction outliers, but the limited applicability of machine learning model may reduce its usefulness and generalization. Therefore, it is important to find a balance between the model accuracy and coverage of test set. There is no need to excessively remove compounds from the test set for a tiny improvement of the model accuracy. To make the machine learning model more widely applicable, it is necessary to use more high-quality training data with structural diversity. At the same time, it would also be useful to develop more advanced AD calculation method.

3.5. Application of the aquatic toxicity models for regulatory purpose

The proposed method was further verified on a dataset from the IECSC, which contained 16,543 compounds that were not included in

the four aquatic datasets. These compounds covered multiple types, such as alcohol, ketone and ester. To further understand the generalization capability of our models, AD was incorporated into the machine learning models to identify reliable predictions from the IECSC dataset. After the AD was introduced, only a small portion of compounds was retained, i.e., 17.88%, 12.37%, 35.95% and 31.22% for IBC50, IGC50, LC50 and LC50-DM models. The relatively low coverage was mainly attributed to the difference between the IECSC data and aquatic toxicity data. For instance, there were a fair number of compounds with high molecular weight (> 266), but the molecular weights of molecules in the aquatic toxicity datasets were generally smaller than 156. To improve the model's generalization capability, more structurally diverse compounds should be incorporated into the aquatic toxicity datasets. As shown in Fig. 6a, in AD range, the prediction results of IBC50 and IGC50 models have high similarity (i.e., similar peak width), and the LC50 and LC50-DM models exhibited similar prediction results. The prediction similarity comes from the species similarity between two datasets such as *Vibrio fischeri* (IBC50 set) and *Tetrahymena pyriformis* (IGC50 set). In addition, the results also showed that LC50 and LC50-DM models generated more compounds with high toxicity than those generated from IBC50 and IGC50 models. This result demonstrated that chemicals exhibited higher toxicity in advanced organisms such as *daphnia magna* and *fathead minnow*.

To further explore the toxicity difference of one chemical towards multiple species, the high-risk compounds were screened out from four

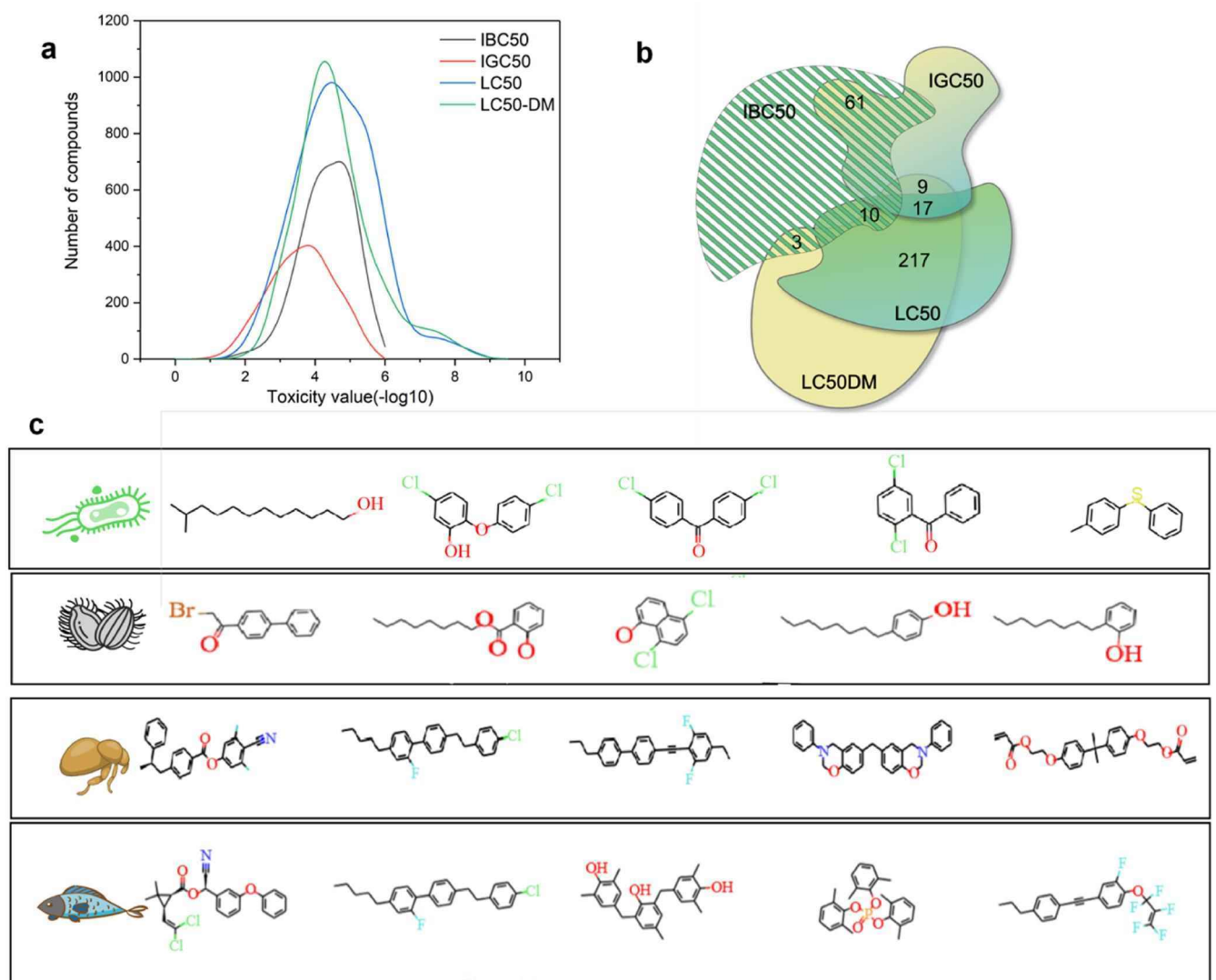


Fig. 6. Aquatic toxicity prediction of IECSC chemicals for regulatory purpose. (a) Distribution of toxicity predictions in four acute aquatic toxicity models. A higher toxicity value indicates a more serious threaten to the aquatic organism. (b) Venn diagram of high-risk chemicals in four acute aquatic toxicity models. (c) Several representative high-risk compounds screened from IECSC.

machine learning models. According to the prediction results, we set compounds with toxicity values in the top 10% as high-risk compounds. Therefore, 399, 209, 204 and 424 compounds from IBC50, IGC50, LC50 and LC50-DM models were determined as high-risk organic substances in the present study in AD range. Similar with the above results, the high-risk compounds screened from IBC50 and IGC50 models also exhibited high similarity (Fig. 6b). 61 organic compounds screened from IGC50 model were screened by IBC50 model at the same time in AD range. Additionally, among the organic compounds screened by LC50-DM model, 217 organic compounds were also screened by LC50 model in AD range. The results showed that the toxicity mechanism of organic compounds to low- and high-level aquatic organisms was quite different. For instance, benzoic acid organics are highly toxic to bacteria and *tetrahymena pyriformis*, but for *daphnia magna* and fish, skin and fat have a strong inhibitory effect on the absorption of ionizable organics in multicellular organism [26].

In addition, several representative compounds with high toxicity were also shown in Fig. 6c. It is worth noting that phenol and its substituents with fluorine or chlorine atoms exhibited lower toxicity to *daphnia magna* and fish. However, these compounds have strong toxic effects on *vibrio fischeri* and *tetrahymena pyriformis*. Phenols and its substituents are reactive compounds, whose mechanism is oxidative

phosphorylation, respiratory uncoupling, and thus have no residual toxic effect on higher aquatic organisms [15]. At the same time, we found that amine, aniline and phenyl ester compounds are prone to exhibit strong toxicity in higher aquatic organisms. The above results demonstrated that the toxicity of chemicals was species sensitivity.

By developing comprehensive toxicity prediction models for different aquatic organisms, we were able to swiftly screen organic compounds in the IECSC database and identify potential toxic substances. This approach could reduce the need for extensive and costly laboratory testing. Furthermore, by conducting comprehensive assessments of compounds based on multiple aquatic toxicity models, it is possible to identify compounds that may have significant impacts on ecosystems. For example, certain halogenated compounds containing benzene moieties, such as those with Cl or Br elements, were identified high-risk compounds towards various aquatic organisms (Fig. 6c). Subsequently, regulatory agencies can develop appropriate measures to mitigate the risks associated with these compounds. Based on the above analysis, it is evident that organic compounds exhibit variations in toxic mechanisms across different aquatic organisms. Regulatory agencies can further classify toxic compounds and establish specific restrictions and standards, such as usage limitations and emission controls, to mitigate their detrimental effects on water environments. Furthermore, more

attention should also be paid to the accumulation of chemical toxicity in the food chain due to the significant increase of toxin concentration as the food chain moves up to higher levels such as human beings.

4. Concluding remarks

Reasonable regulation of chemicals is crucial for mitigating the chemical pollution worldwide. To this end, this study proposed a more reasonable framework to assess aquatic toxicity of chemicals. In the framework, each optimal model was selected from 21 different models that were constructed by five end-to-end deep learning algorithms and all the combination of four kinds of molecular features and four traditional machine learning methods. Multispecies-based toxicity assessment demonstrated that the toxicity of organic substances may be enriched through the food chain, thus increasing hazards to higher organisms. The implementation of AD has further improved the reliability of the optimal prediction model in chemical regulation. Moreover, feature important analysis was successfully applied to identify possible structure alerts, which can be further used for designing environmentally friendly chemicals. Our study provides regulatory authorities an efficient approach for more reasonable regulation of hazardous chemicals. However, the successful implementation of the proposed framework heavily depends on the continuous efforts from experimentalists and modelers; for example, the generation of more high-quality toxicity data, the development of more advanced descriptors and machine learning algorithms. Additionally, more aquatic species should also be added to meet the minimum of the aquatic baseline.

Environmental implication

It is crucial to implement reasonable chemical regulation in terms of mitigating the chemical pollution worldwide. In recent years, the use of non-animal approaches, such as *in silico* or *in vitro* methods for assessing the risks of hazardous chemicals has made significant progress. However, machine learning models based on randomly selected algorithms and a single toxicity endpoint may generate biased prediction. This study could highlight the importance of implementing comprehensive machine learning models and multispecies toxicity assessment to improve chemical regulation.

CRedit authorship contribution statement

Xiliang Yan: Conceived the idea. **Xiliang Yan, Ying He, Guohong Liu:** Designed experiments. **Ying He, Xiaohong Wang, Jianbo Jia:** Collected the data. **Ying He, Guohong Liu, Hongyu Zhou:** Constructed the model. **Xiliang Yan, Ying He, Hongyu Zhou:** Co-wrote the manuscript and all authors discussed and approved the paper.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

The source codes of machine learning models and the corresponding toxicity data can be found at <https://github.com/YanLabAI/AquaTox>.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (22106025, 22036002, 22276042), the Introduced Innovative Research and Development Team Project under the “The Pearl River Talent Recruitment Program” of Guangdong Province (2019ZT08L387), the Basic and Applied Basic Research Foundation of

Guangzhou, China (202201010541).

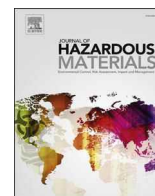
Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at [doi:10.1016/j.jhazmat.2023.131942](https://doi.org/10.1016/j.jhazmat.2023.131942).

References

- [1] Ahmad, W., Tayara, H., Chong, K.T., 2023. Attention-based graph neural network for molecular solubility prediction. *ACS Omega* 8, 3236–3244. <https://doi.org/10.1021/acsomega.2c06702>.
- [2] Ai, H., Wu, X., Zhang, L., Qi, M., Zhao, Y., Zhao, Q., et al., 2019. QSAR modelling study of the bioconcentration factor and toxicity of organic compounds to aquatic organisms using machine learning and ensemble methods. *Ecotoxicol Environ Saf* 179, 71–78. <https://doi.org/10.1016/j.ecoenv.2019.04.035>.
- [3] Asilar, E., Hemmerich, J., Ecker, G.F., 2020. Image based liver toxicity prediction. *J Chem Inf Model* 60, 1111–1121. <https://doi.org/10.1021/acs.jcim.9b00713>.
- [4] Bilbrey, J.A., Marrero, C.O., Sassi, M., Ritzmann, A.M., Henson, N.J., Schram, M., 2020. Tracking the chemical evolution of iodine species using recurrent neural networks. *ACS Omega* 5, 4588–4594. <https://doi.org/10.1021/acsomega.9b04104>.
- [5] Boström, C.E., Gerde, P., Hanberg, A., Jernström, B., Johansson, C., Kyrklund, T., et al., 2002. Cancer risk assessment, indicators, and guidelines for polycyclic aromatic hydrocarbons in the ambient air. *Environ Health Perspect* 110, 451–488. <https://doi.org/10.1289/ehp.110-1241197>.
- [6] Bo, T., Lin, Y., Han, J., Hao, Z., Liu, J., 2023. Machine learning-assisted data filtering and QSAR models for prediction of chemical acute toxicity on rat and mouse. *J Hazard Mater* 452, 131344. <https://doi.org/10.1016/j.jhazmat.2023.131344>.
- [7] Cai, C., Wang, S., Xu, Y., Zhang, W., Tang, K., Ouyang, Q., et al., 2020. Transfer learning for drug discovery. *J Med Chem* 63, 8683–8694. <https://doi.org/10.1021/acs.jmedchem.9b02147>.
- [8] Cortés-Ciriano, I., Bender, A., 2019. KekuleScope: prediction of cancer cell line sensitivity and compound potency using convolutional neural networks trained on compound images. *J Cheminform* 11, 41. <https://doi.org/10.1186/s13321-019-0364-5>.
- [9] Das, R.N., Sintra, T.E., Coutinho, J.A.P., Ventura, S.P.M., Roy, K., Popelier, P.L.A., 2016. Development of predictive QSAR models for Vibrio fischeri toxicity of ionic liquids and their true external and experimental validation tests. *Toxicol Res* 5, 1388–1399. <https://doi.org/10.1039/c6tx00180g>.
- [10] De Sá, A.G.C., Long, Y., Portelli, S., Pires, D.E.V., Ascher, D.B., 2022. toxCSM: comprehensive prediction of small molecule toxicity profiles. *Brief Bioinform* 23, 4827–4836. <https://doi.org/10.1093/bib/bbac337>.
- [11] Downing, J.A., Polasky, S., Olmstead, S.M., Newbold, S.C., 2021. Protecting local water quality has global benefits. *Nat Commun* 12, 2709. <https://doi.org/10.1038/s41467-021-22836-3>.
- [12] Duan, W., Chen, G., Ye, Q., Chen, Q., 2011. The situation of hazardous chemical accidents in China between 2000 and 2006. *J Hazard Mater* 186, 1489–1494. <https://doi.org/10.1016/j.jhazmat.2010.12.029>.
- [13] Eriksson, L., Jaworska, J., Worth, A.P., Cronin, M.T., McDowell, R.M., Gramatica, P., 2003. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. *Environ Health Perspect* 111, 1361–1375. <https://doi.org/10.1289/ehp.5758>.
- [14] Gabelova, A., 2020. 7H-Dibenzof[c,g]carbazole: Metabolic pathways and toxicity. *Chem Biol Interact* 323, 109077. <https://doi.org/10.1016/j.cbi.2020.109077>.
- [15] Ge, T., Han, J., Qi, Y., Gu, X., Ma, L., Zhang, C., et al., 2017. The toxic effects of chlorophenols and associated mechanisms in fish. *Aquat Toxicol* 184, 78–93. <https://doi.org/10.1016/j.aquatox.2017.01.005>.
- [16] Grimm, D., 2019. Ready to pounce. *Science* 364, 522–525. <https://doi.org/10.1126/science.364.6440.522>.
- [17] He, L., Xiao, K., Zhou, C., Li, G., Yang, H., Li, Z., et al., 2019. Insights into pesticide toxicity against aquatic organism: QSTR models on Daphnia Magna. *Ecotoxicol Environ Saf* 173, 285–292. <https://doi.org/10.1016/j.ecoenv.2019.02.014>.
- [18] He, Y., Liu, G., Li, C., Yan, X., 2022. Reaching the full potential of machine learning in mitigating environmental impacts of functional materials. *Rev Environ Contam T* 260, 21. <https://doi.org/10.1007/s44169-022-00024-8>.
- [19] Helma, C., Cramer, T., Kramer, S., De Raedt, L., 2004. Data mining and machine learning techniques for the identification of mutagenicity inducing substructures and structure activity relationships of noncongeneric compounds. *J Chem Inf Comp Sci* 44, 1402–1411. <https://doi.org/10.1021/ci034254q>.
- [20] Hu, S., Liu, G., Zhang, J., Yan, J., Zhou, H., Yan, X., 2022. Linking electron ionization mass spectra of organic chemicals to toxicity endpoints through machine learning and experimentation. *J Hazard Mater* 431, 128558. <https://doi.org/10.1016/j.jhazmat.2022.128558>.
- [21] Ike, I.A., Karanfil, T., Cho, J., Hur, J., 2019. Oxidation byproducts from the degradation of dissolved organic matter by advanced oxidation processes-A critical review. *Water Res* 164, 114929. <https://doi.org/10.1016/j.watres.2019.114929>.
- [22] Kang, B., García, D., Lijffijt, J., Santos-Rodríguez, R., De, B.T., 2021. Conditional t-SNE: more informative t-SNE embeddings. *Mach Learn* 110, 2905–2940. <https://doi.org/10.1007/s10994-020-05917-0>.
- [23] Kang, C., Lai, Y.K., Wu, Y.X., Martin, R., Hu, S.M., 2014. Automatic semantic modeling of indoor scenes from low-quality RGB-D data using contextual

- information. In: ACM. T. Graphic, 33, pp. 1–12. (<https://dl.acm.org/doi/10.1145/2661229.2661239>).
- [24] Keiser, D.A., Shapiro, J.S., 2019. US water pollution regulation over the past half century: burning waters to crystal springs. ? J Econ Perspect 33, 51–75. <https://doi.org/10.1257/jep.33.4.51>.
- [25] Kosnik, M.B., Hauschild, M.Z., Fantke, P., 2022. Toward assessing absolute environmental sustainability of chemical pollution. Environ Sci Technol 56, 4776–4787. <https://doi.org/10.1021/acs.est.1c06098>.
- [26] Lee, P.Y., Chen, C.Y., 2009. Toxicity and quantitative structure-activity relationships of benzoic acids to *Pseudokirchneriella subcapitata*. J Hazard Mater 165, 156–161. <https://doi.org/10.1016/j.jhazmat.2008.09.086>.
- [27] Li, X., Fourches, D., 2020. Inductive transfer learning for molecular activity prediction: Next-Gen QSAR Models with MolPMoFIT. J Cheminform 12, 27. <https://doi.org/10.1186/s13321-020-00430-x>.
- [28] Liu, H., Li, Z., Shen, R., Li, Z., Yang, Y., Yuan, Q., 2021. Point-of-care pathogen testing using photonic crystals and machine vision for diagnosis of urinary tract infections. Nano Lett 21, 2854–2860. <https://doi.org/10.1021/acs.nanolett.0c04942>.
- [29] Lv, M., Xie, Y., Yu, H., Sun, T., Song, L., Wang, F., 2022. Effects of perfluoroalkyl substances on soil respiration and enzymatic activity: differences in carbon chain-length dependence. J Environ Sci Health B 57, 284–296. <https://doi.org/10.1080/03601234.2022.2047563>.
- [30] Mangold-Döring, A., Grimard, C., Green, D., Petersen, S., Nichols, J.W., Hogan, N., et al., 2021. A novel multispecies toxicokinetic modeling approach in support of chemical risk assessment. Environ Sci Technol 55, 9109–9118. <https://doi.org/10.1021/acs.est.1c02055>.
- [31] Mayr, A., Klambauer, G., Unterthiner, T., Hochreiter, S., 2015. DeepTox: toxicity prediction using deep learning. Front Environ Sci 3, 569. <https://doi.org/10.1016/j.toxlet.2017.07.175>.
- [32] Muhire, J., Li, B.Q., Zhai, H.L., Li, S.S., Mi, J.Y., 2020. A simple approach to the toxicity prediction of anilines and phenols towards aquatic organisms. Arch Environ Contam Toxicol 78, 545–554. <https://doi.org/10.1007/s00244-019-00703-z>.
- [33] Mukaidaisi, M., Vu, A., Grantham, K., Tchagang, A., Li, Y., 2022. Multi-objective drug design based on graph-fragment molecular representation and deep evolutionary learning. Front Pharmacol 13, 920747. <https://doi.org/10.3389/fphar.2022.920747>.
- [34] Nepali, K., Lee, H.Y., Liou, J.P., 2019. Nitro-group-containing drugs. J Med Chem 62, 2851–2893. <https://doi.org/10.1021/acs.jmedchem.8b00147>.
- [35] Offeman, R.D., Franqui-Espiet, D., Cline, J.L., Robertson, G.H., Orts, W.J., 2008. Extraction of ethanol with higher alcohol solvents and their toxicity to yeast. Sep Purif Technol 72, 180–185. <https://doi.org/10.1016/j.seppur.2010.02.004>.
- [36] Oliveira, M.R.O., Matheus, M., Oliveira, Ronney J.Dervanoski, AdrianaFranceschi, EltonEgues, Silvia M.De. Conto, Juliana, F., 2019. Amine-modified silica surface applied as adsorbent in the phenol adsorption assisted by ultrasound. Chem Eng Commun 206, 1554–1569. <https://doi.org/10.1080/00986445.2019.1615467>.
- [37] Ortega-Calvo, J.-J., Harmsen, J., Parsons, J.R., Semple, K.T., Aitken, M.D., Ajao, C., et al., 2015. From bioavailability science to regulation of organic chemicals. Environ Sci Technol 49, 10255–10264. <https://doi.org/10.1021/acs.est.5b02412>.
- [38] Popova, M., Isayev, O., Tropsha, A., 2018. Deep reinforcement learning for de novo drug design. Sci Adv 4, eaap7885. <https://doi.org/10.1126/sciadv.aap7885>.
- [39] Qin, Y., Huttlin, E.L., Winsnes, C.F., Gosztyla, M.L., Wacheul, L., Kelly, M.R., et al., 2021. A multi-scale map of cell structure fusing protein images and interactions. Nature 600, 536–542. <https://doi.org/10.1038/s41586-021-04115-9>.
- [40] Seller, C., Honti, M., Singer, H., Fenner, K., 2020. Biotransformation of chemicals in water-sediment suspensions: influencing factors and implications for persistence assessment. Environ Sci Technol Lett 7, 854–860. <https://doi.org/10.1021/acs.estlett.0c00725>.
- [41] Seth, A., Roy, K., 2020. QSAR modeling of algal low level toxicity values of different phenol and aniline derivatives using 2D descriptors. Aquat Toxicol 228, 105627. <https://doi.org/10.1016/j.aquatox.2020.105627>.
- [42] Tugcu, G., Ertürk, M.D., Saçan, M.T., 2017. On the aquatic toxicity of substituted phenols to *Chlorella vulgaris*: QSTR with an extended novel data set and interspecies models. J Hazard Mater 339, 122–130. <https://doi.org/10.1016/j.jhazmat.2017.06.027>.
- [43] Tavakoli, M., Mood, A., Van Vranken, D., Baldi, P., 2022. Quantum mechanics and machine learning synergies: graph attention neural networks to predict chemical reactivity. J Chem Inf Model 62, 2121–2132. <https://doi.org/10.1021/acs.jcim.1c01400>.
- [44] Vilar, S., Cozza, G., Moro, S., 2008. Medicinal chemistry and the molecular operating environment (MOE): application of QSAR and molecular docking to drug discovery. Curr Top Med Chem 8, 1555–1572. <https://doi.org/10.2174/156802608786786624>.
- [45] Wang, Z., Gao, Y., Wang, S., Fang, H., Xu, D., Zhang, F., 2016. Impacts of low-molecular-weight organic acids on aquatic behavior of graphene nanoplatelets and their induced algal toxicity and antioxidant capacity. Environ Sci Pollut R 23, 10938–10945. <https://doi.org/10.1007/s11356-016-6290-4>.
- [46] Wang, W., Kim, M.T., Sedykh, A., Zhu, H., 2015. Developing enhanced blood-brain barrier permeability models: integrating external bio-assay data in QSAR modeling. Pharm Res 32, 3055–3065. <https://doi.org/10.1007/s11095-015-1687-1>.
- [47] Wang, H., Wang, Z., Chen, J., Liu, W., 2022. Graph attention network model with defined applicability domains for screening PBT chemicals. Environ Sci Technol 56, 6774–6785. <https://doi.org/10.1021/acs.est.2c00765>.
- [48] Wang, L., 2019. CAS reaches 150 millionth substance. Cen Glob Enterp 97. <https://doi.org/10.1021/cen-09722-acnews1>.
- [49] Wang, X., Liang, D., Wang, Y., Peijnenburg, W.J.G.M., Monikh, F.A., Zhao, X., et al., 2022. A critical review on the biological impact of natural organic matter on nanomaterials in the aquatic environment. CARR 1, 1–18. <https://doi.org/10.1007/s44246-022-00013-5>.
- [50] Xiong, Z., Wang, D., Liu, X., Zhong, F., Wan, X., Li, X., et al., 2020. Pushing the Boundaries of Molecular Representation for Drug Discovery with the Graph Attention Mechanism. J Med Chem 63, 8749–8760. <https://doi.org/10.1021/acs.jmedchem.9b00959>.
- [51] Xu, L., Xu, L., Chen, Y., Zhang, Y., Yang, J., 2022. Accurate classification of algae using deep convolutional neural network with a small database. ACS Est Water 2, 1921–1928. <https://doi.org/10.1021/acsestwater.1c00466>.
- [52] Yan, X., Sedykh, A., Wang, W., Yan, B., Zhu, H., 2020. Construction of a web-based nanomaterial database by big data curation and modeling friendly nanostructure annotations. Nat Commun 11, 1–10. <https://doi.org/10.1038/s41467-020-16413-3>.
- [53] Yan, X., Zhang, J., Russo, D.P., Zhu, H., Yan, B., 2020. Prediction of nano-bio interactions through convolutional neural network analysis of nanostructure images. ACS Sustain Chem Eng 8, 19096–19104. <https://doi.org/10.1021/acssuschemeng.0c07453>.
- [54] Zhang, H.D., Zheng, X.P., 2012. Characteristics of hazardous chemical accidents in China: A statistical investigation. J Loss Prev Proc 25, 686–693. <https://doi.org/10.1016/j.jlpp.2012.03.001>.
- [55] Zhang, K., Zhang, H., 2022. Machine learning modeling of environmentally relevant chemical reactions for organic compounds. ACS Est Water 8, 1555–1572. <https://doi.org/10.1021/acsestwater.2c00193>.
- [56] Zheng, L., Fan, J., Mu, Y., 2019. OnionNet: a multiple-layer intermolecular-contact-based convolutional neural network for protein-ligand binding affinity prediction. ACS Omega 4, 15956–15965. <https://doi.org/10.1021/acsomega.9b01997>.
- [57] Zhang, S., Wang, N., Su, L., Xu, X., Li, C., Qin, W., et al., 2020. MOA-based linear and nonlinear QSAR models for predicting the toxicity of organic chemicals to *Vibrio fischeri*. Environ Sci Pollut R 27, 9114–9125. <https://doi.org/10.1007/s11356-019-06681-y>.
- [58] Zhang, R., Wu, Q., Qi, X., Wang, X., Zhang, X., Song, C., et al., 2021. Using in vitro and machine learning approaches to determine species-specific dioxin-like potency and congener-specific relative sensitivity among birds for brominated dioxin analogues. Environ Sci Technol 55, 16056–16066. <https://doi.org/10.1021/acs.est.1c05951>.
- [59] Zhong, S., Zhang, K., Bagheri, M., Burken, J.G., Gu, A., Li, B., et al., 2021. Machine learning: new ideas and tools in environmental science and engineering. Environ Sci Technol 55, 12741–12754. <https://doi.org/10.1021/acs.est.1c01339>.



Unraveling the ecotoxicity of micro(nano)plastics loaded with environmental pollutants using ensemble machine learning

Jing Zhang^{a,b,1}, Xiaofang Li^{b,1}, Hanle Chen^b, Guohong Liu^c, Xiliang Yan^{a,*}, Bing Yan^b

^a College of Animal Science, South China Agricultural University, Guangzhou 510642, China

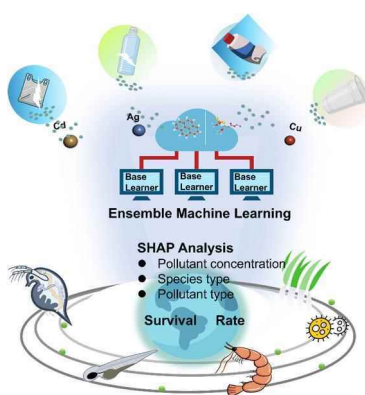
^b Institute of Environmental Research at Greater Bay Area, Key Laboratory for Water Quality and Conservation of the Pearl River Delta, Ministry of Education, Guangzhou University, Guangzhou 510006, China

^c School of Health, Guangzhou Vocational University of Science and Technology, Guangzhou 510555, China

HIGHLIGHTS

- A combined toxicity dataset was manually curated from scientific literature.
- The ensemble machine learning model outperformed individual methods.
- Pollutant concentration and species type mainly influenced the combined toxicity.
- Adsorbed pollutants enhanced the membrane penetration of micro(nano) plastics.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:

Computational toxicology
Emerging pollutant
Combined toxicity
Ensemble machine learning
Nano-bio interaction

ABSTRACT

Micro(nano)plastics are ubiquitous and pose a severe threat to the environment and human health. Despite increasing research, most existing studies have focused on the toxicity of micro(nano)plastics as individual pollutants. Furthermore, fragmented knowledge obtained from separate studies may introduce cognitive biases. Inspired by this, we developed an ensemble machine learning algorithm to predict the combined toxicity (in terms of survival rate) of micro(nano)plastics and environmental pollutants across multiple species. Based on our findings, the following conclusions were drawn: (1) The ensemble machine learning model accurately unraveled the quantitative property–toxicity relationships, achieving strong performance in both cross validation and external validation, with $R^2 > 0.84$. (2) Interpretations from the ensemble machine learning model indicated that the combined toxicity is primarily influenced by factors such as pollutant concentration, species, pollutant type, and plastic diameter. (3) Molecular dynamics (MD) simulations further revealed that micro(nano)plastics, after adsorbing the pollutant BDE-47 (2,2',4,4'-tetrabromodiphenyl ether), interact with the cell membranes through van der Waals interactions (less than -200 kJ/mol^{-1}), ultimately leading to increased membrane damage and deformation. Our modeling results provide an additional avenue for understanding the environmental and health

* Corresponding author.

E-mail address: yanxiliang1991@scau.edu.cn (X. Yan).

¹ The authors contributed equally to this work.

risks associated with micro(nano)plastics and other pollutants, complementing and extending insights gained via experimental methods.

1. Introduction

In the context of global plastic pollution, micro(nano)plastics are now considered emerging contaminants [1], being identified in virtually all environmental compartments, such as soil [2], air [3], and water bodies [4]. Microplastics and nanoplastics are plastic particles differentiated by size: microplastics range from 100 nm to 0.5 mm, whereas nanoplastics measure ≤ 100 nm [5,6]. An increasing body of evidence suggest that micro(nano)plastics have adverse effects on the environment and human health [7]. It has been demonstrated that micro(nano)plastics could induce toxicity in plants [8], fish [9], and microbes [10]. Furthermore, these tiny plastic particles can enter the human bodies through oral inhalation, ingestion, and dermal contact [11]. Recently, micro(nano)plastics have been detected in human blood and feces [12,13]. After absorbing into the human body, they can be transported by the circulatory system and eventually deposited in various organs such as kidneys, intestines, and liver [14], causing oxidative stress, inflammatory damage, and increased particle toxicity due to internalization or displacement of tissues [13,15]. Although research on micro(nano)plastics has been steadily growing, the existing studies have focused on their toxic effects as individual pollutants [16]. In fact, owing to their large specific surface area and excellent hydrophobicity, they easily adsorb other pollutants in the environment, leading to combined toxicity [17,18]. Some studies have indicated that the presence of other pollutants may enhance the toxic effects of micro(nano)plastics when they coexist [19,20]. Therefore, studying micro(nano)plastics alone is insufficient to fully assess their complex environmental impacts.

The combined toxicity of micro(nano)plastics with environmental pollutants is influenced by multiple factors, such as physicochemical properties, experimental conditions, and types of other pollutants [21–23]. However, traditional toxicological experiments are time-consuming and labor-intensive, making it challenging to identify key influencing factors in high-dimensional parameter spaces. More importantly, knowledge from a few individual studies may foster cognitive bias by creating overreliance on isolated findings. In addition, experimental methods alone struggle to capture the molecular interactions between micro(nano)plastics (solo or in mixtures) and biological systems [24]. Therefore, further in-depth and systematic exploration of combined pollution is needed to better understand and assess their comprehensive risks to the environment and human health.

Recently, machine learning has become a powerful tool to mine critical information from big data and has also achieved many encouraging results in predicting the potential adverse effects of materials such as nanomaterials [25–27]. Machine learning approaches excel at unraveling the quantitative relationships between the structures/physicochemical properties of chemicals and their toxicities or bioactivities [28,29]. For example, the quantitative analysis of machine learning models indicated that dendritic cell activation was jointly determined by the core of gold nanoparticles and their surface ligands [30]. As a significant and widely used method, ensemble machine learning combines the predictions of multiple base learners and optimizes model performance through a meta-learner, thereby improving prediction stability and generalization ability [31,32]. Another widely used computational approach, molecular dynamics (MD) simulations, enables the investigation of pollutant adsorption mechanisms on materials [33] and elucidates intermolecular interactions between pollutants and biomolecules at the atomic level [34]. This method helps understanding the molecular mechanisms underlying the combined toxicity of micro(nano)plastics loaded with environmental pollutants.

This study aims to explore the key factors influencing the combined toxicity of micro(nano)plastics and environmental pollutants using

machine learning and MD simulations, which traditional toxicological experiments cannot address. As shown in Fig. 1, a structured workflow was developed with four key components: literature mining, machine learning modeling, model interpretation, and MD simulations. We first conducted comprehensive literature mining, compiling combined toxicity data from 41 studies. The survival rate of various species was used as a representative toxicity endpoint. Predictive models were then built using algorithms such as random forest, extreme gradient boosting, and gradient boosted decision trees, with performance evaluated by 5-fold cross validation and external validation. Among them, ensemble machine learning, integrating these base learners, achieved the best performance in terms of accuracy, stability, and generalizability. To improve interpretability, shapley additive explanations (SHAP) values were used to assess feature importance. The ensemble machine learning model revealed that the combined toxicity is strongly associated with pollutant concentration, species, pollutant type, and plastic diameter. Furthermore, MD simulations indicated that micro(nano)plastics can adsorb pollutant molecules, forming complexes that are subsequently attracted to cell membranes through strong van der Waals interactions, thereby promoting the interaction between the plastic particles and the membrane. Overall, this study comprehensively analyzes the mechanisms of the combined toxicity of micro(nano)plastics with environmental pollutants, providing theoretical support to guide subsequent pollution management and control.

2. Materials and methods

2.1. Data curation

First, rigorous literature mining was performed to integrate toxicity data for subsequent computational modeling. The search used the following terms: “tox* AND (nanoplastic* OR microplastic*)”, specifically focusing on the combined toxicity reported in the literature. The word with an asterisk is a wildcard operator used for database keyword searches. For example, the “tox*” can represent all terms beginning with “tox” (e.g., toxicity, toxicology, toxic). Over 9092 peer-reviewed articles were identified in the Web of Science database until August 2024. Literature screening was conducted to improve the integrity of the data collected from various studies. Each publication was then manually evaluated based on the following criteria: (1) Full text can be acquired, (2) The study focuses on the combined toxicity of micro(nano)plastics and environmental pollutants, (3) the physicochemical properties of both micro(nano)plastics and pollutants have been characterized, (4) experimental conditions have been clearly described, and (5) the publication is not a review article. Here, we focused on the combined ecotoxicity, and the survival rate was selected as the toxicity endpoint. The selection of the survival rate as the toxicity endpoint was based on two main considerations. First, the survival rate is a critical and widely recognized biological indicator for evaluating the ecotoxicological effects of micro(nano)plastics and environmental pollutants on organisms. Second, the survival rate provides a standardized, quantitative output that is both unambiguous and computationally tractable.

We systematically screened the literature by first downloading all titles and abstracts of 9092 articles in batches, excluding nonrelevant studies through title/abstract screening, followed by full-text retrieval of remaining articles and multi-reviewer manual verification to ensure alignment with our research scope. This consistency ensures reproducibility and broader applicability for machine learning modeling. Ultimately, 41 studies were selected for further data extraction. A full list of the selected 41 papers is provided in Table S1. Herein, we aimed to identify the critical factors influencing the toxic effects caused by micro

(nano)plastics after adsorbing environmental pollutants. The selected publications were used to extract the physicochemical properties of micro(nano)plastics along with the corresponding experimental conditions. These input features included the material type, particle shape, and size, surface modifications, exposure concentration of micro(nano) plastics, exposure time, and types and exposure concentrations of environmental pollutants. Detailed information can be found in Table S2. Toxic effects characterized by survival rates of different species were used as output targets. The survival rates were extracted from relevant figures using the digitization tool in Origin2022.

2.2. Data processing

Raw data were further processed to improve usability due to the variability across different studies. First, all values for physicochemical properties and experimental conditions were standardized to consistent units, such as particle size (nm), exposure concentration ($\mu\text{g/mL}$), and exposure time (hours). Outliers and missing values were removed to enhance data integrity and reliability. In addition, one-hot encoding was employed to convert categorical variables into numerical values of 0 or 1 [35]. One-hot encoding removes potential ordering relationships of category values and ensures that different categories are all considered equally in the model. This technique allows the model to capture the properties of different categories more accurately and thus improves model performance. Furthermore, z-score standardization was applied to minimize the negative impact of feature dimensionality on model performance, which converts data of different magnitudes to a uniform

scale, expressing values with a mean of 0 and a standard deviation of 1. This method enables direct comparison of values with differing units or magnitudes such as varying exposure concentrations.

2.3. Construction of interpretable machine learning models

This study used various machine learning methods for modeling to identify optimal results. The modeling methods include random forest, extreme gradient boosting, gradient boosted decision trees, decision trees, k-nearest neighbors, artificial neural network, support vector machine. These machine learning methods have been widely used in some studies and have different characteristics and advantages [36,37]. For example, the random forest model has strong overfitting resistance by integrating multiple decision trees and provides insights into feature importance. The artificial neural network model is highly effective at recognizing patterns in complex and high-dimensional data. All traditional machine learning models were constructed using the Scikit-learn v0.24.1 in Python. The artificial neural network model was implemented in PyTorch v1.0.2 of Python. In addition, an ensemble machine learning model was constructed through a stacking strategy. As shown in Fig. S1, the stacking ensemble machine learning model combines the results of multiple base models to obtain a final prediction. The top three performing individual models served as base learners, and linear regression was used for the final prediction. Ensemble predictions were generated from the weighted averages of the base learners. The primary goal of stacking is to leverage the strengths of different models to create a single robust model.

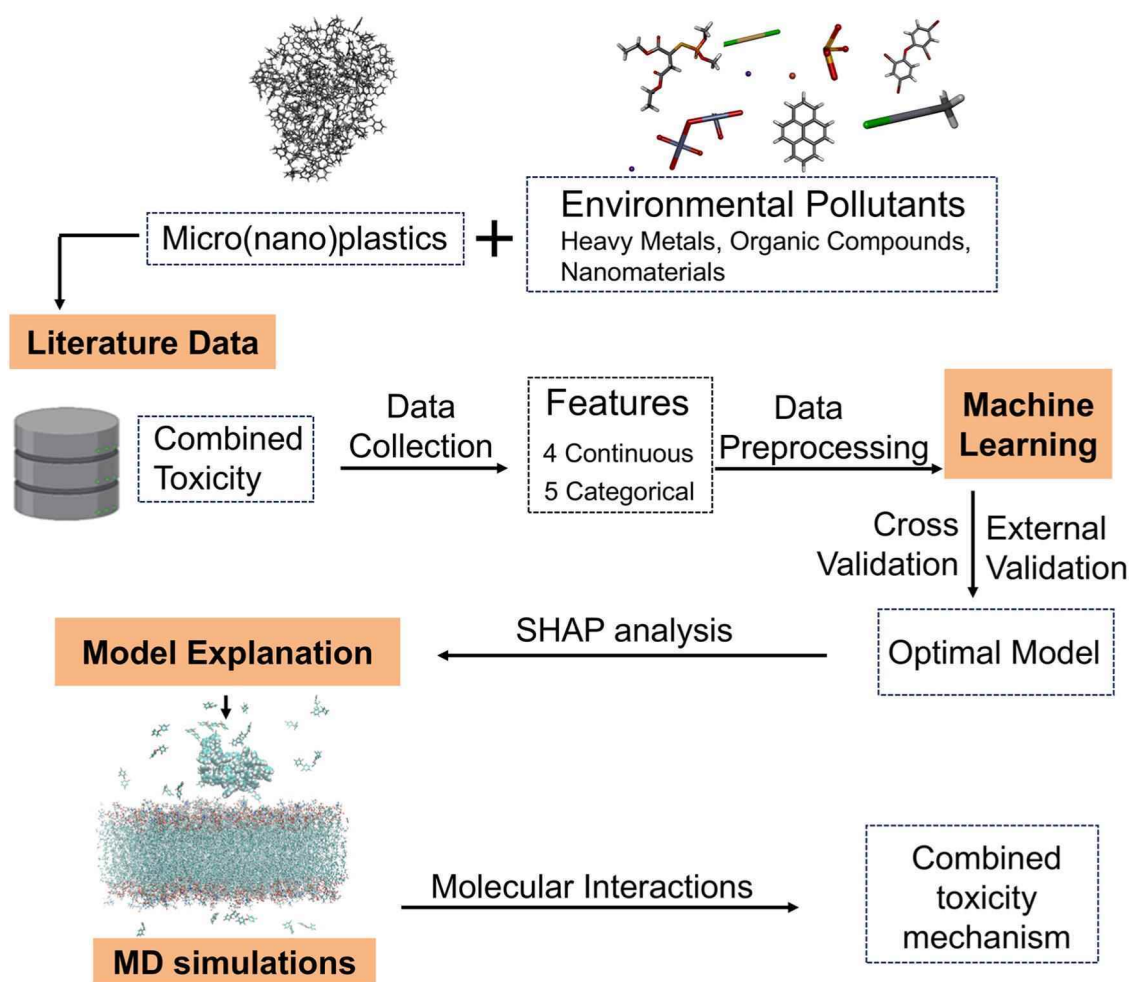


Fig. 1. Workflow of the combined toxicity prediction of micro(nano)plastics with environmental pollutants. In particular, the workflow contained four main steps, i. e., literature data mining, machine learning modeling, model interpretation, and MD simulations.

The whole dataset was randomly divided into training and test sets. Grid search was used for hyperparameter tuning. The key benefit of grid search lies in the comprehensive exploration within a defined parameter space, ensuring the identification of the optimal hyperparameter set. Due to the large number of tunable parameters, it is neither feasible nor unnecessary to evaluate all possible combinations of parameters. Therefore, we restricted parameter optimization to a reasonable range and assessed the performance of the relevant models. The main parameters and their ranges are shown in Table S3. Furthermore, 5-fold cross validation was employed to reduce the risk of overfitting. Model performance was evaluated using standard regression metrics, namely, the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE), which help quantify the differences between predictions and actual values. In addition, we used SHAP values to analyze and interpret the feature importance of the constructed machine learning model. The SHAP value explains how much each feature contributes to the final prediction, helps understand how the model makes decisions, and improves the transparency of the model [38]. In addition, a feature ablation approach was used to evaluate the impact of basic features on model performance. In particular, feature ablation removes individual input features from the model and assesses the effect of features on model predictions.

2.4. MD simulations

The interactions between micro(nano)plastics and BDE-47 (2,2',4,4'-tetrabromodiphenyl ether) with cell membranes were simulated using MD methods. The topology and force field parameters of micro(nano)plastics and BDE-47 were generated using SwissParam [39] based on the Charmm36 force field. The parameters of the polystyrene (PS) and BDE-47 molecules are provided in Table S4 and Table S5. The CHARMM36 force field and its complementary CGenFF have been rigorously optimized and validated for lipid and ligand parameters, rendering them particularly well-suited for simulating biological systems involving cell membranes [40,41]. The interactions between the PS+BDE-47 complex and the cell membrane were visualized using the Multiwfn version 3.8 (dev, updated on May 5, 2025) [42,43]. The micro (nano)plastic model, self-assembled from five PS chains (polymerization degree = 10), exhibited a diameter of ~4 nm. We employed a simplified mixed lipid bilayer that was composed of neutral charged dioleoyl phosphatidylcholine and negatively charged dioleoyl phosphatidylglycerol molecules in a 7:3 ratio, which has been extensively utilized in previous simulations and *in vitro* experiments [44]. Water molecules were simulated using the extended simple point charge model, and sodium ions were added to neutralize the system. The initial size of the simulation box was 10 nm × 10 nm × 11 nm. Initially, a modified PS sphere was placed above the equilibrated lipid bilayer, and 40 BDE-47 molecules were randomly inserted. The system energy was minimized using the steepest descent method until the maximum force was less than 600 kJ/mol. Then, the system was equilibrated under the NVT and NPT ensemble, respectively. In the NVT ensemble, the number of atoms (N), volume (V), and temperature (T) of the system were conserved; in the NPT ensemble, the number of atoms (N), pressure (P), and temperature (T) were conserved. After that, a 100-ns production MD simulation was performed based on the equilibrated system under the NPT ensemble. The temperature was fixed at 300 K using the V-rescale method with a coupling constant of 0.1 ps. The pressure was kept constant at 1 bar using a semi-isotropic Berendsen barostat with a coupling constant of 1 ps. The Lennard-Jones parameters for non-bonded interactions were determined using the conventional Lorentz-Bertelot combination rules. All non-bonded interactions were truncated at a cutoff of 1.2 nm, and the long-range electrostatic interactions were calculated using the particle-mesh-Ewald method. The covalent bonds were constrained using the Lincs algorithm. All simulations were performed with a time step of 1 fs using GROMACS v2020.5 [45]. Periodic boundary conditions were considered in all three directions of the

simulation system. The simulation snapshots were drawn using VMD 1.9.4 [46]. More details of MD simulations can be found in Method S1.

3. Results and discussion

3.1. Overview of the toxicity dataset

Through comprehensive and rigorous literature mining up to August 2024, we compiled an extensive dataset on the combined toxicity of environmental pollutants adsorbed by micro(nano)plastics from 41 peer-reviewed articles. The dataset comprises 351 data points, which contain 9 input features describing the physicochemical properties of relevant micro(nano)plastics and environmental pollutants as well as the experimental conditions. The survival rate was used as the output target. As shown in Fig. 2a, the dataset includes the commonly used PS and polyethylene, with PS micro(nano)plastics accounting for the vast majority at 83.06 %. Meanwhile, other types have a much smaller proportion. Consistent with the current findings, PS is one of the most widely used plastics [47]. Currently, most studies focus on spherical particles (93.16 %), with non-spherical particles accounting for only 6.84 %. Although such spherical particles benefit synthesis and toxicity mechanism analysis, they do not fully represent the high particle heterogeneity caused by different sources and environmental conditions. In addition, most toxicological studies have focused on commercially available particles. To enhance our understanding of micro(nano)plastic effects, it is crucial to perform the micro(nano)plastic exposure studies under environmentally realistic conditions [48]. The particle size (diameter) ranges from 20 nm to 500 μ m (Fig. S2a), reflecting that the tested materials cover a broad scale from nanometer to micrometer levels. Microplastics and nanoplastics comprised 54.7 % and 45.3 % of the dataset, respectively. The plastic particles of different sizes may lead to different common toxicities from pollutants. Previous studies have shown that triphenyl phosphate alone can stimulate the liver and gonad of fish by 1.25–2.12 times, and the presence of nanoplastics further aggravates this stimulation by 1.23–2.84 times, while microplastics cannot [49]. Additionally, the majority of micro(nano)plastics are unmodified on the surface (94.87 %), while a small portion of particles are modified with amino groups (NH_2 , 2.56 %) or carboxyl groups (COOH , 2.56 %). Fig. S2 depicts the detailed distributions of physicochemical properties and experimental conditions. It can be found that most of the exposure concentrations in our dataset were within 5 μ g/mL or even 1 μ g/mL (Figs. S2b and S2c), which can be used to evaluate the safety of micro(nano)plastics at low concentrations. Regarding exposure time, most measurements were performed within 48 h after exposure to micro (nano)plastics (Fig. S2d). Such a short period can meet the high-throughput screening of toxicology experiments.

Fig. 2b shows the combined toxicity studies involving micro(nano) plastics and environmental pollutants across different species. The test species encompass various ecological categories, namely, rotifers, plants, crustaceans, nematodes, mollusks, teleost fish, and microorganisms. Aquatic organisms (84 %), such as zebrafish (34 %) and *Daphnia magna* (13 %), are most frequently used in previous studies. Zebrafish are excellent models for investigating the ecotoxicity of environmental pollutants due to their rapid development, transparency during early stages, genetic similarity to humans, and cost effectiveness, and high-throughput screening capabilities [50]. Furthermore, the survival rate exhibited a wide distribution ranging from 0 % to 100 %, with a mean (\pm standard deviation) of 60.34 ± 35.78 (Fig. 2c). This wide distribution ensures better representation of the target variable, reducing bias, improving generalization, and enhancing the ability of the model to learn patterns and make accurate predictions across diverse scenarios.

In addition, we analyzed the distribution of different types of pollutants and compared their toxicity (Figs. S3 and S4). The environmental pollutants adsorbed by micro(nano)plastics primarily include heavy metals (52 %), organic compounds (38 %), and nanomaterials (10 %). Some heavy metal ions, such as silver ions (Ag^+ , 7 %), typically exhibit

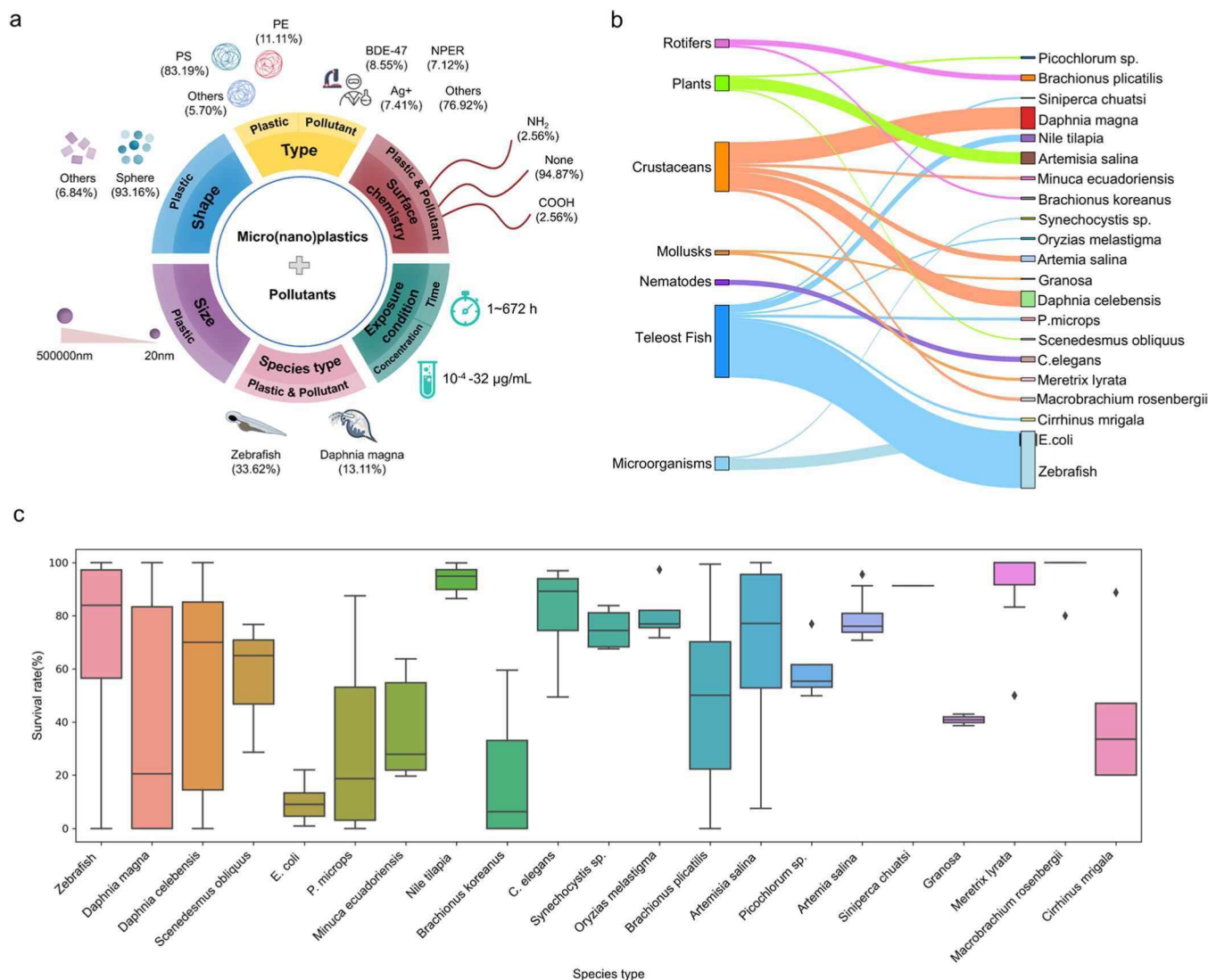


Fig. 2. Overview of the curated toxicity dataset. (a) Physicochemical properties of micro(nano)plastics and adsorbed pollutants, and the corresponding experimental conditions. (b) Species are used for assessing the combined toxicity of micro(nano)plastics and environmental pollutants. (c) Survival rate distribution of different species.

significant toxic effects on various biological groups [51]. Silver ions can generate reactive oxygen species, inducing oxidative stress in cells. This causes oxidative damage to cellular components, including the cell membrane, mitochondria, lysosomes, and nucleus, ultimately leading to cell apoptosis or necrosis [52,53]. The present dataset involved more than 19 types of organic pollutants, including plasticizers, flame retardants, insecticides, and antibiotics. Although these chemicals improve the quality of our lives, they also pose a significant threat to the environment and human health. Furthermore, micro(nano)plastics can function as vectors of other environmental pollutants and thus significantly increase their potential hazards [54,55]. Our dataset also contained four nanomaterials, i.e., TiO₂ nanoparticles, graphene oxide [56, 57], silver nanoparticles, and nanosized permethrin. These nanomaterials have been commonly used in previous toxicology studies and demonstrated to induce various adverse outcomes in biological systems [58,59].

In summary, the constructed combined toxicity dataset contains sufficient high-quality data samples, encompassing diverse input features and output targets. This strongly supports identifying key structure (property)–toxicity relationships and lays a solid foundation for the subsequent machine learning research.

3.2. Modeling quantitative property–toxicity relationships

After constructing the available dataset, the logical next step is to mine enough scientific evidence to support toxicological studies through computational modeling. This study constructed eight machine learning models (random forest, extreme gradient boosting, gradient boosted decision trees, decision trees, k-nearest neighbors, artificial neural network, support vector machine, and ensemble machine learning) to establish the quantitative property–toxicity relationships under the co-exposure of micro(nano)plastics and environmental pollutants. As the input variables of machine learning models, the data quality of physicochemical properties and environmental conditions directly influences the performance and interpretability of models. Irrelevant features may introduce noise, while redundant features can increase the risk of overfitting and computational complexity. Therefore, we first evaluated the quality of input features through principal component analysis (PCA) and Pearson correlation coefficient (PCC). Fig. 3a shows a three-dimensional scatter plot based on PCA, which provides an intuitive analysis of the distribution of the training set and test set in the chemical space. It can be found that the data points are relatively dispersed, with some overlap between the training and test sets, indicating similarity in certain features but noticeable differences in specific dimensions. Fig. 3b

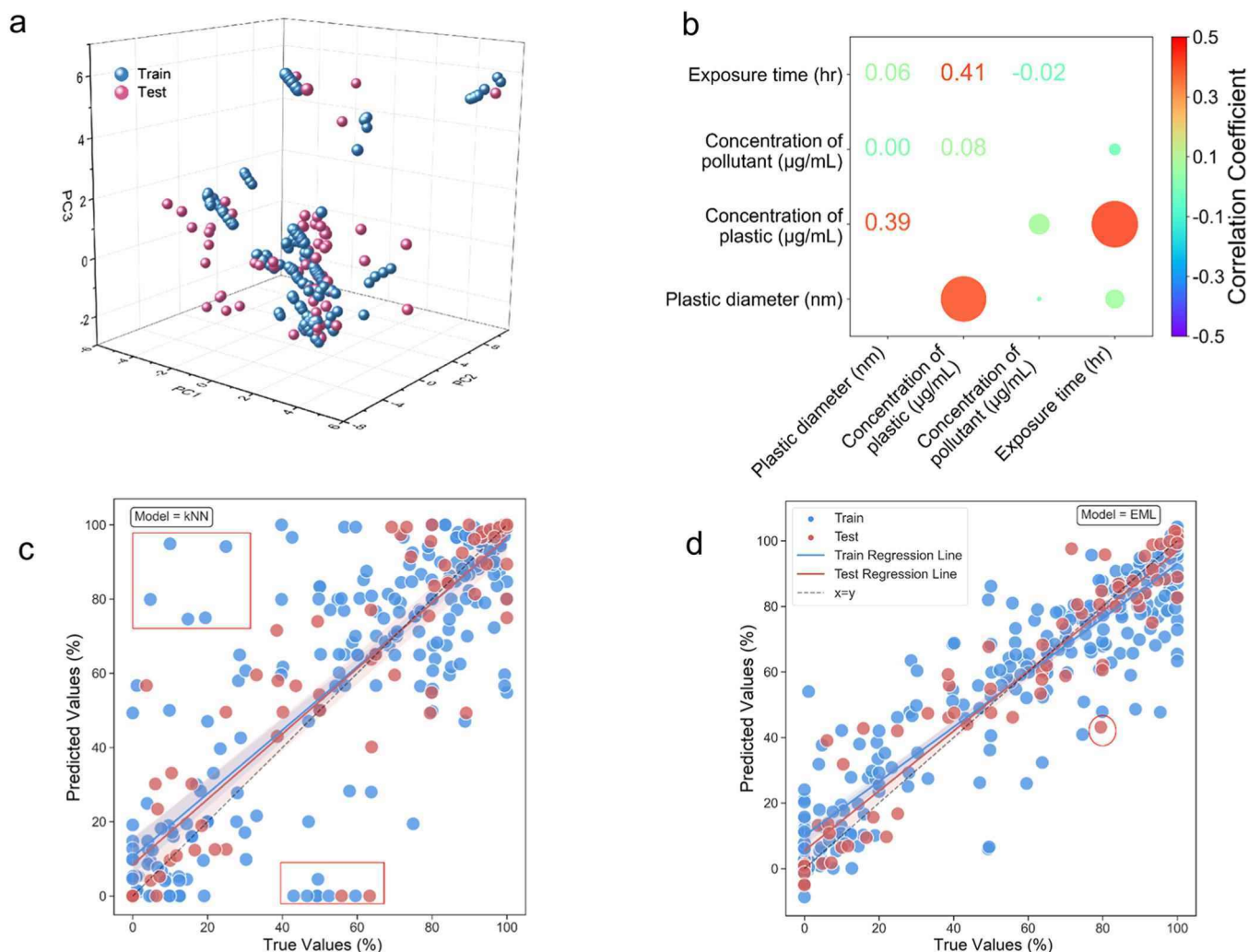


Fig. 3. Analysis of the input features and model performance. (a) PCA of all input features used in the current dataset. (b) Heatmap of the PCC matrix between continuous input features. The bubble size represents the magnitude of the correlation. (c) Correlations between experimental values and the predictions from the k-nearest neighbors. (d) Correlations between experimental values and the predictions from the ensemble machine learning model. Red boxes highlight the prediction outliers.

depicts the PCC values among four continuous variables: exposure time, pollutant concentration, plastic concentration, and plastic diameter. The PCC measures the degree of the linear correlation between two variables. It can be observed that the continuous variables exhibit a moderate ($0.3 < \text{PCC} < 0.49$) or low ($\text{PCC} < 0.29$) degree of correlation. Overall, these results indicate that the input features provide sufficient usefulness, but not redundant information, to differentiate between data points.

Under the optimal parameter configuration, almost all machine learning models achieve an acceptable prediction performance ($R^2 > 0.5$, Table S6) [60]. The ensemble machine learning model performed the best, with the highest R^2 values and lowest RMSE and MAE values. In addition, both 5-fold cross validation and external validation have an R^2 of > 0.84 , indicating the robustness and generalization of the ensemble machine learning model. Furthermore, we employ the y-randomization test to ensure the reliability of the model results. As a commonly used model validation method, y-randomization test randomly rearranges the output targets while keeping the input features the same. As shown in Fig. S5, y-randomization leads to a sharp decrease in model performance, indicating that our original machine learning model was not obtained by chance. By integrating multiple base learners (such as random forest, extreme gradient boosting, and gradient boosted decision trees), the ensemble machine learning model effectively reduced the

bias and variance of the individual models, significantly improving the overall predictive accuracy and stability. Notably, the k-nearest neighbors model performed worst in 5-fold cross validation and external validation. This may be due to its reliance on distance metrics (such as the Euclidean distance), which can become unstable when feature scales or distributions are inconsistent, causing some features to be overly weighted and negatively affecting model performance. The relatively poor performance of the artificial neural network model was likely due to the limited sample size, which prevented the network from effectively learning key patterns, resulting in insufficient generalization. The poor performance of the support vector machine model may be attributed to its limited ability to capture complex relationships between features, which in turn affected its performance in the current dataset.

To intuitively visualize the predictive performance of the constructed machine learning models, we investigated the correlations between the experimentally measured survival rates and predictions from machine learning models. As shown in Fig. 3c, the worst performing k-nearest neighbors model produced noticeable outliers that deviated from the $y = x$ line (highlighted by red boxes). However, these outliers have largely disappeared in the ensemble machine learning model (Fig. 3d), indicating that the ensemble machine learning model achieves better fitting on both the training and test sets. Even so, the prediction performance of the ensemble machine learning model needs to be

further improved. As shown in Fig. 3d, the red circle highlights the discrepancy between the predicted survival rate (43.14 %) by the model and the experimentally measured survival rate (79.49 %) of *Daphnia magna* after 24 h co-exposure to polyethylene microplastics (10 µg/mL) and K₂Cr₂O₇ (0.5 µg/mL). This discrepancy may be due to the sudden decrease in the survival rate from 100 % to 79.49 % observed in the experimental data when the pollutant concentration increased from 0.4 to 0.5 µg/mL, which the model failed to accurately capture due to insufficient median values in the training data. In the future, we still need to expand the diversity of dataset to further improve our model generalization performance. In addition, the development of more advanced molecular descriptors (e.g., tetrahedral descriptors) and machine learning architectures (e.g., transfer learning) can enhance predictive performance by better capturing intricate structure features and nonlinear relationships [25,61]. Furthermore, machine learning models should incorporate a broader range of experimental variables, such as environmental pH levels, to enhance predictive accuracy.

A user interaction system (accessible at <https://www.pubvinas.com/index.php?s=predict&c=pollutant>) was developed to facilitate public access to the predictive model. As illustrated in Fig. S6, users can input feature data into designated fields, and the system evaluates the survival rate of various biological organisms using the pre-trained ensemble machine learning model. This online platform offers a rapid method for risk assessment of micro(nano)plastics with other pollutants in the environment. The system will be continuously updated to incorporate

new experimental data, further refining its accuracy and applicability in ecotoxicological studies.

3.3. Key factors influencing the combined toxicity

According to the Organization for Economic Cooperation and Development AI principles, machine learning models should be interpretable. Their interpretation allows us to identify key physicochemical properties or experimental conditions related to the combined effects of micro(nano)plastics and environmental pollutants, which can be used to elucidate the potential toxicity mechanisms and mitigate the adverse effects. This study uses SHAP values to determine the degree of that impact features have on the combined toxicity. As shown in Fig. 4a, the key contribution of features to the ensemble machine learning model is reflected in their high ranking. In general, pollutant exposure concentration, species type, and pollutant type were the top three critical characteristics affecting the combined toxicity. Next, we performed feature ablation to intuitively understand the influence of these essential features on the performance of the model. Feature ablation removes individual input features from a model and evaluates the impact on model predictions. Generally, the prediction performance of the ensemble machine learning model gradually decreases as the input features are successively eliminated (Fig. 4b). In particular, the R² value dropped sharply from 0.83 to 0.44 after the top three features (concentration of pollutant, species type, and pollutant type) were removed,

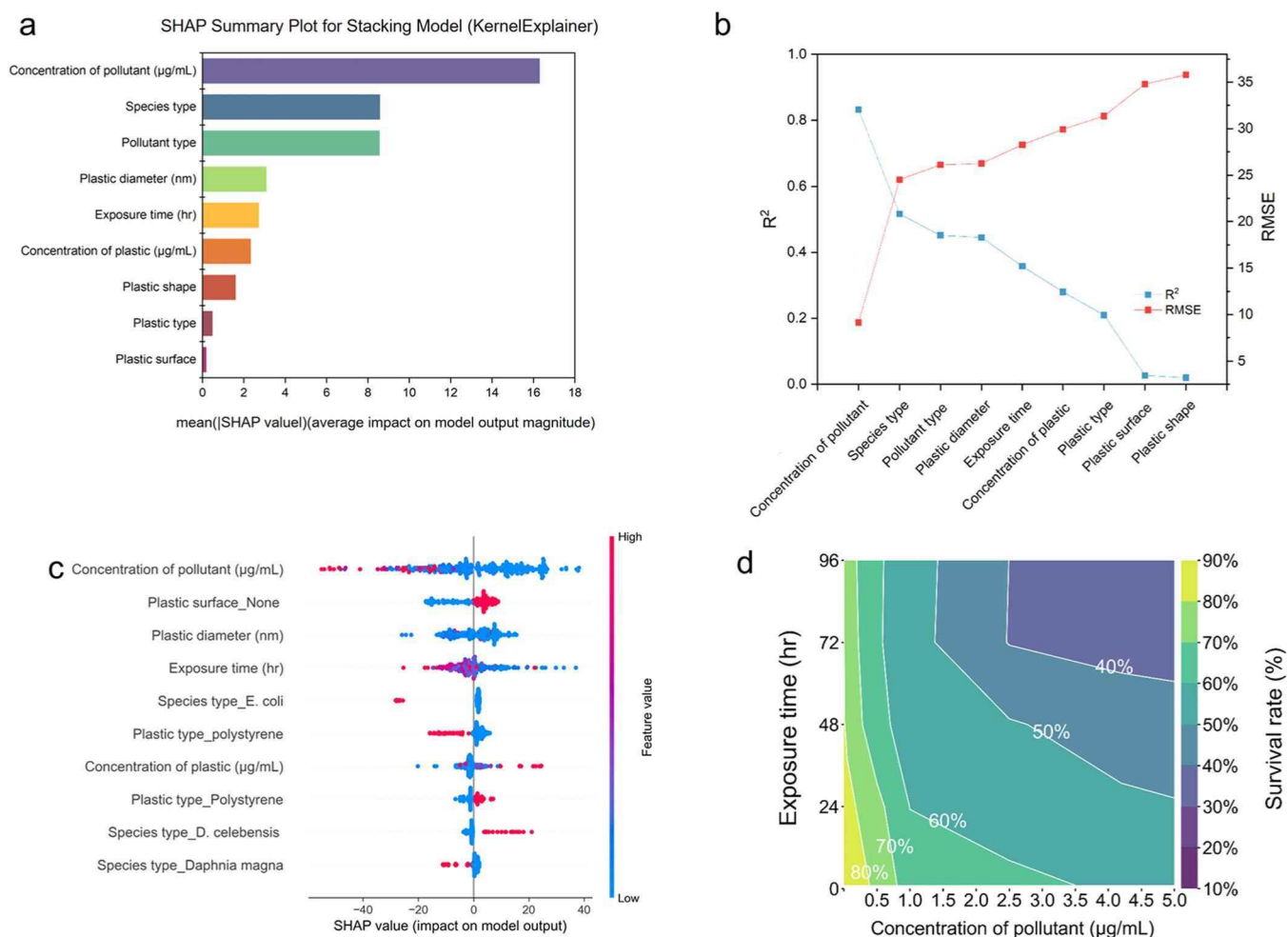


Fig. 4. Interpretation of the machine learning models. (a) SHAP-based feature importance interpretation of the ensemble machine learning model. (b) Prediction accuracy of 5-fold cross validation under feature ablation testing. The features were progressively removed in order of their importance. (c) Local interpretability based on SHAP values. (d) Interaction effects of pollutant concentration and exposure time on model prediction through two-dimensional partial dependence plots analysis.

further demonstrating the importance of these features to the model performance.

We further investigated the contribution of each feature to a given data sample using SHAP local explanation. As shown in Fig. 4c, higher contaminant concentration and shorter exposure time are generally associated with lower survival. For example, at low concentrations, nanosized PS adsorbed polychlorinated biphenyls, decreased its freely dissolved fraction, and ultimately reduced toxicity. However, at high concentrations, the PS itself became a major source of toxicity, potentially causing mortality through physical entrapment or chemical effects [62]. Typically, it involves complex nonlinear interactions between features and the target variable. Therefore, we further applied two-dimensional partial dependence plots to analyze the interaction effect of pollutant concentration and exposure time on model prediction (Fig. 4d). Here, pollutant concentration and exposure time were selected as an example analysis because of their high ranking importance in predictive modeling and easily understood associations with toxic effects on organisms. In addition, the partial dependence plots method is more suitable for analyzing the relationship between continuous variables. Results show that toxicity gradually intensifies as pollutant concentration and exposure time increase, leading to decreased organism survival. At higher concentrations ($>3 \mu\text{g/mL}$) and longer exposure times ($>72 \text{ h}$), survival rates decrease below 40 %, likely due to cumulative effects exacerbating physiological damage. Conversely, at low concentrations ($<1 \mu\text{g/mL}$) and short exposure times ($<24 \text{ h}$), survival rates remain high, approaching 80 %–90 %. Notably, even with prolonged exposure, if pollutant concentration remains low ($<1 \mu\text{g/mL}$), survival rates can remain above 70 %, suggesting that pollutant concentration may have a more significant impact on survival than exposure time.

Particle diameter was prioritized as the most critical micro(nano) plastic property governing the combined toxicity in the feature importance analysis. The size of micro(nano)plastics affects their uptake, distribution, and biocompatibility in living organisms. Smaller nanoplastics are more likely to be absorbed by cells, potentially triggering cytotoxic reactions [63]. However, other micro(nano)plastic properties (e.g., plastic type and surface property) closely associated with toxicity are ranked low (Fig. 4a), highlighting the limitation of the current machine learning models. For example, surface modification can alter the hydrophobicity and charge of micro(nano)plastics, thereby affecting their interaction with organisms and toxicity [64]. The reduced contribution of the variables is primarily due to their limited diversity, i.e., they have the same value for most data points. For example, over 90 % nanoplastics used in previous toxicology studies are unmodified sphere particles (Fig. 2a). Therefore, these variables do not provide much new information for the model. Machine learning models rely on the variability of variables to distinguish between different patterns or classes. If a variable has low variability, it cannot effectively help the model make decisions. In turn, the model might assign less importance to this feature because it does not contribute much to the prediction. More structurally diverse micro(nano)plastics should be incorporated in future toxicology studies to address current data limitations and to ensure comprehensive environmental risk evaluations.

3.4. Interaction mechanisms revealed by molecular dynamics simulations

We conducted MD simulations to further explore the molecular-level toxicity mechanisms of micro(nano)plastics after adsorbing environmental pollutants. The feature importance analysis identified pollutant concentration as a key factor influencing the combined toxicity. Therefore, we designed two scenarios for comparison, i.e., the cross-membrane dynamics of micro(nano)plastics with or without adsorbed pollutants. Here, we selected a phospholipid bilayer as the cell membrane model in the simulation system because of its simple structure and ability to effectively represent the fundamental physicochemical properties of real biological membranes, providing a reliable theoretical

basis for studying the interaction between micro(nano)plastics and cell membranes.

The constructed dataset primarily included heavy metals and organic pollutants. Compared with heavy metals, organic compounds are better characterized in existing force fields, enabling more reliable MD simulations. As the most prevalent organic pollutant identified in the current dataset, BDE-47 has raised significant concerns about its ecological and health risks [65]. Therefore, this study intentionally narrows its focus to BDE-47 as a representative case study to deeply explore interaction mechanisms. The focused investigation enables systematic methodological refinement, which can be adapted to study a broader range of pollutants in future research. Herein, all-atom MD simulations were performed. The initial simulation systems are shown in Fig. 5a and b. As depicted in Fig. 5c and d, PS molecules exhibit different molecular interaction mechanisms with cell membranes in the presence or absence of BDE-47. In the simulation system without BDE-47, during the 100 ns simulation, PS molecules neither interact with the cell membrane, nor do they tend to penetrate the membrane, indicating that single micro(nano)plastics exhibit low toxicity. In another system, we added BDE-47, allowing it to coexist with PS in the simulation system. The simulation results demonstrated the effects of PS particles adsorbing BDE-47 on the cell membrane at different time points. At 0 ns, PS particles and BDE-47 molecules were distributed near the surface of the cell membrane. By 25 ns, BDE-47 molecules began to aggregate and gradually adsorbed onto the surface of PS particles, while the PS particles started to interact with the cell membrane. At 50 ns, PS particles were partially inserted into the cell membrane, and some BDE-47 molecules also penetrated the membrane. Finally, at 100 ns, PS particles were further inserted into the cell membrane, causing membrane deformation. In summary, the adsorption of BDE-47 may alter the interaction mechanism between the PS particles and the cell membrane, facilitating the insertion of PS particles into the membrane and significantly enhancing their toxic effects.

Fig. 5e shows the interaction energy distribution between the PS particles and the cell membrane (blue bars) and between the PS+BDE-47 system and the cell membrane (green bars). Results indicate that the Lennard-Jones interaction energy (van der Waals interaction) of the PS+BDE-47 system is significantly higher than that of the PS-only system, suggesting that the adsorption of BDE-47 enhances the van der Waals interactions. The visualization analysis further highlights the strong van der Waals interactions between the PS+BDE-47 complex and the cell membrane, as depicted by the large and widespread green iso-surfaces in Fig. S7. In addition, localized blue regions suggest the presence of attractive interactions, such as hydrogen bonding or halogen bonding. These findings indicate that the PS+BDE-47 complex can establish stable and diverse interactions on the membrane surface, which may potentially disrupt membrane integrity and exert toxic effects on biological systems.

Overall, MD simulations show that after adsorbing the pollutant BDE-47, the PS particles exhibit enhanced interactions with cell membranes, potentially accelerating the transmembrane process of the complex. This process may induce the generation of reactive oxygen species, activate inflammatory pathways such as NF- κ B, and promote the release of inflammatory factors, ultimately leading to cell apoptosis or necrosis and exacerbating chronic toxicity effects [66–68]. These mechanisms collectively reveal the enhanced toxicity of micro(nano) plastics in combined pollution scenarios, providing theoretical support for further research on their environmental risks and biological safety. While our MD simulations provide valuable insights into the interaction mechanisms between environmental pollutants and simplified lipid bilayers, certain limitations must be acknowledged. First, the lipid bilayer cannot fully replicate the structural and functional complexity of *in vivo* biological membranes (e.g., lipid diversity, embedded proteins, or membrane asymmetry). Second, despite efforts to align parameters with experimental conditions, inherent computational constraints, such as truncated timescales and limited spatial resolution, may not fully

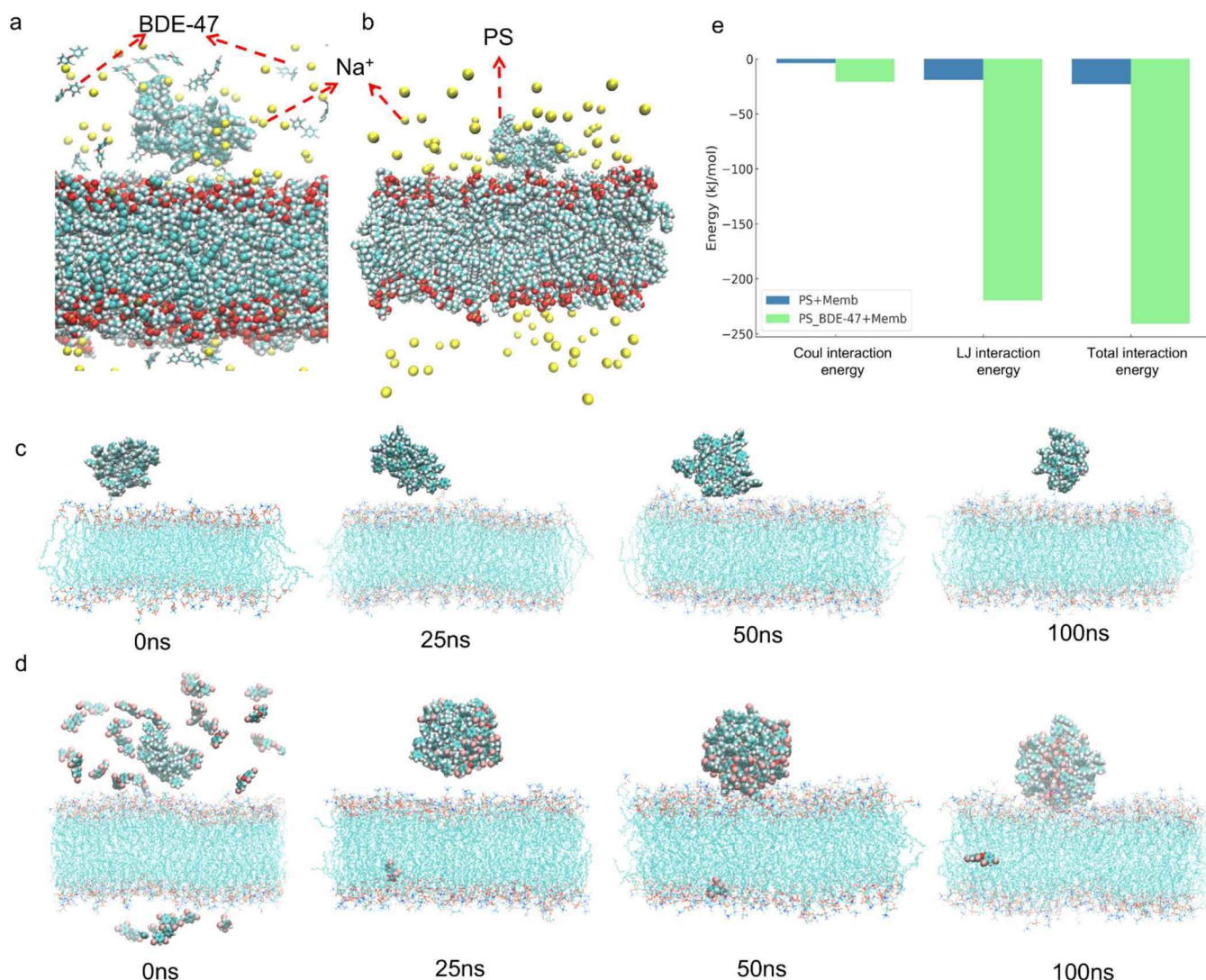


Fig. 5. MD simulations showing interactions between PS, BDE-47, and cell membrane. (a) Initial system with pollutants: BDE-47 molecules and Na⁺ ions (yellow). (b) Initial system without pollutants: only PS and cell membrane. (c) Time series of interactions between the PS and the membrane showing minimal interaction. (d) Time series of interactions between PS, BDE-47, and membrane showing enhanced PS adsorption and gradual insertion into the cell membrane. (e) Interaction energies between the cell membrane and the PS particles with (green bars) or without BDE-47 (blue bars).

capture the dynamic heterogeneity of real-world cellular environments. These simplifications highlight the need for complementary experimental validation to bridge the gap between simulation predictions and biological reality.

4. Conclusion

Micro(nano)plastics, as emerging pollutants, may synergize with adsorbed environmental contaminants to amplify ecological and human health risks. However, understanding their combined toxicity and mechanisms remains limited, compounded by fragmented research and data. To bridge this gap, we integrated machine learning and MD simulations to systematically decode toxicity patterns and interaction mechanisms. The main conclusions are as follows: (1) The ensemble machine learning model outperformed single-model frameworks, achieving robust predictions of quantitative property–toxicity relationships. (2) MD simulations unveiled atomistic insights into the membrane penetration mechanisms of micro(nano)plastic–pollutant complexes. (3) The online predictive system transforms heterogeneous data into an actionable tool for policymakers, enabling evidence-based safety thresholds for micro(nano)plastics in environmental matrices. Current

limitations include dataset diversity, necessitating collaborations between experimentalists and modelers to expand data quality and coverage. Future work should also incorporate multi-omics data and bioinformatics approaches to elucidate *in vivo* toxicity mechanisms, offering stronger support for the risk assessment of micro(nano)plastic–pollutant complexes.

Environmental implication

This study integrates global toxicity data and ensemble machine learning to decode combined toxicity mechanisms, providing a critical foundation for targeted mitigation strategies and regulatory decision-making. For instance, identifying high-risk micro(nano)plastic–pollutant pairs enables regulators to prioritize bans or restrictions on specific chemical products based on mechanistic evidence. Furthermore, understanding how the combined toxicity varies across species types is essential for setting stricter micro(nano)plastics limits in habitats housing high-sensitivity or endangered species. Our findings also support evidence-based interventions, such as updating water quality standards and ecological risk assessment frameworks to account for micro(nano)plastic–pollutant complexes.

CRediT authorship contribution statement

Xiliang Yan: Writing – review & editing, Writing – original draft, Resources, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization. **Jing Zhang:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Data curation. **Xiaofang Li:** Writing – review & editing, Writing – original draft, Methodology, Investigation. **Hanle Chen:** Writing – review & editing, Methodology. **Guohong Liu:** Writing – review & editing, Investigation, Funding acquisition. **Bing Yan:** Writing – review & editing, Funding acquisition, Conceptualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by the National Key R&D Program of China (2023YFA0915101), the National Natural Science Foundation of China (22106025, 22476056, and 22036002), the Specific University Discipline Construction Project (2023B10564001), the Introduced Innovative R&D Team Project under the “The Pearl River Talent Recruitment Program” of Guangdong Province (2019ZT08L387), and the Guangdong Basic and Applied Basic Research Foundation (2022A1515111082).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at [doi:10.1016/j.jhazmat.2025.138911](https://doi.org/10.1016/j.jhazmat.2025.138911).

Data availability

The source codes and data can be found at GitHub.

References

- Mitrano, D.M., Wick, P., Nowack, B., 2021. Placing nanoplastics in the context of global plastic pollution. *Nat Nanotechnol.* 16, 491–500. <https://doi.org/10.1038/s41565-021-00888-2>.
- Pérez-Reverón, R., Álvarez-Méndez, S.J., González-Sálamo, J., Socas-Hernández, C., Díaz-Peña, F.J., Hernández-Sánchez, C., Hernández-Borges, J., 2023. Nanoplastics in the soil environment: analytical methods, occurrence, fate and ecological implications. *Environ Pollut* 317, 120788. <https://doi.org/10.1016/j.envpol.2022.120788>.
- Allen, S., Allen, D., Phoenix, V.R., Le Roux, G., Durántez Jiménez, P., Simonneau, A., Binet, S., Galop, D., 2019. Atmospheric transport and deposition of microplastics in a remote mountain catchment. *Nat Geosci* 12, 339–344. <https://doi.org/10.1038/s41561-019-0335-5>.
- Mason, S.A., Garneau, D., Sutton, R., Chu, Y., Ehmann, K., Barnes, J., Fink, P., Papazissimos, D., Rogers, D.L., 2016. Microplastic pollution is widely detected in US municipal wastewater treatment plant effluent. *Environ Pollut* 218, 1045–1054. <https://doi.org/10.1016/j.envpol.2016.08.056>.
- Li, J., Liu, H., Chen, J.P., 2018. Microplastics in freshwater systems: a review on occurrence, environmental effects, and methods for microplastics detection. *Water Res* 137, 362–374. <https://doi.org/10.1016/j.watres.2017.12.056>.
- Amobonye, A., Bhagwat, P., Raveendran, S., Singh, S., Pillai, S., 2021. Environmental impacts of microplastics and nanoplastics: A current overview. *Front Microbiol* 12, 768297. <https://doi.org/10.3389/fmicb.2021.768297>.
- Shen, M., Huang, W., Chen, M., Song, B., Zeng, G., Zhang, Y., 2020. Microplastic crisis: Un-ignorable contribution to global greenhouse gas emissions and climate change. *J Clean Prod* 254, 120138. <https://doi.org/10.1016/j.jclepro.2020.120138>.
- Liu, Y., Guo, R., Zhang, S., Sun, Y., Wang, F., 2022. Uptake and translocation of nano/microplastics by rice seedlings: evidence from a hydroponic experiment. *J Hazard Mater* 421, 126700. <https://doi.org/10.1016/j.jhazmat.2021.126700>.
- Jovanović, B., 2017. Ingestion of microplastics by fish and its potential consequences from a physical perspective. *Integr Environ Assess Manag* 13, 510–515. <https://doi.org/10.1002/ieam.1913>.
- Song, K., Xue, Y., Li, L., Deng, M., Zhao, X., 2022. Impact and microbial mechanism of continuous nanoplastics exposure on the urban wastewater treatment process. *Water Res* 223, 119017. <https://doi.org/10.1016/j.watres.2022.119017>.
- Yee, M.S., Hii, L.W., Looi, C.K., Lim, W.M., Wong, S.F., Kok, Y.Y., Tan, B.K., Wong, C.Y., Leong, C.O., 2021. Impact of Microplastics and Nanoplastics on Human Health. *Nanomaterials* 11, 496. <https://doi.org/10.3390/nano11020496>.
- Leslie, H.A., van Velzen, M.J.M., Brandsma, S.H., Vethaak, A.D., Garcia-Vallejo, J. J., Lamoree, M.H., 2022. Discovery and quantification of plastic particle pollution in human blood. *Environ Int* 163, 107199. <https://doi.org/10.1016/j.envint.2022.107199>.
- Yan, Z., Liu, Y., Zhang, T., Zhang, F., Ren, H., Zhang, Y., 2022. Analysis of Microplastics in Human Feces Reveals a Correlation between Fecal Microplastics and Inflammatory Bowel Disease Status. *Environ Sci Technol* 56, 414–421. <https://doi.org/10.1021/acs.est.1c03924>.
- Deng, Y., Zhang, Y., Lemos, B., Ren, H., 2017. Tissue accumulation of microplastics in mice and biomarker responses suggest widespread health risks of exposure. *Sci Rep* 7, 46687. <https://doi.org/10.1038/srep46687>.
- Llorca, M., Farré, M., 2021. Current insights into potential effects of micro-nanoplastics on human health by in-vitro tests. *Front Toxicol* 3, 752140. <https://doi.org/10.3389/ftox.2021.752140>.
- Yan, X., Yue, T., Winkler, D.A., Yin, Y., Zhu, H., Jiang, G., Yan, B., 2023. Converting nanotoxicity data to information using artificial intelligence and simulation. *Chem Rev* 123, 8575–8637. <https://doi.org/10.1021/acs.chemrev.3c00070>.
- Shi, W., Han, Y., Sun, S., Tang, Y., Zhou, W., Du, X., Liu, G., 2020. Immunotoxicities of microplastics and sertraline, alone and in combination, to a bivalve species: size-dependent interaction and potential toxication mechanism. *J Hazard Mater* 396, 122603. <https://doi.org/10.1016/j.jhazmat.2020.122603>.
- Lu, L., Huang, W., Han, Y., Tong, D., Sun, S., Yu, Y., Liu, G., Shi, W., 2023. Toxicity of microplastics and triclosan, alone and in combination, to the fertilisation success of a broadcast spawning bivalve *Tegillarca granosa*. *Environ Toxicol Pharm* 101, 104208. <https://doi.org/10.1016/j.etap.2023.104208>.
- Hong, A.R., Kim, J.S., 2024. Biological hazards of micro- and nanoplastic with adsorbents and additives. *Front Public Health* 12, 1458727. <https://doi.org/10.3389/fpubh.2024.1458727>.
- Shukla, S., Khanna, S., Khanna, K., 2025. Unveiling the toxicity of micro-nanoplastics: a systematic exploration of understanding environmental and health implications. *Toxicol Rep* 14, 101844. <https://doi.org/10.1016/j.toxrep.2024.101844>.
- Avio, C.G., Gorbi, S., Milan, M., Benedetti, M., Fattorini, D., d'Errico, Pauletto, G., Bargelloni, M., Regoli, F., 2015. Pollutants bioavailability and toxicological risk from microplastics to marine mussels. *Environ Pollut* 198, 211–222. <https://doi.org/10.1016/j.envpol.2014.12.021>.
- Browne, M.A., Dissanayake, A., Galloway, T.S., Lowe, D.M., Thompson, R.C., 2008. Ingested microscopic plastic translocates to the circulatory system of the Mussel, *Mytilus edulis* (L.). *Environ Sci Technol* 42, 5026–5031. <https://doi.org/10.1021/es800249a>.
- Cole, M., Lindeque, P., Fileman, E., Halsband, C., Galloway, T.S., 2015. The impact of polystyrene microplastics on feeding, function and fecundity in the marine copepod *Calanus helgolandicus*. *Environ Sci Technol* 49, 1130–1137. <https://doi.org/10.1021/es504525u>.
- Kögel, T., Bjørøy, Ø., Toto, B., Bienfait, A.M., Sanden, M., 2020. Micro- and nanoplastic toxicity on aquatic life: Determining factors. *Sci Total Environ* 709, 136050. <https://doi.org/10.1016/j.scitotenv.2019.136050>.
- Yan, X., Sedykh, A., Wang, W., Zhao, X., Yan, B., Zhu, H., 2019. In silico profiling nanoparticles: predictive nanomodeling using universal nanodescriptors and various machine learning approaches. *Nanoscale* 11, 8352–8362. <https://doi.org/10.1039/C9NR00844F>.
- Yan, X., Zhang, J., Russo, D.P., Zhu, H., Yan, B., 2020. Prediction of nano-bio interactions through convolutional neural network analysis of nanostructure images. *ACS Sustain Chem Eng* 8, 19096–19104. <https://doi.org/10.1021/acssuschemeng.0c07453>.
- Dang, F., Wang, Q., Yan, X., Zhang, Y., Yan, J., Zhong, H., Zhou, D., Luo, Y., Zhu, Y.-G., Xing, B., Wang, Y., 2022. Threats to terrestrial plants from emerging nanoplastics. *ACS Nano* 16, 17157–17167. <https://doi.org/10.1021/acsnano.2c07627>.
- Bai, X., Wang, S., Yan, X., Zhou, H., Zhan, J., Liu, S., Sharma, V.K., Jiang, G., Zhu, H., Yan, B., 2020. Regulation of cell uptake and cytotoxicity by nanoparticle core under the controlled shape, size, and surface chemistries. *ACS Nano* 14, 289–302. <https://doi.org/10.1021/acsnano.9b04407>.
- Yan, X., Sedykh, A., Wang, W., Yan, B., Zhu, H., 2020. Construction of a web-based nanomaterial database by big data curation and modeling friendly nanostructure annotations. *Nat Commun* 11, 2519. <https://doi.org/10.1038/s41467-020-16413-3>.
- Ma, J., Wang, S., Zhao, C., Yan, X., Ren, Q., Dong, Z., Qiu, J., Liu, Y., Shan, Qe, Xu, M., Yan, B., Liu, S., 2023. Computer-Aided Discovery of Potent Broad-Spectrum Vaccine Adjuvants. *Angew Chem Int Ed* 62, e202301059. <https://doi.org/10.1002/anie.202301059>.
- Cao, Y., Geddes, T.A., Yang, J.Y.H., Yang, P., 2020. Ensemble deep learning in bioinformatics. *Nat Mach Intell* 2, 500–508. <https://doi.org/10.1038/s42256-020-0217-y>.
- Zounemat-Kermani, M., Batelaan, O., Fadaee, M., Hinkelmann, R., 2021. Ensemble machine learning paradigms in hydrology: A review. *J Hydrol* 598, 126266. <https://doi.org/10.1016/j.jhydrol.2021.126266>.
- Zhu, S., Xu, H., Khan, M.S., Xia, M., Wang, F., Chen, Y., 2025. Enhanced removal of Ni²⁺ and Co²⁺ from wastewater using a novel 2-hydroxyphosphonoacetic acid modified Mg/Fe-LDH composite adsorbent. *Water Res* 272, 122997. <https://doi.org/10.1016/j.watres.2024.122997>.

- [34] Liu, G., Yan, X., Li, C., Hu, S., Yan, J., Yan, B., 2023. Unraveling the joint toxicity of transition-metal dichalcogenides and per- and polyfluoroalkyl substances in aqueous mediums by experimentation, machine learning and molecular dynamics. *J Hazard Mater* 443, 130303. <https://doi.org/10.1016/j.jhazmat.2022.130303>.
- [35] Qiao, Y., Yang, X., Wu, E., 2019. The research of BP Neural Network based on One-Hot Encoding and Principle Component Analysis in determining the therapeutic effect of diabetes mellitus. *IOP Conf Ser Earth Environ Sci* 267, 042178. <https://doi.org/10.1088/1755-1315/267/4/042178>.
- [36] Zhang, L., Tan, J., Han, D., Zhu, H., 2017. From machine learning to deep learning: progress in machine intelligence for rational drug discovery. *Drug Discov Today* 22, 1680–1685. <https://doi.org/10.1016/j.drudis.2017.08.010>.
- [37] Lavecchia, Antonio, 2015. Machine-learning approaches in drug discovery: methods and applications. *Drug Discov Today* 20 (3), 318–331. <https://doi.org/10.1016/j.drudis.2014.10.012>.
- [38] Strumbelj, E., Kononenko, I., 2014. An efficient explanation of individual classifications using game theory. *J Mach Learn Res* 15, 1–18. <https://doi.org/10.1145/1756006.1756007>.
- [39] Zoete, V., Cuendet, M.A., Grosdidier, A., Michielin, O., 2011. SwissParam: A fast force field generation tool for small organic molecules. *J Comput Chem* 32, 2359–2368. <https://doi.org/10.1002/jcc.21816>.
- [40] Vanommeslaeghe, K., Hatcher, E., Acharya, C., Kundu, S., Zhong, S., Shim, J., Darian, E., Guvench, O., Lopes, P., Vorobyov, I., Mackerell, A.D., 2010. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J Comput Chem* 31, 671–690. <https://doi.org/10.1002/jcc.21367>.
- [41] Hollingsworth, S., Dror, R., 2018. Molecular dynamics simulation for all. *Neuron* 99, 129–143. <https://doi.org/10.1016/j.neuron.2018.08.011>.
- [42] Lu, T., Chen, F., 2012. Multiwfn: a multifunctional wavefunction analyzer. *J Comput Chem* 33, 580–592. <https://doi.org/10.1002/jcc.22885>.
- [43] Lu, T., 2024. A comprehensive electron wavefunction analysis toolbox for chemists, Multiwfn. *J Chem Phys* 161, 082503. <https://doi.org/10.1063/5.0216272>.
- [44] Pan, X., Li, L., Huang, H.-H., Wu, J., Zhou, X., Yan, X., Xia, J., Yue, T., Chu, Y.-H., Yan, B., 2022. Biosafety-inspired structural optimization of triazolium ionic liquids based on structure-toxicity relationships. *J Hazard Mater* 424, 127521. <https://doi.org/10.1016/j.jhazmat.2021.127521>.
- [45] Lindahl, E., Abraham, M.J., Hess, B., van der Spoel, D., 2021. GROMACS 2020.5 Source Code. Zenodo. <https://doi.org/10.5281/zenodo.4420785>.
- [46] Humphrey, W., Dalke, A., Schulten, K., 1996. VMD: Visual molecular dynamics. *J Mol Graph* 14, 33–38. [https://doi.org/10.1016/0263-7855\(96\)00018-5](https://doi.org/10.1016/0263-7855(96)00018-5).
- [47] Rodrigues, M.O., Abrantes, N., Gonçalves, F.J.M., Nogueira, H., Marques, J.C., Gonçalves, A.M.M., 2019. Impacts of plastic products used in daily life on the environment and human health: What is known? *Environ Toxicol Pharm* 72, 103239. <https://doi.org/10.1016/j.etap.2019.103239>.
- [48] Abdolapur Monikh, F., Baun, A., Hartmann, N.B., Kortet, R., Akkanen, J., Lee, J.-S., Shi, H., Lahive, E., Uurasjärvi, E., Tufenkji, N., Altmann, K., Wiesner, Y., Grossart, H.P., Peijnenburg, W., Kukkonen, J.V.K., 2023. Exposure protocol for ecotoxicity testing of microplastics and nanoplastics. *Nat Protoc* 18, 3534–3564. <https://doi.org/10.1038/s41596-023-00886-9>.
- [49] He, J., Yang, X., Liu, H., 2021. Enhanced toxicity of triphenyl phosphate to zebrafish in the presence of micro- and nano-plastics. *Sci Total Environ* 756, 143986. <https://doi.org/10.1016/j.scitotenv.2020.143986>.
- [50] Bambino, K., Chu, J., 2017. Zebrafish in Toxicology and Environmental Health. *Curr Top Dev Biol* 124, 331–367. <https://doi.org/10.1016/bbs.ctdb.2016.10.007>.
- [51] Tortella, G.R., Rubilar, O., Durán, N., Díez, M.C., Martínez, M., Parada, J., Seabra, A.B., 2020. Silver nanoparticles: Toxicity in model organisms as an overview of its hazard for human health and the environment. *J Hazard Mater* 390, 121974. <https://doi.org/10.1016/j.jhazmat.2019.121974>.
- [52] Akter, M., Sikder, M.T., Rahman, M.M., Ullah, A.K.M.A., Hossain, K.F.B., Banik, S., Hosokawa, T., Saito, T., Kurasaki, M., 2018. A systematic review on silver nanoparticles-induced cytotoxicity: Physicochemical properties and perspectives. *J Adv Res* 9, 1–16. <https://doi.org/10.1016/j.jare.2017.10.008>.
- [53] Gaillot, S., Rouanet, J.-M., 2015. Silver nanoparticles: their potential toxic effects after oral exposure and underlying mechanisms – a review. *Food Chem Toxicol* 77, 58–63. <https://doi.org/10.1016/j.fct.2014.12.019>.
- [54] Zambrano-Pinto, M.V., Tinizaray-Castillo, R., Riera, M.A., Maddela, N.R., Luque, R., Díaz, J.M.R., 2024. Microplastics as vectors of other contaminants: Analytical determination techniques and remediation methods. *Sci Total Environ* 908, 168244. <https://doi.org/10.1016/j.scitotenv.2023.168244>.
- [55] Ziani, K., Ioniță-Mindrican, C.B., Mititelu, M., Neacșu, S.M., Negrei, C., Moroșan, E., Drăgănescu, D., Preda, O.T., 2023. Microplastics: A Real Global Threat for Environment and Food Safety: A State of the Art Review. *Nutrients* 15, 617. <https://doi.org/10.3390/nu15030617>.
- [56] Wang, X., Khan, M.A., Xia, M., 2019. Synthesis of RGO and g-C₃N₄ hybrid with WO₃/Bi₂WO₆ to boost degradation of nitroguanidine under visible light irradiation. *J Mater Sci Mater Electron* 30, 5503–5515. <https://doi.org/10.1007/s10854-019-00844-w>.
- [57] Bano, Z., Ali, N.Z., Khan, M.A., Mutahir, S., Zhu, S., Wang, F., Xia, M., 2022. Synthesis, characterization and applications of 3D porous graphene hierarchical structure by direct carbonization of maleic acid. *Ceram Int* 48 (6), 8409–8416. <https://doi.org/10.1016/j.ceramint.2021.12.048>.
- [58] Gojznikar, J., Zdravković, B., Vidak, M., Leskošek, B., Ferk, P., 2022. TiO₂ Nanoparticles and Their Effects on Eukaryotic Cells: A Double-Edged Sword. *Int J Mol Sci* 23, 12353. <https://doi.org/10.3390/ijms230212353>.
- [59] Ghulam, A.N., dos Santos, O.A.L., Hazeem, L., Pizzorno Backx, B., Bououdina, M., Bellucci, S., 2022. Graphene Oxide (GO) Materials—Applications and Toxicity on Living Organisms and Environment. *J Funct Biomater* 13, 77. <https://doi.org/10.3390/jfb13020077>.
- [60] Eriksson, L., Jaworska, J., Worth Andrew, P., Cronin Mark, T.D., McDowell Robert, M., Gramatica, P., 2003. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. *Environ Health Perspect* 111, 1361–1375. <https://doi.org/10.1289/ehp.5758>.
- [61] Weiss, K., Khoshgoftaar, T.M., Wang, D., 2016. A survey of transfer learning. *J Big Data* 3, 9. <https://doi.org/10.1186/s40537-016-0043-6>.
- [62] Lin, W., Jiang, R.F., Xiong, Y.X., Wu, J.Y., Xu, J.Q., Zheng, J., Zhu, F., Ouyang, G. F., 2019. Quantification of the combined toxic effect of polychlorinated biphenyls and nano-sized polystyrene on *Daphnia magna*. *J Hazard Mater* 364, 531–536. <https://doi.org/10.1016/j.jhazmat.2018.10.056>.
- [63] Ding, R., Chen, Y., Shi, X., Li, Y., Yu, Y., Sun, Z., Duan, J., 2024. Size-dependent toxicity of polystyrene microplastics on the gastrointestinal tract: Oxidative stress related-DNA damage and potential carcinogenicity. *Sci Total Environ* 912. <https://doi.org/10.1016/j.scitotenv.2023.169514>.
- [64] Shi, X., Wang, X., Huang, R., Tang, C., Hu, C., Ning, P., Wang, F., 2022. Cytotoxicity and Genotoxicity of Polystyrene Micro- and Nanoplastics with Different Size and Surface Modification in A549 Cells. *Int J Nanomed* 17, 4509–4523. <https://doi.org/10.2147/IJN.S381776>.
- [65] Liu, H., Tang, S., Zheng, X., Zhu, Y., Ma, Z., 2015. Bioaccumulation, Biotransformation, and Toxicity of BDE-47, 6-OH-BDE-47, and 6-MeO-BDE-47 in Early Life-Stages of Zebrafish (*Danio rerio*). *Environ Sci Technol* 49 (3), 1823–1833. <https://doi.org/10.1021/es503833q>.
- [66] Donisi, I., Colloca, A., Anastasio, C., Balestrieri, M.L., D'Onofrio, N., 2024. Micro (nano)plastics: an Emerging Burden for Human Health. *Int J Biol Sci* 20 (14), 5779–5792. <https://doi.org/10.7150/ijbs.99556>.
- [67] Xu, T., Cui, J., Xu, R., Cao, J., Guo, M.Y., 2023. Microplastics induced inflammation and apoptosis via ferroptosis and the NF-κB pathway in carp. *Aquat Toxicol* 262, 106659. <https://doi.org/10.1016/j.aquatox.2023.106659>.
- [68] Liu, T.J., Yang, J., Wu, J.W., Sun, X.R., Gao, X.J., 2024. Polyethylene microplastics induced inflammation via the miR-21/TRAK4/NF-κB axis resulting to endoplasmic reticulum stress and apoptosis in muscle of carp. *Fish Shellfish Immunol* 145, 109375. <https://doi.org/10.1016/j.fsi.2024.109375>.

Advanced Mass-Spectra-Based Machine Learning for Predicting the Toxicity of Traditional Chinese Medicines

Chen Jia,[#] Xiaofang Li,[#] Song Hu, Guohong Liu, Jiansong Fang,^{*} Xiaoxia Zhou,^{*} Xiliang Yan,^{*} and Bing Yan



Cite This: *Anal. Chem.* 2025, 97, 783–792



Read Online

ACCESS |



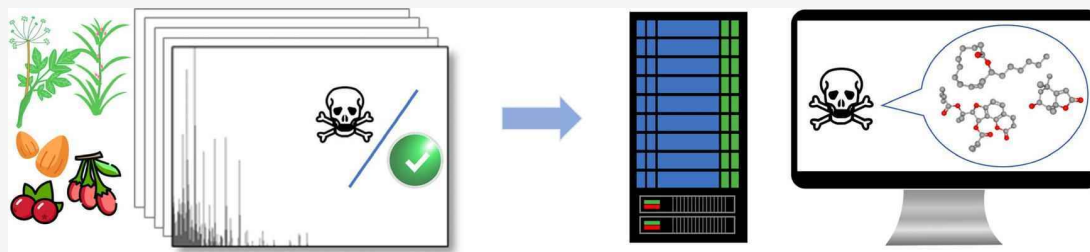
Metrics & More



Article Recommendations



Supporting Information



ABSTRACT: Traditional Chinese medicine (TCM) has been a cornerstone of health care for centuries, valued for its preventive and therapeutic properties. However, recent decades have revealed significant toxicological concerns associated with TCMs due to their complex chemical compositions. Traditional QSAR (quantitative structure–activity relationships) models, which predict toxicity based on chemical structures, face challenges with the intricate nature of TCM compounds. In this study, we effectively resolved this issue by correlating the toxicity of TCMs with advanced analytical descriptors from electron ionization mass spectra (EI-MS) data. The optimal classification model achieved a balanced accuracy of over 0.74. Through interpretable machine learning models, we identified specific toxic components, such as 13-hexyloxacyclotridec-10-en-2-one and loliolide. We applied molecular dynamics (MD) simulations to explore the interactions of identified toxic components with crucial protein targets, using hepatic cytochrome P450 3A4 as an example. This novel approach not only enhances our understanding of the toxicological profiles of TCMs but also maximizes their therapeutic benefits while minimizing adverse effects. More importantly, our findings support the application of analytical descriptor-based machine learning in predicting the toxicity of unknown mixtures in the real environment.

INTRODUCTION

Traditional Chinese medicine (TCM) has been utilized for thousands of years as a comprehensive medical system to prevent and treat diseases.^{1,2} Notably, in 2015, the Nobel Prize in Physiology or Medicine recognized the discovery of Artemisinin, an antimalarial compound derived from *Artemisia annua*, a plant used in TCM.³ However, recent decades have seen growing concerns over the toxicological risks associated with TCM, including the presence of heavy metals and drug-induced liver injuries.^{4,5} These concerns underscore the crucial need for rigorous toxicological assessments to ensure the safe and sustainable development of TCM and related research. Traditional toxicological methods, primarily reliant on labor-intensive and time-consuming experiments, struggle to keep pace with the rapid advancement of modern drug development.^{6,7} Furthermore, these methods often fall short when evaluating compounds from virtual libraries that have yet to be synthesized.⁸

Advancements in computational toxicology, such as Quantitative Structure–Activity Relationship (QSAR) modeling, have shown promise in predicting the toxicity of chemical compounds using machine learning techniques.⁹ Despite the

maturity of QSAR models, they require precise chemical structures for accurate toxicity prediction, and their application to complex mixtures remains challenging.¹⁰ Traditional machine learning approaches, including Random Forest (RF) and Extreme Gradient Boosting (XGBoost), depend on structural descriptors or physicochemical properties as inputs,^{11–13} while deep learning models, like recurrent neural networks and graph neural networks, convert molecular structure information into machine-readable formats such as SMILES or molecular graphs.^{14,15} However, accurately identifying all molecular structures within TCM, which can contain hundreds to thousands of constituents, is nearly impossible.

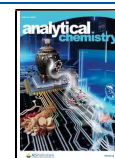
Building on our previous work, which successfully combined Electrospray Ionization Mass Spectrometry (EI-MS) and

Received: September 30, 2024

Revised: December 6, 2024

Accepted: December 10, 2024

Published: December 20, 2024



ACS Publications

© 2024 American Chemical Society

783

218

<https://doi.org/10.1021/acs.analchem.4c05311>
Anal. Chem. 2025, 97, 783–792

machine learning to decipher compound toxicity, we propose a novel approach that bypasses the need for known molecular structures.¹⁶ Initially validated only on chemicals with well-defined structures, our current study explores the applicability of this method to real-world mixtures encountered in daily life. By correlating the toxicity profiles reported for TCMs with their EI-MS spectra, we have developed a machine-learning model demonstrating high predictive accuracy in identifying toxic constituents within TCMs. This study is the first to link EI-MS data of TCMs with their toxicity, showcasing significant potential for predicting the toxicity of mixtures. Our method provides a more precise and realistic evaluation of potential risks associated with combined chemical exposures, offering valuable insights for reducing health risks associated with TCMs.

MATERIALS AND METHODS

Collection of Toxicity Data of TCMs. In the present study, all the TCMs used were purchased from Beijing Tong Ren Tang Chinese Medicine Co., Ltd. The toxicity data for the TCMs were sourced from the Pharmacopoeia of the People's Republic of China/the Chinese Pharmacopoeia (<https://ydz.chp.org.cn/>), the official compendium of drug standards in China, which contains detailed information on various medicinal substances. According to the reports from the Chinese Pharmacopoeia, toxic TCMs were classified into three categories, i.e., highly toxic, toxic, and slightly toxic. In current study, the slightly toxic TCMs were considered toxic because they have been demonstrated to cause adverse health effects in previous studies (Table S1). The toxicity of TCMs was attributed to the specific sites described in the Chinese Pharmacopoeia (Table S2). The toxicity data were integrated with the EI-MS data to establish a comprehensive data set of TCMs. A total of 101 types of TCMs were analyzed using EI-MS. To balance the data set, 40 nontoxic TCMs were randomly selected and paired with the 40 toxic TCMs, forming the final data set used for modeling. The Latin names and toxicity information on the 80 TCMs used for modeling are presented in Table S3.

Spectral Data Acquisition and Preprocessing. According to the methods outlined in the Pharmacopoeia of the People's Republic of China, we prepared the test solutions for the TCM samples. The preparation methods for the test solutions of all 101 TCMs are provided in Table S2. EI-MS data were then acquired using single quadrupole liquid chromatography–mass spectrometry (LC-MS, Agilent 1260), obtaining both positive and negative ion mass spectra for each TCM over a mass-to-charge ratio (m/z) range of 0–1200. Despite being classified as low-resolution mass spectrometry, with a mass accuracy of better than 0.1 amu across the full mass range, it adequately meets our data acquisition needs for rapid screening of samples. For clarity, a “-” will be added before the m/z values in the negative ion spectra in the following text.

Before applying machine learning, the spectral data underwent preprocessing. The EI-MS data consists of mass-to-charge ratios and their corresponding intensities. Initially, interpolation was performed to fill in the missing values within the EI-MS data. We employed zero interpolation to avoid introducing irrelevant noise peaks. Then, we removed the feature columns that were zero for all samples, resulting in a total of 20,429 features used for modeling. Finally, the EI-MS data for each TCM were standardized using zero-mean normalization based

on the mean (\bar{x}) and standard deviation (σ) of the relative intensity (x) within each m/z (eq 1):

$$x' = \frac{x - \bar{x}}{\sigma} \quad (1)$$

Description of Machine Learning Methods. In this study, we employed four classic machine learning methods to build the prediction models: RF, XGBoost, Support Vector Machine (SVM), and k -Nearest Neighbor (k NN).^{17–20} To achieve optimal model performance, the grid search algorithm was used to iterate through the predefined hyperparameters on the training set and store a model for each combination. Then, the model with the best combination of hyperparameters was retained and used for prediction. Grid search is a systematic method for hyperparameter optimization that exhaustively explores predefined combinations of parameter values. Each combination is used to train and evaluate the model, and the configuration with the best performance is selected. Due to the vast parameter space, we restricted the hyperparameter tuning to a reasonable range. The main parameters and their ranges were listed in Table S4. The primary advantage of grid search lies in its comprehensiveness and ability to find a global optimum. It explores all parameter combinations simultaneously and captures the interaction effects between parameters, leading to the identification of the globally optimal configuration. All machine learning models were constructed using Python3.9.19's scikit-learn v1.4.2 package.

Development of QSAR Models. To evaluate the predictive capabilities of classification models, we utilize the F1 score, accuracy, balanced accuracy, and ROC (Receiver Operating Characteristic) curve.^{21,22} The F1 score is the harmonic mean of precision and recall (eq 2), with accuracy measuring the proportion of correctly classified TCMs relative to the total number of TCMs (eq 3). Balanced accuracy (BACC) calculates the average of the true positive rate and the true negative rate for each class (eq 4). These metrics range from 0 to 1, with values closer to 1 indicating stronger predictive performance. The ROC curve is plotted with the false positive rate (FPR) on the x -axis and the true positive rate (TPR) on the y -axis. FPR (eq 5) represents the proportion of false positives among all negative samples, while TPR (eq 6) measures the proportion of true positives among all positive samples. A ROC curve closer to the upper left corner signifies higher model accuracy. All models were evaluated using 5-fold cross-validation.

$$\begin{aligned} \text{F1 score} &= 2 \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \\ &= \frac{\text{TP}}{\text{TP} + 0.5 \times (\text{FP} + \text{TN})} \end{aligned} \quad (2)$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \quad (3)$$

$$\text{Balance accuracy} = \frac{\frac{\text{TP}}{\text{TP} + \text{FN}} + \frac{\text{TN}}{\text{TN} + \text{FP}}}{2} \quad (4)$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \quad (5)$$

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (6)$$

Here, true positive (TP) and true negative (TN) represent cases correctly predicted as positive and negative, respectively. Conversely, false positive (FP) and false negative (FN) denote instances erroneously predicted as positive and negative, respectively.

Identification of Possible Toxic Components in TCMs.

In this study, we used the “feature_importances_” attribute from the RF model and SHAP (SHapley Additive exPlanations) values to identify the key features (i.e., m/z values) influencing the toxicity of TCMs. The “feature_importances_” attribute quantifies the reduction in impurity brought by a feature across all splits where it is used in the RF model.^{17,23} The SHAP analysis generally determines the contribution of each feature to individual predictions, providing a local interpretation of feature effects. Additionally, we performed ablation experiments to further elucidate the identified key features. An ablation study aims to determine the contribution of a feature to the machine learning model by removing the feature. After determining the importance of each m/z value, we first identified toxic TCMs that exhibited signal responses at m/z values with high importance. Next, we searched the Encyclopedia of Traditional Chinese Medicine (ETCM) and relevant literature for potential toxic components within these toxic TCMs.²⁴ Finally, we examined whether these toxic components, either as whole molecules or fragments, matched the high-importance m/z values. A match was considered valid if the error was less than 0.5 m/z .

Molecular Docking. In the present study, the docking software utilized is Autodock Vina.²⁵ A semiflexible docking approach is adopted, wherein the conformation of the receptor macromolecule is fixed while the conformation of the ligand molecule undergoes a certain degree of variation. The geometric center of the receptor molecule was set as the center of the docking box, and the docking was performed using the AutoDock Vina scoring function. The dimensions of the docking box were set to 48.0 Å × 76.0 Å × 70.0 Å (x , y , and z directions). The exhaustiveness of the global search was set to 8. The maximum number of binding modes was set to 9. The maximum energy difference between modes was 3 kcal/mol. Hepatotoxicity of TCMs is frequently reported as a common form of drug toxicity.^{26,27} Therefore, we selected hepatic enzymes as the receptor proteins for molecular docking and MD simulations. The selected protein molecule is CYP3A4, which is one of the most important drug-metabolizing enzymes in humans.²⁸ Consistent CYP3A4 activity inhibition can lead to increased drug toxicity.²⁹ The molecular conformation file was obtained from the Protein Data Bank (PDB), a collaborative structural bioinformatics research laboratory (<https://www.rcsb.org/>).³⁰ The specific CYP3A4 structure chosen for this study is identified by PDB ID 6UNE.³¹

MD Simulation. MD simulation is designed to explain the conformational changes of the protein–ligand complex during the binding process and to provide evidence for the binding mode of small ligand molecules to receptor proteins.^{32,33} The optimal conformation results from molecular docking were used for MD simulations to elucidate the conformational changes of the complex during the binding process. Simulations were conducted using the Gromacs package (version 19.5), with CHARMM force field for the protein fields.^{34,35} The topology files for the small molecules used in the simulations were obtained from CGenFF (<https://cgenff.com/>).³⁶ Each system was placed in a cubic box filled with

TIP3P water molecules, ensuring a minimum distance of 1.5 nm between any solute atom and the edge of the periodic box. To neutralize the total charge of the system, water molecules were replaced with the appropriate number of charged ions, specifically Cl^- and Na^+ , based on the system's charge and electrical properties. After energy minimization, the system was equilibrated in two steps: (1) NVT (constant Number of particles, Volume, and Temperature) equilibration for 0.2 ns and (2) NPT (constant Number of particles, Pressure, and Temperature) equilibration for 0.5 ns. A 100 ns MD simulation was then performed, maintaining a pressure of 1 atm and a temperature of 300 K using the Berendsen thermostat. The simulation used a 2 fs time step, recording atomic coordinates every 10 ps for analysis. Initial velocities were set at 300 K based on a Maxwell distribution, and the water compressibility was set to $4.5 \times 10^{-5} \text{ bar}^{-1}$. After the MD simulations, RMSD (Root Mean Square Deviation), Rg (Radius of gyration), and SASA (Solvent-Accessible Surface Area) of the complex was analyzed using GROMACS (version 19.5). The visualization analysis of both molecular docking and MD simulations was performed using PyMOL.³⁷

Molecular mechanics/Poisson–Boltzmann (or Generalized Born) surface area (MM/PBSA or MM/GBSA) is one of the most commonly used methods for estimating binding free energy. In this study, we employed the gmx_MMPBSA calculation program, a tool that performs end point free energy calculations based on GROMACS molecular dynamics trajectories.^{38,39} The binding free energy (ΔG_{bind}) can be estimated using the following equations (eqs 7–9):

$$\Delta G_{\text{bind}} = \Delta H - T\Delta S \quad (7)$$

$$\Delta H = \Delta E_{\text{MM}} + \Delta G_{\text{polar}} + \Delta G_{\text{nonpolar}} \quad (8)$$

$$\Delta E_{\text{MM}} = \Delta E_{\text{vdw}} + \Delta E_{\text{ele}} \quad (9)$$

where ΔH represents the binding enthalpy, $-T\Delta S$ denotes the conformational entropy of the system after ligand binding, ΔE_{MM} refers to the molecular mechanics interaction energy, including van der Waals interaction energy (ΔE_{vdw}) and electrostatic interaction energy (ΔE_{ele}). ΔG_{polar} represents the polar solvation energy, and $\Delta G_{\text{nonpolar}}$ corresponds to the nonpolar solvation energy. When the entropy term is omitted, the calculated result is referred to as the effective free energy, which is usually sufficient for comparing the relative binding free energies of related ligands.^{40,41}

RESULTS AND DISCUSSION

The objective of this study is to predict the toxicity of TCMs using their EI-MS profiles. The experimental workflow, as illustrated in Figure 1, includes EI-MS measurement, data processing, model building, model interpretation, identification of toxic components, and MD simulations. The TCMs used for EI-MS measurement were commercially purchased, and the toxicity information for these TCMs was sourced from the Pharmacopoeia of the People's Republic of China. The measured data were processed to ensure compatibility with machine learning models. Four machine learning models were selected to find the optimal model and minimize prediction randomness. Once the models were built, feature importance was utilized to identify potential toxic components. Finally, MD simulations were employed to analyze the molecular mechanisms of toxicity.

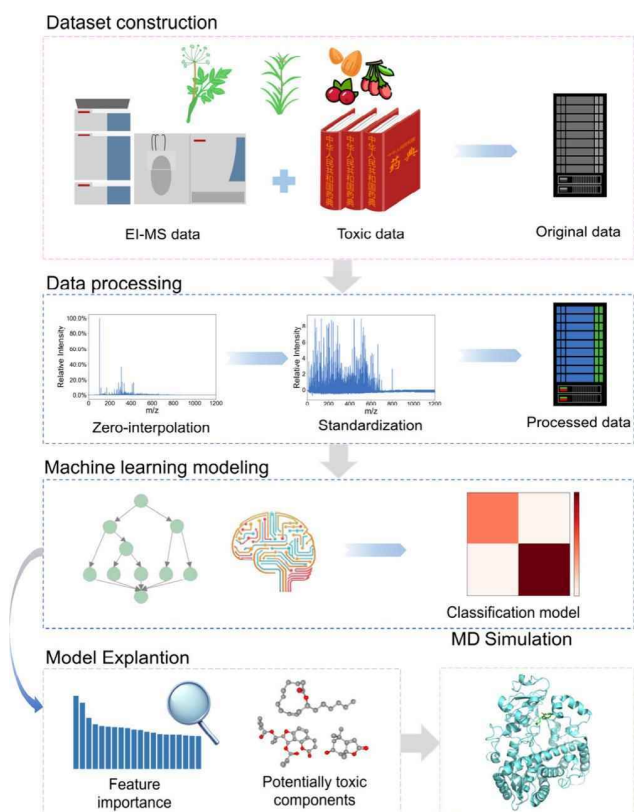


Figure 1. Workflow for predicting the toxicity of TCMs using EI-MS. Specifically, the workflow consists of four main steps: data set construction, data processing, machine learning modeling, and MD simulation. First, EI-MS measurements are taken for various TCMs to construct an EI-MS and toxicity data set. The processed data set is then used to develop machine-learning models. Next, feature importance analysis is conducted to identify potential toxic components. Finally, MD simulations are conducted to explore the interaction mechanisms between the identified toxic components and proteins.

TCMs Exhibit a Wide-Ranging Structural Distribution and a Balanced Toxicity Distribution. Before developing the machine learning models, we analyzed the chemical structure diversity and toxicity distribution within the data set. As shown in Figure 2a and 2b, the m/z values for all the TCMs fall within the range of 0 to 1200; to facilitate differentiation, the m/z values for negative ions are represented as negative numbers. By observing the distribution of m/z values for 40 safe and 40 toxic TCMs (Figure 2a and 2b), the differences in the components and structures between safe and toxic TCMs can be roughly discerned. Specifically, safe TCMs exhibit higher peak values in certain m/z regions within the 0 to 1200 range, while toxic TCMs show higher peaks in certain regions within the 0 to -1200 m/z range. To better visualize the EI-MS differences between the two groups of TCMs, the m/z values were segmented into 50 equal intervals (bins). For each interval, the mean count of toxic and safe TCMs was calculated, and an independent sample t -test was performed to compute p -values. As shown in Figure S1a and S1b, the distribution of toxic and safe TCMs is similar in some m/z intervals, but significant differences ($p < 0.05$) are observed in certain intervals, such as 0–500 and 750–1100 in positive ion mass spectra, and 650–750 and 850–1100 in negative ion mass spectra. These regions with significant differences may

indicate variations in the chemical components of the two groups of TCMs. These peaks in toxic TCMs may correspond to specific chemical components associated with toxicity. This distinct distribution suggests significant chemical diversity within different m/z ranges of TCMs, possibly due to the presence of specific molecules or clusters in higher concentrations. These components may be closely related to the pharmacological or toxicological properties of the TCMs.

For our preprocessed EI-MS data set, we performed Principal Component Analysis (PCA).⁴² The data set was reduced to three principal components, enabling us to visualize the distribution of TCMs in a three-dimensional space. Analyzing this space revealed significant differences in the chemical structures of the compounds, with a wide distribution of structures (Figure 2c). This diversity may be attributed to the complex chemical compositions of TCMs. From the Figure 2c, it can be seen that the green squares representing toxic TCMs and the red squares representing safe TCMs exhibit a certain separation trend in the three-dimensional PCA space. Although there is some overlap between the two types of medicines in certain regions, overall, the distribution of toxic TCMs is relatively dispersed and concentrated in specific areas of the space. This distribution difference suggests that toxic TCMs may contain certain distinctive chemical components, which form unique clusters in the PCA-reduced feature space.

Generally, the balance of the data set significantly impacts the predictive performance of classification models. The balanced data sets tend to exhibit better predictive performance than imbalanced ones in classification models.⁴³ Therefore, we balanced the data set by selecting 40 toxic and 40 nontoxic TCMs for model construction. This approach not only improves the training efficiency of the model but also more accurately reflects the impact of the chemical diversity of TCMs on their toxicity characteristics. The analysis results show significant differences in the m/z value distribution between safe and toxic TCMs, with toxic TCMs exhibiting distinct peaks in specific m/z ranges, suggesting the presence of chemical components related to toxicity. PCA further confirmed that these chemical components form distinct separations and clustering trends in three-dimensional space, reflecting the complexity and diversity of TCM compositions.

Comparison of Different Machine Learning Models in Predicting the Toxicity of TCMs. We used four machine learning methods—RF, XGBoost, k NN, and SVM—to link the EI-MS data of TCMs to their toxicity. To enhance the predictive performance of the models, we employed grid search to fine-tune certain parameters. Given the vast parameter space, we could not optimize the models across the entire parameter space; hence, we selected a range of parameters for optimization. The final parameters used in the study are listed in Table S5.

After optimizing the models using grid search, the prediction results of the four models are presented in Table 1. An accuracy >0.7 was adopted as the reference standard for the performance of individual models.⁴⁴ It is evident from the balanced accuracy that the RF model performed the best (with a balanced accuracy >0.75). The XGBoost and SVM models demonstrated acceptable performance, while the k NN model exhibited relatively poor performance (balanced accuracy <0.6). A higher F1 score further validates the model's classification performance. The RF model achieved not only the highest balanced accuracy but also the highest F1 score. Comparing the ROC curves of the four models (Figure 3a), it

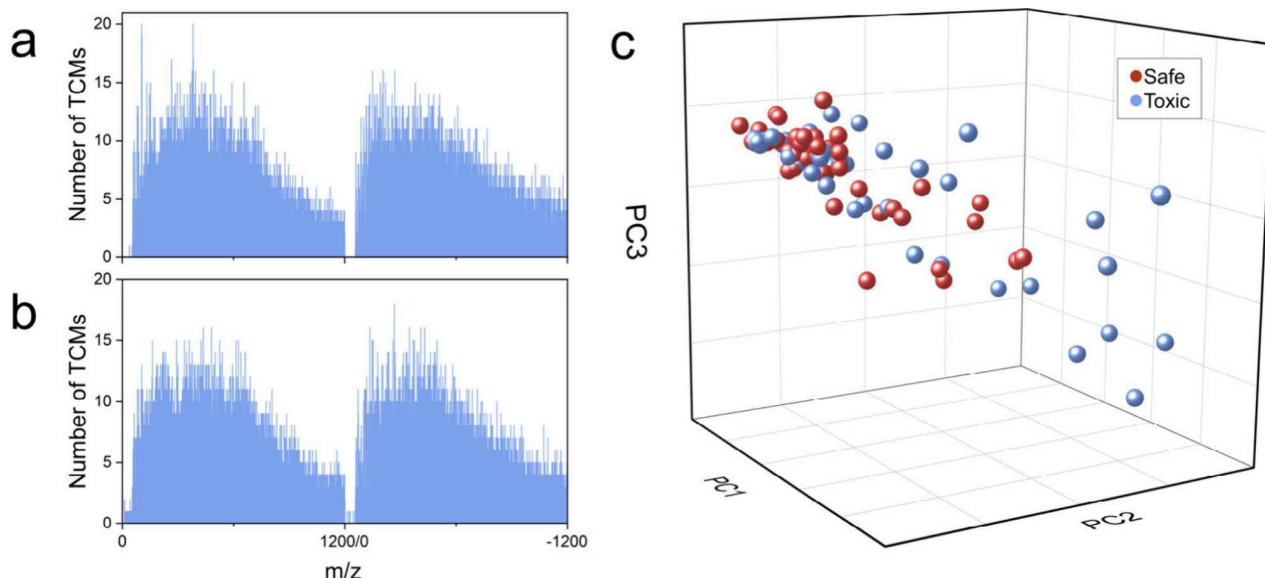


Figure 2. Visualization of chemical structure diversity and toxicity value distribution: (a) distribution of m/z values in safe TCMs; (b) distribution of m/z values in toxic TCMs; (c) the principal component analysis results of molecular features generated from EI-MS in the data set. In (a) and (b), the range from 0 to 1200 m/z represents the positive ion mass spectra, while the range from 0 to -1200 m/z represents the negative ion mass spectra.

Table 1. Performance of Various Machine Learning Models for Toxicity Classification

| Model | RF | XGBoost | kNN | SVM |
|----------|------|---------|------|------|
| BACC | 0.76 | 0.74 | 0.59 | 0.68 |
| ACC | 0.76 | 0.73 | 0.59 | 0.69 |
| F1 Score | 0.75 | 0.72 | 0.71 | 0.67 |

is observed that the ROC accuracy for RF, XGBoost, and SVM all exceed 0.7, while kNN's ROC accuracy is only 0.59. This indicates that RF, XGBoost, and SVM possess reasonable discriminative abilities, whereas kNN's discriminative ability is comparatively weaker. The RF algorithm excels in predictive performance due to its ensemble learning approach, effectively

reducing overfitting, handling complex data relationships, and providing robustness to noise and outliers.

Based on the highest accuracy and F1 score achieved by the RF model, along with its reasonable ROC accuracy, we selected the RF model for subsequent simulation and interpretation tasks. Figure 3b shows the prediction results of the RF model: 29 TCMs were correctly predicted as toxic, 32 TCMs were correctly predicted as nontoxic, 11 toxic TCMs were predicted as nontoxic, and 8 nontoxic TCMs were predicted as toxic. These results demonstrate that the optimized RF model has a good classification capability for TCM toxicity, while kNN exhibits the poorest classification performance. In conclusion, we used four machine learning methods (RF, XGBoost, kNN, and SVM) to link the EI-MS data of TCMs to their toxicity. After optimizing the model

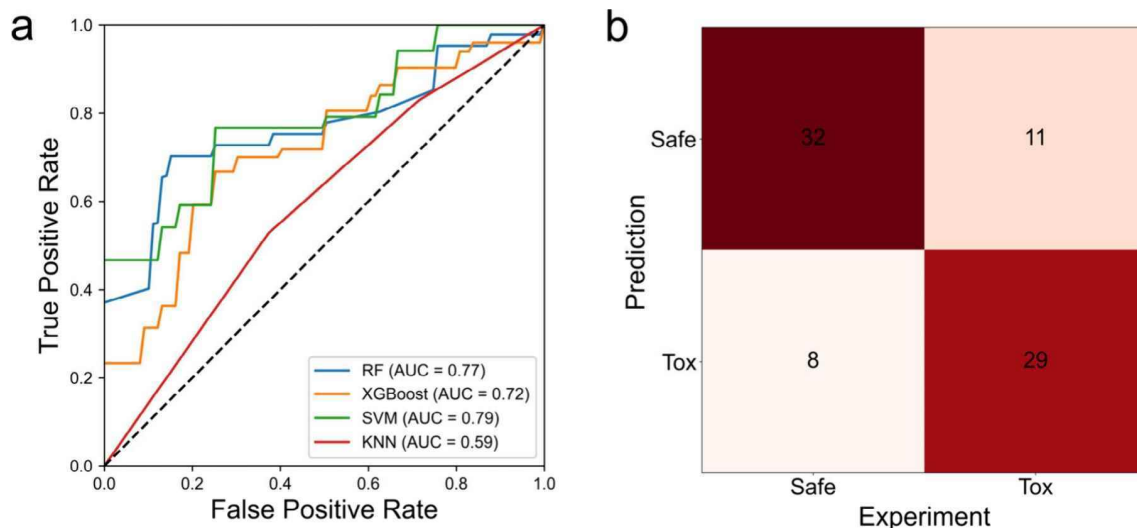


Figure 3. Performance evaluation of different machine learning models: (a) ROC curves of four models; (b) confusion matrix of the RF model for toxicity classification.

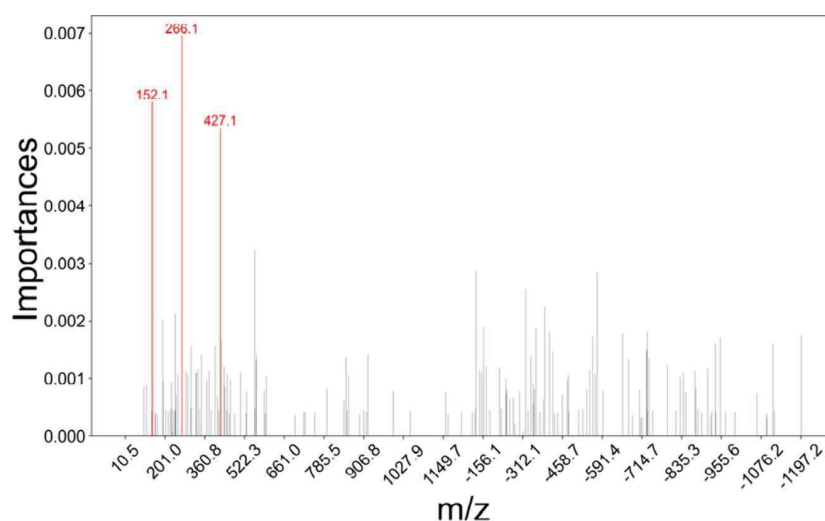


Figure 4. Importance analysis of spectra features. The feature importance was calculated from the optimal model (RF model). The top 3 ranked m/z values are marked in red. The range from 0 to 1200 m/z represents the positive ion mass spectra, while the range from 0 to -1200 m/z represents the negative ion mass spectra.

Table 2. Possible Toxic Components in TCMs^a

| Latin name of TCM | Toxic component | Molecular 2D Structure | Chemical formula | Molar mass (g·mol ⁻¹) |
|------------------------------------|------------------------------------|------------------------|--|-----------------------------------|
| <i>Ricinus communis</i> L. | 13-hexyloxacyclotridec-10-en-2-one | | C ₁₈ H ₃₂ O ₂ | 280.45 |
| <i>Melia azedarach</i> L. | Loliolide | | C ₁₁ H ₁₆ O ₃ | 196.24 |
| <i>Cnidium monnieri</i> (L.) Cuss. | Archangelicin | | C ₂₄ H ₂₆ O ₇ | 426.46 |

^aThe red dashed lines in the 2D structure map represent the potential cleavage sites of molecules.

parameters using grid search, the RF model performed the best, with the highest balanced accuracy and F1 score, and its ROC curve also outperformed the other models. To ensure the robustness and reliability of our constructed machine learning model, we conducted a y -randomization test on the selected RF model. Specifically, we kept the input features unchanged and randomly shuffled the toxicity of TCMs 100 times to build 100 new models. As shown in Figure S2, the predictive performance of the RF model was significantly better than that of all the random models, indicating that the predictive results were not obtained by chance.

Additionally, the best-performing model was externally validated on the 21 TCMs that were excluded for data balance. Detailed information about the 21 TCMs can be found in Table S6. As shown in Figure S3, the external validation showed an accuracy of 0.90, a balanced accuracy of 0.85, and

an F1-score of 0.75. These results prove the generalizability of constructed RF model.⁴⁵

Identification of Possible Chemicals Responsible for Toxicity. According to the OECD (Organization for Economic Co-operation and Development) principles, machine learning models are considered interpretable. Analyzing these models allows us to identify highly influential descriptors, which help elucidate the potential mechanisms underlying the toxicity of TCMs.

In Figure 4, we presented the spectral feature importance analysis results calculated from the optimal model (RF model). This figure reveals the importance of different m/z values in the TCM toxicity classification model. For easier visualization, the top 3 m/z values are marked in red. The figure highlights the top 3 m/z values, the m/z value with the highest importance is 266.1, with an importance score of 0.0070, indicating its critical role in the model. This suggests that m/z

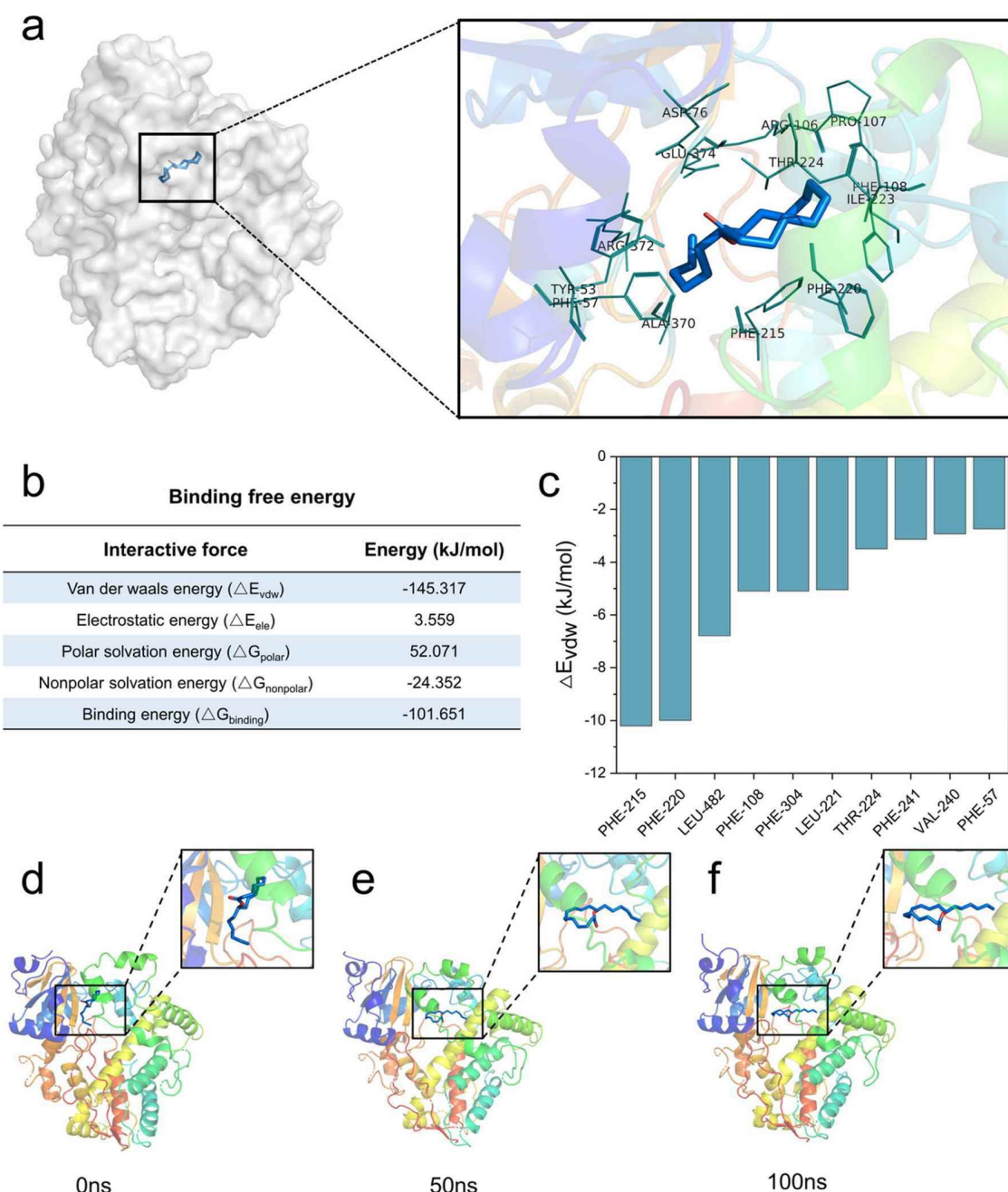


Figure 5. Interaction of 13-hexyloxacyclotridec-10-en-2-one with CYP3A4 through molecular docking and MD simulation: (a) the molecular docking model of 13-hexyloxacyclotridec-10-en-2-one with CYP3A4, showing the binding interactions in detail (magnified panel); (b) the binding free energy between 13-hexyloxacyclotridec-10-en-2-one and CYP3A4; (c) van der Waals interactions between 13-hexyloxacyclotridec-10-en-2-one and specific residues of CYP3A4; (d–f) conformational changes of the 13-hexyloxacyclotridec-10-en-2-one and CYP3A4 complex during MD simulation.

= 266.1 may correspond to a chemical fragment closely related to toxicity or biological activity. Additionally, the *m/z* values 266.1, 152.1, and 427.1 are the only features with an importance score greater than 0.005, indicating their significant contribution to the model's predictions. These values may represent specific toxicity-related compounds or their characteristic fragments. Table S7 displays the top 50 spectral descriptors. The cumulative importance of the top 3 descriptors is 1.81%. Considering that there are a total of 20429 features, the top 3 descriptors are evidently more important than the others. We then analyzed the mean SHAP-

value of the key features. As shown in Figure S4, the features are shown in the order of their global feature importance, the first one being the most important and the latter being less important. Similar with feature importance analysis from RF model, the SHAP values also identified *m/z* values 266.1, 152.1, and 427.1 as the top three most important features. Figure S5 illustrates the change in model performance as features were sequentially removed based on their importance. It is evident that the balanced accuracy of the model significantly decreased (from 0.76 to 0.53) upon the removal

of the top three features, further highlighting the importance of m/z values 266.1, 152.1, and 427.1.

To further explore why these spectral features play a significant role in machine learning models, we linked the toxic molecules in TCMs with their corresponding m/z values. In EI-MS, a single peak can represent several fragment ions. To identify molecules that may influence the toxicity of TCMs, we examined the presence of peaks at $m/z = 266.1$, 152.1, and 427.1 in 40 toxic TCMs. We then analyzed the components of these TCMs from literature or ETCM to match the molecules corresponding to these m/z values. In the present study, we presented only the three molecules with an m/z error range of less than 0.5. The identified molecules are listed in Table 2, and we have marked the potential cleavage sites in the 2D structural diagrams of the molecules. As a bioactive ingredient in *Ricinus communis* L., the compound 13-hexyloxacyclotridec-10-en-2-one could generate a fragment ion at $m/z = 266.1$.⁴⁶ Similarly, *Melia azedarach* L. shows a peak at $m/z = 152.1$, with the molecule Lolilolide, referenced in ETCM, producing a fragment ion at $m/z = 152.1$. *Cnidium monnieri* (L.) Cuss. has a peak at $m/z = 427.1$, where the molecule Archangelicin, also referenced in ETCM, can produce a peak at $m/z = 427.1$.

In summary, by identifying these critical m/z values, we can further focus on the corresponding compounds and explore their specific roles in the toxicity of TCMs. Understanding how these fragments interact with biological targets not only aids in the safety evaluation and risk management of TCMs but also provides deeper insights into the complex mechanisms of TCM toxicity. The importance analysis of spectral features in Figure 4 helps to identify key m/z values as well as holds significant importance for enhancing the safety research of TCMs.

Molecular Docking and Molecular Simulation Validation. To investigate the underlying molecular mechanisms of TCM toxicity, we conducted molecular docking and MD simulation of identified toxic components with the target protein. Here, the hepatic enzyme (CYP3A4) was selected since the hepatotoxicity is one of primary reasons for drugs or herbal compounds being withdrawn from the market. The compound 13-hexyloxacyclotridec-10-en-2-one was used for a case study, and the hepatotoxicity of *Ricinus communis* L. was also widely reported in previous studies.^{47,48} Molecular docking was first carried out between 13-hexyloxacyclotridec-10-en-2-one and the liver enzyme CYP3A4 to predict the optimal binding conformation. We conducted 9 docking attempts and evaluated the binding affinity based on the binding energies of the docking results. Figure 5a shows the docking site and interacting residues, which include Tyr53, Asp76, Arg106, Pro107, Ile223, Thr224, Ala370, phenylalanine residues (Phe57, Phe108, Phe215, Phe220), and arginine residues (Arg106, Arg372). Previous related studies have also identified Phe108, Phe215, and Ala370 as key interacting residues, suggesting that these residues may play a crucial role in the binding specificity of CYP3A4.⁴⁹ In this study, these residues were again confirmed as significant interaction points, reinforcing their importance in the binding affinity and stability of 13-hexyloxacyclotridec-10-en-2-one with CYP3A4. These results demonstrate the existence of a stable interaction between 13-hexyloxacyclotridec-10-en-2-one and CYP3A4 through specific amino acid residues. These findings provide new insights into its potential toxicity mechanism and lay the groundwork for further MD simulation.

The optimal molecular conformation was retained for MD simulation to analyze the dynamic interaction process further. To explore the primary reasons behind the binding of 13-hexyloxacyclotridec-10-en-2-one with CYP3A4, the binding free energy was decomposed into electrostatic potential, van der Waals energy, polar energy, and nonpolar energy. As shown in Figure 5b, the binding free energy (-101.651 kJ/mol) indicates a tight binding between the ligand and the protein. Among the various binding energies, the van der Waals energy (ΔE_{vdw}) was -145.317 kJ/mol, suggesting that van der Waals interactions played the most significant role in the binding mechanism. Figure 5c lists the top ten residues contributing to ΔE_{vdw} . Comparing the molecular docking results (Figure 5a), it can be observed that the phenylalanine cluster residues (Phe57, Phe108, Phe215, Phe220), have high contributions to ΔE_{vdw} , are also the main interacting residues in molecular docking, further validating the accuracy of the molecular docking. Additionally, the RMSF data analysis of the protein residues (Figure S6a) shows that phenylalanine residues (such as Phe215 and Phe220) exhibit significant fluctuations during the simulation, further supporting the above conclusion. Similarly, our simulation results are highly consistent with previous reports, which highlighted the importance of the phenylalanine cluster residues at the active site in CYP3A4-ligand interaction.⁵⁰

Based on our hypothesis, the exposure of 13-hexyloxacyclotridec-10-en-2-one induces conformational changes in CYP3A4, thereby inhibiting its activity. To validate this hypothesis, we analyzed the conformational changes of CYP3A4 during the MD simulation. The RMSD data (Figure S6b) indicated that the system stabilized after 60 ns, suggesting that the simulation had reached equilibrium. Furthermore, the Rg (Figure S6c) and SASA (Figure S6d) results confirmed the structural stability of the complex once equilibrium was achieved. Given this, we selected the CYP3A4 conformations at 0, 50, and 100 ns for analysis. Observing the three conformations (Figure 5d–f), we found that the binding region of CYP3A4 underwent significant conformational changes during the simulation. These findings demonstrate that 13-hexyloxacyclotridec-10-en-2-one can induce three-dimensional structural changes in the binding region of CYP3A4, leading to reduced CYP3A4 activity.

CONCLUSIONS

Although TCM has been used for centuries to treat a wide range of ailments, its herbal compounds' complexity and variability pose challenges in ensuring consistent safety. The toxicity of TCMs may come from specific unidentified chemical components or combinations. Therefore, predicting TCM toxicity using traditional chemical structure-based modeling such as QSAR is almost impossible. In this study, we combined EI-MS data with machine learning models for the first time to enhance evidence-based toxicity evaluation of TCMs. The optimized RF model obtained acceptable prediction performance and effectively identified crucial toxic compounds from m/z values. Subsequent molecular docking and MD simulations revealed that the toxicity may be induced by the van der Waals interactions between the protein targets and toxic components in TCMs. Against the widespread use of TCMs and increasing concern about their safety, our approach provides a powerful tool to transform complex chemical data into actionable knowledge. It supports the development of reliable, evidence-based strategies for toxicity assessment of

complex mixtures, and helps establish more comprehensive usage guidelines.

■ ASSOCIATED CONTENT

Data Availability Statement

The source codes and data can be found at <https://github.com/YanLabAI/TCMEIMSTox>.

SI Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.analchem.4c05311>.

Adverse health effects of slightly toxic TCMs (Table S1), preparation methods of TCM test samples (Table S2), toxicity of TCMs used for modeling (Table S3), main parameters used for modeling (Table S4), final hyperparameters for machine learning (Table S5), toxicity of TCMs used for external validation (Table S6), top 50 spectral features (Table S7), *t*-test of the EI-MS differences between the two groups of TCMs (Figure S1), comparison of RF model and random models (Figure S2), confusion matrix of the RF model on external validation (Figure S3), SHAP analysis plot based on RF model (Figure S4), ablation experiments on key features of RF model (Figure S5), and the result of MD simulation (Figure S6) (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Jiansong Fang — Science and Technology Innovation Center, Guangzhou University of Chinese Medicine, Guangzhou 510405, China; orcid.org/0000-0002-6998-5384; Email: fangjs@gzucm.edu.cn

Xiaoxia Zhou — National-Regional Joint Engineering Research Center for Soil Pollution Control and Remediation in South China, Guangdong Key Laboratory of Integrated Agro-Environmental Pollution Control and Management, Institute of Eco-Environmental and Soil Sciences, Guangdong Academy of Sciences, Guangzhou 510650, China; orcid.org/0000-0001-6508-8750; Email: xiaoxiazhou89@126.com

Xiliang Yan — Institute of Environmental Research at Greater Bay Area, Key Laboratory for Water Quality and Conservation of the Pearl River Delta, Ministry of Education, Guangzhou University, Guangzhou 510006, China; College of Animal Science, South China Agricultural University, Guangzhou 510642, China; orcid.org/0000-0003-4173-6228; Email: yanxiliang1991@scau.edu.cn

Authors

Chen Jia — Institute of Environmental Research at Greater Bay Area, Key Laboratory for Water Quality and Conservation of the Pearl River Delta, Ministry of Education, Guangzhou University, Guangzhou 510006, China

Xiaofang Li — Institute of Environmental Research at Greater Bay Area, Key Laboratory for Water Quality and Conservation of the Pearl River Delta, Ministry of Education, Guangzhou University, Guangzhou 510006, China

Song Hu — Institute of Environmental Research at Greater Bay Area, Key Laboratory for Water Quality and Conservation of the Pearl River Delta, Ministry of Education, Guangzhou University, Guangzhou 510006, China

Guohong Liu — School of Health, Guangzhou Vocational and Technical University of Science and Technology, Guangzhou 510555, China

Bing Yan — Institute of Environmental Research at Greater Bay Area, Key Laboratory for Water Quality and Conservation of the Pearl River Delta, Ministry of Education, Guangzhou University, Guangzhou 510006, China; orcid.org/0000-0002-7970-6764

Complete contact information is available at: <https://pubs.acs.org/doi/10.1021/acs.analchem.4c05311>

Author Contributions

*C.J. and X.L. contributed equally to this work.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the National Key R&D Program of China (Grant 2023YFA0915101), the National Natural Science Foundation of China (Grants 22106025, 22476056, and 22036002), the Introduced Innovative R&D Team Project under the “The Pearl River Talent Recruitment Program” of Guangdong Province (Grant 2019ZT08L387), the Guangdong Basic and Applied Basic Research Foundation (Grants 2022A1515111082 and 2023A1515012148), and the Special Foundation of Guangdong Educational Committee (Grant 2021ZDZX2001), GDAS’ Project of Science and Technology Development (Grants 2023GDASZH-2023010103 and 2023GDASQNR-0105), and Guangdong Foundation for Program of Science and Technology Research (Grant 2023B1212060044).

■ REFERENCES

- (1) Wang, J.; Wong, Y. K.; Liao, F. *Expert Rev. Mol. Med.* **2018**, *20*, 20.
- (2) Wang, Y.; Zhang, Q.; Chen, Y.; Liang, C. L.; Liu, H.; Qiu, F.; Dai, Z. *Biomed. Pharmacother.* **2020**, *121*, 109570.
- (3) Tu, Y. *Angew. Chem., Int. Ed.* **2016**, *55* (35), 10210–10226.
- (4) Lord, G. M.; Tagore, R.; Cook, T.; Gower, P.; Pusey, C. D. *Lancet* **1999**, *354* (9177), 481–482.
- (5) Lee, C. H.; Wang, J. D.; Chen, P. C. *PLoS One* **2011**, *6* (1), No. e16064.
- (6) Yuen, M. F.; Tam, S.; Fung, J.; Wong, D. K.; Wong, B. C.; Lai, C. L. *Aliment. Pharmacol. Ther.* **2006**, *24* (8), 1179–1186.
- (7) Teschke, R.; Zhang, L.; Long, H.; Schwarzenboeck, A.; Schmidt-Taenzer, W.; Genthner, A.; Wolff, A.; Frenzel, C.; Schulze, J.; Eickhoff, A. *Ann. Hepatol* **2015**, *14* (1), 7–19.
- (8) Dearden, J. C. *J. Comput. Aided Mol. Des.* **2003**, *17* (2–4), 119–127.
- (9) Huang, T.; Sun, G.; Zhao, L.; Zhang, N.; Zhong, R.; Peng, Y. *Int. J. Mol. Sci.* **2021**, *22* (16), 8557.
- (10) Muratov, E. N.; Bajorath, J.; Sheridan, R. P.; Tetko, I. V.; Filimonov, D.; Poroikov, V.; Oprea, T. I.; Baskin, I. I.; Varnek, A.; Roitberg, A.; et al. *Chem. Soc. Rev.* **2020**, *49* (11), 3525–3564.
- (11) Fang, Z.; Yu, X.; Zeng, Q. *Toxicology* **2022**, *480*, 153325.
- (12) Yan, J.; Yan, X.; Hu, S.; Zhu, H.; Yan, B. *Environ. Sci. Technol.* **2021**, *55* (21), 14720–14731.
- (13) Kurosaki, K.; Wu, R.; Uesawa, Y. *Int. J. Mol. Sci.* **2020**, *21* (21), 7853.
- (14) Bertinetto, C.; Duce, C.; Micheli, A.; Solaro, R.; Tiné, M. R. *AIP Conf. Proc.* **2012**, *1504* (1), 721–724.
- (15) Ma, H.; An, W.; Wang, Y.; Sun, H.; Huang, R.; Huang, J. *Chem. Res. Toxicol.* **2021**, *34* (2), 495–506.
- (16) Hu, S.; Liu, G.; Zhang, J.; Yan, J.; Zhou, H.; Yan, X. *J. Hazard. Mater.* **2022**, *431*, 128558.
- (17) Breiman, L. *Mach Learn* **2001**, *45* (1), 5–32.
- (18) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International*

- Conference on Knowledge Discovery and Data Mining; ACM SIGKDD, 2016; pp 785–794, DOI: 10.1145/2939672.2939785.
- (19) Awad, M.; Khanna, R. Support Vector Machines for Classification. In *Efficient Learning Machines*; Apress: Berkeley, CA, USA, 2015; pp 39–66, DOI: 10.1007/978-1-4302-5990-9_3.
- (20) Zhang, Z. *Ann. Transl. Med.* **2016**, 4, 218–218.
- (21) Fawcett, T. *Pattern Recognit Lett.* **2006**, 27 (8), 861–874.
- (22) Powers, D. J. *Mach. Learn. Technol.* **2011**, 2, 2229–3981.
- (23) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. *J. Mach. Learn. Res.* **2011**, 12 (85), 2825–2830.
- (24) Xu, H. Y.; Zhang, Y. Q.; Liu, Z. M.; Chen, T.; Lv, C. Y.; Tang, S. H.; Zhang, X. B.; Zhang, W.; Li, Z. Y.; Zhou, R. R.; et al. *Nucleic Acids Res.* **2019**, 47 (D1), D976–d982.
- (25) Trott, O.; Olson, A. J. *J. Comput. Chem.* **2010**, 31 (2), 455–461.
- (26) Amadi, C. N.; Orisakwe, O. E. *Toxics* **2018**, 6 (2), 24.
- (27) Jing, J.; Teschke, R. *J. Clin. Transl. Hepatol.* **2018**, 6 (1), 57–68.
- (28) Wrighton, S. A.; Schuetz, E. G.; Thummel, K. E.; Shen, D. D.; Korzekwa, K. R.; Watkins, P. B. *Drug Metab. Rev.* **2000**, 32 (3–4), 339–361.
- (29) Sevrioukova, I. F.; Poulos, T. L. *Dalton Trans.* **2013**, 42 (9), 3116–3126.
- (30) Berman, H.; Henrick, K.; Nakamura, H. *Nat. Struct. Mol. Biol.* **2003**, 10 (12), 980–980.
- (31) Samuels, E. R.; Sevrioukova, I. F. *Bioorg. Med. Chem.* **2020**, 28 (6), 115349.
- (32) Latour, R. A. *Biointerphases* **2008**, 3 (3), FC2–12.
- (33) Pan, F.; Zhao, L.; Cai, S.; Tang, X.; Mehmood, A.; Alnadari, F.; Tuersuntuoheti, T.; Zhou, N.; Ai, X. *Food Chem.* **2022**, 367, 130677.
- (34) Lindahl, E.; Abraham, M. J.; Hess, B.; van der Spoel, D. *GROMACS 2019.5 Manual*; 2019; DOI: 10.5281/zenodo.3577988.
- (35) Hess, B. *J. Chem. Theory Comput.* **2008**, 4 (1), 116–122.
- (36) Vanommeslaeghe, K.; MacKerell, A. D., Jr. *J. Chem. Inf Model* **2012**, 52 (12), 3144–3154.
- (37) Delano, W. L. *The PyMOL Molecular Graphics System*; 2002.
- (38) Miller, B. R.; McGee, T. D.; Swails, J. M.; Homeyer, N.; Gohlke, H.; Roitberg, A. E. *J. Chem. Theory Comput.* **2012**, 8 (9), 3314–3321.
- (39) Valdés-Tresanco, M. S.; Valdés-Tresanco, M. E.; Valiente, P. A.; Moreno, E. *J. Chem. Theory Comput.* **2021**, 17 (10), 6281–6291.
- (40) Srinivasan, J.; Cheatham, T. E.; Cieplak, P.; Kollman, P. A.; Case, D. A. *J. Am. Chem. Soc.* **1998**, 120 (37), 9401–9409.
- (41) Wang, E.; Sun, H.; Wang, J.; Wang, Z.; Liu, H.; Zhang, J. Z. H.; Hou, T. *Chem. Rev.* **2019**, 119 (16), 9478–9508.
- (42) Lever, J.; Krzywinski, M.; Altman, N. *Nat. Methods* **2017**, 14 (7), 641–642.
- (43) He, H.; Garcia, E. A. *IEEE Trans. Knowl. Data Eng.* **2009**, 21, 1263–1284.
- (44) Eriksson, L.; Jaworska, J.; Worth, A. P.; Cronin, M. T. D.; McDowell, R. M.; Gramatica, P. *Environ. Health Perspect.* **2003**, 111 (10), 1361–1375.
- (45) Clift, A. K.; Dodwell, D.; Lord, S.; Petrou, S.; Brady, M.; Collins, G. S.; Hippisley-Cox, J. *BMJ* **2023**, e073800.
- (46) Sogan, N.; Kapoor, N.; Singh, H.; Kala, S.; Nayak, A.; Nagpal, B. N. *J. Vector Borne Dis.* **2018**, 55 (4), 282–290.
- (47) Worbs, S.; Köhler, K.; Pauly, D.; Avondet, M.-A.; Schaer, M.; Dorner, M. B.; Dorner, B. G. *Toxins* **2011**, 3 (10), 1332–1372.
- (48) Franke, H.; Scholl, R.; Aigner, A. *Naunyn Schmiedeberg Arch. Pharmacol.* **2019**, 392 (10), 1181–1208.
- (49) Tao, Y.; Fan, Y.; Wang, M.; Wang, S.; Cui, J. J.; Lian, D.; Lu, S.; Li, L. *Luminescence* **2023**, 38 (9), 1654–1667.
- (50) Kiani, Y. S.; Ranaghan, K. E.; Jabeen, I.; Mulholland, A. J. *Int. J. Mol. Sci.* **2019**, 20 (18), 4468.

ILTox: A Curated Toxicity Database for Machine Learning and Design of Environmentally Friendly Ionic Liquids

Jiachen Yan, Guohong Liu, Hanle Chen, Song Hu, Xiaohong Wang, Bing Yan,* and Xiliang Yan*



Cite This: *Environ. Sci. Technol. Lett.* 2023, 10, 983–988



Read Online

ACCESS |



Metrics & More



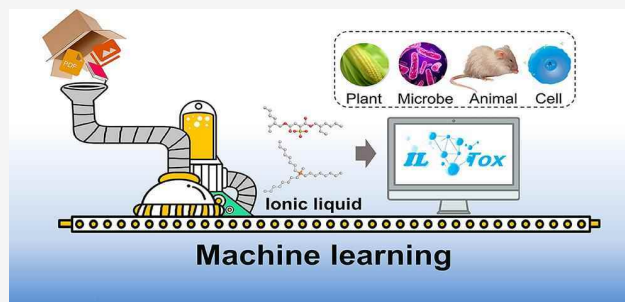
Article Recommendations



Supporting Information

ABSTRACT: A comprehensive online database on the toxicity of ionic liquids (ILs) is urgently needed to facilitate machine learning to design environmentally friendly ILs. In this direction, we present ILTox, a manually curated database of 1183 ILs with over 6700 pieces of toxicity data across different living organisms, including mammalian cells, bacteria, and plants. All the toxicity values and structural information on ILs have been rigorously assessed to ensure data quality. Using this database, various machine learning models have been constructed to quantitatively analyze the relationship between the ILs' structures and their toxicities. Furthermore, the optimized models were used for a virtual screening of desired properties from 8 million ILs. Our results demonstrated that the ILTox database could accelerate the transformation of toxicity data into critical structure–toxicity relationship knowledge. As far as we know, ILTox is the only available database on IL toxicity and is now openly accessible at <http://www.iltox.com>.

KEYWORDS: IL toxicity database, machine learning, toxicity prediction of chemicals, design of biosafety ILs



1. INTRODUCTION

Due to the wide applications and increasing use of ILs,^{1–3} their potential hazards have attracted much attention during the past decades.^{4–8} It has been reported that ILs have different degrees of toxic effects on fish,⁹ plants,¹⁰ microorganisms,¹¹ and cells.¹² However, due to their sheer volume, it is impossible to explore all possible hazards of ILs by simply utilizing experimental methods. As an emerging tool for revealing the structure–activity relationships of chemicals, machine learning holds great promise to predict the toxicity of ILs.^{13,14} The predictive ability of machine learning is mainly driven by enough high-quality data from a structured database such as DrugBank,¹⁵ Toxcast,¹⁶ or ChEMBL.¹⁷ In these databases, chemical structures and their bioactivity information were stored in a standard electronic file format such as a structure data file (SDF). However, most IL toxicity data are stored in the literature in unstructured formats including tables, figures, and even supporting material, which cannot be directly fed into machine learning models. Previous efforts have created several useful IL-related databases (Table S1), but none of them is available for toxicity prediction. For example, IPE Ionic Liquid and ILThermo databases only provide the physicochemical properties of ILs. A previous toxicity-related database, UFT/Merck ILs Biological Effects Database, is unavailable for researchers. Therefore, there is an urgent need to develop a new comprehensive IL toxicity database to facilitate the application of machine learning. Moreover, the rapid develop-

ment of high-throughput screening (HTS) technology has generated large amounts of IL toxicity data that build a foundation for developing such a comprehensive database.^{18–20}

To fill the gap of lacking IL toxicity databases, we present ILTox, a manually curated database specially designed for the toxicity evaluation of ILs. The ILTox database was integrated from comprehensive and meticulous literature data mining. It currently contains 1183 ILs with over 6700 pieces of toxicity data across different living organisms, including mammalian cells, bacteria, and plants. All the toxicity values and structure information on ILs have been rigorously assessed to ensure data quality. Using the ILTox database, a series of machine learning models have been constructed to quantitatively analyze the relationship between the ILs' structures and their toxicities. As the most extensive and only available database for IL toxicity investigations, ILTox is now freely accessible to global users via www.iltox.com. In summary, the ILTox database provides an integrated platform to apply artificial intelligence approaches to significantly accelerate the toxicity

Special Issue: Data Science for Advancing Environmental Science, Engineering, and Technology

Received: February 12, 2023

Revised: March 16, 2023

Accepted: March 17, 2023

Published: March 21, 2023



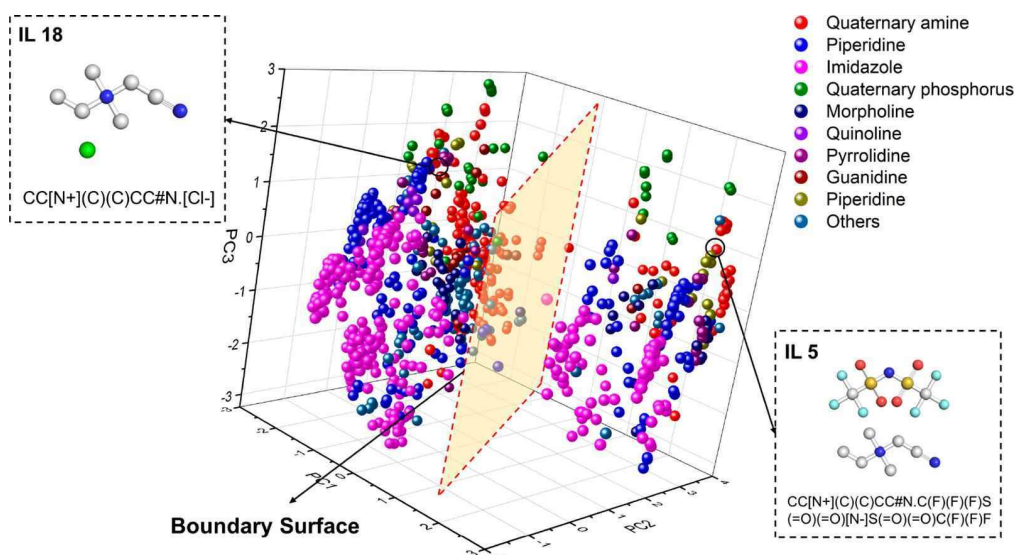


Figure 1. Principal component analysis of 1183 ILs based on MACCS descriptors. ILs containing the same type of cations were labeled the same color. PC1, PC2, and PC3 were the top three principal components and accounted for 42% of the total descriptor variance. Each principal component was a derived variable formed as the linear combination of the MACCS descriptors. A marked boundary surface divides the ILs into two groups. Two representative ILs with the same cations but different anions are also shown.

evaluation of novel ILs and their safety and sustainability by design.

2. MATERIALS AND METHODS

All toxicity data were manually obtained from the published articles in the ISI Web of Knowledge database; the articles before April 30, 2022, were retrieved and collected. Initially, 3369 articles were retrieved from Web of Science using the following search formula: TS = ionic liquids and toxicity. Given the heterogeneity of the toxicity data, each obtained publication was then filtered by the following conditions: (1) The full text was available. (2) The topic was about the toxicity of an IL. (3) At least one toxicity end point was included. (4) The specific structure of an IL was provided. (5) The primary IL characterization data and experimental conditions were provided. Ultimately, 249 papers were identified to establish the toxicity database. For each IL, we mainly focused on the following information: name, CAS (chemical abstract service) register number, molecular formula, molecular weight, SMILES (simplified molecular input line entry system) representation, chemical structure, toxicity value, experimental conditions, and literature information. All the information was integrated into the ILTox database for user convenience. Additionally, to verify the availability of the ILTox database, several toxicity data were selected to construct machine learning models. Detailed information about the data sets and model development can be found in [Methods S1 and S2](#).

3. RESULTS AND DISCUSSION

Currently, there is a total of 1183 ILs with 6726 toxicity data points in the ILTox database. These ILs can be classified into 10 main types based on cationic structures ([Figure S1](#)). Using MACCS descriptors, we performed principal component analysis (PCA) and used the top three principal components, which account for 42% of the total descriptor variance, to show the occupations of all ILs in 3D chemical space ([Figure 1](#)). All the ILs were structurally diverse and occupied most of this chemical space. Typically, similar ILs are classified into the

same cluster. Additionally, there is also an obvious boundary that separates the ILs into two main groups ([Figure 1](#)). The structural difference of ILs in the two groups mainly comes from their anions, such as IL18 and IL5 containing the same cation but different anions. The above results indicated that most ILs can be effectively distinguished by the calculated descriptors, and the toxicities of ILs may be affected by their anions. Typically, a PC was the linear combination of the original descriptors with different coefficients, i.e., loading values. Therefore, we further calculated these loading values to identify descriptors significant in terms of the top three PCs. As shown in [Figure S2](#), several molecular fragments (e.g., carbon–oxygen bond) were identified as key features driving the cluster of ILs in chemical space and thus influencing their toxicities. In addition, it can be concluded that the induced toxicity mechanisms are different in case of ILs in different clusters. SMILES of 1183 ILs are listed in the [Supporting Information Excel file](#).

A common assumption of all QSAR (quantitative structure–activity relationships) and other relevant modeling studies is that similar molecules should exhibit similar bioactivities. To quantitatively study the structural similarity among ILs, we use the MACCS fingerprints to calculate the pairwise Tanimoto coefficients (Tc) of all ILs. A total of 699,153 distances were generated among each two of the 1183 ILs. The distribution of values ranged from 0.033 to 1 with an average of 0.423 ([Figure S3](#)). Two molecules are typically considered similar if their Tc is higher than 0.5. In this database, some ILs are deemed significantly dissimilar. For example, the Tc between IL261 and IL653 is 0.033. IL261 and IL653 with Tc values near zero are considered structurally dissimilar because they have almost completely different cations and anions ([Figure S3](#)). Such structural dissimilarities are sometimes useful for designing novel ILs, e.g., from high cytotoxicity to low cytotoxicity. On the other hand, some ILs are deemed completely similar (Tc = 1), but their structures are actually different such as IL363 and IL364, IL1174 and IL1175 ([Figure S3](#)). IL363 and IL364 are chiral molecules, while IL1174 and IL1175 have carbon chains

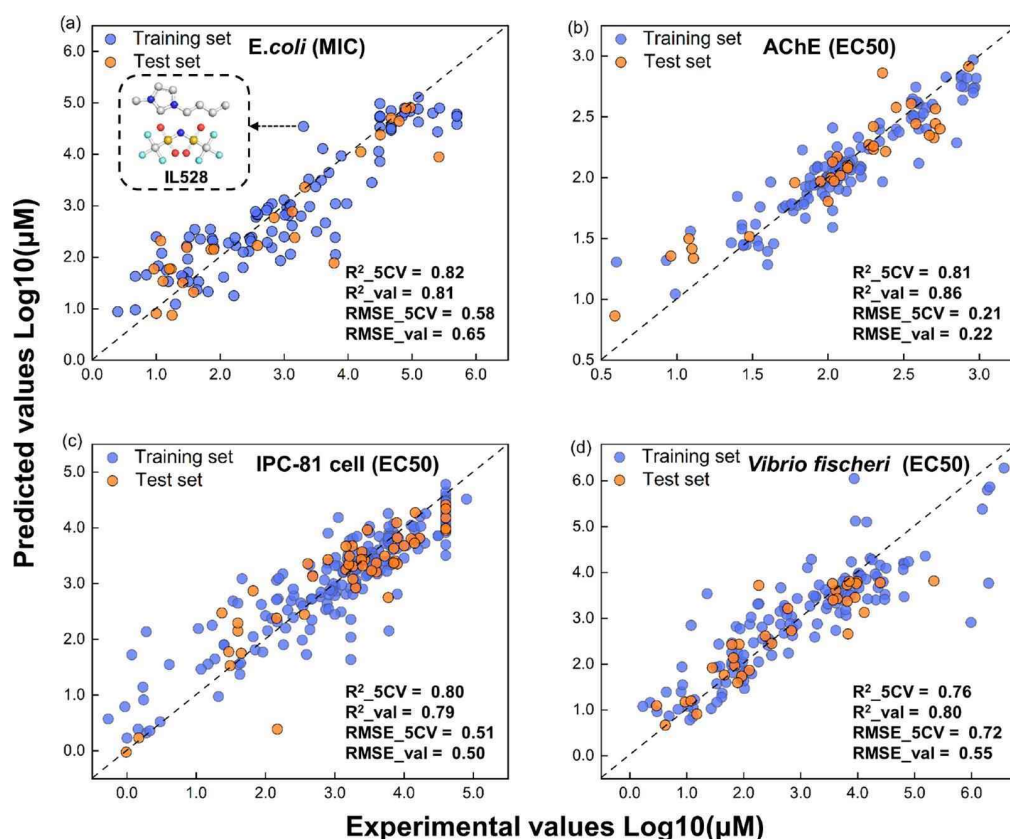


Figure 2. Performances of the consensus machine learning models based on ECFP descriptors. Correlations between experimental results and consensus model predictions for (a) *E. coli*, (b) AChE, (c) Rat IPC-81 cell, and (d) *Vibrio fischeri* toxicity. Blue dots are ILs in the training set, and red dots are those in the test set. The R^2 (coefficient of determination) and RMSE (root-mean-square error) values from consensus modeling results are also shown.

of different lengths. The results indicated that current molecular fingerprints cannot differentiate some special structures or minor differences.

Figure S4 is an overview of the toxicity data curated in this study, involving a variety of IL–bio interfaces such as biomolecules, cells, microorganisms, and plants. Among this, the cells and microorganisms are the most studied species since they are readily available for HTS assays. In addition, aquatic organisms and plants are also widely used for IL toxicity evaluations, which are mainly due to the nature of ILs. Although the low vapor pressure of ILs made them safe to the atmosphere, ILs have posed a persistent threat to environmental waters and soils due to the high stability and excellent miscibility with most media. The toxicity end points were expressed as the concentration of an IL molecule to produce a particular effect on a group of tested species. Figure S5 shows the distribution of the half maximal effective concentration collected in the database. It can be seen that the toxicity values have a very extensive distribution range spanning from -4.21 to 7.35 in units of $\text{Log}_{10}(\mu\text{M})$, indicating the bioactivity diversity of ILs. The structure and bioactivity diversities of ILs enable machine learning algorithms to build predictive models with high generalizability. To share the collected structure and toxicity information, we developed an online database portal (<http://www.ilttox.com>) that currently can be used to search/download the curated data, visualize the IL structures, and upload new data. A full-time computer systems administrator is responsible for maintaining the portal.

Using toxicity data from the database, we constructed machine learning models to identify quantitative relationships between the ILs' structures and their toxicities. The model performances indicated by R^2 and RMSE are shown in Table S2. Overall, R^2 and RMSE for 5-fold cross-validation and external validation were in the same order of magnitude, indicating the robustness and generalizability of the constructed machine learning models. All determination coefficients were above 0.61, meaning that all machine learning models successfully predicted the relationships between the IL structures and related toxicities. By contrast, the consensus models achieved almost the same predictive ability as the optimal performance of all individual models for both cross-validation and external validation. The predictive results also demonstrated that three types of molecular fingerprints could well represent the structure features of ILs. In addition, deep learning models did not exhibit better predictive ability in comparison to traditional machine learning models, which was mainly attributed to the relatively small data sets currently used.

The correlations between experimentations and predicted values of the consensus models are shown in Figure 2. Despite the consensus models' high R^2 and low RMSE, some prediction outliers were also noticeable. For example, the experimental toxicity value of IL528 was $3.30 \mu\text{M}$, while the consensus prediction was $4.54 \mu\text{M}$ (Figure 2). Generally, the predicted values of ILs were inferred from their structural analogs. In the *E. coli* data set, the closest structural analog of

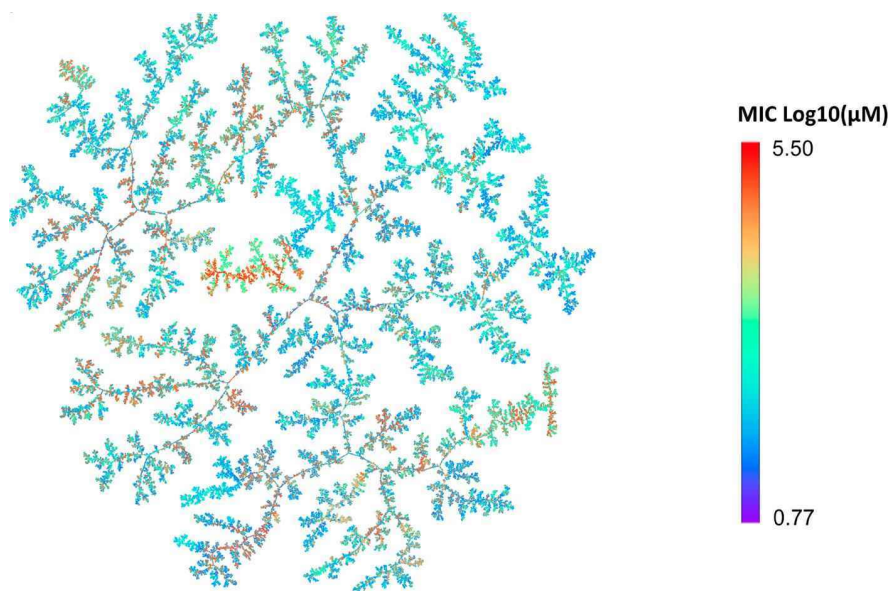


Figure 3. Chemical space of over 8 million virtual ILs. The tree diagram was created by a novel dimensionality reduction technique, tree manifold approximation, and projection (TMAP), which could represent up to millions of data points as a two-dimensional tree.²⁴ In the tree diagram, each data point represents an IL, and the tree structure shows the chemical similarity of all ILs. The data points are colored according to the predicted antimicrobial activity (MIC) of different ILs against *E. coli*, blue (high antimicrobial activity) and red (low antimicrobial activity).

IL528 was considered IL570, but the experimental toxicity values of the two ILs were quite different (3.30 vs 4.80 μM). This issue suggested that more advanced descriptors should be developed to distinguish molecules with similar structures. For user convenience, the constructed machine learning models were also integrated into the ILTox database.

As stated in the OECD principles of QSAR, a machine learning model should be explainable. Herein, we applied the built-in function to estimate feature importance scores of random forest models. These can help us identify several structure features responsible for the corresponding toxicity and thus guide safe IL design. As shown in Figure S6, the top five ranking structural features were identified from the best machine learning models. The high ranking of a descriptor indicated its essential contribution to the final models and its crucial role in the corresponding toxicity. The descriptor, which represents the carbon skeleton of the cation, was ranked the highest in three machine learning models (Figure S6). Previous results also proved that the toxicity of ILs was mainly attributed to their cations. In addition, an increase in the alkyl chain length can enhance membrane and lipidomic perturbations and thus lead to higher toxicity.²¹ As for the AChE data set, the descriptor representing a substructure (piperidine) of the cation core was ranked the highest (Figure S6b). A previous study demonstrated that the toxicity of ILs showed a trend with cation type of morpholine < imidazolium < pyrrolidine < piperidine.²² In addition, the fluorine atom also played an essential role in determining the cytotoxicity of ILs (Figure S6c). In an aqueous solution, some anions (e.g., [BF₄]⁻ and [PF₆]⁻) can release fluorinated anions that potently inhibit Na⁺-K⁺-ATPase located at the cell surface and thus negatively influence essential processes within the cell.²³ In summary, the toxicity of ILs was mainly determined by the carbon chain length and core type of cations. In addition, some organic anions also had a great impact on the toxicity of ILs. The specific examples can be seen in Figure S7.

Our ultimate goal is to prioritize ILs of concern and design environmentally friendly ILs. Thus, we applied the constructed machine learning models to predict the potential hazards of 8 million virtually designed ILs. These ILs consisted of 219,216 cations and 38 anions, which were collected from a previous study.²⁵ As shown in Figure 3, these ILs covered a broad chemical space, indicating the structural diversity of designed ILs. Structural diversity is essential for HTS that identifies molecules with desired properties, as compounds with similar structural features are likely to exhibit similar biological activities. Thus, the predicted toxicity values also show a broad range in the present study (Figure 3 and Figure S8a–c). The structure and biological activity diversity were well illustrated in the branch of the tree diagram. In these areas, data points in blue show molecules with high toxicity (low toxicity values), while data points in red indicate molecules with low toxicity. Details about the tree diagram can be seen in Method S3. Although we expect to obtain low toxicity ILs in most cases, the high toxicity ILs are sometimes desired. For instance, ILs with higher AChE toxicity can be candidates used to treat nervous system disorders. The ILs that severely inhibit bacterial activity can also be used as antibacterial drugs. Herein, we further applied molecular docking to validate our prediction results and elucidate the corresponding toxicity mechanism. Figure S9a–d shows the molecular interactions between AChE and four representative ILs (i.e., two high toxicity and two low toxicity ILs). Compared with ILs with low toxicity, the predicted high toxicity ILs exhibited higher binding affinity with AChE. Further, the binding mode analysis revealed that the inhibition of enzyme activity was associated with a series of noncovalent interactions, such as pi–pi interactions and hydrogen bonds. Together, these results indicated that the toxicity database could facilitate the application of computational methods (e.g., machine learning and molecular simulation) in transforming big data into critical knowledge. Detailed information about molecular docking can be seen in Method S4.

4. ENVIRONMENTAL IMPLICATIONS

A large amount of evidence has pointed to a broad range of deleterious effects of ILs on various organisms. In the recent past, machine learning has been expected to facilitate the risk assessment of ILs and generalize reliable, evidence-based policy actions. However, the application of machine learning is greatly hindered by the lack for more high-quality data. We contribute to filling this knowledge gap by developing a comprehensive IL toxicity database. More than 6700 pieces of IL toxicity data were stored in the database, and all IL structures are in a uniform format, which can be directly fed into machine learning models for unraveling the quantitative relationships between IL structures and toxicities. These relationships play a critical role in driving the progress of risk assessment of ILs and establishing effective policymaking decisions in the future. Furthermore, establishing QSAR models enables property optimization of ILs from a multidimensional parameter space to enhance beneficial properties while filtering out detrimental properties. In the era of big data, combining toxicity databases and machine learning can facilitate the transformation of massive data into critical knowledge.

■ ASSOCIATED CONTENT

Data Availability Statement

The source codes and data can be found at <https://github.com/YanLabAI/ILTox>.

SI Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.estlett.3c00106>.

Additional experimental details, Methods S1–S4, Figures S1–S10, and Tables S1–S3 (PDF)

SMILES of 1183 ILS and model parameters (XLSX)

■ AUTHOR INFORMATION

Corresponding Authors

Xiliang Yan – Institute of Environmental Research at Greater Bay Area, Key Laboratory for Water Quality and Conservation of the Pearl River Delta, Ministry of Education, Guangzhou University, Guangzhou 510006, China;

orcid.org/0000-0003-4173-6228;

Email: yanxiliang1991@gzhu.edu.cn

Bing Yan – Institute of Environmental Research at Greater Bay Area, Key Laboratory for Water Quality and Conservation of the Pearl River Delta, Ministry of Education, Guangzhou University, Guangzhou 510006, China;

orcid.org/0000-0002-7970-6764; Email: drbingyan@gzhu.edu.cn

Authors

Jiachen Yan – Institute of Environmental Research at Greater Bay Area, Key Laboratory for Water Quality and Conservation of the Pearl River Delta, Ministry of Education, Guangzhou University, Guangzhou 510006, China

Guohong Liu – Institute of Environmental Research at Greater Bay Area, Key Laboratory for Water Quality and Conservation of the Pearl River Delta, Ministry of Education, Guangzhou University, Guangzhou 510006, China

Hanle Chen – Institute of Environmental Research at Greater Bay Area, Key Laboratory for Water Quality and Conservation of the Pearl River Delta, Ministry of Education, Guangzhou University, Guangzhou 510006, China

Song Hu – School of Environmental Science and Engineering, Shandong University, Qingdao 266237, China

Xiaohong Wang – Institute of Environmental Research at Greater Bay Area, Key Laboratory for Water Quality and Conservation of the Pearl River Delta, Ministry of Education, Guangzhou University, Guangzhou 510006, China

Complete contact information is available at:

<https://pubs.acs.org/doi/10.1021/acs.estlett.3c00106>

Notes

The authors declare no competing financial interest.

J. Yan and G. Liu contributed equally to this work.

■ ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (22106025, 22036002), the Introduced Innovative R&D Team Project under the “The Pearl River Talent Recruitment Program” of Guangdong Province (2019ZT08L387), and the Basic and Applied Basic Research Foundation of Guangzhou, China (202201010541).

■ REFERENCES

- (1) Lei, Z.; Chen, B.; Koo, Y. M.; Macfarlane, D. R. Introduction: Ionic Liquids. *Chem. Rev.* **2017**, *117*, 6633–6635.
- (2) Gomes, J. M.; Silva, S. S.; Reis, R. L. Biocompatible Ionic Liquids: Fundamental Behaviours and Applications. *Chem. Soc. Rev.* **2019**, *48*, 4317–4335.
- (3) Egorova, K. S.; Gordeev, E. G.; Ananikov, V. P. Biological Activity of Ionic Liquids and Their Application in Pharmaceuticals and Medicine. *Chem. Rev.* **2017**, *117*, 7132–7189.
- (4) Wei, P.; Pan, X.; Chen, C. Y.; Li, H. Y.; Yan, X.; Li, C.; Chu, Y. H.; Yan, B. Emerging Impacts of Ionic Liquids on Eco-Environmental Safety and Human Health. *Chem. Soc. Rev.* **2021**, *50*, 13609–13627.
- (5) Cho, C. W.; Pham, T. P. T.; Zhao, Y.; Stolte, S.; Yun, Y. S. Review of the Toxic Effects of Ionic Liquids. *Sci. Total Environ.* **2021**, *786*, 147309.
- (6) Petkovic, M.; Seddon, K. R.; Rebelo, L. P. N.; Silva Pereira, C. Ionic Liquids: A Pathway to Environmental Acceptability. *Chem. Soc. Rev.* **2011**, *40*, 1383–1403.
- (7) Thuy Pham, T. P.; Cho, C.-W.; Yun, Y.-S. Environmental Fate and Toxicity of Ionic Liquids: A Review. *Water Res.* **2010**, *44*, 352–372.
- (8) Zhao, D.; Liao, Y.; Zhang, Z. D. Toxicity of Ionic Liquids. *Clean - Soil, Air, Water* **2007**, *35*, 42–48.
- (9) Ruokonen, S. K.; Sanwald, C.; Sundvik, M.; Polnick, S.; Vyavaharkar, K.; Duša, F.; Holding, A. J.; King, A. W. T.; Kilpeläinen, I.; Lämmerhofer, M.; Panula, P.; Wiedmer, S. K. Effect of Ionic Liquids on Zebrafish (*Danio Rerio*) Viability, Behavior, and Histology; Correlation between Toxicity and Ionic Liquid Aggregation. *Environ. Sci. Technol.* **2016**, *50*, 7116–7125.
- (10) Pawłowska, B.; Telesiński, A.; Biczak, R. Phytotoxicity of Ionic Liquids. *Chemosphere* **2019**, *237*, 124436.
- (11) Vieira, N. S. M.; Stolte, S.; Araújo, J. M. M.; Rebelo, L. P. N.; Pereira, A. B.; Markiewicz, M. Acute Aquatic Toxicity and Biodegradability of Fluorinated Ionic Liquids. *ACS Sustain. Chem. Eng.* **2019**, *7*, 3733–3741.
- (12) Dzhemileva, L. U.; D'Yakov, V. A.; Seitkalieva, M. M.; Kulikovskaya, N. S.; Egorova, K. S.; Ananikov, V. P. A Large-Scale Study of Ionic Liquids Employed in Chemistry and Energy Research to Reveal Cytotoxicity Mechanisms and to Develop a Safe Design Guide. *Green Chem.* **2021**, *23*, 6414–6430.
- (13) Cao, L.; Zhu, P.; Zhao, Y.; Zhao, J. Using Machine Learning and Quantum Chemistry Descriptors to Predict the Toxicity of Ionic Liquids. *J. Hazard. Mater.* **2018**, *352*, 17–26.
- (14) Yan, J.; Yan, X.; Hu, S.; Zhu, H.; Yan, B. Comprehensive Interrogation on Acetylcholinesterase Inhibition by Ionic Liquids

Using Machine Learning and Molecular Modeling. *Environ. Sci. Technol.* **2021**, *55*, 14720–14731.

(15) Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; Assempour, N.; Iynkkaran, I.; Liu, Y.; MacLejewski, A.; Gale, N.; Wilson, A.; Chin, L.; Cummings, R.; Le, D.; Pon, A.; Knox, C.; Wilson, M. DrugBank 5.0: A Major Update to the DrugBank Database for 2018. *Nucleic Acids Res.* **2018**, *46*, D1074–D1082.

(16) Kavlock, R.; Chandler, K.; Houck, K.; Hunter, S.; Judson, R.; Kleinstreuer, N.; Knudsen, T.; Martin, M.; Padilla, S.; Reif, D.; Richard, A.; Rotroff, D.; Sipes, N.; Dix, D. Update on EPA's ToxCast Program: Providing High Throughput Decision Support Tools for Chemical Risk Management. *Chem. Res. Toxicol.* **2012**, *25*, 1287–1302.

(17) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012**, *40*, D1100–D1107.

(18) Rebros, M.; Gunaratne, H. Q. N.; Ferguson, J.; Seddon, K. R.; Stephens, G. A High Throughput Screen to Test the Biocompatibility of Water-Miscible Ionic Liquids. *Green Chem.* **2009**, *11*, 402–440.

(19) Pinto, P. C. A. G.; Costa, S. P. F.; Lima, J. L. F. C.; Saraiva, M. L. M. F. S. Automated High-Throughput *Vibrio Fischeri* Assay for (Eco)Toxicity Screening: Application to Ionic Liquids. *Ecotoxicol. Environ. Saf.* **2012**, *80*, 97–102.

(20) Tether, A. L.; Laverty, G.; Puga, A. V.; Seddon, K. R.; Gilmore, B. F.; Kelly, S. A. High-Throughput Toxicity Screening of Novel Azepanium and 3-Methylpiperidinium Ionic Liquids. *RSC Adv.* **2020**, *10*, 22864–22870.

(21) Pan, X.; Li, L.; Huang, H. H.; Wu, J.; Zhou, X.; Yan, X.; Jia, J.; Yue, T.; Chu, Y. H.; Yan, B. Biosafety-Inspired Structural Optimization of Triazolium Ionic Liquids Based on Structure-Toxicity Relationships. *J. Hazard. Mater.* **2022**, *424*, 127521.

(22) Basant, N.; Gupta, S.; Singh, K. P. Predicting Acetyl Cholinesterase Enzyme Inhibition Potential of Ionic Liquids Using Machine Learning Approaches: An Aid to Green Chemicals Designing. *J. Mol. Liq.* **2015**, *209*, 404–412.

(23) Weyhing-Zerrer, N.; Gundolf, T.; Kalb, R.; Oßmer, R.; Rossmann, P.; Mester, P. Predictability of Ionic Liquid Toxicity from a SAR Study on Different Systematic Levels of Pathogenic Bacteria. *Ecotoxicol. Environ. Saf.* **2017**, *139*, 394–403.

(24) Probst, D.; Reymond, J. L. Visualization of Very Large High-Dimensional Data Sets as Minimum Spanning Trees. *J. Cheminform.* **2020**, *12*, 1–13.

(25) Venkatraman, V.; Evjen, S.; Chellappan Lethesh, K. The Ionic Liquid Property Explorer: An Extensive Library of Task-Specific Solvents. *Data* **2019**, *4*, 88.

De novo Design of Biocompatible Nanomaterials Using Quasi-SMILES and Recurrent Neural Networks

Ying He,[#] Fang Liu,[#] Weicui Min, Guohong Liu, Yinbao Wu, Yan Wang, Xiliang Yan,^{*} and Bing YanCite This: *ACS Appl. Mater. Interfaces* 2024, 16, 66367–66376

Read Online

ACCESS |



Metrics & More



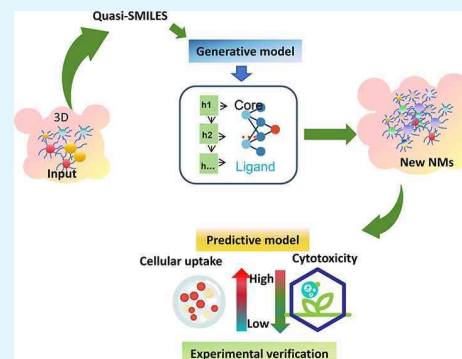
Article Recommendations



Supporting Information

ABSTRACT: Screening nanomaterials (NMs) with desired properties from the extensive chemical space presents significant challenges. The potential toxicity of NMs further limits their applications in biological systems. Traditional methods struggle with these complexities, but generative models offer a possible solution to producing new molecules without prior knowledge. However, converting complex 3D nanostructures into computer-readable formats remains a critical prerequisite. To overcome these challenges, we proposed an innovative deep-learning framework for the *de novo* design of biocompatible NMs. This framework comprises two predictive models and a generative model, utilizing a Quasi-SMILES representation to encode three-dimensional structural information on NMs. Our generative model successfully created 289 new NMs not previously seen in the training set. The predictive models identified a particularly promising NM characterized by high cellular uptake and low toxicity. This NM was successfully synthesized, and its predicted properties were experimentally validated. Our approach advances the application of artificial intelligence in NM design and provides a practical solution for balancing functionality and toxicity in NMs.

KEYWORDS: nano-QSAR, nanotoxicity, nanobio interactions, nanodescriptors, nanocombinatorial chemistry



1. INTRODUCTION

Nanomaterials (NMs) have profoundly impacted virtually every sector of human society, driving significant innovations due to their unique physical and chemical properties derived from the nanoscale effect.^{1–4} The ability to vary elements, sizes, shapes, and surface modifications dramatically enhances the potential for the design of new NMs. However, identifying desirable NMs within the vast chemical space remains a significant challenge. Traditional methods, such as trial-and-error and virtual screening, rely heavily on existing chemical knowledge and the expertise of chemists, evaluating molecular properties sequentially.^{5,6} Recently, generative models have emerged as solutions for automatically generating new molecules.^{7,8} These models, often based on neural networks, can generate novel molecular structures by learning patterns from large data sets of existing molecules. By encoding the structural and chemical features of molecules, these models help in exploring the vast chemical space to identify promising candidates for drug discovery, material science, or other applications.⁹ Common types of generative models used in molecular design include variational autoencoders (VAEs),¹⁰ recurrent neural networks (RNNs),¹¹ generative adversarial networks (GANs),¹² and reinforcement learning-based methods.¹³ Typically, the generative models require a simplified machine-readable format of molecules as inputs such as SMILES.

Compared with small organic molecules, NMs exhibit a higher level of structure complexity due to their diversity in particle size, shape, and surface chemistry. Therefore, it is extremely difficult to encode the complete structural information on NMs with canonical SMILES mainly designed for linear representations of small molecules. Recently, an advanced representation method, Quasi-SMILES, was proposed to capture more complex geometric and topological features.^{14,15} For example, a 20 nm Al₂O₃ nanoparticle (NP) with a surface charge of 40 mV can be represented as [aAl₂O₃][b20][c40]. More importantly, Quasi-SMILES can be used as the inputs of generative models in a very compact data format. However, Quasi-SMILES has not yet been used in generative models for NM design, especially for surface-modified functional NMs.

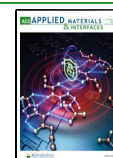
Despite their promising applications in medicine, energy,¹⁶ and environmental sectors, the potential toxicity of NMs poses significant market entry barriers.¹⁷ The nanoscale and unique structures of NMs lead to complex and unpredictable interactions with biological systems. For instance, NPs such

Received: September 12, 2024

Revised: October 31, 2024

Accepted: November 13, 2024

Published: November 20, 2024



as TiO_2 , CuO , and Ag_2O can induce oxidative damage to cell membranes at certain concentrations, potentially leading to apoptosis, inflammation, and adverse effects on microbial communities.^{18,19} The exposure of aquatic organisms to silver, titanium, and zinc oxide NPs has been shown to negatively affect their growth and survival.^{20–22} Additionally, studies indicate that 30–99% of NPs, once inside the body, accumulate primarily in the liver.²³ Traditional toxicology experiments, which are time-intensive and resource-heavy, are further complicated by ethical and legal constraints on animal testing.²⁴ In this context, artificial intelligence (AI) offers a promising alternative, enabling the rapid transformation of nanotoxicity data into actionable insights and facilitating the design of NMs with reduced toxicity through analyzing extensive data sets and identifying key nanostructure–activity relationships.²⁵

In our work, we focus on the design of low-toxicity NMs with high cellular uptake. We introduce an RNN-based generative model framework that effectively explores the NM design space. Our generative model trains on diverse NMs (i.e., Au, Ag, Pt, and Pd), learns the intricate relationships encoded in Quasi-SMILES representations, and generates hypothetical yet viable nanostructures. Subsequently, two predictive models are employed to screen the newly generated chemical space for optimal NMs. Through this approach, we identified and synthesized a novel NM demonstrating high cellular uptake and low cytotoxicity. The effectiveness of the selected NM was confirmed through *in vitro* experiments on human lung adenocarcinoma cells (A549). Our method is more efficient and practical than traditional exhaustive enumeration, offering a faster and more reliable means of designing the desired NMs from the sampled chemical component space. The results provide a new paradigm for designing novel nanocarriers in cancer therapy and can also be extended to the design of biocompatible NMs in medicine, agriculture, energy, and various other fields.

2. EXPERIMENTAL SECTION

2.1. Data Collection. Our proposed *de novo* design framework has two separate components: the generative model and the predictive models, both trained with data from the PubVINAS database.²⁶ PubVINAS is a modeling-friendly NM database based on nanostructure annotation. Currently, PubVINAS contains more than 1300 nanostructures involving over 15 material types such as metal and metal oxide NPs, nanoplastics, carbon nanotubes, and two-dimensional NMs. All nanostructures were generated from our VINAS toolbox by entering the structure parameters of the NMs. Additionally, the nanobio interactions covered multiple biological systems such as biomacromolecules, cells, aquatic organisms, plants, and mammals. The generative model was trained with Quasi-SMILES of 455 structurally diverse NMs covering four core materials (Au, Ag, Pt, and Pd) and 244 small molecular ligands. Two predictive models were used to evaluate the cellular uptake activity and cytotoxicity of the NMs. The cellular uptake data set consists of experimental test values for 71 gold NPs coated with 45 small molecules. The cellular uptake values range from 0.74 to 22.81 (10^{-11} g Au cell⁻¹). The cytotoxicity data set consists of EC_{70} values for A549 cells, covering 24 NMs, including two core materials (Au and Pd) and 6 small molecular ligands. The cytotoxicity values ranged from 0.8 to 351.5 nmol/L.

2.2. Generative Model and Predictive Models.

Generative model: The first step in the generative model involves transforming nanostructures into a machine-readable format. Inspired by SMILES for small molecules, Quasi-SMILES was used to encode the structural information on NMs, including core types, shapes, particle sizes, ligand structures, and numbers.²⁷ The stack-RNN was used to construct the generative models, which can learn the syntax of nanostructure representation in terms of Quasi-SMILES.²⁸ The stack-RNN model added 512 units in the stacked expansion layer, and the learning rate was set at 0.001. Additionally, a comprehensive synthetic accessibility score (SA score) was applied to evaluate the synthesizability of generated surface ligands of NMs. Ligands with high SA score values (typically above 6) are considered challenging to synthesize, while those with low SA score values are easier to synthesize.²⁹

Predictive models: Four traditional machine learning methods, i.e., random forest (RF), support vector machine (SVM), *k*-nearest neighbor regression (*k*NN), and extreme gradient boosting (XGBoost), were used to construct the predictive models. The cellular uptake activity and cytotoxicity of NMs were regarded as output targets, and tetrahedral descriptors calculated from virtual nanostructures were used for input variables.³⁰ All virtual nanostructures were generated from our VINAS toolbox.³¹ Each data set was divided into a training set (80% of the whole set) and a test set (20% of the entire set). Hyperparameter optimization was performed using a grid search algorithm, and the model with the best combination of hyperparameters was retained and used for prediction on the test set.³² All models were evaluated by the *R*-square (R^2), root-mean-square error (RMSE), and mean absolute error (MAE) and validated using 5-fold cross-validation. The constructed optimal models were finally used to screen the desired NMs from the newly generated nanostructure spaces.

Additionally, multiple model interpretation methods were used to understand how the machine learning models make decisions. These methods included RF feature importance analysis, SHAP (Shapley Additive Interpretation) values, and permutation importance analysis. RF feature importance is a built-in method to determine the importance of features based on the decrease in impurity. Permutation feature importance measures the increase in the prediction error of the model when a feature is permuted. SHAP analysis typically quantifies the contribution of each feature to individual predictions, providing a local interpretation of feature effects.

2.3. NP Synthesis and Characterization. The finally selected NP was then experimentally synthesized as follows: First, 10 mg of thiotamide was dissolved in 10 mL of dimethylformamide (DMF), and while stirring, 1.25 mL of 50 mmol/L $\text{HAuCl}_4 \cdot 3\text{H}_2\text{O}$ was then added. Next, 10 mL of deionized water was added and stirred for 30 min at room temperature until the solution became transparent. Further separation is carried out to obtain the desired NPs. Moreover, transmission electron microscopy (TEM) was carried out with JEM-2100F (JEOL, Japan) operated at 200 kV. Typically, 5 μL aliquots of NP aqueous solution were dropped onto the carbon-coated TEM grid and dried at room temperature in a vacuum drier.

2.4. Cellular Uptake Assay. A549 cells were cultured in 24-well plates at a density of 100000 cells/mL. After incubating for 24 h, the cells were rinsed with PBS.³³ Next, the NP solution was introduced into the RPMI-1640 cell culture

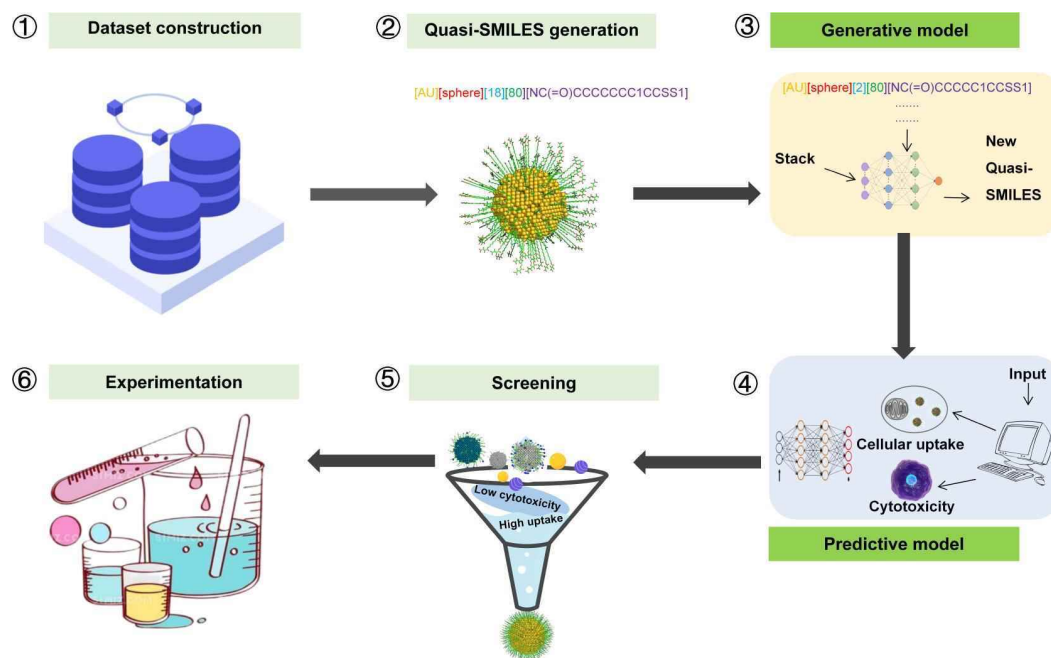


Figure 1. Workflow of our proposed *de novo* design framework. The entire process commences with (1) the collection of nanobioactivity data from PubVINAS, followed by (2) the conversion of nanostructures into Quasi-SMILES representations. Subsequently, (3) the generative model is employed to discern patterns and fabricate new NMs. The calculated tetrahedron descriptors of newly generated nanostructures (4) are then fed into the predictive models for cellular uptake and cytotoxicity evaluation. During the (5) screening phase, many criteria, such as synthetic feasibility and novelty, are utilized to assess the new NMs further. Ultimately, the shortlisted NM (6) is subjected to experimental validation.

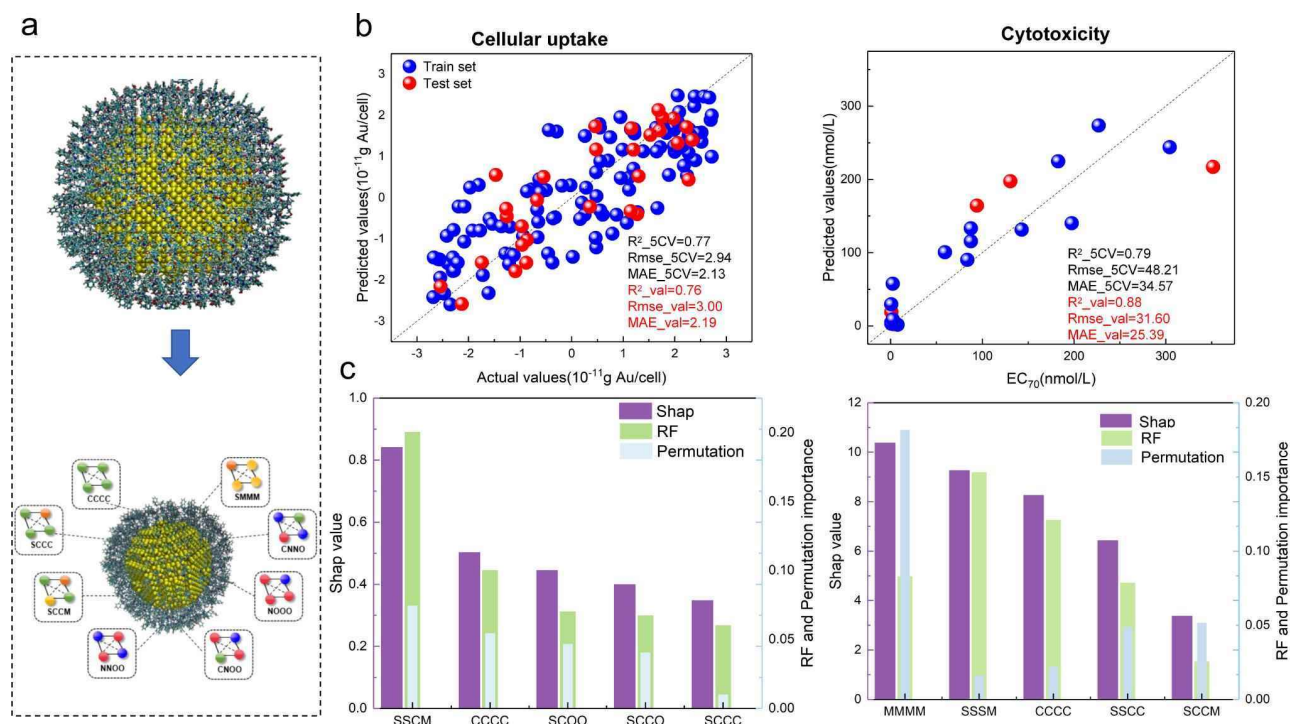


Figure 2. Predictive performance and interpretation of the machine learning models. (a) Virtual nanostructure generation and geometrical nanodescriptor calculation. The virtual nanostructures were generated from the VINAS toolbox, and the nanodescriptors were calculated by atomic properties and tetrahedron fragments. (b) Correlations between the predictions from RF models and the experimental results. R^2 , RMSE, and MAE values are also depicted. (c) Feature importance analysis of two predictive models. Three model interpretation methods were used to calculate the critical descriptors, including RF feature importance analysis, SHAP value analysis, and permutation importance analysis. The top 5 important features are shown in the figure.

medium at a concentration of 50 μ g/mL. Following a 24 h incubation, the sample was washed in PBS to eliminate any

excess NPs in wells. Subsequently, the cells were detached by using a trypsin-EDTA solution (containing 0.25% trypsin and

1 mM EDTA) and quantified. The separated cells were placed in aqua regia to dissolve the NPs completely. The cellular uptake of NPs was determined using inductively coupled plasma mass spectrometry (ICP-MS).

2.5. Cytotoxicity Assay. The viability of A549 cells exposed to NPs was conducted using CellTiter-Glo kits obtained from Promega Corporation (Madison, WI, USA).³⁴ A549 cells were seeded in 96-well plates at a density of 60000 cells/mL. After 24 h, the cells were treated with NPs at different concentrations for another 48 h.³⁵ Cell lysis and luminescence analysis were performed using the CellTiter-Glo assay and a microplate reader.

3. RESULTS

3.1. A *de novo* design Framework Was Proposed to Generate NMs with Desired Properties. The primary objective of this study is to utilize a novel deep-learning framework for designing NMs with high cellular uptake and low cytotoxicity. This framework encompasses several vital steps: data set construction, Quasi-SMILES generation, construction of generative and predictive models, virtual screening, and experimental validation (Figure 1). The core of this deep learning framework involves one generative model and two parallel predictive models, which were heavily reliant on the quality of the data used for training. A critical first step is, therefore, to acquire enough high-quality data from PubVINAS, a modeling-friendly NM database designed for machine learning and molecular simulation.²⁶ As a prerequisite for the generative model, the three-dimensional nanostructural information (e.g., core materials, particle sizes, and ligand structures) was then encoded into Quasi-SMILES. Figure S1 shows the distribution of the nanostructure data used in the current study. In the *de novo* design process, the generative model learned the syntax in Quasi-SMILES to produce novel nanostructures. The predictive models were constructed by various machine learning methods and tetrahedral nano-descriptors, and the optimal models were used for virtual screening from the newly generated NMs. Two parallel predictive models were employed to ensure the NMs with high cellular uptake and low toxicity. Furthermore, the novelty and synthesizability of NMs were also considered during the screening process. The finally selected NM was synthesized, and its predicted bioactivities were experimentally validated.

3.2. The Predictive Models Demonstrated Outstanding Efficacy for Predicting the Nanobioactivities. The virtual nanostructures and tetrahedral descriptors of NMs were generated before the predictive models were constructed (Figure 2a). The virtual nanostructures were automatically generated from our VINAS toolbox by inputting the corresponding physicochemical parameters of NPs, such as core materials, particle sizes, and ligand structures. The tetrahedral descriptors were calculated by combining the nanostructure fragments and the atomic electronegativity values.³⁰ The predictive models employed RF, XGBoost, SVM, and kNN methods. Among them, the RF method demonstrated the best results for cellular uptake, cytotoxicity, and LogP prediction (Table S1). RF enhances the model's generalization capability by constructing multiple decision trees and aggregating their predictions. RF can improve the model's robustness and generalization performance in scenarios with limited data through random sampling and feature selection.³⁶ As shown in Figure 2b, the predictive models exhibited superior performance for both 5-fold cross-

validation and external validation, indicated by high R^2 and low RMSE and MAE. The R^2 values of both predictive models exceeded 0.75, revealing that the machine-learning models successfully captured the relationships between nanostructures and their associated cellular uptake and cytotoxicity.

Model interpretation allows us to understand how the predictive models make decisions. This analysis helps identify crucial structures or physicochemical properties of NMs associated with the nanobioactivities, such as cellular uptake and cytotoxicity. The tetrahedron descriptors were ranked based on three model interpretation methods, i.e., the SHAP values, the built-in feature importance of the RF model, and the permutation importance analysis, as shown in Figure 2c. The high ranking of a molecular descriptor indicates its pivotal role in the final predictive model. The results from the analysis using three different model interpretation methods are broadly consistent. The findings indicate that five tetrahedral structural descriptors (SSCM, CCCC, SCOO, SCCO, and SCCC) significantly impact the cellular uptake of NMs. As described from the above definition of the tetrahedral nanodescriptors, the SSCM represents the core material-related nanostructures, while the other four nanodescriptors (CCCC, SCOO, SCCO, and SCCC) indicate the structural features of the surface ligands. The above results revealed that the core material and surface ligands both contributed significantly to determining the cellular uptake of NMs.³⁵ In addition to univariate analysis, we investigated the impact of pairwise interactions between input features on the model predictions. From Figure S2a, it can be seen that cellular uptake is driven by a combination of CCCC, SCOO, SCCO, and SCCC. The ranking of feature interaction values closely aligns with individual feature importance, indicating that cellular uptake is influenced not only by dominant features but also by interactions between features. By leveraging multiple interpretation methods, we can achieve a more balanced, detailed, and actionable understanding of model predictions, which could significantly improve the interpretability of the results and guide future optimization strategies.

Regarding the cytotoxicity model, the top five tetrahedral structure descriptors are MMMM, SSSM, CCCC, SCCC, and SSCM. Similar to the cellular uptake model results, both the core materials (MMMM, SSSM, and SSCM) and surface ligands (CCCC and SCCC) played significant roles in manifesting cytotoxicity. The NMs with different metal cores have unique physical and chemical properties that cause different biochemical activities and exhibit different toxicities. These NMs penetrate cell membranes, interact with protein molecules or subcellular organelle structures, and ultimately lead to various adverse effects such as neurotoxicity, immunotoxicity, and genotoxicity.³⁷ In addition, the cytotoxicity of NMs is significantly influenced by surface modification. NMs modified with different surface ligands induce varying levels of cytotoxicity.³⁸ By regulating surface chemistry, it affects the interaction of NMs with cell membranes and the intracellular responses triggered after penetration.³⁹ As shown in Figure S2b, the SHAP interaction values further demonstrated that the cytotoxicity of NMs was also influenced by potential synergistic or antagonistic effects among critical features.

Notably, the carbon-skeleton-related descriptor (CCCC) significantly influences cellular uptake and cytotoxicity. The ligand with aliphatic/aromatic carbons contributed substantially to the NMs' hydrophobicity, thus determining their

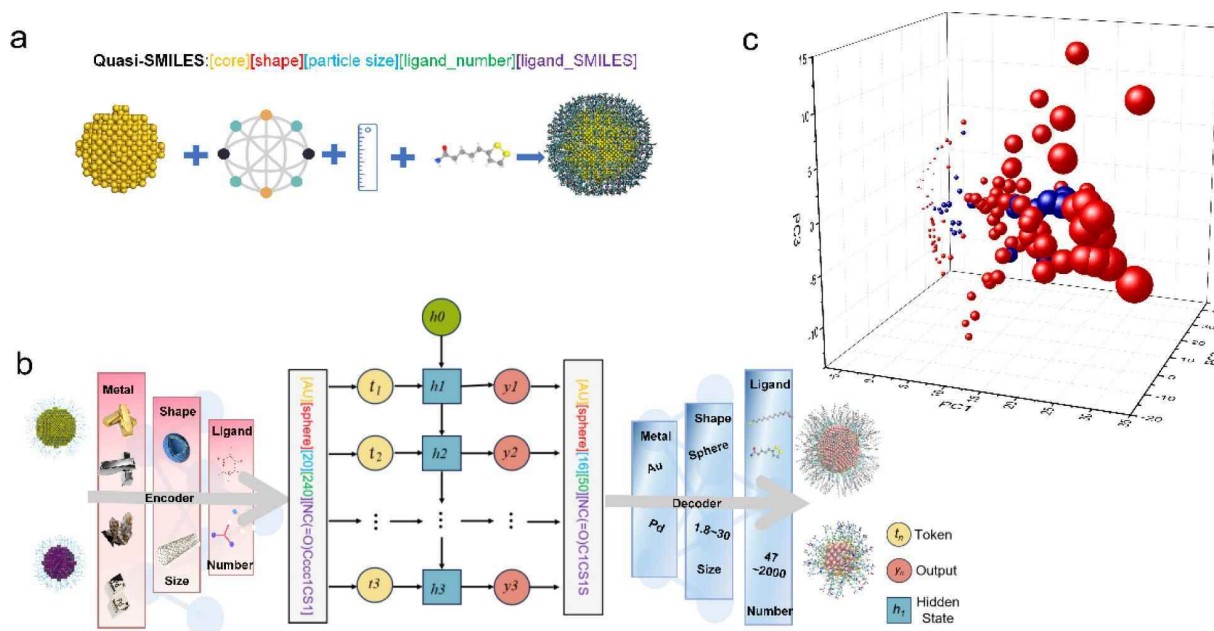


Figure 3. The generative model created novel NPs with structure diversity. (a) Example of representing nanostructures with the Quasi-SMILES format. The Quasi-SMILES model encodes detailed information about core materials, particle sizes, and surface ligands. (b) The proposed generative model is based on encoder-decoder architecture. (c) PCA diagram of NPs. The red spheres represent the NPs from PubVINAS, while the blue spheres represent the newly generated NPs. The diameter of the sphere is set according to the diameter of NPs.

cellular uptake capacity and cytotoxicity.³⁵ To verify the above results, an extra machine learning model was constructed to predict the hydrophobicity (indicated by LogP) of the NMs. As shown in Figure S3, the LogP prediction model also reveals the combined effects of core materials and surface ligands (especially the carbon skeleton). NMs with high lipophilicity traverse cell membranes more efficiently due to the hydrophobic nature of phospholipid tails and thus result in various cellular responses such as cytotoxicity. The result further demonstrates the significant role of LogP in determining the NP's cellular uptake and cytotoxicity.

3.3. Our Generative Model Successfully Produced a Variety of Innovative NMs with Structure Diversity.

Parallely, we developed a generative model to create novel NMs for subsequent virtual screening. The generative model employs a stack-RNN to learn the construction rules of Quasi-SMILES representations and then generates many novel NMs. As shown in Figure 3a, the Quasi-SMILES representations encode detailed information about material types, shapes, particle sizes, ligand numbers, and ligand structures.¹⁵ The generative model, comprising an encoder-decoder architecture, efficiently transforms input Quasi-SMILES sequences into molecular latent vectors decoded back into new Quasi-SMILES sequences (Figure 3b). These new Quasi-SMILES sequences were further converted to 3D nanostructures for tetrahedral descriptor calculation and visualization analysis, as described above. The model underwent training throughout 1000000 epochs, with the training loss exhibiting convergence (Figure S4), indicative of effective generation. Finally, 289 novel NMs with diverse chemical properties were generated and demonstrated by the detailed Quasi-SMILES representations, available in Table S2. The newly generated NMs contained three types of core materials (Au, Ag, Pd) and 153 ligands, none of which had been found in the input data. The principal component analysis (PCA) further demonstrated the

structural diversity and novelty of the newly generated NMs (Figure 3c).

As described above, the bioactivities of NMs are typically dependent on their material composition and ligand structure. However, the diversity of our newly generated NMs was mainly affected by their ligand structure since there are few material types, i.e., Au, Ag, and Pd. Therefore, we next investigated the novelty of NM ligands by calculating the Tanimoto coefficient. Between the 153 newly generated ligands and the 239 ligands in the original data set, a total of 36567 distances were computed, with values ranging from 0 to 1. If the value between two molecules is smaller than 0.5, they are generally considered dissimilar.⁴⁰ As shown in Figure S5, 65.9% of the Tanimoto coefficients are less than 0.5, indicating the high novelty of the newly generated ligands. Furthermore, the diversity of these freshly generated molecules was reflected by the wide distribution of their calculated molecular weights and LogP values (Figure S6). The novelty and diversity of surface ligands further demonstrate the generative model's effectiveness, ensuring that desired NMs can be selected from the newly generated chemical space.

Afterward, virtual screening was performed to pick up NMs with desired bioactivities. A diagram of the screening process is shown in Figure S7. Besides the novelty of the chemical structure, the feasibility of synthesis is also an essential criterion for assessing the quality of newly designed NMs. Factors such as the synthesis process, cost, and reaction conditions play crucial roles in determining the overall synthesis of NMs. In the current study, the core materials are commonly used, and we have established systematic methods for synthesizing these materials.³⁵ Therefore, the feasibility of synthesizing NMs is closely linked to the complexity of ligand synthesis, and we use the SA score method to assess the synthesizability of newly generated ligand molecules. In the future, a more comprehensive evaluation of

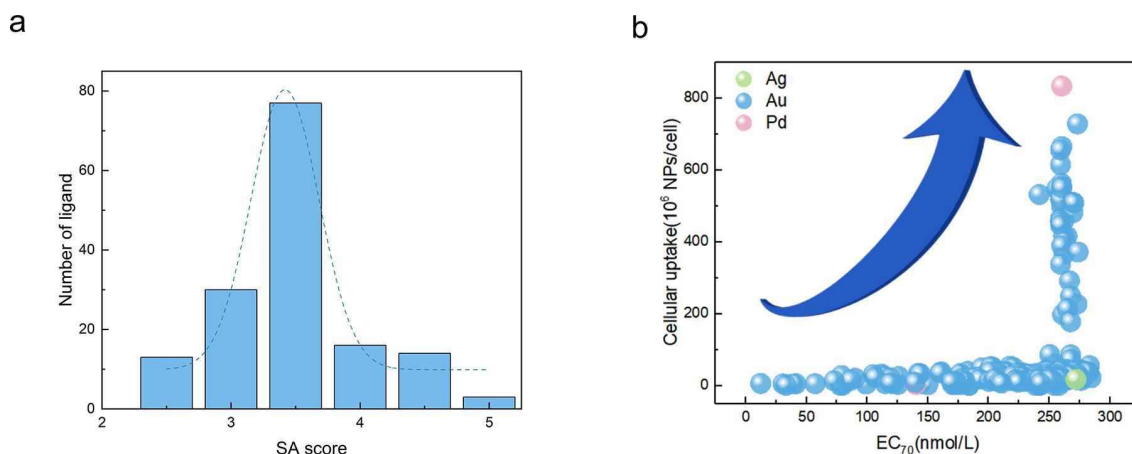


Figure 4. Synthesis analysis and virtual screening of NMs. (a) SA score distribution of 153 newly generated ligands. (b) Virtual screening of NMs with desired bioactivities. The arrow points to the NMs exhibiting high cellular uptake activity or low cytotoxicity.

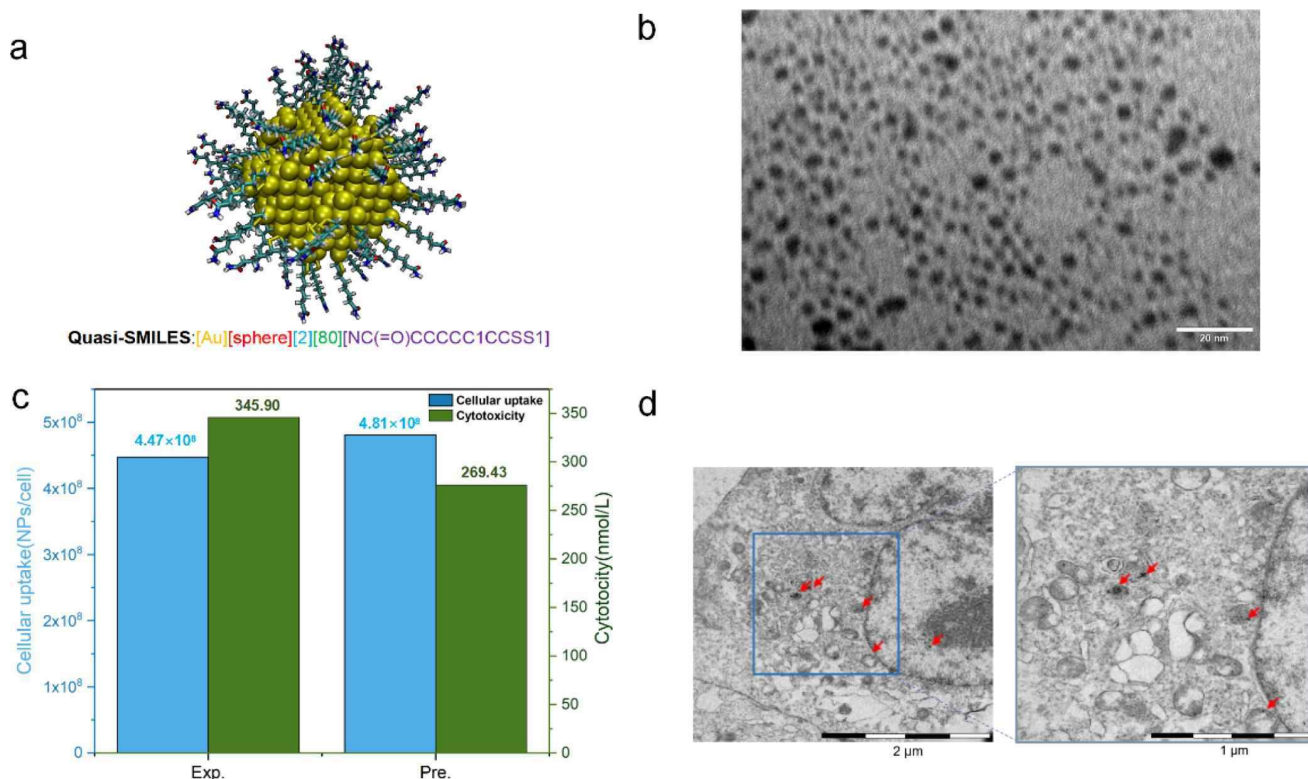


Figure 5. Experimental validation results of NP38 with high cellular uptake and low cytotoxicity. (a) The three-dimensional structure and Quasi-SMILES of the screened NP are shown. The 3D structure is constructed using the VINAS tool and rendered by the VMD software. (b) TEM image of the synthesized NPs. (c) Comparison of predicted and experimental values of NP38. (d) TEM images of cells with internalized NPs.

the synthetic challenges associated with the entire NMs is required. For example, it will be helpful to develop a quantitative indicator to assess the reproducibility and synthesizability of desired NMs through comprehensively considering factors such as material properties, synthesis methods, experimental conditions, and economic costs. Possible alternatives should be considered if the proposed materials have proven difficult or impossible to synthesize. As shown in Figure 4a, the SA scores of the ligands ranged from 2.02 to 4.90, with an average of 3.25. Lower SA scores indicate a higher possibility of obtaining a synthetic route for a compound, and molecules with SA scores above 6 are generally

considered infeasible to synthesize.⁴¹ Thus, these newly generated ligands are easy to synthesize. Overall, the NMs generated by stack-RNN possessed high structural diversity, which increased the probability of discovering novel NPs with high cellular uptake activity and low cytotoxicity. Additionally, the newly generated ligands are easy to synthesize, which improves the practicality of NMs.

We calculated the tetrahedral descriptors of these newly generated NMs and used optimal predictive models to evaluate their cellular uptake and cytotoxicity. As shown in Figure 4b, the predicted cellular uptake values range from 0.26 to 833.19 (10^6 NPs/cell), while the expected values for cell toxicity

(EC₇₀) are from 12.13 to 284.42 (nmol/L). The wide range of predicted values indicated that our newly generated NPs exhibited a high diversity of biological activities, demonstrating our generative model's effectiveness. However, it should also be noted that our current model may be mainly applicable to these metal NMs due to the limitations of the data set's diversity.

3.4. An NP with High Cellular Uptake and Low Cytotoxicity Was Identified and Experimentally Validated. According to our prediction results, we selected an NM with a high cellular uptake and low cytotoxicity for synthesis and experimental validation. Specifically, the predicted cellular uptake of our selected NP (NP38) is 481.11×10^6 NPs/cell, and the EC₇₀ cytotoxicity for A549 cells is 269.43 nmol/L. These findings indicate that NP38 performs exceptionally well in cellular uptake while maintaining low toxicity, making it a promising candidate for further use as a drug carrier or in other medical applications. Figure 5a shows the three-dimensional structure and Quasi-SMILES of NP38, which has a spherical gold core and is coated with 80 lipoamide ligands.

In addition, by synthesis of NP38, the morphologies of the NPs were characterized by TEM. NP38 exhibits a uniform spherical and three-dimensional structure (Figure 5b). The size of the NP38 particles is very close to the design. NP38 is monodispersed, facilitating its entrance into the cells. Experiments were conducted to verify the accuracy of the model further. Figure 5c shows the predicted and experimental cellular uptake and cytotoxicity values.

We next investigated the cellular uptake of NP38 in A549 cells by TEM (Figure 5d). TEM images revealed that the screened NPs were extensively internalized and entered cellular organelles, such as the nucleus and lysosomes (Figure S8). As observed by TEM images, the internalization of NPs into cells did not induce mitochondrial swelling, endoplasmic reticulum dilation, or lysosomal proliferation, indicating no apparent cellular stress responses.⁴² No evidence of plasma membrane disruption or organelle disintegration was observed, suggesting that the synthesized NPs exhibit minimal cytotoxicity. Obviously, the difference between the predicted results of the model and the actual experimental values is small, and the new NPs generated by the model have characteristics of high cellular uptake and low cytotoxicity. However, the biological activities of synthesized NP were only validated through simple *in vitro* experiments, which may differ from those in more complex biological environments. For example, intercellular interactions and microenvironmental factors collectively determine the behavior and toxicity of NMs within biological systems.^{43,44} Additionally, successful translation from the laboratory to the clinic requires elaborate and exhaustive investigations of the newly designed NMs including their *in vivo* biodistribution and clearance. These factors suggest that we should consider a broader biological context when assessing the biocompatibility and functionality of NMs to expand their practical applications.

4. DISCUSSION

The configuration space of NMs is vast, and their performance hinges on structural features such as shape, size, and surface modification.⁴⁵ Traditional experiments or machine-learning-assisted material design heavily requires significant resources or relies on extensive prior knowledge. Recently, the generative model provides a new research paradigm for material science

by automatically creating novel materials with desired properties.⁴⁶ Translating complex nanostructural information into computer-recognizable inputs poses significant challenges. The environmental and *in vivo* health safety of NMs is crucial for promoting the sustainable development of nanotechnology. Here, we propose and validate a generative-model-based method for the *de novo* design of novel NMs. This method employed the Quasi-SMILES representations to encode the structural information on NMs, and these representations were imported into the generative model to create an entirely novel NM space. Additionally, two parallel predictive models were constructed to ensure the high functionality and low toxicity of the newly generated NMs. Our proposed proof-of-concept method successfully designed a new NM with high cellular uptake and low cytotoxicity in cancer cells, and it is broadly transferable to other material design tasks.

Compared with small molecules, digitalizing the complex chemical structures is a great challenge.⁴⁷ This work demonstrates that the Quasi-SMILES representations of NMs serve as an input to the generative model, enabling the computer to swiftly recognize the material type, shape, particle size, and ligand information. As a kind of chemical language, the Quasi-SMILES representation is quite compact but can encode enough structural information for NMs. On the other hand, the Quasi-SMILES can, in turn, be converted into virtual nanostructures using our developed VINAS tool. Importantly, these virtual nanostructures can be further used for visualization analysis,⁴⁸ nanodescriptor calculation,^{14,49} and molecular simulations.⁵⁰ As a machine learning-aided material design prerequisite, the molecular descriptors significantly affected the quantitative structure–activity (toxicity) modeling.²⁷ In this study, virtual screening and experimental validation further demonstrated the practicality and versatility of the tetrahedral descriptors. We anticipate that a better understanding of NM representations could aid in the discovery and design of additional, much-needed classes of NMs and promote the sustainable development of nanoscience and nanotechnology.⁵¹

Evidence shows that NMs exert their toxic effects on various organisms, including cells,⁵² microorganisms,⁵³ aquatic animals,^{54,55} and mammals.⁵⁶ It is a complex and vital issue to balance the functionality and toxicity of NMs before they are brought to the market. During the past two decades, the rapid development of nanotoxicology has led to the accumulation of large quantities of nanotoxicity data. However, there is a considerable gap between the nanotoxicity data and critical information, which traditional experimental methods cannot completely bridge.²⁵ The emergence of AI enables the extraction of insights from these nanotoxicity big data. Typically, AI benefits nanotoxicology in the following ways: 1) unraveling the quantitative structure (property)–toxicity relationships, 2) predicting the potential hazards of new NMs, and 3) designing novel NMs with low toxicity. Here, the AI-driven toxicology model was introduced to decipher the critical structural features influencing nanobioactivities and was used to mitigate the adverse effects of our newly designed NMs.⁵⁷ This approach can be broadly extended to various materials to balance their biocompatibility and functionality.

Our results validate the potential application of Quasi-SMILES and generative models in designing biocompatible NMs with the desired functionality. However, the applicability of our proposed method may warrant further investigation and validation in future studies. The current data set needs to be

more extensive in size and diversity, particularly regarding the types of NM cores. For instance, the polymer and lipid NPs that are widely preferred in biological applications should be taken into consideration. Additionally, integrating data from multiple end points enables the model to capture a wider range of biological responses and complex nanobio interactions across different biological systems, such as other cell lines, aquatic organisms, and mammals. Sufficient, high-quality, and chemically diverse data form the foundation of machine learning modeling. By ensuring the accuracy and diversity of the data, we can enhance the model's learning capabilities and predictive performance, thereby advancing research and applications in related fields.⁵⁸ Additionally, by converting the characteristics of NMs into a standardized and machine-readable format, such as PDB (protein data bank) files and Quasi-SMILES, we can facilitate the integration of data from various sources. This approach can potentially develop a comprehensive and user-friendly database that encompasses the properties of NMs and critical nanobioactivity data.

Furthermore, there is an urgent need to develop a more extensive set of universal descriptors that capture the physical, chemical, and biological properties of NMs.⁴⁷ For instance, descriptors that combine structural features and multiple biological assays (e.g., oxidative stress, inflammation, and DNA damage) can increase the predictive performance of machine learning models and improve the toxicity/activity mechanism insights.⁵⁹ The nanodescriptors should also be explainable and easy for calculation. The development of more comprehensive nanodescriptors not only aids in rapidly improving model performance but also guides material design and optimization. Currently, the Quasi-SMILES representations created from generative models need to be manually input into the VINAS toolbox for nanostructure visualization and nanodescriptor calculation. An interactive online platform integrating a deep learning framework and VINAS toolbox would improve the accessibility for users and facilitate the discovery of novel NMs. Finally, establishing a feedback loop between experimental results and machine learning models will enable the ongoing refinement of data-driven material design. This requires ongoing efforts from stakeholders throughout the project lifecycle. We anticipate that addressing these challenges will be critical for advancing the field and ensuring successful application of our proposed method in real-world scenarios.

5. CONCLUSIONS

In this work, we propose a novel deep learning framework to accelerate the design of functional NMs with desired properties. Our method permits the exploration of expansive chemical space, autonomously generating NMs with RNN and Quasi-SMILES. More importantly, we have ensured that the designed NMs are not just effective for their intended applications but also have a minimized human health risk. This dual focus on performance and safety represents a crucial advancement in the sustainable development of next-generation NMs. Our method can also be applied to the design of other functional materials in different fields. By providing a faster and more reliable method than traditional exhaustive enumeration, our framework significantly advances the application of AI in material design, offering a practical solution for balancing the functionality and toxicity of materials.

■ ASSOCIATED CONTENT

Data Availability Statement

The source codes and data can be found at https://github.com/YanLabAI/NM_de-novo_design.

Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acsami.4c15600>.

Overview of physicochemical properties of NPs used in the current study (Figure S1); the SHAP interaction values of top 10 features for cellular uptake and cytotoxicity (Figure S2); the predictive performance and feature importance analysis of LogP machine learning model (Figure S3); training loss of the generative model; loss is recorded every 10 epochs, with a total of 200000 epochs (Figure S4); the Tanimoto similarity between ligands of the newly generated NPs and ligands of NPs in the PubVINAS database (Figure S5); molecular weight and LogP distribution of newly generated ligands (Figure S6); a diagram for screening NMs (Figure S7); TEM images of A549 cells after treatment with culture media alone and Au NP (Figure S8); the performance of various machine learning models for predicting cellular uptake, cytotoxicity and LogP (Table S1); Quasi-SMILES information for newly generated NPs (Table S2) (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Xiliang Yan – Institute of Environmental Research at Greater Bay Area, Key Laboratory for Water Quality and Conservation of the Pearl River Delta, Ministry of Education, Guangzhou University, Guangzhou 510006, China; College of Animal Science, South China Agricultural University, Guangzhou 510642, China; orcid.org/0000-0003-4173-6228; Email: yanxiliang1991@scau.edu.cn

Authors

Ying He – Institute of Environmental Research at Greater Bay Area, Key Laboratory for Water Quality and Conservation of the Pearl River Delta, Ministry of Education, Guangzhou University, Guangzhou 510006, China

Fang Liu – Department of Plastic Surgery, The First Affiliated Hospital of Shandong First Medical University and Shandong Provincial Qianfoshan Hospital, Jinan, Shandong 250014, PR China; Jinan Clinical Research Center for Tissue Engineering Skin Regeneration and Wound Repair, Jinan, Shandong 250014, PR China

Weicui Min – Institute of Environmental Research at Greater Bay Area, Key Laboratory for Water Quality and Conservation of the Pearl River Delta, Ministry of Education, Guangzhou University, Guangzhou 510006, China

Guohong Liu – School of Health, Guangzhou Vocational University of Science and Technology, Guangzhou 510555, China

Yinbao Wu – College of Animal Science, South China Agricultural University, Guangzhou 510642, China; orcid.org/0000-0002-3235-444X

Yan Wang – College of Animal Science, South China Agricultural University, Guangzhou 510642, China

Bing Yan – Institute of Environmental Research at Greater Bay Area, Key Laboratory for Water Quality and Conservation of the Pearl River Delta, Ministry of Education,

Guangzhou University, Guangzhou 510006, China;

orcid.org/0000-0002-7970-6764

Complete contact information is available at:
<https://pubs.acs.org/10.1021/acsami.4c15600>

Author Contributions

*Y.H. and F.L. contributed equally to this work. X.Y. conceived this study. B.Y. supervised the project, provided resources, and revised the manuscript. Y.H. and F.L. wrote the initial draft of the paper. X.Y., G.L., and Y.H. developed the machine learning models. Y.H. and G.L. analyzed the data. F.L. and G.H. performed the wet experiments. All authors have reviewed and approved the final version of the paper.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (nos. 22106025, 22476056, and 22036002), the Introduced Innovative R&D Team Project under “The Pearl River Talent Recruitment Program” of Guangdong Province (no. 2019ZT08L387), and the Guangdong Basic and Applied Basic Research Foundation (no. 2022A1515111082).

REFERENCES

- (1) Gu, M.; Zhang, Q.; Lamon, S. Nanomaterials for optical data storage. *Nat. Rev. Mater.* **2016**, *1* (12), 16070.
- (2) Keller, A. A.; Ehrens, A.; Zheng, Y.; Nowack, B. Developing trends in nanomaterials and their environmental implications. *Nat. Nanotechnol.* **2023**, *18* (8), 834–837.
- (3) Ramya, M.; Senthil Kumar, P.; Rangasamy, G.; Uma Shankar, V.; Rajesh, G.; Nirmala, K.; Saravanan, A.; Krishnapandi, A. A recent advancement on the applications of nanomaterials in electrochemical sensors and biosensors. *Chemosphere* **2022**, *308* (Pt2), 136416.
- (4) Mahmoudi, M. The need for robust characterization of nanomaterials for nanomedicine applications. *Nat. Commun.* **2021**, *12* (1), 5246.
- (5) Chen, J.; Cross, S. R.; Miara, L. J.; Cho, J. J.; Wang, Y.; Sun, W. Navigating phase diagram complexity to guide robotic inorganic materials synthesis. *Nat. Synth.* **2024**, *3*, 606–614.
- (6) Huang, G.; Guo, Y.; Chen, Y.; Nie, Z. Application of Machine Learning in Material Synthesis and Property Prediction. *Materials* **2023**, *16* (17), 5977.
- (7) Anstine, D. M.; Isayev, O. Generative Models as an Emerging Paradigm in the Chemical Sciences. *J. Am. Chem. Soc.* **2023**, *145* (16), 8736–8750.
- (8) Pollice, R.; Dos Passos Gomes, G.; Aldeghi, M.; Hickman, R. J.; Krenn, M.; Lavigne, C.; Lindner-D’Addario, M.; Nigam, A.; Ser, C. T.; Yao, Z.; Aspuru-Guzik, A. Data-Driven Strategies for Accelerated Materials Design. *Acc. Chem. Res.* **2021**, *54* (4), 849–860.
- (9) Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science* **2018**, *361* (6400), 360–365.
- (10) Oliveira, A. F.; Da Silva, J. L. F.; Quiles, M. G. Molecular Property Prediction and Molecular Design Using a Supervised Grammar Variational Autoencoder. *J. Chem. Inf. Model.* **2022**, *62* (4), 817–828.
- (11) Kotsias, P. C.; Arús-Pous, J.; Chen, H.; Engkvist, O.; Tyrchan, C.; Bjerrum, E. J. Direct steering of de novo molecular generation with descriptor conditional recurrent neural networks. *Nat. Mach. Intell.* **2020**, *2* (5), 254–265.
- (12) Prykhodko, O.; Johansson, S. V.; Kotsias, P. C.; Arús-Pous, J.; Bjerrum, E. J.; Engkvist, O.; Chen, H. A de novo molecular generation method using latent vector based generative adversarial network. *J. Cheminform.* **2019**, *11* (1), 74.
- (13) Wang, J.; Hsieh, C. Y.; Wang, M.; Wang, X.; Wu, Z.; Jiang, D.; Liao, B.; Zhang, X.; Yang, B.; He, Q.; Cao, D.; Chen, X.; Hou, T. Multi-constraint molecular generation based on conditional transformer, knowledge distillation and reinforcement learning. *Nat. Mach. Intell.* **2021**, *3* (10), 914–922.
- (14) Toropova, A. P.; Toropov, A. A.; Leszczynski, J.; Sizochenko, N. Using quasi-SMILES for the predictive modeling of the safety of 574 metal oxide nanoparticles measured in different experimental conditions. *Environ. Toxicol. Pharmacol.* **2021**, *86*, 103665.
- (15) Toropova, A. P.; Toropov, A. A. Quasi-SMILES as a basis to build up models of endpoints for nanomaterials. *Environ. Technol.* **2023**, *44* (28), 4460–4467.
- (16) Gao, H.; Li, Y.; Xie, Y.; Liang, D.; Li, J.; Wang, Y.; Xiao, Z.; Wang, H.; Gan, W.; Pattelli, L.; Xu, H. Optical wood with switchable solar transmittance for all-round thermal management. *Compos. B: Eng.* **2024**, *275*, 111287.
- (17) Malakar, A.; Kanel, S. R.; Ray, C.; Snow, D. D.; Nadagouda, M. N. Nanomaterials in the environment, human exposure pathway, and health effects: A review. *Sci. Total Environ.* **2021**, *759*, 143470.
- (18) Baek, H. S.; Park, M. K.; Kim, H. M.; Im, J. M.; Seo, H. S.; Park, H. J.; Nah, S. S. Reproductive and developmental toxicity screening test of new TiO₂ GST in Sprague-Dawley rats. *Environ. Anal. Health. Toxicol.* **2022**, *37* (3), No. e2022018.
- (19) Zhang, L.; Li, Q. X.; Li, X.; Yoza, B.; Zhou, L. Toxicity of Nanoparticles of AgO, La₂O₃, CuO, AgO–Fe₃O₄, Ag-Graphene, and GO-Cu-AgO to the Fungus *Moniliella wahieum* Y12(T) Isolated from Degraded Biodiesel and the Bacterium *Escherichia coli*. *J. Biomed. Nanotechnol.* **2022**, *18* (3), 928–938.
- (20) Scanlan, L. D.; Reed, R. B.; Loguinov, A. V.; Antczak, P.; Tagmount, A.; Aloni, S.; Nowinski, D. T.; Luong, P.; Tran, C.; Karunaratne, N.; Pham, D.; Lin, X. X.; Falciani, F.; Higgins, C. P.; Ranville, J. F.; Vulpe, C. D.; Gilbert, B. Silver nanowire exposure results in internalization and toxicity to *Daphnia magna*. *ACS Nano* **2013**, *7* (12), 10681–10694.
- (21) Zhou, Y.; Lei, L.; Chen, P.; Guo, W.; Guo, Y.; Yang, L.; Han, J.; Hu, B.; Zhou, B. Effects of nano-TiO₂ on the bioavailability and toxicity of bis(2-ethylhexyl)-2,3,4,5-tetrabromophthalate (TBPH) in developing zebrafish. *Chemosphere* **2022**, *295*, 133862.
- (22) Li, J.; Chen, Z.; Huang, R.; Miao, Z.; Cai, L.; Du, Q. Toxicity assessment and histopathological analysis of nano-ZnO against marine fish (*Mugilogobius chulae*) embryos. *J. Environ. Sci.* **2018**, *73*, 78–88.
- (23) Zhang, Y. N.; Poon, W.; Tavares, A. J.; McGilvray, I. D.; Chan, W. C. Nanoparticle-liver interactions: Cellular uptake and hepatobiliary elimination. *J. Control. Release* **2016**, *240*, 332–348.
- (24) Pastorino, P.; Prearo, M.; Barceló, D. Ethical principles and scientific advancements: In vitro, in silico, and non-vertebrate animal approaches for a green ecotoxicology. *Green. Anal. Chem.* **2024**, *8*, 100096.
- (25) Yan, X.; Yue, T.; Winkler, D. A.; Yin, Y.; Zhu, H.; Jiang, G.; Yan, B. Converting Nanotoxicity Data to Information Using Artificial Intelligence and Simulation. *Chem. Rev.* **2023**, *123* (13), 8575–8637.
- (26) Yan, X.; Sedykh, A.; Wang, W.; Yan, B.; Zhu, H. Construction of a web-based nanomaterial database by big data curation and modeling friendly nanostructure annotations. *Nat. Commun.* **2020**, *11* (1), 2519.
- (27) Trinh, T. X.; Choi, J. S.; Jeon, H.; Byun, H. G.; Yoon, T. H.; Kim, J. Quasi-SMILES-Based Nano-Quantitative Structure–Activity Relationship Model to Predict the Cytotoxicity of Multiwalled Carbon Nanotubes to Human Lung Cells. *Chem. Res. Toxicol.* **2018**, *31* (3), 183–190.
- (28) Hu, P.; Zou, J.; Yu, J.; Shi, S. De novo drug design based on Stack-RNN with multi-objective reward-weighted sum and reinforcement learning. *J. Mol. Model.* **2023**, *29* (4), 121.
- (29) Ertl, P.; Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminform.* **2009**, *1* (1), 8.
- (30) Yan, X.; Sedykh, A.; Wang, W.; Zhao, X.; Yan, B.; Zhu, H. In silico profiling nanoparticles: predictive nanomodeling using universal

nanodescriptors and various machine learning approaches. *Nanoscale* **2019**, *11* (17), 8352–8362.

(31) Wang, W.; Sedykh, A.; Sun, H.; Zhao, L.; Russo, D. P.; Zhou, H.; Yan, B.; Zhu, H. Predicting Nano–Bio Interactions by Integrating Nanoparticle Libraries and Quantitative Nanostructure Activity Relationship Modeling. *ACS Nano* **2017**, *11* (12), 12641–12649.

(32) Bhattacharjee, A.; Murugan, R.; Soni, B.; Goel, T. Ada-GridRF: A Fast and Automated Adaptive Boost Based Grid Search Optimized Random Forest Ensemble model for Lung Cancer Detection. *Phys. Eng. Sci. Med.* **2022**, *45* (3), 981–994.

(33) Bacova, J.; Knotek, P.; Kopecka, K.; Hromadko, L.; Capek, J.; Nyvltova, P.; Bruckova, L.; Schröterova, L.; Sestakova, B.; Palarcik, J.; Motola, M.; Cizkova, D.; Bezrouk, A.; Handl, J.; Fiala, Z.; Rudolf, E.; Bilkova, Z.; Macak, J. M.; Rousar, T. Evaluating the Use of TiO₂ Nanoparticles for Toxicity Testing in Pulmonary A549 Cells. *Int. J. Nanomed.* **2022**, *17*, 4211–4225.

(34) Akita, T.; Horiguchi, M.; Ozawa, C.; Terada, H.; Yamashita, C. The Effect of a Retinoic Acid Derivative on Cell-Growth Inhibition in a Pulmonary Carcinoma Cell Line. *Biol. Pharm. Bull.* **2016**, *39* (3), 308–312.

(35) Bai, X.; Wang, S.; Yan, X.; Zhou, H.; Zhan, J.; Liu, S.; Sharma, V. K.; Jiang, G.; Zhu, H.; Yan, B. Regulation of Cell Uptake and Cytotoxicity by Nanoparticle Core under the Controlled Shape, Size, and Surface Chemistries. *ACS Nano* **2020**, *14* (1), 289–302.

(36) He, Y.; Liu, G.; Hu, S.; Wang, X.; Jia, J.; Zhou, H.; Yan, X. Implementing comprehensive machine learning models of multi-species toxicity assessment to improve regulation of organic compounds. *J. Hazard. Mater.* **2023**, *458*, 131942.

(37) Xiong, P.; Huang, X.; Ye, N.; Lu, Q.; Zhang, G.; Peng, S.; Wang, H.; Liu, Y. Cytotoxicity of Metal-Based Nanoparticles: From Mechanisms and Methods of Evaluation to Pathological Manifestations. *Adv. Sci.* **2022**, *9* (16), 2106049.

(38) Sanchez-Cano, C.; Carril, M. Recent Developments in the Design of Non-Biofouling Coatings for Nanoparticles and Surfaces. *Int. J. Mol. Sci.* **2020**, *21* (3), 1007.

(39) Zhu, X. M.; Fang, C.; Jia, H.; Huang, Y.; Cheng, C. H.; Ko, C. H.; Chen, Z.; Wang, J.; Wang, Y. X. Cellular uptake behaviour, photothermal therapy performance, and cytotoxicity of gold nanorods with various coatings. *Nanoscale* **2014**, *6* (19), 11462–11472.

(40) Matter, H. Selecting optimally diverse compounds from structure databases: a validation study of two-dimensional and three-dimensional molecular descriptors. *J. Med. Chem.* **1997**, *40* (8), 1219–1229.

(41) Bilodeau, C.; Jin, W.; Xu, H.; Emerson, J. A.; Mukhopadhyay, S.; Kalantar, T. H.; Jaakkola, T.; Barzilay, R.; Jensen, K. F. Generating molecules with optimized aqueous solubility using iterative graph translation. *React. Chem. Eng.* **2022**, *7* (2), 297–309.

(42) Augustine, R.; Hasan, A.; Primavera, R.; Wilson, R. J.; Thakor, A. S.; Kevadiya, B. D. Cellular uptake and retention of nanoparticles: Insights on particle properties and interaction with cellular components. *Mater. Today Commun.* **2020**, *25*, 101692.

(43) Ji, Y.; Wang, Y.; Wang, X.; Lv, C.; Zhou, Q.; Jiang, G.; Yan, B.; Chen, L. Beyond the promise: Exploring the complex interactions of nanoparticles within biological systems. *J. Hazard. Mater.* **2024**, *468*, 133800.

(44) Seo, S.; Lee, J. E.; Lee, K.; Kim, H. N. Effects of microenvironmental factors on assessing nanoparticle toxicity. *Environ. Sci. Nano* **2022**, *9* (2), 454–476.

(45) Reimer, M.; Niemeier, S.; Laumann, D.; Denz, C.; Heusler, S. An acoustic teaching model illustrating the principles of dynamic mode magnetic force microscopy. *Nanotechnol. Rev.* **2017**, *6* (2), 221–232.

(46) Schütt, K. T.; Arbabzadah, F.; Chmiela, S.; Müller, K. R.; Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *Nat. Commun.* **2017**, *8* (1), 13890.

(47) Wyrzykowska, E.; Mikołajczyk, A.; Lynch, I.; Jeliazkova, N.; Kochev, N.; Sarimveis, H.; Doganis, P.; Karatzas, P.; Afantitis, A.; Melagraki, G.; Serra, A.; Greco, D.; Subbotina, J.; Lobaskin, V.; Bañares, M. A.; Valsami-Jones, E.; Jagiello, K.; Puzyn, T. Representing

and describing nanomaterials in predictive nanoinformatics. *Nat. Nanotechnol.* **2022**, *17* (9), 924–932.

(48) Toropova, A. P.; Toropov, A. A. Nanomaterials: Quasi-SMILES as a flexible basis for regulation and environmental risk assessment. *Sci. Total Environ.* **2022**, *823*, 153747.

(49) Tang, H.; Jiang, J. Active learning boosted computational discovery of covalent–organic frameworks for ultrahigh CH₄ storage. *Nanomaterials* **2022**, *1014* (68), No. e17856.

(50) Meneses, J.; González-Durruthy, M.; Fernandez-de-Gortari, E.; Toropova, A. P.; Toropov, A. A.; Alfaro-Moreno, E. A Nano-QSTR model to predict nano-cytotoxicity: an approach using human lung cells data. *Part. Fibre Toxicol.* **2023**, *20* (1), 21.

(51) Ding, Z.; Pattelli, L.; Xu, H.; Sun, W.; Li, X.; Pan, L.; Zhao, J.; Wang, C.; Zhang, X.; Song, Y.; et al. Iridescent Daytime Radiative Cooling with No Absorption Peaks in the Visible Range. *Small* **2022**, *18* (25), 2202400.

(52) Botha, T. L.; Elemike, E. E.; Horn, S.; Onwudiwe, D. C.; Giesy, J. P.; Wepener, V. Cytotoxicity of Ag, Au and Ag-Au bimetallic nanoparticles prepared using golden rod (*Solidago canadensis*) plant extract. *Sci. Rep.* **2019**, *9* (1), 4169.

(53) Metch, J. W.; Burrows, N. D.; Murphy, C. J.; Pruden, A.; Vikesland, P. J. Metagenomic analysis of microbial communities yields insight into impacts of nanoparticle design. *Nat. Nanotechnol.* **2018**, *13* (3), 253–259.

(54) Magro, M.; De Liguoro, M.; Franzago, E.; Baratella, D.; Vianello, F. The surface reactivity of iron oxide nanoparticles as a potential hazard for aquatic environments: A study on *Daphnia magna* adults and embryos. *Sci. Rep.* **2018**, *8* (1), 13017.

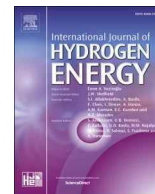
(55) Gong, H.; Li, R.; Li, F.; Guo, X.; Xu, L.; Gan, L.; Yan, M.; Wang, J. Toxicity of nanoplastics to aquatic organisms: Genotoxicity, cytotoxicity, individual level and beyond individual level. *J. Hazard. Mater.* **2023**, *443* (PtB), 130266.

(56) Boyes, W. K.; van Thriel, C. Neurotoxicology of Nanomaterials. *Chem. Res. Toxicol.* **2020**, *33* (5), 1121–1144.

(57) Ding, Z.; Li, X.; Ji, Q.; Zhang, Y.; Li, H.; Zhang, H.; Pattelli, L.; Li, Y.; Xu, H.; Zhao, J. Machine-Learning-Assisted Design of a Robust Biomimetic Radiative Cooling Metamaterial. *ACS Mater. Lett.* **2024**, *6* (6), 2416–2424.

(58) Greener, J. G.; Kandathil, S. M.; Moffat, L.; Jones, D. T. A guide to machine learning for biologists. *Nat. Rev. Mol. Cell Biol.* **2022**, *23* (1), 40–55.

(59) Jia, X.; Wen, X.; Russo, D. P.; Aleksunes, L. M.; Zhu, H. Mechanism-driven modeling of chemical hepatotoxicity using structural alerts and an in vitro screening assay. *J. Hazard. Mater.* **2022**, *436*, 129193.



Effect of different strategies for modifying graphene on the adsorption and gas sensing of trimethylamine: Insights from DFT study

Yuanchao Li^a, Xiliang Yan^{a,b,*}

^a Institute of Environmental Research at Greater Bay Area, Key Laboratory for Water Quality and Conservation of the Pearl River Delta, Ministry of Education, Guangzhou University, Guangzhou, 510006, China

^b College of Animal Science, South China Agricultural University, Guangzhou, 510642, China

ARTICLE INFO

Handling Editor: I Tolj

Keywords:

Trimethylamine
Graphene-based sensor
Sensing mechanism
Electric field
Density functional theory

ABSTRACT

Graphene materials have shown a great promise for gas sensors, however, the effect of different modification strategies on gas sensing mechanisms has not been studied. In this research, the adsorption and sensitivity properties of trimethylamine (TMA) on pristine (PG), Zn-decorated (Zn-G1), Zn-doped (Zn-G2), ZnN₄-embedded (ZnN₄-G), Zn-defect combined (ZnC₄-G) graphene are carefully discussed by density functional theory (DFT) calculations. Among five substrates, Zn-G2, ZnN₄-G, and ZnC₄-G exhibit high sensitivity and selectivity toward TMA due to the shorter adsorption distances, large amount of charge transfer and stronger adsorption strengths. Especially for Zn-G2, energy gap and work function are obviously altered after interaction with the TMA molecule. The recovery time of ZnN₄-G and ZnC₄-G are 269 and 6 s at 398 K, respectively. When the conditional temperature reaches 498 K, the recovery time of Zn-G2 only takes 91 s by applying negative electric field of 8×10^6 a.u. Such short recovery time makes Zn-G2, ZnN₄-G, and ZnC₄-G as a reversible gas sensor. Moreover, the applied positive electric fields can further enhance the selectivity and sensitivity of Zn-G2, ZnN₄-G, and ZnC₄-G. In light of these findings may provide theoretical guidance for developing high selectivity and sensitivity graphene-based gas sensor.

1. Introduction

Trimethylamine (TMA) is a harmful volatile organic amine gas, which can be released from fertilizer wastewater and livestock waste [1]. Prolonged exposure to TMA can cause several health problems such as headaches, eye irritation, nausea, breathing difficulty, and even death [2,3]. The allowable exposure time for humans in TMA environment is 15 ppm for 15 min [4]. In addition, TMA can be regarded as a signaling molecule for patients with chronic kidney disease when the concentration is greater than 0.2 ppm [5]. Meanwhile, TMA gas can also be generated by spoiled seafood (e.g., fish). The concentration of TMA can be used to evaluate the freshness of seafood (fresh, <10 ppm; preliminary rot, 10–50 ppm; corruption, >60 ppm) [6]. Therefore, rapid and effective detection of TMA gas is of great significance for environmental monitoring, disease diagnosis, and seafood quality control. To date, various techniques for TMA detection have been widely applied, such as calorimetric analysis, mass spectrometry, ion mobility spectrometry and PH test. However, expensive precise instruments, skilled

technicians, slow detection speed and long pre-experiment process make them unsuitable for real-time and on-site detection. In contrast, gas sensors provide an ideal platform for TMA detection with unique advantages of easy fabrication, high sensitivity, low cost, fast response speed, small size, etc.

Graphene as sensing material has drawn remarkable attention in the field of gas sensors due to its huge specific surface area, high-thermal stability, fascinating charge transfer properties and high surface reactivity [7]. Importantly, graphene exhibits excellent conductivity at room temperature, which is beneficial for reducing the operating temperature of the gas sensor. The gas sensing mechanism depends on the change in conductivity caused by the charge transfer between target molecule and graphene. Previous experimental and theoretical investigations have confirmed that surface modification of graphene can effectively enhance the surface reactivity and result in excellent sensing performance. Fan et al. fabricated NO₂ sensor based Au nanoparticles decorated graphene, showing fast response, good reversibility, and high sensitivity at room temperature [8]. Odey et al. theoretically predicted that doping of Ni

* Corresponding author. Institute of Environmental Research at Greater Bay Area, Key Laboratory for Water Quality and Conservation of the Pearl River Delta, Ministry of Education, Guangzhou University, Guangzhou, 510006, China.

E-mail address: yanxiliang1991@gzhu.edu.cn (X. Yan).

<https://doi.org/10.1016/j.ijhydene.2024.02.345>

Received 26 October 2023; Received in revised form 4 February 2024; Accepted 27 February 2024

Available online 9 March 2024

0360-3199/© 2024 Hydrogen Energy Publications LLC. Published by Elsevier Ltd. All rights reserved.

atom on graphene can enhance the selectivity and sensing properties to COCl_2 molecule [9]. Zhou et al. reported that the introduction of the defects in graphene greatly enhanced the gas adsorption stability [10]. Bo et al. revealed that adsorption properties of NO_2 molecule is significantly increased via decorating graphene with transition metal (Cu, Ni, Pd and Pt) [11,12]. Qu et al. studied the adsorption characteristics of SO_2 on the Pt-doping and C-vacancy co-modulated graphene by density functional theory (DFT) calculations [13]. They reported that this dopant-defect combination effect can enhance the sensitivity of graphene to gas molecules. Recently, the fascinating materials constructed by four-nitrogen coordinated transition metal (such as FeN_4 , NiN_4 , MnN_4 , etc.) embedded graphene have been studied from a theoretical perspective and proposed as promising substrate material for detecting harmful gases [14–17]. These achievements have highlighted the significance of different modification strategies (such as introducing defects and/or heteroatoms etc) on promoting the sensing performance of graphene-based sensor. However, the effect of different modification methods on the sensing mechanism of TMA has rarely been reported. As a consequence, a thorough insight into the adsorption characteristics of doped, decorated, four-nitrogen coordinated transition metal embedded, dopant-defect combined graphene is very necessary.

In this work, zinc (Zn) is selected as dopant due to its excellent electronic properties, which exhibits high binding affinity for the detection of harmful gases based on the many experimental and theoretical studies [18–21]. The adsorption behavior and sensitivity properties of TMA molecule on pristine (PG), Zn-decorated (Zn-G1), Zn-doped (Zn-G2), ZnN_4 -embedded (ZnN_4 -G), Zn-defect combined (ZnC_4 -G) graphene have been investigated from a theoretical perspective. Some important parameters, including adsorption energy, natural bond orbital charge, energy gap, work function, charge density differential, recovery time and so on are systematically analyzed to understand sensing mechanism. Moreover, the influence of electric field, background gas, and temperature are also analyzed to further explore the sensing properties in actual conditions. We hope that this study can provide some new ideas for developing high selectivity and sensitivity graphene-based gas sensor.

2. Computational method and model

The geometry optimizations and electrical properties of TMA molecule and five graphene substrates are performed using the B3LYP functional along with the LANL2DZ basis set for Zn atom and 6-311G(d, p) basis set for H, C, and N atoms. This method is widely used in theoretical research and provides reliable results [22–24]. ω B97XD is a hybrid long-range separated empirical-corrected dispersion functional, which can accurately evaluate the interaction energy [25–27]. Therefore, the adsorption system is calculated at the ω B97XD level of theory. The adsorption energy (E_{ads}) can be evaluated by:

$$E_{\text{ads}} = E_{\text{Complex}} - (E_{\text{Substrate}} + E_{\text{TMA}}) + E_{\text{BSSE}} \quad (1)$$

where E_{Complex} , $E_{\text{Substrate}}$ and E_{TMA} represent the total energy of adsorption system, functionalized graphene and TMA molecule, respectively. E_{BSSE} is the basis set superposition error corrected.

To evaluate the stability of doping systems, the cohesive energy (E_{cho}) and binding energy (E_{b}) are defined as:

$$E_{\text{cho}} = [E_{\text{Substrate}} - E_{\text{Zn}} - 22E_{\text{H}} - nE_{\text{C}} - mE_{\text{N}}] / (23 + n + m) \quad (2)$$

$$E_{\text{b}} = E_{\text{Substrate}} - E_{\text{Zn}} - E_{\text{Def-substrate}} \quad (3)$$

where $E_{\text{Def-substrate}}$, E_{Zn} , E_{H} , E_{C} and E_{N} stand for the total energy of Zn free substrate, isolated zinc, hydrogen, carbon and nitrogen atom, respectively.

Natural bond orbital (NBO) is utilized to analysis amount of charge transfer between TMA molecule and functionalized graphene. The electron density accumulation and depletion regions during the

adsorption process can be visualized through charge density difference (CDD), as follows:

$$\Delta\rho = \rho_{\text{Complex}} - \rho_{\text{Substrate}} - \rho_{\text{TMA}} \quad (4)$$

where ρ_{Complex} , $\rho_{\text{Substrate}}$ and ρ_{TMA} denote the charge densities of the adsorption system, functionalized graphene and TMA molecule, respectively. In addition, electrostatic potential (ESP), quantum theory of atoms in molecules (QTAIM) and independent gradient model based on Hirshfeld partition (IGMH) are carried out via Multiwfn 3.8 code [28]. All DFT calculations are performed using the Gaussian 16 program [29].

In this study, the model of pristine graphene (PG) with 80 carbon atoms is selected to explore surface adsorption characteristics, and the dangling bonds of the edge carbon atoms are saturated by hydrogen atoms, as shown in Fig. 1. This model provides sufficient surface reaction sites and is widely used in theoretical research [30–32]. For Zn-decorated graphene (Zn-G1), there are three different decorated positions, including carbon atom site, the bond site, and ring site. The model of Zn-doped graphene (Zn-G2) is obtained by replacing one carbon atom with one Zn atom. Zn-defect combined graphene (ZnC_4 -G) can be obtained by following two steps; (1) two neighboring carbon atoms are removed from the center of PG model, (2) a single Zn atom is placed at the defect center. ZnN_4 -embedded graphene (ZnN_4 -G) is obtained by replacing the four dangling carbon atoms of Zn-defect-G with four nitrogen atoms.

3. Results and discussion

3.1. Geometric and electronic properties of five carbon substrates

The optimized geometries of five carbon substrates (PG, Zn-G1, Zn-G2, ZnN_4 -G, and ZnC_4 -G) are shown in Fig. 1. It can be seen that C–C bond for PG has a length value of 1.42 Å, which in agreement with previous reported values [33,34]. After full optimization of Zn-G1, the Zn atom moves above the center of the hexagonal ring with the value of 3.08 Å. For Zn-G2, two equivalent Zn–C bonds (2.18 Å) and a shorter Zn–C bond (2.01 Å) are formed. In addition, Zn atom exhibits the larger atomic radius than that of C atom, resulting in the Zn atom protruding out of the graphene plane (Fig. 1 (c)). For ZnN_4 -G and ZnC_4 -G, the Zn–N and Zn–C bond lengths are almost equal and close to 1.97 Å, respectively, which are smaller than the sum of the covalent radii of the Zn atom, C atom and the N atom [35], implying that Zn–N and Zn–C form a stable covalent bond. The stability of doping systems can also be explained by the cohesive energy (E_{cho}). As shown in Table 1, the E_{cho} of Zn-G2, ZnN_4 -G, and ZnC_4 -G are −7.51, −7.50, and −7.49 eV, respectively, which are comparable with PG (−7.67 eV) and more negative than that of pristine and doped graphdiyne [36], indicating their high structural stability. The E_{b} of Zn-G2, ZnN_4 -G, and ZnC_4 -G are −1.44, −4.05, and −4.59 eV, respectively, suggesting that their acceptable stability. Previous experimental studies have also confirmed the possibility of those doping graphene [37–40].

The effects of different modification strategies on electronic properties are analyzed through total density of state (TDOS), as shown in Fig. 2(a)–(d). It can be seen that the TDOS of PG and Zn-G1 overlaps obviously, implying that their similar band structure. Zn atom doped graphene (Zn-G2) causes the highest occupied molecular orbitals (HOMO) and the lowest unoccupied molecular orbitals (LUMO) energy levels shift to the right (0.11 eV) and left (0.32 eV), respectively. Compared with PG, HOMO (LUMO) of ZnN_4 -G (ZnC_4 -G) shifts to the right (left) with the value of 0.46 eV. Based on the above analysis, the charge of Zn-G2, ZnN_4 -G and ZnC_4 -G is more easily excited, and leading to a larger conductivity of the sensitive material [41].

Electrostatic potential (ESP) can provide charge distribution of the molecule based on electronic density and has been widely used to predict the reaction sites of intermolecular interactions [42,43]. The blue

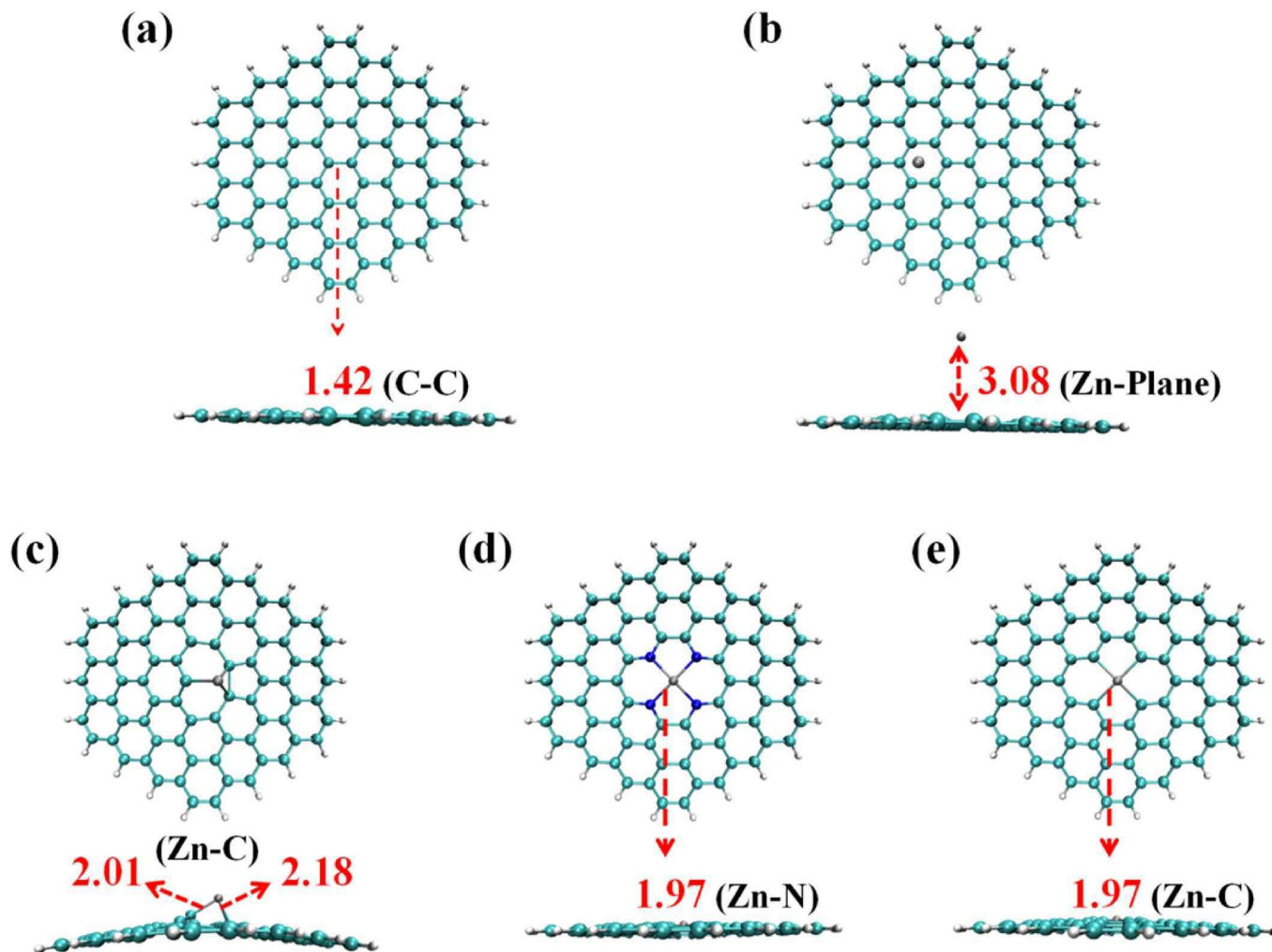


Fig. 1. Optimized configuration of (a) PG, (b) Zn-G1, (c) Zn-G2, (d) ZnN₄-G and (e) ZnC₄-G. The unit of the data is Å.

Table 1

The cohesive energy (E_{cho}) and binding energy (E_b) of doping systems. The unit of the data is eV.

| Substrate | E_{cho} | E_b |
|---------------------|------------------|-------|
| PG | −7.67 | – |
| Zn-G2 | −7.51 | −1.44 |
| ZnN ₄ -G | −7.50 | −4.05 |
| ZnC ₄ -G | −7.49 | −4.59 |

and red on the ESP maps refer to the most positive and negative regions. As can be seen from Fig. 3(a), PG surface shows negative ESP with value of −5.6 kcal/mol. It is worth noting that different strategies for modifying PG surface can induce significant changes in ESP distribution characters, in which a positive potential region appears near the Zn atom. Zn-G2 exhibits the most positive ESP with the global surface maximum of about 49.1 kcal/mol, followed by the ZnN₄-G (41.1 kcal/mol), ZnC₄-G (37.6 kcal/mol), and Zn-G1 (0.7 kcal/mol), respectively. It is predicted based on ESP analysis that Zn atom of Zn-G2, ZnN₄-G and ZnC₄-G can form a strong interaction with N atom of TMA via electrostatic interaction.

3.2. Sensing properties of pristine and modified graphene

Previous investigation has shown that H atom and N atom of organic molecule exhibit positive and negative ESP [44–46]. Hence, N atom of

TMA is placed above Zn atom of Zn-G2, ZnN₄-G and ZnC₄-G, while H atom of TMA is placed above PG and Zn-G1. After full relaxed optimization, the most stable adsorption configuration of TMA gas on pristine and modified graphene are depicted in Fig. 4(a1)–(d1), and the corresponding adsorption parameters are listed in Table 2. For PG/TMA and Zn-G1/TMA complexes, TMA molecule prefers to be adsorbed in a vertical orientation, and the corresponding adsorption distance (D) of about 2.70 Å. For Zn-G2/TMA, ZnN₄-G/TMA and ZnC₄-G/TMA complexes, the N atom of TMA molecule is trapped by the Zn atom of substrate, with the bond lengths of N–Zn measured to be 2.09, 2.20 and 2.23 Å, respectively, which is shorter than that of other 2D materials [47–51]. It is worth noting that H atom of TMA molecule is also trapped by the C atom of Zn-G2, which leads to TMA molecule tends to be obliquely adsorbed on the Zn-G2 surface. The smaller adsorption distance means that TMA molecule is more favorable to adsorb on Zn-G2, ZnN₄-G and ZnC₄-G. The TMA molecule on the Zn-G2 exhibits the most negative adsorption energy (−1.84 eV), followed by ZnN₄-G (−1.16 eV), ZnC₄-G (−1.08 eV), Zn-G1 (−0.29 eV) and PG (−0.06 eV), respectively. These values suggest that the adsorption of TMA molecule on Zn-G2, ZnN₄-G and ZnC₄-G can be identified as strong chemisorption, which are much higher than that of SnS/SnS₂ (−0.60 eV) [50], Nb₂C(OH)₂ (−1.19 eV) [51], stanene nanotube (−0.35 eV) [52] and other functional graphene substrates (−0.01 to −0.76 eV) [47–49], implying that Zn-G2, ZnN₄-G and ZnC₄-G exhibit high sensitivity toward TMA gas. While TMA molecule is weakly physisorbed on PG and Zn-G1 due to their smaller adsorption energy.

The charge density differential (CDD) is used to study charge

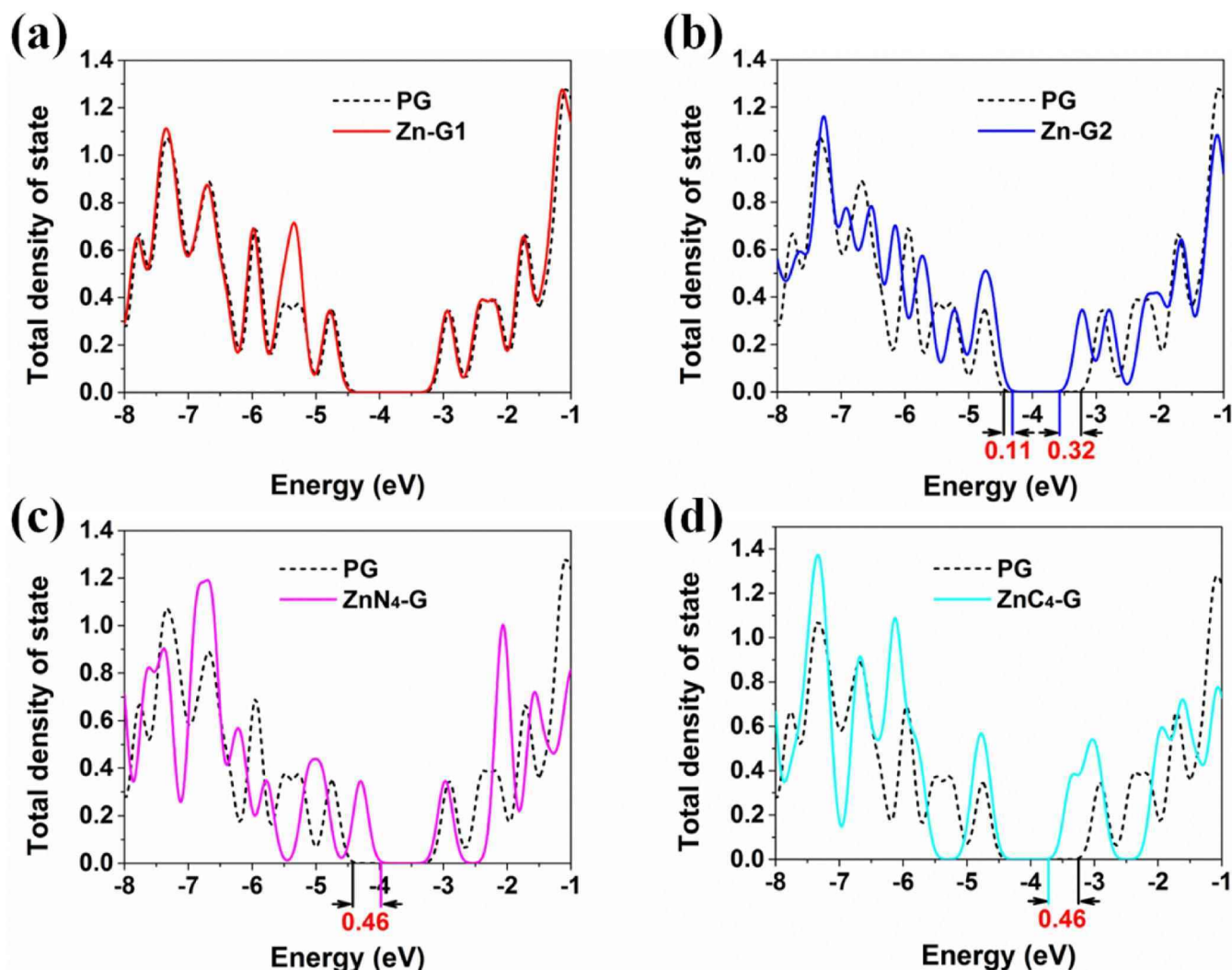


Fig. 2. Total density of state (TDOS) of pristine and various modified graphene.

redistribution during adsorption process, and the corresponding isosurface schematic is shown in Fig. 4(a2)-(e2). Yellow and mauve colors represent electron accumulation and depletion, respectively. TMA molecule exhibits obvious charge transfer with the Zn-G2, ZnN₄-G and ZnC₄-G, in which the electron accumulation area appears around Zn atom and the interface, while depletion locates around TMA molecule. However, there is no overlap between TMA molecule and substrate (PG and Zn-G1), implying their weak interaction, which is one of the reasons for their smaller adsorption energy. In addition, the amount of charge transfer (Q) between TMA molecule and substrate can be evaluated by natural bond orbital (NBO). The results manifest that TMA molecule acts as the donor and transfers largest charges of 0.138 e to ZnC₄-G, followed by the Zn-G2 (0.135 e), ZnN₄-G (0.107 e) and Zn-G1 (0.002 e), respectively. There is obvious charge transfer between Zn-G2, ZnN₄-G, ZnC₄-G and TMA compared with fluorographene (0.010 e) [51], Nb₂C(OH)₂ (0.070 e) [52] and stanene nanotube (0.049 e) [47]. For PG/TMA complex, TMA molecule as electron-acceptor by withdrawing 0.002 e from the PG. Based on the above analysis, the shorter adsorption distance, larger adsorption energy and more charge transfer indicate that Zn-G2, ZnN₄-G and ZnC₄-G have excellent sensitivity to TMA gas. Therefore, the Zn-G2, ZnN₄-G and ZnC₄-G are selected to further explore the effect of background gas, temperature, and applied electric field on their interaction with TMA molecule in following part.

In order to gain a more reliable and deeper insight into the nature of interaction between TMA molecule and substrate, independent gradient

model based on Hirshfeld partition of the molecular density (IGMH) is performed, which exhibits better graphical effect than independent gradient model [53]. The type and region of interaction can be intuitively observed by filling IGMH isosurfaces with different colors, where the blue and green colors denote strong attractive interaction and van der Waals interaction, respectively. As shown in Fig. 4(a3) and Fig. 4(b3), the large light green isosurface can be clearly observed between PG, Zn-G1 substrate and TMA molecule, indicating weak van der Waals interaction, which is consistent with the adsorption energy. Importantly, there is smaller green isosurface exists between the Zn atom of Zn-G1 substrate and H atom of TMA molecule, which leads to its larger adsorption energy than that of PG/TMA complex. From Fig. 4(c3)-(e3), it can be observed that there is dark blue isosurface exists between the Zn atom of substrates and the N atom of TMA molecule, indicating the domination of strong attractive interaction, while there are few light green isosurfaces between the C atom of substrates and H atom of TMA molecule. Those multiple interaction leads to their higher adsorption energy.

More details of the interaction characteristics between TMA molecule and substrate could be obtained from the quantum theory of atoms in molecules (QTAIM) analysis. The bond critical points (BCP) and the bond paths of adsorption system are shown in Fig. 5, and corresponding topological parameters of BCP (3, -1) between TMA molecule and substrate, including electron density ($\rho(r)$), laplacian of electron density ($\nabla^2\rho(r)$), lagrangian kinetic energy ($G(r)$), potential energy density ($V(r)$), eigenvalues of hessian (λ_n), and bond ellipticity index (ϵ) are

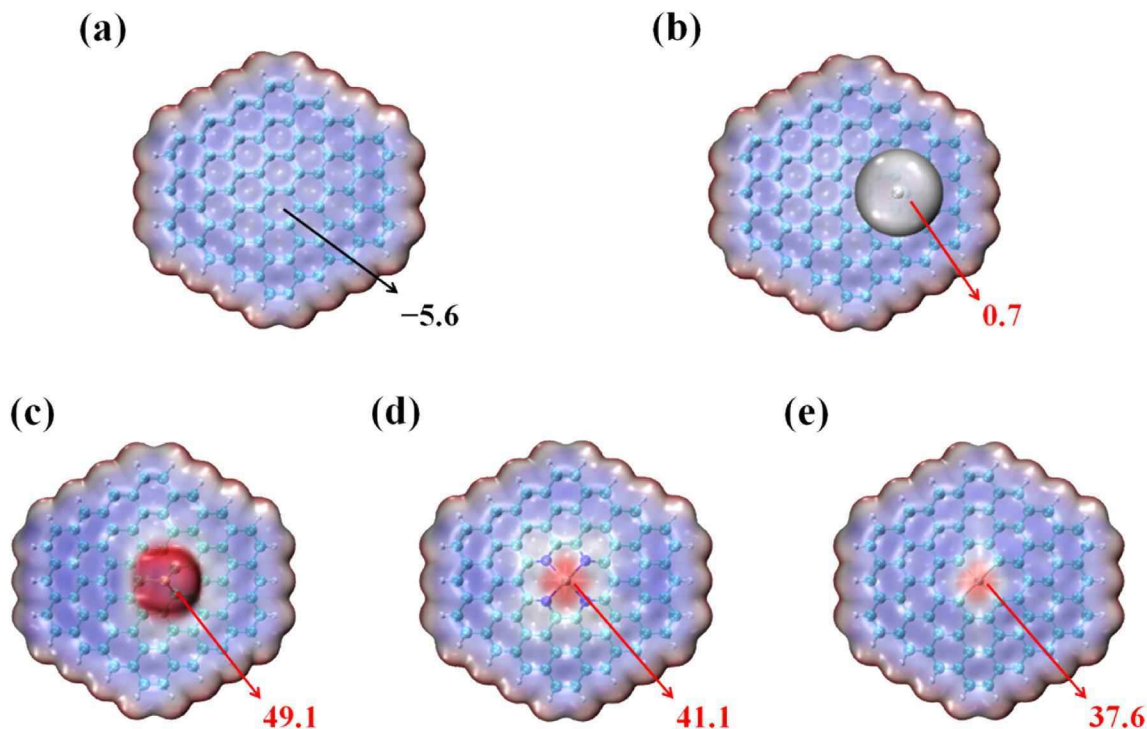


Fig. 3. Electrostatic potential (ESP) surfaces of (a) PG, (b) Zn-G1, (c) Zn-G2, (d) ZnN₄-G and (e) ZnC₄-G, respectively. The unit of the data is kcal/mol.

summarized in Table 3. The larger $\rho(r)$ value represents the higher interaction strength [54]. Zn-G2/TMA, ZnN₄-G/TMA and ZnC₄-G/TMA complexes exhibit larger $\rho(r)_{N...H}$ values compared with $\rho(r)_{H...C}$ of PG/TMA and Zn-G1/TMA complexes, implying remarkable interaction between the Zn atom of substrates and the N atom of TMA molecule. In addition, Zn-G2/TMA complex has another larger $\rho(r)_{H...C}$ value, which leads to its largest adsorption energy among all studied adsorption systems. The positive $\nabla^2\rho(r)$ is associated to ionic bond. The value of $G(r)/|V(r)|$ is less than 0.5, the interaction can be considered as completely covalent, while value greater than 1 represents the purely non-covalent nature of interaction. As shown in Table 3, The $G(r)/|V(r)|$ of N ... Zn in Zn-G2/TMA, ZnN₄-G/TMA and ZnC₄-G/TMA complexes are in the range of 0.5–1 a.u., and positive values for $\nabla^2\rho(r)$ are also found, indicating these interactions include partially covalent characters. For PG/TMA and Zn-G1/TMA complexes, the values of $G(r)/|V(r)|$ are larger than 1 a.u., suggesting purely non-covalent nature of interaction. The structural stability of the adsorption systems can be evaluated through ϵ , the smaller values indicate interaction is stable. The more stability of Zn-G2/TMA, ZnN₄-G/TMA and ZnC₄-G/TMA complexes relative to PG/TMA and Zn-G1/TMA complexes can be inferred from smaller ϵ values.

The electron properties of five carbon substrates upon TMA molecule adsorption are also calculated to further analyze the sensing mechanism, and the results are tabulated in Table 4. The sensitivity of materials is closely related to their conductivity, which is defined as:

$$\sigma \propto \exp\left(\frac{-E_g}{2K_B T}\right) \quad (5)$$

where the σ is the electric conductivity, E_g is the energy gap, K_B is the Boltzmann's constant and T is Kelvin temperature. The obvious variation of electrical conductivity before and after molecule adsorption means higher sensitivity. As shown in Table 4, the energy gap of PG and Zn-G1 remains almost unchanged after TMA adsorption, implying that TMA has no influence on the resistivity of PG and Zn-G1, thus PG and Zn-G1 have poor sensitivity for TMA. Zn-G2 exhibits significant energy gap changes (28.87%) after TMA adsorption, followed by the ZnC₄-G

(10.44%) and ZnN₄-G (8.33%), respectively, which can induce larger conductivity variation. Therefore, Zn-G2, ZnC₄-G and ZnN₄-G have better sensitivity for TMA. Work function (Φ) is another important parameter for analyzing sensitivity of a material, especially for the Φ -type sensor. Gas adsorption on sensitive materials can induce changes in work function, the work function of pristine and modified graphene before and after TMA adsorption is listed in Table 4. It can be noticed that Zn-G2 exhibits largest amount of change in work function (11.23 %) upon the adsorption process, allowing it to be a promising Φ -type sensor.

3.3. Selectivity of Zn-G2, ZnN₄-G and ZnC₄-G

Selectivity is also an important indicator for evaluating gas sensors, the atmosphere environment contains few other elements, such as N₂, O₂, and H₂O, which may affect the sensing properties of the target gas. Based on the above comprehensive analysis, Zn-G2, ZnN₄-G and ZnC₄-G exhibit excellent sensitivity to TMA. Thus, the adsorption characteristics of N₂, O₂, and H₂O on three substrate surfaces have been examined to evaluate the selectivity of TMA. The stable adsorption configuration and corresponding adsorption parameters are displayed in Fig. 6. The adsorption energies of N₂, O₂, and H₂O on three substrate surfaces are smaller than that of TMA adsorption systems, implying that TMA molecule is preferentially adsorbed on the Zn-G2, ZnN₄-G and ZnC₄-G in the presence of N₂, O₂, and H₂O molecule. Therefore, Zn-G2, ZnN₄-G and ZnC₄-G have a high selectivity to TMA molecule.

3.4. Electric field effect

Previous studies have been reported that the applied electric field can significantly influence the gas-sensing performance of graphene based sensors. Therefore, adsorption energy and natural bond orbital charge of TMA molecule on Zn-G2, ZnN₄-G and ZnC₄-G under the applied electric field are discussed, as shown in Fig. 7. The applied electric field is perpendicular to the substrate with the downward (upward) direction is defined as the negative (positive) electric field (the inset in Fig. 7(b)). From Fig. 7 (a), it can be seen that the adsorption

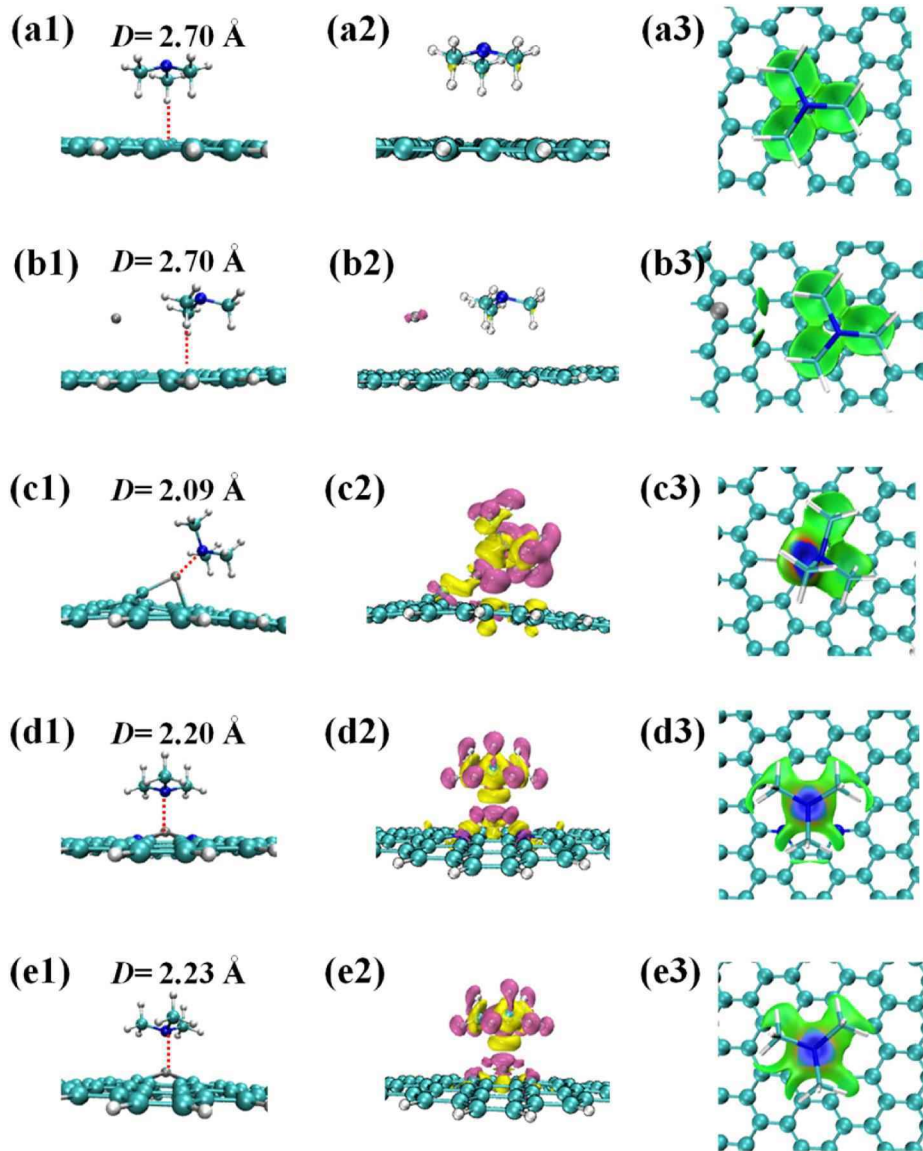


Fig. 4. Adsorption configuration (left column), CDD maps (middle column) and IGMH (right column) of (a1-a3) PG/TMA, (b1-b3) Zn-G1/TMA, (c1-c3) Zn-G2/TMA, (d1-d3) ZnN₄-G/TMA and (e1-e3) ZnC₄-G/TMA complexes, respectively. *D* represents the adsorption distance between Zn atom (or plane) and adsorbed atom (red dashed line). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 2

The adsorption energy (E_{ads}), charge transfer (Q) and interaction distance (D) of TMA molecules on different substrates.

| Substrate | E_{ads} (eV) | Q (e) | D (Å) | Reference |
|------------------------------------|-----------------------|---------|--------------|-----------|
| Fluorographene | −0.27 | 0.010 | 2.54 | [47] |
| Covalent graphene | −0.04 to −0.76 | – | 1.57 to 2.75 | [48] |
| Graphene/carbene | −0.01 to −0.51 | – | 1.56 to 3.12 | [49] |
| SnS/SnS ₂ | −0.60 | – | 2.72 | [50] |
| Nb ₂ C(OH) ₂ | −1.19 | 0.070 | 2.56 | [51] |
| Stanene nanotube | −0.35 | 0.049 | – | [52] |
| PG/TMA | −0.06 | −0.002 | 2.70 | This work |
| Zn-G1/TMA | −0.29 | 0.002 | 2.70 | This work |
| Zn-G2/TMA | −1.84 | 0.135 | 2.09 | This work |
| ZnN ₄ -G/TMA | −1.16 | 0.107 | 2.20 | This work |
| ZnC ₄ -G/TMA | −1.08 | 0.138 | 2.23 | This work |

energy decreases as the applied electric field changes from −0.008 to +0.008 a.u. This indicates that the applied positive electric field can enhance the interaction between TMA and substrate, while TMA can be desorbed from the substrate by applying negative electric field. There is

obvious linear relationship between natural bond orbital charge and electric field (Fig. 7(b)), more electrons would be drawn from TMA molecule to substrate under a increasing positive electric field, while the applied negative electric field can weaken electron transfer between them. Therefore, the applied positive electric field can improve gas sensing performance of Zn-G2, ZnN₄-G and ZnC₄-G to TMA molecule.

3.5. Recovery time

Reusability is another important index to evaluate gas sensors. In order to systematically study the feasibility of Zn-G2, ZnN₄-G and ZnC₄-G as reusable material for TMA sensing, the recovery time (τ) is calculated by following equation:

$$\tau = v_0^{-1} \exp\left(-\frac{E_{\text{ads}}}{k_B T}\right) \quad (6)$$

where v_0 is attempt frequency (10^{12} s^{-1}). The above equation implies that more negative adsorption energy would result in a longer recovery time. Importantly, a shorter recovery time can be achieved by heating at

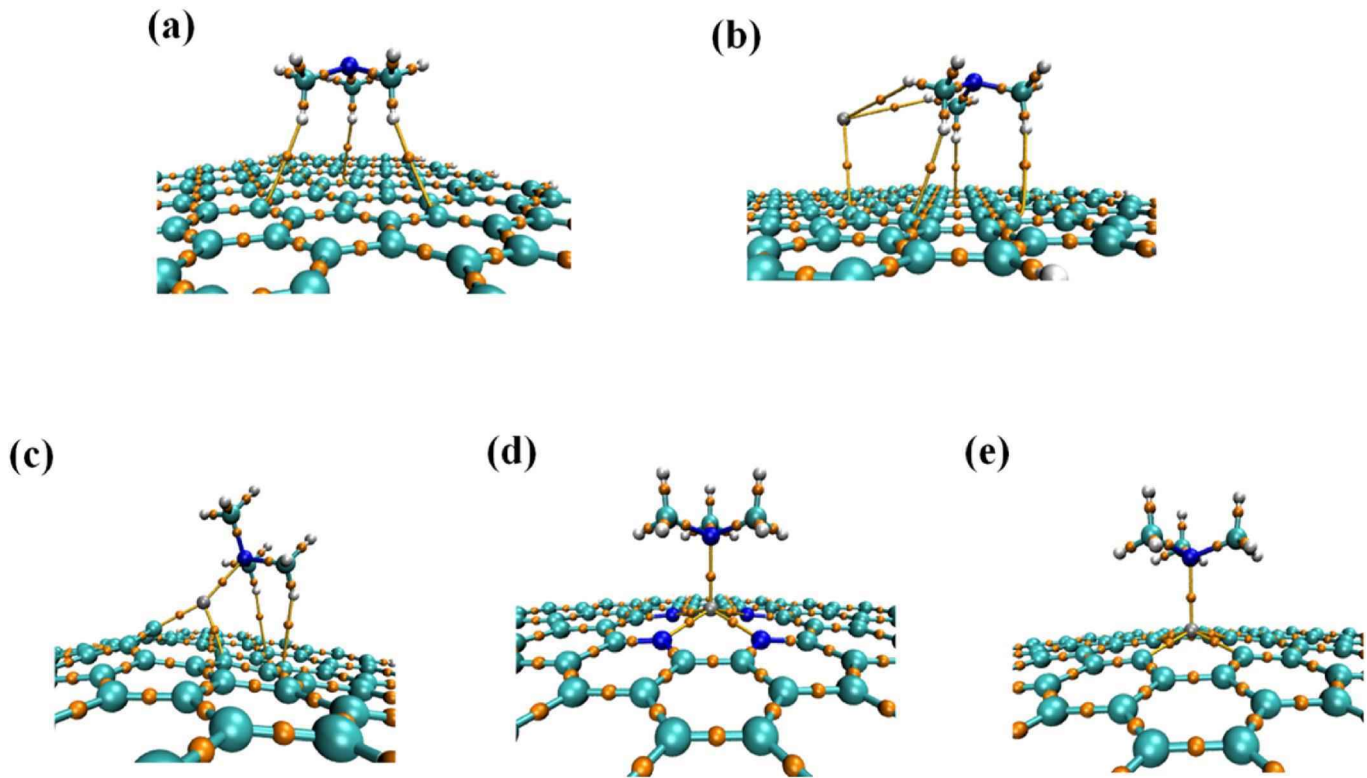


Fig. 5. The bond critical points (BCPs) and bond paths for the adsorption of TMA on the (a) PG, (b) Zn-G1, (c) Zn-G2, (d) ZnN₄-G and (e) ZnC₄-G, respectively.

Table 3
Topological parameters for studied TMA adsorption systems at BCP (3, −1). The unit of the data is a.u.

| Gas | Bond | $\rho(r)$ | $\nabla^2\rho(r)$ | $G(r)$ | $V(r)$ | $G(r)/ V(r) $ | λ_1 | λ_2 | λ_3 | ε |
|---------------------|----------|-----------|-------------------|--------|---------|---------------|-------------|-------------|-------------|---------------|
| PG | H ... C | 0.0063 | 0.0189 | 0.0040 | −0.0033 | 1.2121 | −0.0037 | −0.0004 | 0.0231 | 9.2500 |
| Zn-G1 | H ... C | 0.0066 | 0.0193 | 0.0041 | −0.0033 | 1.2424 | −0.0042 | −0.0013 | 0.0249 | 3.2307 |
| | H ... Zn | 0.0038 | 0.0072 | 0.0015 | −0.0012 | 1.2500 | −0.0019 | −0.0018 | 0.0110 | 1.0555 |
| Zn-G2 | N ... Zn | 0.0693 | 0.2879 | 0.0691 | −0.0846 | 0.8167 | −0.0874 | −0.0861 | 0.4615 | 1.0150 |
| | H ... C | 0.0100 | 0.0300 | 0.0063 | −0.0052 | 1.2115 | −0.0076 | −0.0055 | 0.0432 | 1.3818 |
| ZnN ₄ -G | N ... Zn | 0.0545 | 0.2090 | 0.0536 | −0.0661 | 0.8108 | −0.0627 | −0.0622 | 0.3339 | 1.0080 |
| ZnC ₄ -G | N ... Zn | 0.0505 | 0.1880 | 0.0485 | −0.0598 | 0.8110 | −0.0554 | −0.0549 | 0.2984 | 1.0091 |

Table 4
Electronic properties of the investigated systems. All parameters are in eV.

| Systems | E_H | E_L | E_g | % E_g | Φ | % $\Delta\Phi$ |
|-------------------------|-------|-------|-------|---------|--------|----------------|
| PG | −4.75 | −2.90 | 1.85 | — | 3.83 | — |
| PG/TMA | −4.77 | −2.93 | 1.84 | 0.54 | 3.85 | 0.52 |
| Zn-G1 | −4.78 | −2.93 | 1.85 | — | 3.86 | — |
| Zn-G1/TMA | −4.78 | −2.95 | 1.83 | 1.08 | 3.87 | 0.25 |
| Zn-G2 | −4.64 | −3.22 | 1.42 | — | 3.93 | — |
| Zn-G2/TMA | −4.47 | −2.64 | 1.83 | 28.87 | 3.56 | 11.23 |
| ZnN ₄ -G | −4.29 | −2.97 | 1.32 | — | 3.63 | — |
| ZnN ₄ -G/TMA | −4.06 | −2.85 | 1.21 | 8.33 | 3.46 | 4.68 |
| ZnC ₄ -G | −4.70 | −3.36 | 1.34 | — | 4.03 | — |
| ZnC ₄ -G/TMA | −4.58 | −3.10 | 1.48 | 10.44 | 3.84 | 4.71 |

a high temperature. The calculated recovery time of TMA molecule at three experimental temperatures (298, 398 and 498 K) is calculated and plotted in Fig. 8. Zn-G2, ZnN₄-G and ZnC₄-G have longer recovery time at $T = 298$ K due to their strong interaction with TMA molecule. When the temperature is elevated to 418 K, the recovery time decreases to only about 269 and 6 s for ZnN₄-G and ZnC₄-G, respectively, confirming the potential of the ZnN₄-G and ZnC₄-G to be used as a reusable gas sensor at high temperatures. However, TMA molecule can hardly be desorbed from the Zn-G2 due to its stronger adsorption energy, the recovery time

is 1.2×10^6 even if the temperature reaches 498 K. It is worth noting that the applied negative electric field can decrease adsorption strength between TMA and Zn-G2 (Fig. 7 (a)), which is beneficial for achieving desorption effects. When the conditional temperature reaches 498 K, the desorption of TMA molecule from the Zn-G2 only takes 91 s by applying negative electric field of 8×10^6 a.u. This suggests that applying electric field and increasing temperature is a feasible strategy to achieve gas desorption.

4. Conclusion

In the present work, the adsorption and sensing properties of TMA molecule on pristine, Zn-decorated, Zn-doped, ZnN₄-embedded, Zn-defect combined graphene have been systematically investigated by DFT calculation. The sensing mechanism including stability, sensitivity, selectivity, recovery time, and electric field effect, etc. are discussed, and the core conclusions can be summarized as follow.

- (1) From the cohesive energy and binding energy analysis, it can be noted that Zn-G2, ZnN₄-G, and ZnC₄-G exhibit good structural stability.
- (2) Shorter adsorption distances, larger amount of charge transfer and stronger adsorption strengths demonstrated that Zn-G2,

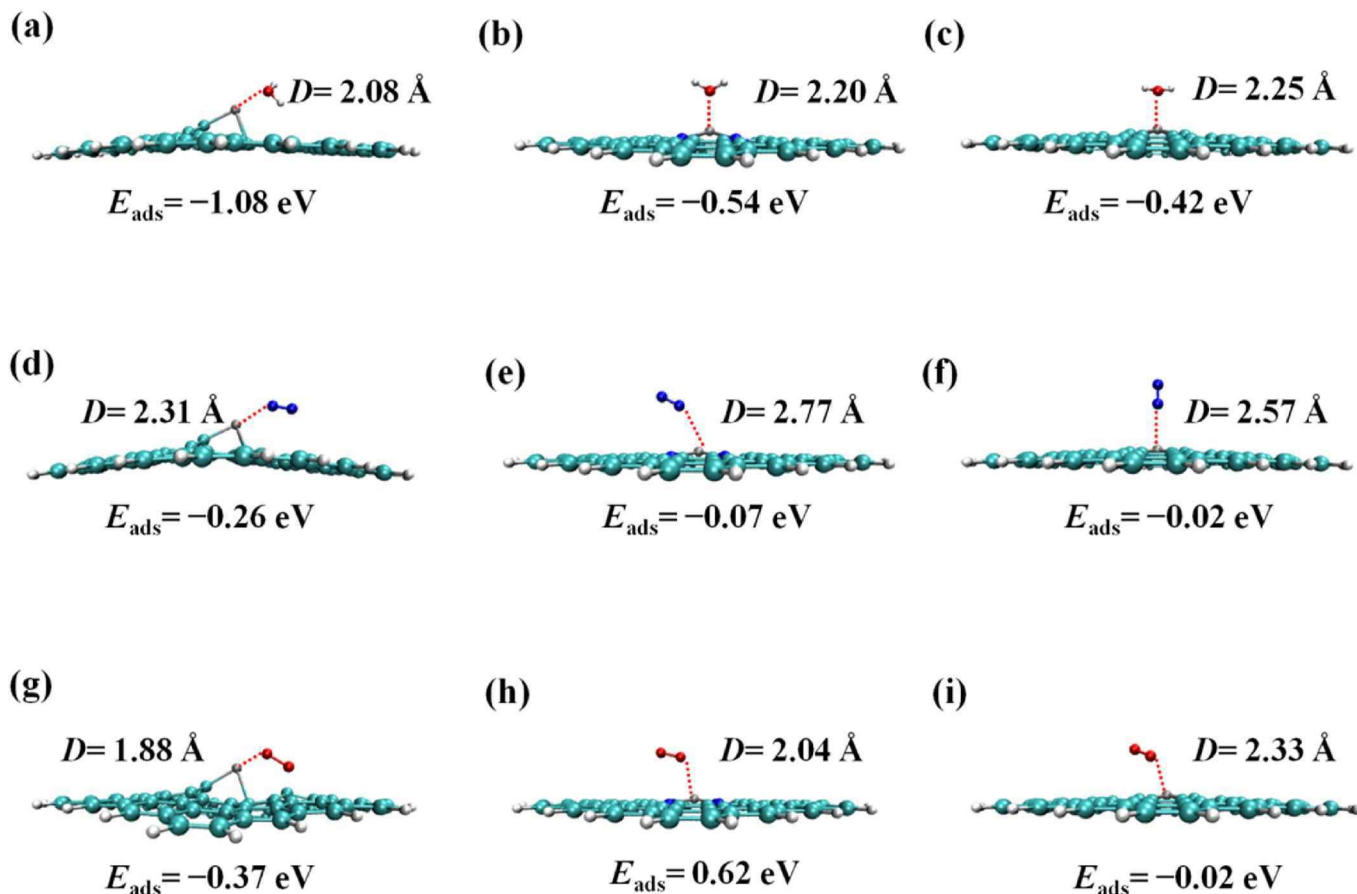


Fig. 6. The stable adsorption configuration of (a) Zn-G2/H₂O, (b) ZnN₄-G/H₂O, (c) ZnC₄-G/H₂O, (d) Zn-G2/N₂, (e) ZnN₄-G/N₂, (f) ZnC₄-G/N₂, (g) Zn-G2/O₂, (h) ZnN₄-G/O₂, (i) ZnC₄-G/O₂, respectively. D represents the adsorption distance between Zn atom and adsorbed atom (red dashed line). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

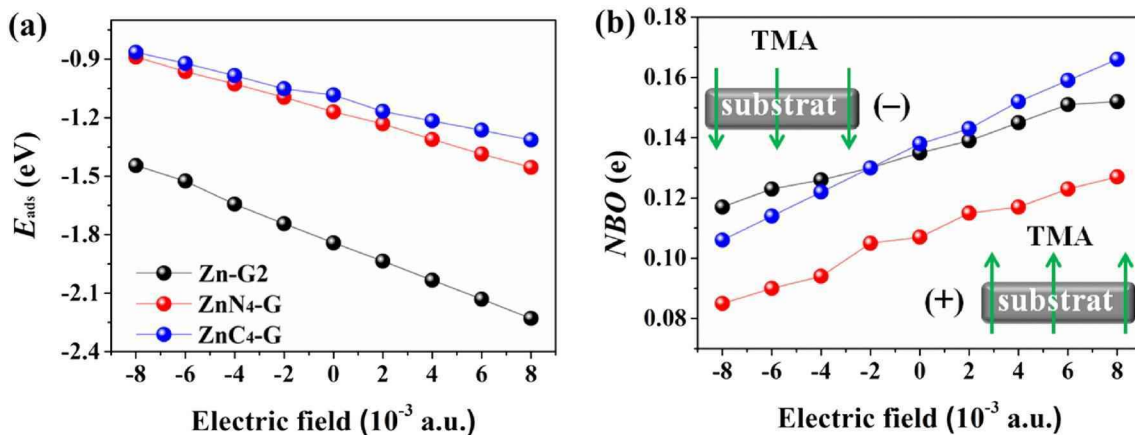


Fig. 7. (a) E_{ads} and (b) NBO charge of TMA adsorption systems under the applied electric field.

ZnN₄-G, and ZnC₄-G are more sensitive to the TMA than the PG and Zn-G1. Especially for Zn-G2, the significant changes in energy gap and work function after TMA adsorption.

(3) Zn-G2, ZnN₄-G, and ZnC₄-G exhibit high selectivity to TMA in the presence of N₂, O₂, and H₂O molecule due to more negative adsorption energy.

(4) Analysis of recovery time suggested that the ZnN₄-G, and ZnC₄-G may utilize as a reversible gas sensor at high temperatures. For Zn-G2, TMA can effectively desorb by applying negative electric field and increasing temperature, thus achieving multi-time uses.

(5) Furthermore, adsorption energy and charge transfer of Zn-G2, ZnN₄-G, and ZnC₄-G could be further improved by applying positive electric field, make their higher sensitivity and selectivity.

In summary, the calculated results show that Zn-G2, ZnN₄-G, and ZnC₄-G could be promising sensing materials to detect TMA, which would be helpful for developing high selectivity and sensitivity graphene-based gas sensor.

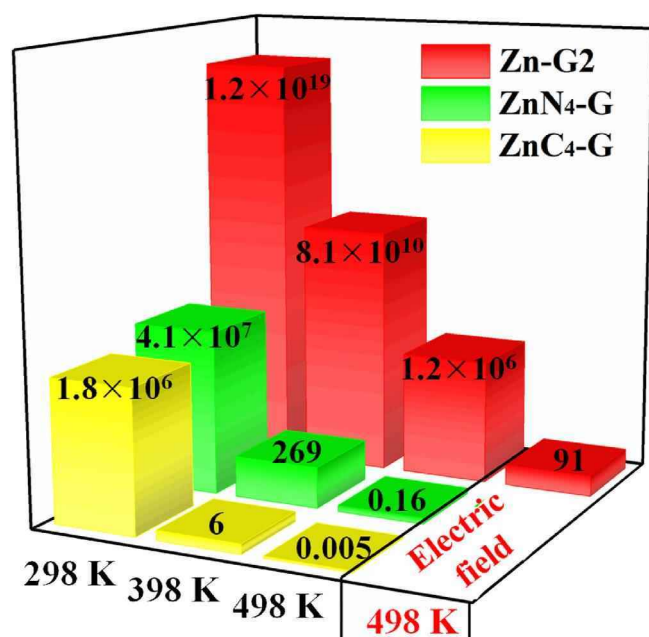


Fig. 8. The recovery time (unit: s) of adsorption systems at different temperatures.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

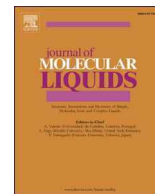
This research is supported by the National Natural Science Foundation of China (22106025), the Basic and Applied Basic Research Foundation of Guangzhou, China (202201010541).

References

- [1] Rappert S, Müller R. Odor compounds in waste gas emissions from agricultural operations and food industries. *Waste Manag* 2005;25:887–907.
- [2] Rianjanu A, Aflaha R, Khamidy NI, Djamel M, Triyana K, Wasisto HS. Room-temperature ppb-level trimethylamine gas sensors functionalized with citric acid-doped polyvinyl acetate nanofibrous mats. *Materials Advances* 2021;2:3705–14.
- [3] Yan W, Xu H, Ling M, Zhou S, Qiu T, Deng Y, et al. MOF-derived porous hollow $\text{Co}_3\text{O}_4/\text{ZnO}$ cages for high-performance MEMS trimethylamine sensors. *ACS Sens* 2021;6:2613–21.
- [4] Zhang F, Dong X, Cheng X, Xu Y, Zhang X, Huo L. Enhanced gas-sensing properties for trimethylamine at low temperature based on $\text{MoO}_3/\text{Bi}_2\text{Mo}_3\text{O}_{12}$ hollow microspheres. *ACS Appl Mater Interfaces* 2019;11:11755–62.
- [5] Chang J, Deng Z, Li M, Wang S, Mi L, Sun Q, et al. Visible light boosting hydrophobic $\text{ZnO}/(\text{Sr}_{0.6}\text{Bi}_{0.305})_2\text{Bi}_2\text{O}_7$ chemiresistor toward ambient trimethylamine. *Sensor Actuator B Chem* 2022;352:131076.
- [6] Li X, Jin L, Ni A, Zhang L, He L, Gao H, et al. Tough and antifreezing MXene@Au hydrogel for low-temperature trimethylamine gas sensing. *ACS Appl Mater Interfaces* 2022;14:30182–91.
- [7] Liang L, Wang J, Lin W, Sumpter BG, Meunier V, Pan M. Electronic bandgap and edge reconstruction in phosphorene materials. *Nano Lett* 2014;14:6400–6.
- [8] Fan Y-Y, Tu H-L, Pang Y, Wei F, Zhao H-B, Yang Y, et al. Au-decorated porous structure graphene with enhanced sensing performance for low-concentration NO_2 detection. *Rare Met* 2020;39:651–8.
- [9] Odey DO, Edet HO, Louis H, Gber TE, Nwagwu AD, Adalikwu SA, et al. Heteroatoms (B, N, and P) doped on nickel-doped graphene for phosgene (COCl_2) adsorption: insight from theoretical calculations. *Mater Today Sustain* 2023;21:100294.
- [10] Zhou Q, Ju W, Su X, Yong Y, Li X. Adsorption behavior of SO_2 on vacancy-defected graphene: a DFT study. *J Phys Chem Solid* 2017;109:40–5.
- [11] Bo Z, Guo X, Wei X, Yang H, Yan J, Cen K. Density functional theory calculations of NO_2 and H_2S adsorption on the group 10 transition metal (Ni, Pd and Pt) decorated graphene. *Phys E Low-dimens Syst Nanostruct* 2019;109:156–63.
- [12] Guo X, Yang H, Zhou M, Wei X, Bo Z, Yan J, et al. Tuning and monitoring of nitrogen dioxide fixation on Cu decorated graphene: a density functional theory study. *J Phys Condens Matter* 2020;32:355001.

- [13] Qu Y, Ding J, Chen H, Hu W, Fan H. The effect of both Pt decoration and the defects on the adsorption of graphene for SO_2 . *Int J Quant Chem* 2022;122:e26888.
- [14] Impeng S, Junkaew A, Maitarad P, Kungwan N, Zhang D, Shi L, et al. A MnN_4 moiety embedded graphene as a magnetic gas sensor for CO detection: a first principle study. *Appl Surf Sci* 2019;473:820–7.
- [15] Khosravi A, Vessally E, Oftadeh M, Behjatmanesh-Ardakani R. Ammonia capture by MN_4 ($\text{M} = \text{Fe}$ and Ni) clusters embedded in graphene. *J Coord Chem* 2018;71:3476–86.
- [16] Luo M, Liang Z, Gouse Peera S, Chen M, Liu C, Yang H, et al. Theoretical study on the adsorption and predictive catalysis of MnN_4 embedded in carbon substrate for gas molecules. *Appl Surf Sci* 2020;525:146480.
- [17] Cai Y, Luo X. First-principles investigation of carbon dioxide adsorption on MN_4 doped graphene. *AIP Adv* 2020;10:125013.
- [18] Ayesh AI. DFT investigation of H_2S and SO_2 adsorption on Zn modified MoSe_2 . *Superlattice Microst* 2022;162:107098.
- [19] Lontio Fomekong R, Tedjieukeng Kamta HM, Ngolui Lambi J, Lahem D, Eloy P, Debliquy M, et al. A sub-ppm level formaldehyde gas sensor based on Zn-doped NiO prepared by a co-precipitation route. *J Alloys Compd* 2018;731:1188–96.
- [20] Hussain S, Chatha SAS, Hussain AI, Hussain R, Yasir Mehboob M, Mansha A, et al. A theoretical framework of zinc-decorated inorganic $\text{Mg}_{12}\text{O}_{12}$ nanoclusters for efficient COCl_2 adsorption: a step forward toward the development of COCl_2 sensing materials. *ACS Omega* 2021;6:19435–44.
- [21] Hassan K, Hossain R, Sahajwalla V. Novel microrecycled ZnO nanoparticles decorated macroporous 3D graphene hybrid aerogel for efficient detection of NO_2 at room temperature. *Sensor Actuator B Chem* 2021;330:129278.
- [22] Syaahiran MA, Mahadi AH, Lim CM, Kooh MRR, Chau Y-FC, Chiang H-P, et al. Theoretical study of CO adsorption interactions with Cr-doped tungsten oxide/graphene composites for gas sensor application. *ACS Omega* 2022;7:528–39.
- [23] Demir S, Fellah MF. A DFT study on Pt doped (4,0) SWCNT: CO adsorption and sensing. *Appl Surf Sci* 2020;504:144141.
- [24] Sadia H, Ullah S, Ullah F, Jadoon T. DFT study about capturing of toxic sulfur gases over cyclic tetrapyrrole. *Computational and Theoretical Chemistry* 2023;1219:113966.
- [25] Asif M, Sajid H, Ullah F, Khan S, Ayub K, Amjad Gilani M, et al. Quantum chemical study on sensing of NH_3 , NF_3 , NCl_3 and NBr_3 by using cyclic tetrapyrrole. *Computational and Theoretical Chemistry* 2021;1199:113221.
- [26] Yuksel N, Kose A, Fellah MF, Pd, Ag and Rh doped (8,0) single-walled carbon nanotubes (SWCNTs): a DFT study on furan adsorption and detection. *Surf Sci* 2022;715:121939.
- [27] Sajid H, Khan S, Ayub K, Amjad Gilani M, Mahmood T, Farooq U, et al. Ab initio study for superior sensitivity of graphyne nanoflake towards nitrogen halides over ammonia. *J Mol Model* 2022;28:161.
- [28] Lu T, Chen F. Multiwfn: a multifunctional wavefunction analyzer. *J Comput Chem* 2012;33:580–92.
- [29] Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, et al. Gaussian 16 rev. C.01. 2016. Wallingford, CT.
- [30] He G, He H. DFT studies on the heterogeneous oxidation of SO_2 by oxygen functional groups on graphene. *Phys Chem Chem Phys* 2016;18:31691–7.
- [31] Chen JL, Zhang RQ. Strong interaction between cyclo[18]Carbon and graphene. *Advanced Theory and Simulations* 2021;4:2100022.
- [32] Mudeda SK, Balamurugan K, Kamaraj M, Subramanian V. Interaction of nucleobases with silicon doped and defective silicon doped graphene and optical properties. *Phys Chem Chem Phys* 2016;18:295–309.
- [33] Tang Y, Chen W, Li C, Pan L, Dai X, Ma D. Adsorption behavior of Co anchored on graphene sheets toward NO , SO_2 , NH_3 , CO and HCN molecules. *Appl Surf Sci* 2015;342:191–9.
- [34] Luo H, Zhang L, Xu S, Shi M, Wu W, Zhang K. NH_3 , PH_3 and AsH_3 adsorption on alkaline earth metal (Be-Sr) doped graphenes: insights from DFT calculations. *Appl Surf Sci* 2021;537:147542.
- [35] Cordero B, Gómez V, Platero-Prats AE, Revés M, Echeverría J, Cremades E, et al. Covalent radii revisited. *Dalton Trans* 2008:2832–8.
- [36] Wu Y, Chen X, Weng K, Arramel, Jiang J, Ong W-J, et al. Highly sensitive and selective gas sensor using heteroatom doping graphdiyne: a DFT study. *Advanced Electronic Materials* 2021;7:2001244.
- [37] Li Z, Wang D, Li H, Ma M, Zhang Y, Yan Z, et al. Single-atom Zn for boosting supercapacitor performance. *Nano Res* 2022;15:1715–24.
- [38] Song K, Feng Y, Zhou X, Qin T, Zou X, Qi Y, et al. Exploiting the trade-offs of electron transfer in MOF-derived single Zn/Co atomic couples for performance-enhanced zinc-air battery. *Appl Catal B Environ* 2022;316:121591.
- [39] Tatrari G, Tewari C, Pathak M, Karakoti M, Bohra BS, Pandey S, et al. Bulk production of zinc doped reduced graphene oxide from tire waste for supercapacitor application: computation and experimental analysis. *J Energy Storage* 2022;53:105098.
- [40] Jogender, Mandeep, Rita K. Recent advances on graphene-based gas sensors. *Russ J Phys Chem A* 2020;94:2115–20.
- [41] Chen J, Jia L, Cui X, Zeng W, Zhou Q. Adsorption and gas-sensing properties of SF_6 decomposition components (SO_2 , SOF_2 and SO_2F_2) on Co or Cr modified GeSe monolayer: a DFT study. *Mater Today Chem* 2023;28:101382.
- [42] Wang A, Cui J, Zhang L, Liang L, Cao Y, Liu Q. Monitoring of COS , SO_2 , H_2S , and CS_2 gases by $\text{Al}_{12}\text{P}_{24}$ nanoclusters: a DFT inspection. *J Mol Model* 2023;29:98.
- [43] Dolmasev S, Yuksel N, Fellah MF, Au, Ag and Cu Doped BNNT for ethylene oxide gas detection: a density functional theory study. *Sensor Actuator Phys* 2023;350:114109.
- [44] Vessally E, Hosseinali M, Poor Heravi MR, Mohammadi B. DFT study of the adsorption of simple organic sulfur gases on $\text{g-C}_3\text{N}_4$: periodic and non-periodic approaches. *J Sulfur Chem* 2023;44:733–50.

- [45] Ema SN, Khaleque MA, Ghosh A, Piya AA, Habiba U, Shamim SUD. Surface adsorption of nitrosourea on pristine and doped (Al, Ga and In) boron nitride nanosheets as anticancer drug carriers: the DFT and COSMO insights. *RSC Adv* 2021;11:36866–83.
- [46] Li Y, Li X, Xu Y. The sensing mechanism of pristine and transition metals doped $\text{Zn}_{12}\text{O}_{12}$, $\text{Sn}_{12}\text{O}_{12}$ and $\text{Ni}_{12}\text{O}_{12}$ nanocages towards NH_3 and PH_3 : a DFT study. *J Mater Chem C* 2021;9:17382–91.
- [47] Rouhani M. Fluoro-functionalized graphene as a promising nanosensor in detection of fish spoilage: a theoretical study. *Chem Phys Lett* 2019;719:91–102.
- [48] Baachaoui S, Aldulaijan S, Raouafi F, Besbes R, Sementa L, Fortunelli A, et al. Pristine graphene covalent functionalization with aromatic aziridines and their application in the sensing of volatile amines – an ab initio investigation. *RSC Adv* 2021;11:7070–7.
- [49] Baachaoui S, Sementa L, Hajlaoui R, Aldulaijan S, Fortunelli A, Dhoub A, et al. Tailoring graphene functionalization with organic residues for selective sensing of nitrogenated compounds: structure and transport properties via QM simulations. *J Phys Chem C* 2023;127:15474–85.
- [50] Zhou Qa, Zheng C, Zhu L, Wang J. Tin sulfides heterostructure modified quartz crystal microbalance sensors with high sensitivity for hazardous trimethylamine gas. *Sensor Actuator B Chem* 2022;371:132520.
- [51] Vovusha H, Bae H, Lee S, Park J, Raza A, Kotmool K, et al. Density functional theory studies of MXene-based nanosensors for detecting volatile organic compounds in meat spoilage assessment. *ACS Appl Nano Mater* 2023;6:18592–601.
- [52] Bhuvaneswari R, Nagarajan V, Chandiramouli R. Adsorption studies of trimethyl amine and n-butyl amine vapors on stanene nanotube molecular device – a first-principles study. *Chem Phys* 2018;501:78–85.
- [53] Lu T, Chen Q. Independent gradient model based on Hirshfeld partition: a new method for visual study of interactions in chemical systems. *J Comput Chem* 2022;43:539–55.
- [54] Wang Y, Liu M, Li J, Wang Q, Ouyang X, Wei H, et al. Exploring competitive inhibition of a family 10 xylanase derived from Hu sheep rumen microbiota by *Oryza sativa* xylanase inhibitor protein: in vitro and in silico perspectives. *Enzym Microb Technol* 2022;160:110082.



MnN₄ embedded zeolite-templated carbon for methylamine and trimethylamine sensing: Insights from DFT study

Yuanchao Li^a, Cuijuan Jiang^c, Xiliang Yan^{a,b,*}

^a Institute of Environmental Research at Greater Bay Area, Key Laboratory for Water Quality and Conservation of the Pearl River Delta, Ministry of Education, Guangzhou University, Guangzhou 510006, China

^b College of Animal Science, South China Agricultural University, Guangzhou 510642, China

^c School of Environmental Science and Engineering, Shandong University, Qingdao 266237, China

ARTICLE INFO

Keywords:

Methylamine
Trimethylamine
Zeolite templated carbon
Sensing mechanism
Density functional theory

ABSTRACT

From the perspective of raw material losses and human health, monitoring fish-based food quality has become more prominent. As a nondestructive technology, gas sensors have received extensive attention for evaluating the freshness of fish. In this study, the sensing and electronic properties of methylamine (MA) and trimethylamine (TMA) molecule adsorption on MnN₄ embedded zeolite templated carbon (MnN₄-C) were studied by density functional theory (DFT) method. The results show that the MnN₄-C exhibits good structural stability, and displays high sensitivity toward MA and TMA gases by analysis of the interaction distance, adsorption energy, charge transfer and sensing response. It is worth noting that the ambient humidity has demonstrated no effect on the gas-sensing properties of MnN₄-C for the studied gases. Meanwhile, applying a positive electric field can increase the adsorption energy and charge transfer, which enhances the MA and TMA gas sensitivity of MnN₄-C. The recovery time reveals that MnN₄-C can be an excellent reusability sensor for MA and TMA gases. This study provides theoretical direction for exploring the sensing application of MnN₄-C in evaluating freshness of fish.

1. Introduction

Fish-based food is widely favored in the human diet due to its rich protein, fatty acids and vitamins [1]. However, fish products are extremely perishable during processing, transportation and storage, and produce offensive off-flavors. Fish spoilage will bring substantial economic losses to the fishermen and the market. In addition, consuming seafood with harmful bacterial contaminants can endanger human health, causing diarrhea, abdominal pain, vomiting, urticaria, and fever [2]. For these reasons, it is necessary to monitor food quality to ensure health and food safety. Freshness is one of the crucial indexes reflecting food quality. In fact, human senses often fail to evaluate fish freshness by simply looking, smelling and touching, especially for packaged fish [3]. The survey results of this method will inevitably change greatly due to the external environment, human factors and other reasons. Thus, developing an effective technology to monitor food quality is an urgent concern for the fish industry and consumers.

It is known that the spoilage of fish can release total volatile basic amines (TVB-A) such as methylamine (MA) and trimethylamine (TMA),

which can be employed as an important indicator to evaluate the freshness of fish. In the recent decades, various methods such as gas chromatography–mass spectrometry and chemiluminescence, have been employed to detect volatile amines [4]. However, these methods require complicated operation, a long time for sample preprocessing and expensive analytical equipment. Recently, gas sensors for detecting and evaluating fish freshness have attracted great attention due to their high sensitivity, low-cost, portability, nondestructive nature and lack of need for sample preparation. Liu *et al.* developed TMA sensor based α -Fe₂O₃ nanoparticles, showing good selectivity and high sensitivity with a response value for 100 ppm TMA gas of 27.8 at 250 °C [5]. Shen *et al.* reported that Au@Pt/ α -Fe₂O₃ hollow nanocubes exhibited a faster response time (5 s) and higher response ($R_a/R_g = 32$) toward 100 ppm TMA gas at 150 °C [6]. Zhao *et al.* prepared a Au/WO₃ nanosheet gas sensor to detect TMA at 300 °C. It exhibited rapid response-recovery time (8 s/6 s) and low detection limit (0.5 ppm) [7]. Numerous experimental studies also show that metal oxide semiconductor based sensors exhibit good sensitivity to TVB-A [8–10]. However, the higher operating temperature limits their multiscene application due to the potential

* Corresponding author at: Institute of Environmental Research at Greater Bay Area, Key Laboratory for Water Quality and Conservation of the Pearl River Delta, Ministry of Education, Guangzhou University, Guangzhou 510006, China.

E-mail address: yanxiliang1991@gzhu.edu.cn (X. Yan).

<https://doi.org/10.1016/j.molliq.2024.124090>

Received 17 September 2023; Received in revised form 14 January 2024; Accepted 16 January 2024

Available online 24 January 2024

0167-7322/© 2024 Elsevier B.V. All rights reserved.

safety hazards, which has triggered the development of sensitive materials that can efficiently detect TVB-A at room temperature.

Previous theoretical studies have shown that single atom doped carbon-based materials such as graphdiyne [11], C₁₈ nanocluster [12], carbon nanotube [13,14], graphene quantum dots [15,16] and fullerene, [17–20] exhibit excellent sensitivity to various organic compounds at room temperature. Recently, a fascinating carbon material, four-nitrogen coordinated manganese (MnN₄) embedded graphene, has been successfully synthesized for catalytic applications [21]. Impeng *et al.* found that MnN₄ embedded graphene is a promising magnetic gas sensor for CO detection from a theoretical perspective [22]. Luo *et al.* reported that the adsorption properties of MnN₄ embedded graphene, carbon nanotubes and fullerene are mainly affected by the curvature of the substrate [23]. Sensitive materials with porous structures usually can promote gas diffusion and exhibit a higher response [24,25]. A novel zeolite templated carbon with microporous ordered nanostructures has been successfully synthesized [26] and can be used in hydrogen storage and supercapacitors [27,28]. Moradi *et al.* theoretically analyzed the potential of CoN₄ and ZnN₄ embedded zeolite templated carbon as an anode material for sodium-ion batteries [29]. However, the potential of zeolite templated carbon as a sensitive material for detecting TVB-A has not been reported.

In the present work, the gas-sensing properties of on MnN₄ embedded zeolite templated carbon (MnN₄-C) toward MA and TMA gases were investigated in detail by analyzing the geometry, adsorption energy, charge density differential, natural bond orbital charge, and recovery time. Furthermore, the effects of the electric field, relative humidity and temperature on the adsorption and desorption behaviors are also studied. The results show that MnN₄-C is a potential candidate material for monitoring fish spoilage at room temperature, which provides novel perspectives for experimentalists to design effective gas sensors.

2. Computational methods and models

The fully relaxed geometries of MA, TMA, MnN₄-C and the corresponding adsorption systems were optimized by applying the ωB97XD [30] method of density functional theory (DFT) with the 6-311G(d,p) [31] basis set for nonmetal (H, C and N) atoms and the LANL2TZ [32] basis set for metal (Mn) atom. The ωB97XD is a hybrid long-range separated empirical-corrected dispersion functional, which can accurately evaluate the covalent and non-covalent interactions between the gas and the substrate [33–35]. ωB97XD functional can also correctly reproduce the geometry observed in the experiment and provide reliable calculation results [12,36–38]. The current base set is large enough and well able to simulate molecular orbitals [39,40]. The frequency calculations have been carried out at the same level to confirm that the all structures are local minima on the potential energy surfaces. The independent gradient model based on Hirshfeld partition (IGMH) and electrostatic potential (ESP) surface were performed with the Multiwfn 3.8 code [41], and corresponding isosurface maps were rendered by means using VMD software [42] based on the cube files exported from Multiwfn. All calculations were performed using Gaussian 16C program [43].

To evaluate the stability of MnN₄ embedded C₃₉H₉, the formation energy (E_{form}) was calculated using the following equation:

$$E_{\text{form}} = (E_{\text{MnN}_4\text{-C}} - 9E_{\text{H}} - 33E_{\text{C}} - 4E_{\text{N}} - E_{\text{Mn}})/47 \quad (1)$$

in which E_{H} , E_{C} , E_{N} , E_{Mn} , and $E_{\text{MnN}_4\text{-C}}$ refer to the total energies of single H, C, N, Mn atoms and MnN₄-C, respectively.

The adsorption energy (E_{ads}) of the gas molecule on MnN₄-C was calculated using the following equation:

$$E_{\text{ads}} = E_{\text{Complex}} - (E_{\text{MnN}_4\text{-C}} + E_{\text{Gas}}) + E_{\text{BSSE}} \quad (2)$$

where E_{Complex} , $E_{\text{MnN}_4\text{-C}}$ and E_{Gas} are the total energy of the gas/MnN₄-C

adsorption system, MnN₄-C and isolated gas, respectively. E_{BSSE} is the basis set superposition error (BSSE) corrected for all adsorption energies. The energy values include zero point energy (ZPE) correction and thermal correction. In addition, thermal enthalpy (H_{ads}) and Gibbs free energy (G_{ads}) were also obtained upon adsorption process.

The amount of charge transfer between MnN₄-C and the gas adsorbate was evaluated through natural bond orbital (NBO) charge and Mulliken analysis. The electron density accumulation and depletion regions of the adsorption system can be observed by the charge density difference, which was obtained from the equation:

$$\Delta\rho = \rho_{\text{Complex}} - \rho_{\text{MnN}_4\text{-C}} - \rho_{\text{Gas}} \quad (3)$$

where ρ_{Complex} , $\rho_{\text{MnN}_4\text{-C}}$ and ρ_{Gas} are the charge densities of the adsorption system, substrate, and the isolated gas molecule, respectively.

The C₃₉H₉ carbon structure which includes three 5-membered rings and ten 6-membered rings is used as a stable unit cell of zeolite templated carbon and is widely used in theoretical research [27,29,44]. MnN₄ embedded C₃₉H₉ (MnN₄-C) is obtained by removing two neighboring carbon atoms of the central hexagonal ring, and then four C atoms around the defect site are replaced by N atoms. At the same time, the Mn atom incorporates in the center of the divacancy, as shown in Fig. 1. This strategy is widely used to construct a four-nitrogen coordinated transition-metal embedded carbon substrate [45–47].

3. Results and discussion

The optimized geometries of MA, TMA, and MnN₄-C are shown in Fig. 1, and the corresponding key parameters are also presented. MA and TMA exhibit similar lengths and angles, such as the C–N bond and ∠H–C–H. For MnN₄-C, curved surfaces result in different bond lengths between Mn and N atoms, and the corresponding values are 1.88, 1.97, 1.83 and 1.84 Å, respectively, implying that Mn–N forms a stable covalent bond. The stability of the substrate can be evaluated by the cohesive energy (E_{form}). The E_{form} of MnN₄-C is –7.32 eV, which is close to that of the pristine C₃₉H₉ carbon structure (–7.33 eV), suggesting its high structural stability. The structural stability can also be identified by the IR spectroscopy. Fig. S1 shows that there is no imaginary frequency for MnN₄-C, which indicates that MnN₄-C is local minima on the potential energy surfaces and is a dynamically stable structure.

Electrostatic potential (ESP) analysis can predict possible interaction sites by analyzing charge distribution characteristics [48,49]. Red, white and blue represent the positive, neutral and negative regions, respectively. Orange and ochre spheres refer to the local minima and maxima of the ESP, respectively. As depicted in Fig. 2, the most negative region of the MA molecule is located near the N atom with a global surface minimum of –41.2 kcal/mol and a local surface maximum of 24.1 kcal/mol over the H atom of –NH₂. The TMA molecule has strong negative and positive potentials near the N and H atoms, and the corresponding values are –33.6 and 9.7 kcal/mol, respectively. A strong positive potential region appears on the MnN₄-C surface with a local surface maximum of 47.3 kcal/mol over the Fe atom and a negative ESP value in the range of –15 to –19 kcal/mol, suggesting that the MnN₄-C substrate is conducive to adsorbing molecules with relatively negative potential. Hence, MA and TMA molecules with obvious negative potential can form a strong interaction with MnN₄-C through electrostatic attraction. In addition, the area distribution of different ESP intervals shows that the MnN₄-C substrate has uniform positive and negative potential areas.

Based on the ESP results, MA and TMA molecules are placed near the Mn atom of MnN₄-C, and the most stable adsorption configurations are graphically presented in Fig. 3. The adsorption parameters of MnN₄-C and other 2D materials toward MA and TMA molecules are listed in Table 1 and Table S1. The absence of negative frequency confirms that all adsorption systems are local minima on the potential energy surfaces (Fig. S1). The IR peaks of MnN₄-C were at 1218 cm^{–1}, 1519 cm^{–1} and

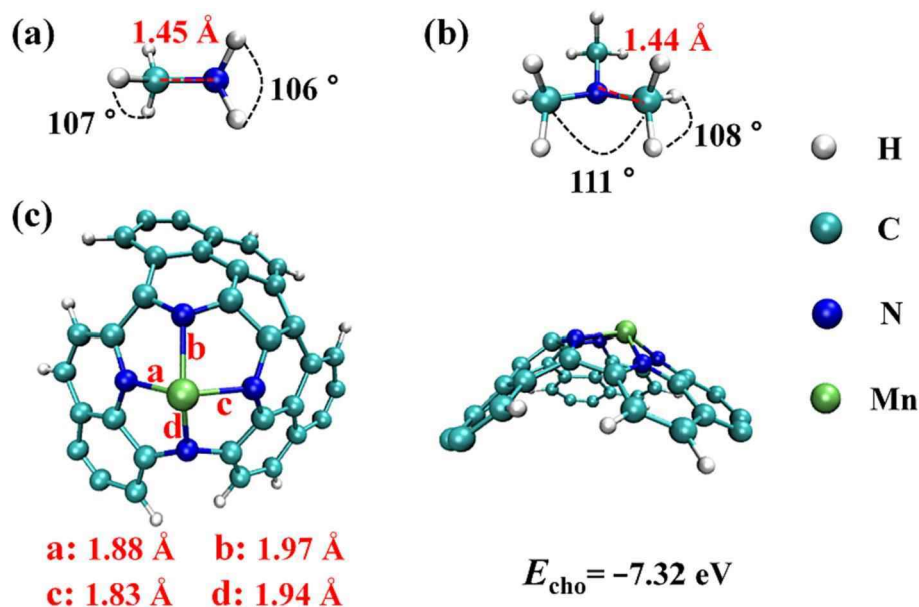


Fig. 1. The optimized structures of (a) MA, (b) TMA, and (c) $\text{MnN}_4\text{-C}$ (top and side views).

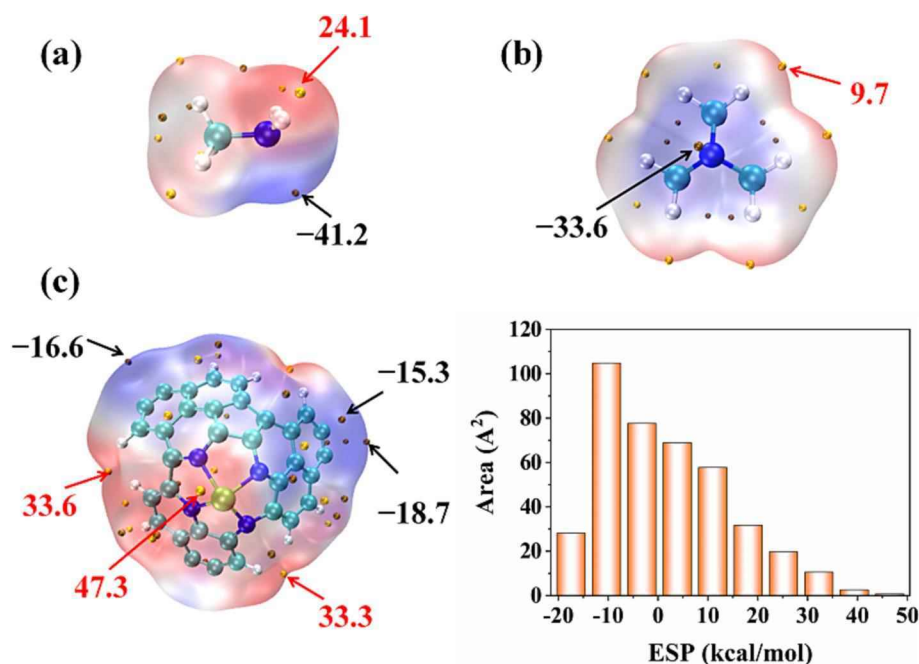


Fig. 2. Electrostatic potential (ESP) surfaces of (a) MA, (b) TMA, and (c) $\text{MnN}_4\text{-C}$. The area distribution of different ESP intervals for $\text{MnN}_4\text{-C}$ is plotted as histograms. The unit of the data is kcal/mol.

1568 cm^{-1} , respectively. After MA gas adsorption, there is no significant shift in the three main IR peaks, which are located at 1168 cm^{-1} , 1512 cm^{-1} and 1568 cm^{-1} , respectively. It is worth noting that some new IR peaks have appeared in the range of 3000 cm^{-1} to 3600 cm^{-1} . After TMA gas adsorption, it shifts to 1508 cm^{-1} , 1564 cm^{-1} and 1612 cm^{-1} , respectively. Importantly, there is new strong IR peak appeared at 3036 cm^{-1} .

As shown in Fig. 3, the N atom of MA and TMA is close to the Fe atom of $\text{MnN}_4\text{-C}$ with the nearest intermolecular distance (D) of 2.09 and 2.10 Å, respectively, which is shorter than that of other 2D materials [50–53]. This means that MA and TMA are more favorable to adsorb on the $\text{MnN}_4\text{-C}$ surface. As shown in shown in Table S1, the adsorption energy (E_{ads}) of $\text{MnN}_4\text{-C}/\text{MA}$ was -1.41 eV with enthalpy (H_{ads}) and free adsorption

energy (G_{ads}) values of -1.43 and -0.93 eV , respectively. The adsorption of TMA on the $\text{MnN}_4\text{-C}$ exhibited an E_{ads} value of -1.47 eV , possessing individual H_{ads} and G_{ads} values of -1.49 and -0.97 eV , respectively. The negative values suggest that the adsorption process of MA and TMA on the $\text{MnN}_4\text{-C}$ is favorable and spontaneous. The high adsorption energy of MA and TMA on the $\text{MnN}_4\text{-C}$ can be related to chemisorptions, which is much higher than that of the other 2D materials (see Table 1), such as covalent graphene (-0.09 to -0.57 for MA, -0.04 to -0.76 for TMA) [50], $\text{WS}_2/\text{MWCNTs}$ (-0.35 for MA, -0.59 for TMA) [51], fluorographene (-0.33 for MA, -0.27 for TMA) [52], implying that $\text{MnN}_4\text{-C}$ has high sensitivity toward MA and TMA gases. Generally, fish food-based are accompanied by high humidity. According to experimental reports, the actual sensitive performance of the gas

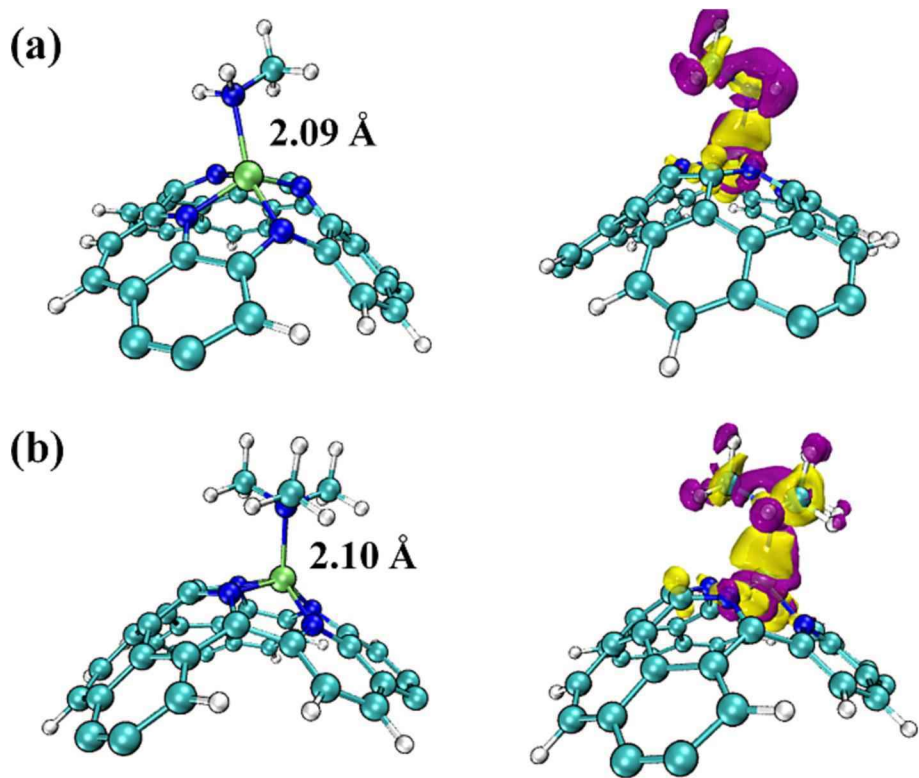


Fig. 3. Adsorption configuration (left) charge density difference (right) plots of (a) MA and (b) TMA adsorption on MnN₄-C. Yellow and purple represent electron accumulation and depletion, respectively. The isosurface value is set to 0.002 e/Å³. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
The adsorption energy (E_{ads}), interaction distance (D), NBO (Mulliken) charge transfer (Q) of MA and TMA molecules on different substrates.

| Gas | substrate | E_{ads} (eV) | Q (e) | D (Å) | Reference |
|------------------|-------------------------|-----------------------|------------------|---------|-----------|
| MA | Covalent | −0.09 to − | – | 1.60 to | [50] |
| | graphene | 0.16 | – | 2.59 | |
| | WS ₂ /MWCNTs | −0.35 | – | 2.69 | [51] |
| | Fluorographene | −0.33 | 0.015 | 2.19 | [52] |
| | MnN ₄ -C | −1.41 | 0.285 (0.224) | 2.09 | This work |
| TMA | Covalent | −0.04 to − | – | 1.57 to | [50] |
| | graphene | 0.76 | – | 2.75 | |
| | WS ₂ /MWCNTs | −0.59 | – | 2.78 | [51] |
| | Fluorographene | −0.27 | 0.010 | 2.54 | [52] |
| | Phosphorene | −0.28 | −0.189 | – | [53] |
| | MnN ₄ -C | −1.47 | 0.290 (0.234) | 2.10 | This work |
| H ₂ O | MnN ₄ -C | −0.91 | – | 2.07 | This work |

sensor will always be affected by relative humidity [54–56], which is due to a large number of H₂O molecules on the surface of sensitive materials. Therefore, the effect of humidity on the sensing characteristics of the gas sensor was studied by placing H₂O molecule on the MnN₄-C surface. Table 1 shows that the E_{ads} of H₂O on the MnN₄-C is −0.91 eV, which is smaller than that of MnN₄-C/MA and MnN₄-C/TMA complexes, implying that MA and TMA gases are preferentially adsorbed, so as to achieve adequate detection in a humid environment. The charge transfer between the gas and the substrate is another important indicator to NBO and Mulliken charge. As shown in Table 1, the NBO (Mulliken) charge analysis shows that MA and TMA gases act as charge donors and transfer 0.285 (0.224) and 0.290 (0.234) e to MnN₄-C, respectively, much larger than that of fluorographene and phosphorene as substrates [52,53]. This can induce a sharp change in the device’s impedance to achieve effective detection. The phenomenon of charge redistribution was visualized

using the charge density difference (CDD), as shown in Fig. 3. There are obvious electron depletion regions around MA and TMA gases. In contrast, an electron accumulation area appears around and MnN₄-C, which is consistent with the NBO charge analysis that MA and TMA gases act as an electron donors. It can be seen that the large charge density overlap occurs between gases and MnN₄-C, suggesting a strong interaction between them. Based on the analysis of the interaction distance, adsorption energy and charge transfer, MnN₄-C is a potential novel sensing material for monitoring fish spoilage.

The density of state spectra for MnN₄-C and MnN₄-C/gas complexes is displayed in Fig. 4. After adsorption of gases, the HOMO and LUMO levels have been shifted to the higher energy region, leading to a slight increase in the energy gap. In other words, the adsorption of MA and

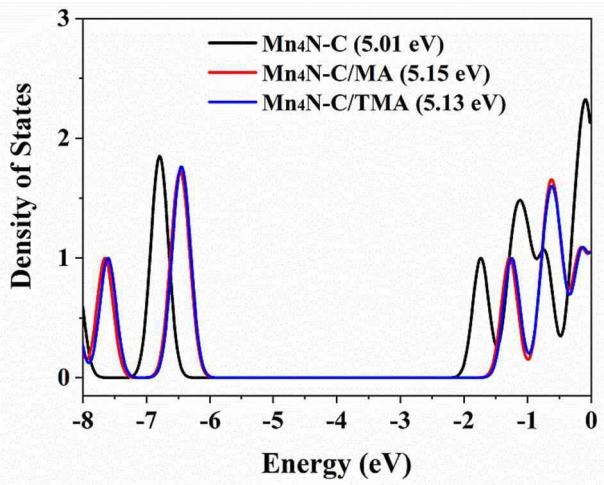


Fig. 4. The density of state plot for MnN₄-C and MnN₄-C/gas complexes.

TMA gases on $\text{MnN}_4\text{-C}$ could change the electrical properties of the substrate, which can be used as a sign in designing the electrical sensor. Fig. S2 presents HOMO and LUMO electron densities after MA and TMA adsorption over $\text{MnN}_4\text{-C}$. There are few orbital surrounding the adsorbed MA and TMA. $\text{MnN}_4\text{-C/MA}$ and $\text{MnN}_4\text{-C/TMA}$ exhibit similar frontier molecular orbital, which is consistent with the density of state results.

Recently, an independent gradient model based on the Hirshfeld partition (IGMH) method has been proposed to analyze intermolecular and intramolecular interactions. This method has a more rigorous physics background and can provide a markedly better graphical effect than an independent gradient model. The region and the type of the interaction can be visualized through the IGMH isosurface map in different colors, in which the blue area represents strong attraction, the green area indicates weak attraction such as the van der Waals (vdW) interaction, and the red area denotes the steric hindrance effect. As shown in Fig. 5a, a deep blue patch appears between gases (MA/TMA) and $\text{MnN}_4\text{-C}$, representing a strong attraction interaction, corresponding to a “spike” located at approximately -0.72 a.u.. In addition, there is also a green isosurface in the adsorption system, which refers to the vdW interaction. $\text{MnN}_4\text{-C/TMA}$ exhibits a larger $\text{sign}(\lambda_2)\rho$ (-0.12 a.u.) than that of $\text{MnN}_4\text{-C/MA}$ (-0.08 a.u.), which is caused by the mutual attraction between more methyl groups in the TMA molecule and $\text{MnN}_4\text{-C}$. Therefore, there is a strong attractive interaction between the TMA molecule and $\text{MnN}_4\text{-C}$ compared to the MA molecule, which is in good agreement with the results of adsorption energy.

The quantum theory of atoms in molecules (QTAIM) can provide more information about the mechanism of the interaction through the topological analysis of the electron density [57]. The topological properties of the bond critical point (BCP) (3, -1), such as electron density ($\rho(r)$), Laplacian of electron density ($\nabla^2\rho(r)$), Lagrangian kinetic energy ($G(r)$), potential energy density ($V(r)$), eigenvalues of Hessian (λ_n), and bond ellipticity index (ϵ), can be used to reveal the nature of bonding.

Fig. 5 depicts the BCPs and bond paths of $\text{MnN}_4\text{-C/MA}$ and $\text{MnN}_4\text{-C/TMA}$ complexes, and the corresponding topological parameters are summarized in Table 2. Larger values for $\rho(r)$ represent a stronger bonding interaction [58,59]. $\rho(r)$ values of H...C and N...Mn in $\text{MnN}_4\text{-C/TMA}$ complexes are 0.0072 and 0.0732, which are larger than those of $\text{MnN}_4\text{-C/MA}$ (0.0069 and 0.0709) complexes. This indicates that there is a higher interaction strength between $\text{MnN}_4\text{-C}$ and TMA, which is consistent with the result of adsorption energy results. The positive values of $\nabla^2\rho(r)$ indicate closed-shell interactions (i.e., ionic bonds, hydrogen bonds or vdW interactions). As shown in Table 2, $\text{MnN}_4\text{-C/MA}$ and $\text{MnN}_4\text{-C/TMA}$ complexes exhibit positive $\nabla^2\rho(r)$ values, suggesting closed-shell interactions. Interaction characteristics can be categorized by a balancing between $G(r)$ and $V(r)$ [60]. $G(r)/|V(r)| < 0.5$ a.u. means that the interaction is characterized by a purely covalent bond, and the dominant character of the bond interaction is noncovalent when the value of $G(r)/|V(r)| > 1$ a.u.. As indicated in Table 2, the $G(r)/|V(r)|$ values of N...Mn in $\text{MnN}_4\text{-C/MA}$ (0.8716) and $\text{MnN}_4\text{-C/TMA}$ (0.8504) complexes are located in the range of $0.5 \sim 1$ a.u., suggesting partial covalent interactions. The interaction type of H...C is noncovalent due to larger $G(r)/|V(r)|$ values. ϵ can be used to evaluate the stability of adsorption systems, and a small value of ϵ corresponds to the excellent stability. The partially covalent interaction is dominant in the current adsorption systems, and the greater stability of $\text{MnN}_4\text{-C/TMA}$ (0.3650) complexes relative to the $\text{MnN}_4\text{-C/MA}$ (0.4256) case can be inferred from smaller ϵ values.

For resistance-type gas sensor, the electrical conductivity (σ) will change before and after gas adsorption on the substrate. Based on the literature [61,62], the energy gap (E_g) is a good indicator of a sensor's electrical conductivity for a gas molecule. The relationship between the electrical conductivity and energy gap can be expressed as follows:

$$\sigma \propto \exp\left(\frac{-E_g}{2KT}\right) \quad (4)$$

where K is the Boltzmann's constant, T is working temperature.

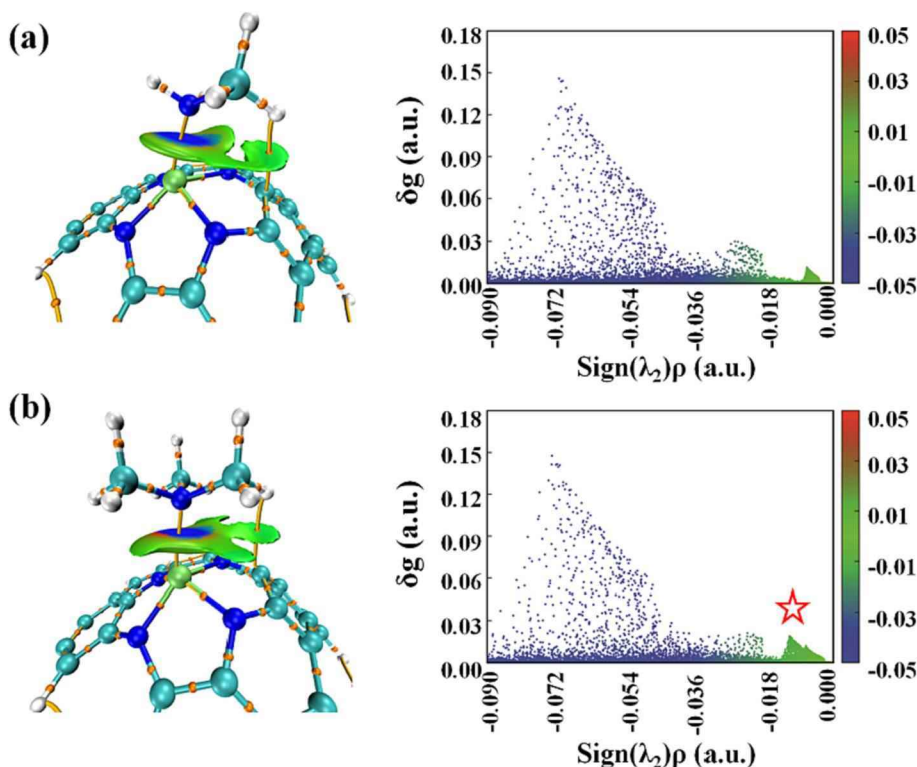


Fig. 5. Independent gradient model based on Hirshfeld partition isosurfaces (left) and scatter graphs (right) of (a) $\text{MnN}_4\text{-C/MA}$ and (b) $\text{MnN}_4\text{-C/TMA}$, respectively. Isosurfaces of $\delta g^{\text{inter}}(\rho) = 0.007$ a.u.. Orange spheres and yellow lines between residue pairs denote bond critical points (3, -1) and bond paths, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2

Topological parameters of MnN₄-C/MA and MnN₄-C/TMA complexes at BCP in key residue pairs. The unit of the data is a.u.

| Gas | Bond | $\rho(r)$ | $\nabla^2\rho(r)$ | $G(r)$ | $V(r)$ | $G(r)/ V(r) $ | ϵ |
|-----|--------|-----------|-------------------|--------|---------|---------------|------------|
| MA | H...C | 0.0069 | 0.0208 | 0.0043 | -0.0035 | 1.2285 | 0.9130 |
| | N...Mn | 0.0709 | 0.3014 | 0.0883 | -0.1013 | 0.8716 | 0.4256 |
| TMA | H...C | 0.0072 | 0.0227 | 0.0047 | -0.0038 | 1.2368 | 2.2142 |
| | N...Mn | 0.0732 | 0.2790 | 0.0847 | -0.0996 | 0.8504 | 0.3650 |

Based on the equation (4), the sensing response (S) can be calculated as:

$$S = \left(\frac{1}{\sigma_{\text{Complex}}} - \frac{1}{\sigma_{\text{MnN}_4-\text{C}}} \right) / \frac{1}{\sigma_{\text{MnN}_4-\text{C}}} \quad (5)$$

where $\sigma_{\text{MnN}_4-\text{C}}$ and σ_{Complex} are electrical conductivity of the clean and gas adsorbed substrate, respectively. As shown in Table S1, the sensing response (S) of MA and TMA are 14 and 9, respectively, which are greater than the threshold of 0.5 [63], implying their higher sensitivity.

Theoretical and experimental studies have confirmed that the gas sensing performance of the device will be affected by an applied electric field [64,65]. The charge transfer and adsorption energy variation of MA and TMA on MnN₄-C was discussed by applying electric fields ranging from -0.008 to +0.008 a.u., as shown in Fig. 6. The applied electric field is perpendicular to the MnN₄-C surface with the upward (downward) arrows representing the positive (negative) electric field (the inset in Fig. 6a). Fig. 6a shows that the electric field exhibits an apparent influence on E_{ads} , and the values are more negative when the electric field changes from -0.008 to +0.008 a.u., suggesting that applying a positive electric field is beneficial to enhance the gas adsorption. This obvious linear relationship also shows that the adsorption and desorption behaviors of MA and TMA on MnN₄-C can be regulated by applying a positive and negative electric field, which provides a new option to develop recyclable MA and TMA gas sensors. From Fig. 6b, the amount of charge transfer toward MnN₄-C increases gradually as the increase of electric field intensity from -0.008 to +0.008 a.u., which indicates that the MA and TMA gas sensitivity is increased by applying a positive electric field. Based on the analysis of E_{ads} and NBO charge, applying a positive electric field can improve the sensing performance of MnN₄-C for MA and TMA gas detection.

The effective desorption of gas molecules from sensitive materials is also very important for a good gas sensor. A short recovery time (τ) is required to achieve sustainable utilization of the gas sensor, which can be expressed as:

$$\tau = v_0^{-1} \exp\left(-\frac{E_{\text{ads}}}{KT}\right) \quad (6)$$

where v_0 is the attempt frequency. According to the equation (6), it is clear that strong interactions prevent spontaneous desorption. However, the fast desorption process of the gas molecules can be easily realized through the heating process. Hence, three temperatures including 298, 398 and 498 K are considered to fully understand the desorption performance of the gas sensor, as shown in Fig. 7. At the room temperature

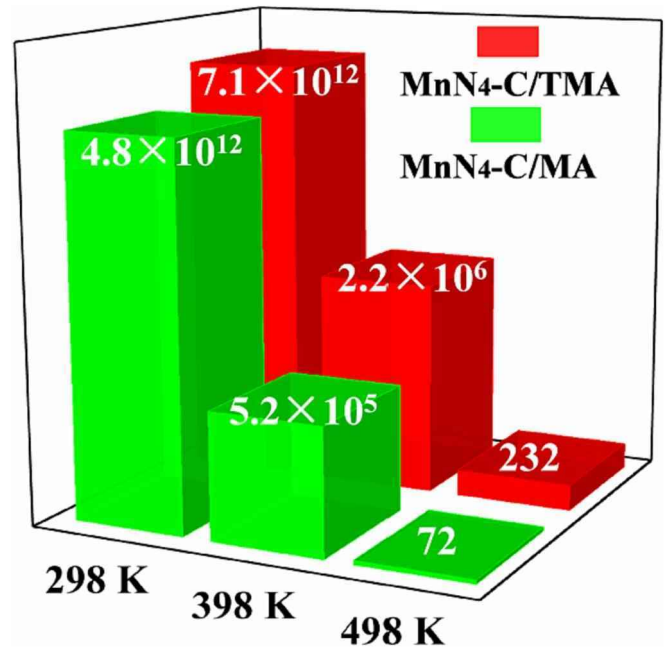


Fig. 7. The recovery time of MA and TMA adsorption on MnN₄-C at 298, 398 and 498 K. The unit of the data is s.

(298 K), MA and TMA gas desorption on MnN₄-C is very difficult due to their strong interactions. It is noted that MA and TMA gases could be easily desorbed from MnN₄-C within 72 and 232 s by increasing the temperature to 498 K. Therefore, MnN₄-C could be a reusable sensor for the detection of MA and TMA gases.

4. Conclusion

In the present work, MnN₄-C as a potential candidate for monitoring fish spoilage is systematically investigated by DFT calculations. The NBO (Mulliken) charge analysis shows that MA and TMA gases act as charge donors and transfer 0.285 (0.224) and 0.290 (0.234) e to MnN₄-C, respectively, which is in good agreement with the CDD results. The obtained adsorption energies indicate that there is a strong interaction between gas molecules (MA and TMA) and MnN₄-C. The results of

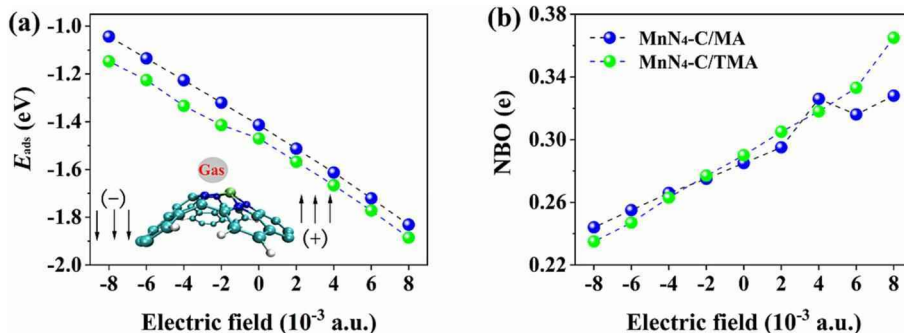


Fig. 6. (a) E_{ads} and (b) NBO charge of adsorption systems under the applied electric field.

QTAIM and IGMH analysis identified the nature of intermolecular interactions is the coexistence of non-covalent and covalent interactions during the adsorption process. Sensing response analysis can also approve that MnN₄-C has high sensitivity toward MA and TMA gases compared with other 2D materials. Furthermore, the MnN₄-C sensor exhibits stable gas-sensing properties under ambient humidity. The electric field has a significant influence on the adsorption energy and charge transfer of the adsorption system. In other words, the MA and TMA gas sensitivity of MnN₄-C can be improved by applying a positive electric field. The studied gases can be desorbed from the surface of MnN₄-C through heating process. As a result, MnN₄-C can be regarded as a recyclable sensitive material for MA and TMA detection.

CRedit authorship contribution statement

Yuanchao Li: Conceptualization, Data curation, Investigation, Writing – original draft. **Cuijuan Jiang:** Formal analysis, Methodology, Visualization. **Xiliang Yan:** Funding acquisition, Software, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (22106025), the Basic and Applied Basic Research Foundation of Guangzhou, China (202201010541).

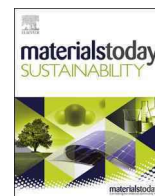
Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.molliq.2024.124090>.

References

- [1] D. Birch, J. Memery, N. Johns, M. Musarskaya, *Journal of International Food & Agribusiness Marketing* 30 (2018) 61.
- [2] G. Jairath, P.K. Singh, R.S. Dabur, M. Rani, M. Chaudhari, *J. Food Sci. Technol.* 52 (2015) 6835.
- [3] E.M. Postma, C. De Graaf, S. Boesveldt, *Food Quality and Preference* 79 (2020) 103771.
- [4] L. Franceschelli, A. Berardinelli, S. Dabbou, L. Ragni, M. Tartagni, *Sensors*, 2021.
- [5] L. Liu, S. Fu, X. Lv, L. Yue, L. Fan, H. Yu, X. Gao, W. Zhu, W. Zhang, X. Li, W. Zhu, *Frontiers in Bioengineering and Biotechnology* 8 (2020).
- [6] J. Shen, S. Xu, C. Zhao, X. Qiao, H. Liu, Y. Zhao, J. Wei, Y. Zhu, *ACS Applied Materials & Interfaces* 13 (2021) 57597.
- [7] C. Zhao, J. Shen, S. Xu, J. Wei, H. Liu, S. Xie, Y. Pan, Y. Zhao, Y. Zhu, *Food Chem.* 392 (2022) 133318.
- [8] Y. Chen, Y. Li, B. Feng, Y. Wu, Y. Zhu, J. Wei, *Sens. Actuators B* 360 (2022) 131662.
- [9] Q. Ma, S. Chu, H. Li, J. Guo, Q. Zhang, Z. Lin, *Appl. Surf. Sci.* 569 (2021) 151074.
- [10] J. Zhang, B. Zhang, S. Yao, H. Li, C. Chen, H. Bala, Z. Zhang, *J. Materiomics* 8 (2022) 518.
- [11] Y. Wu, X. Chen, K. Weng, J. Arramel, W.-J. Jiang, P. Ong, X. Zhang, N.L. Zhao, *Adv. Electron. Mater.* 7 (2021) 2001244.
- [12] S. Vadalkar, D. Chodvadiya, N.N. Som, K.N. Vyas, P.K. Jha, B. Chakraborty, *ChemistrySelect* 7 (2022) e202103874.
- [13] M. Doust Mohammadi, M. Hamzehloo, *Comput. Theor. Chem.* (2018, 1144,) 26.
- [14] N. Yuksel, A. Kose, M.F. Fellah, *Surf. Sci.* 715 (2022) 121939.
- [15] H. Louis, K. Chukwuemeka, E.C. Agwamba, H.Y. Abdullah, A.M.S. Pembere, *J. Mol. Graph. Model.* 124 (2023) 108551.
- [16] A. Kanzariya, S. Vadalkar, S.K. Jana, L.K. Saini, P.K. Jha, *J. Phys. Chem. Solid* 186 (2024) 111799.
- [17] M. Doust Mohammadi, H.Y. Abdullah, *SILICON* 14 (2022) 6075.
- [18] N.A. Tukadiya, S.K. Jana, B. Chakraborty, P.K. Jha, *Surf. Interfaces* 41 (2023) 103220.
- [19] M. Doust Mohammadi, H.Y. Abdullah, *J. Mol. Model.* 27 (2021) 330.
- [20] M.D. Mohammadi, I.H. Salih, H.Y. Abdullah, *Journal of Computational Biophysics and Chemistry* 20 (2020) 23.
- [21] Y.-C. Lin, P.-Y. Teng, C.-H. Yeh, M. Koshino, P.-W. Chiu, K. Suenaga, *Nano Lett.* 15 (2015) 7408.
- [22] S. Impeng, A. Junkaew, P. Maitarad, N. Kungwan, D. Zhang, L. Shi, S. Namuangruk, *Appl. Surf. Sci.* 473 (2019) 820.
- [23] M. Luo, Z. Liang, S. Ghouse Peera, M. Chen, C. Liu, H. Yang, J. Liu, U. Pramod Kumar, T. Liang, *Applied Surface Science* 525 (2020) 146480.
- [24] F. Chen, M. Yang, X. Wang, Y. Song, L. Guo, N. Xie, X. Kou, X. Xu, Y. Sun, G. Lu, *Sens. Actuators B* 290 (2019) 459.
- [25] H. Cai, X. Qiao, M. Chen, D. Feng, A.A. Alghamdi, F.A. Alharthi, Y. Pan, Y. Zhao, Y. Zhu, Y. Deng, *Chin. Chem. Lett.* 32 (2021) 1502.
- [26] H. Nishihara, Q.-H. Yang, P.-X. Hou, M. Unno, S. Yamauchi, R. Saito, J.I. Paredes, A. Martínez-Alonso, J.M.D. Tascón, Y. Sato, M. Terauchi, T. Kyotani, *Carbon* 47 (2009) 1220.
- [27] F.J. Isidro-Ortega, J.H. Pacheco-Sánchez, R. Alejo, L.A. Desales-Guzmán, J. S. Arellano, *Int. J. Hydrogen Energy* 44 (2019) 6437.
- [28] H. Wang, Q. He, S. Liang, Y. Li, X. Zhao, L. Mao, F. Zhan, L. Chen, *Energy Storage Mater.* 43 (2021) 531.
- [29] M. Moradi, M. Nangir, A. Massoudi, *Appl. Surf. Sci.* 562 (2021) 150156.
- [30] J.-D. Chai, M. Head-Gordon, *PCCP* 10 (2008) 6615.
- [31] A.D. McLean, G.S. Chandler, *J. Chem. Phys.* 72 (2008) 5639.
- [32] L.E. Roy, P.J. Hay, R.L. Martin, *J. Chem. Theory Comput.* 4 (2008) 1029.
- [33] M.D. Mohammadi, H.Y. Abdullah, G. Biskos, S. Bhowmick, *Bull. Mater. Sci.* 44 (2021) 198.
- [34] M. Asif, H. Sajid, F. Ullah, S. Khan, K. Ayub, M. Amjad Gilani, M. Arshad, M. Salim Akhtar, T. Mahmood, *Comput. Theor. Chem.* 1199 (2021) 113221.
- [35] M. Doust Mohammadi, H.Y. Abdullah, V. Kalamse, A. Chaudhari, *Comput. Theor. Chem.* 1212 (2022) 113699.
- [36] R.F. de Menezes, F. Pirani, C. Coletti, L.G.M. de Macedo, R. Gargano, *Mater. Today Commun.* 31 (2022) 103426.
- [37] C. John, M. Rajeevan, R.S. Swathi, *Chemistry – an Asian Journal* 17 (2022) e202200625.
- [38] C.L. Radford, P.D. Mudiyansele, A.L. Stevens, T.L. Kelly, *ACS Energy Lett.* 7 (2022) 1635.
- [39] M.D. Mohammadi, H.Y. Abdullah, A. Suvitha, *Iranian Journal of Science and Technology, Transactions a: Science* 45 (2021) 1287.
- [40] M.D. Mohammadi, H.Y. Abdullah, *Journal of Computational Biophysics and Chemistry* 20 (2021) 765.
- [41] T. Lu, F. Chen, *J. Comput. Chem.* 33 (2012) 580.
- [42] W. Humphrey, A. Dalke, K. Schulten, *J. Mol. Graph.* 14 (1996) 33.
- [43] M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, G. Scalmani, V. Barone, G.A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A.V. Marenich, J. Bloino, B.G. Janesko, R. Gomperts, B. Mennucci, H.P. Hratchian, J.V. Ortiz, A.F. Izmaylov, J.L. Sonnenberg, Williams, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V.G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J.A. Montgomery Jr., J.E. Peralta, F. Ogliaro, M.J. Bearpark, J.J. Heyd, E.N. Brothers, K.N. Kudin, V.N. Staroverov, T.A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A.P. Rendell, J.C. Burant, S.S. Iyengar, J. Tomasi, M. Cossi, J.M. Millam, M. Klene, C. Adamo, R. Cammi, J.W. Ochterski, R.L. Martin, K. Morokuma, O. Farkas, J.B. Foresman, D.J. Fox, Wallingford, CT, 2016.
- [44] F.J. Isidro-Ortega, J.H. Pacheco-Sánchez, L.A. Desales-Guzmán, *Int. J. Hydrogen Energy* 42 (2017) 30704.
- [45] F. Yang, P. Song, X. Liu, B. Mei, W. Xing, Z. Jiang, L. Gu, W. Xu, *Angew. Chem.* 130 (2018) 12483.
- [46] A. Khosravi, E. Vessally, M. Oftadeh, R. Behjatmanesh-Ardakani, *J. Coord. Chem.* 71 (2018) 3476.
- [47] J. Zhang, J. Yang, Y. Wang, H. Lu, M. Zhang, *Int. J. Energy Res.* 45 (2021) 10858.
- [48] A.M. Alsbaiyel, S. Alshehri, R.M. Alzhrani, A.D. Alatawi, M. Ahmed Algarni, M. H. Abduljabbar, S.M. Alshahrani, N. Begum Mohammed, *J. Mol. Liq.* 369 (2023) 120855.
- [49] Y. Cao, A. Khan, E. Tazikeh-Lemeski, M. Javan, M.T. Baei, M. Ramezani Taghartapeh, H. Mighani, A. Soltani, M. Pishnamazi, A. Nouri, A.B. Albadarin, *J. Mol. Liq.* 340 (2021) 116845.
- [50] S. Baachaoui, S. Aldulajjan, F. Raouafi, R. Besbes, L. Sementa, A. Fortunelli, N. Raouafi, A. Dhoubi, *RSC Adv.* 11 (2021) 7070.
- [51] Q. Zhou, L. Zhu, C. Zheng, J. Wang, *ACS Appl. Mater. Interfaces* 13 (2021) 41339.
- [52] M. Rouhani, *Chem. Phys. Lett.* 719 (2019) 91.
- [53] S. Saravanan, V. Nagarajan, R. Chandiramouli, *Mater. Res. Express* 6 (2019) 105518.
- [54] J. He, B. Liang, X. Yan, F. Liu, J. Wang, Z. Yang, R. You, C. Wang, P. Sun, X. Yan, H. Lin, B. Kang, Y. Wang, G. Lu, *Sens. Actuators B* 327 (2021) 128940.
- [55] J. Yu, F. Tsow, S.J. Mora, V.V. Tipparaju, X. Xian, *Sens. Actuators B* 345 (2021) 130404.
- [56] V.T. Duong, C.T. Nguyen, H.B. Luong, D.C. Nguyen, H.L. Nguyen, *Solid State Sci.* 113 (2021) 106534.
- [57] A. Imojara, J.E. Ishigbe, H. Abdullah, H.O. Edet, T.E. Gber, M.-B.-A. Eba, A.M. S. Pembere, H. Louis, *Chemical Physics Impact* 7 (2023) 100348.
- [58] Y. Wang, M. Liu, J. Li, Q. Wang, X. Ouyang, H. Wei, K. Zhang, *Enzyme Microb. Technol.* 160 (2022) 110082.
- [59] B. Bankiewicz, P. Matczak, M. Palusiak, *Chem. A Eur. J.* 116 (2012) 452.
- [60] E. Nemati-Kande, R. Karimian, V. Goodarzi, E. Ghazizadeh, *Appl. Surf. Sci.* 510 (2020) 145490.

- [61] S. Kanti Jana, N.N. Som, P.K. Jha, J. Mol. Liq. 383 (2023) 122084.
- [62] S.K. Jana, D. Chodvadiya, N.N. Som, P.K. Jha, Diam. Relat. Mater. 129 (2022) 109305.
- [63] X. Wan, W. Yu, A. Wang, X. Wang, J. Robertson, Z. Zhang, Y. Guo, ACS Sensors 8 (2023) 2319.
- [64] O.G. Agbonlahor, M. Muruganathan, T. Imamura, H. Mizuta, ACS Sensors 5 (2020) 2003.
- [65] H.-B. Li, Y.-T. Feng, Z.-G. Shao, C.-L. Wang, L. Yang, Appl. Surf. Sci. 586 (2022) 152749.



DFT perspective of gas sensing properties of metal oxide nanocages toward trimethylamine: Effects of humidity, temperature and electric field

Yuanchao Li^{a,*}, Jing Sun^{c,*}, Cuijuan Jiang^d, Xiliang Yan^{a,b,**}

^a Institute of Environmental Research at Greater Bay Area, Key Laboratory for Water Quality and Conservation of the Pearl River Delta, Ministry of Education, Guangzhou University, Guangzhou, 510006, China

^b College of Animal Science, South China Agricultural University, Guangzhou, 510642, China

^c College of Engineering and Technology, Northeast Forestry University, Harbin, 150040, China

^d School of Environmental Science and Engineering, Shandong University, Qingdao, 266237, China

ARTICLE INFO

Keywords:

Gas sensors

Nanocages

Trimethylamine adsorption behavior

Sensing mechanism

Density functional theory

ABSTRACT

Developing highly efficient sensitive materials for detecting trimethylamine gas is crucial for analyzing food quality and protecting human health. In this paper, the adsorption and sensing properties of trimethylamine on pristine and doped nanocages have been analyzed based on density functional theory. The results show that Be₁₂O₁₂ with higher electronegativity alkaline metal atoms exhibits the shorter adsorption distance (1.75 Å), the larger adsorption energy (−1.39 eV) and more charge transfer (0.155 e) compared with Mg₁₂O₁₂ and Ca₁₂O₁₂, indicating Be₁₂O₁₂ is more sensitive toward trimethylamine gas. Fe and Zn-doped Be₁₂O₁₂ exhibit superior sensitivity and selectivity towards trimethylamine even in the humid environments and the presence of interfere gas (N₂ and O₂). Additionally, trimethylamine could be desorbed from Fe and Zn-doped Be₁₂O₁₂ through heating. Meanwhile, applying positive electric fields can further enhance and sensitivity of Fe and Zn-doped Be₁₂O₁₂ to trimethylamine. The analyses of sensitive response and recovery time reveal the potentials of Zn-doped Be₁₂O₁₂ as reusability resistance-type gas sensors with comparable performances. This theoretical investigation will provide valuable information for experimentalists to design and synthesis novel sensitive materials for detecting trimethylamine.

1. Introduction

Trimethylamine (TMA), as a tertiary amine, will be produced by the decomposition of seafood during processing, transportation and storage, which can be used as an effective indicator to evaluate fish/shrimp/shellfish freshness [1]. The concentration of TMA is closely related to the seafood quality (fresh, <10 ppm; preliminary decay, 10–50 ppm; corruption, >60 ppm) [2]. Eating spoiled food can cause some diseases, such as diarrhea and fever. In addition, TMA is also toxic gas that can cause intensively damage to the human respiratory system [3]. The National Institute for Occupational Safety and Health of the United States reported that the safe exposure time in the TMA environment should be less than 10 h for 10 ppm [4]. Therefore, effective detection of TMA is highly required to analyze food quality and protect human health. To date, various methods such as fluorescent probes, ion

mobility spectrometry, colorimetric test strips, calorimetric analysis, and mass spectrometry etc., have been widely applied to detect TMA. However, long pre-experiment process, expensive precise instruments, professional technical operation and destructive testing are not suitable for on-site assessment.

Gas sensor, as a nondestructive technology, provides a promising platform for detecting TMA with advantages in terms of easy portability, low cost, simple operation, on-site, fast response speed and real-time detection. Sensing material, as the core part of gas sensor, directly determines the performance of the devices [5–7]. In recent years, massive researchers have been devoted to developing high-performance sensing materials [8,9]. Inorganic nanomaterials with fullerene like nanocages have attracted great attention due to their fascinating electronic, optical, and stability properties [10–12]. This unique nanocage structure has high symmetry and surface area, providing sufficient sites for molecular

* Corresponding author

** Corresponding author Institute of Environmental Research at Greater Bay Area, Key Laboratory for Water Quality and Conservation of the Pearl River Delta, Ministry of Education, Guangzhou University, Guangzhou, 510006, China.

E-mail addresses: sunjing@nefu.edu.cn (J. Sun), yanxiliang1991@gzhu.edu.cn (X. Yan).

<https://doi.org/10.1016/j.mtsust.2024.100668>

Received 22 November 2023; Received in revised form 26 December 2023; Accepted 12 January 2024

Available online 13 January 2024

2589-2347/© 2024 Elsevier Ltd. All rights reserved.

adsorption. Numerous researches have shown that nanoclusters exhibit excellent adsorption performance for toxic and harmful gases [13–15]. Among various nanocage clusters, $\text{Be}_{12}\text{O}_{12}$, $\text{Mg}_{12}\text{O}_{12}$, and $\text{Ca}_{12}\text{O}_{12}$ exhibit excellent semiconducting properties, and could be a potential adsorbent or gas sensitive material [16,17]. Zhao et al. reported that $\text{Ca}_{12}\text{O}_{12}$ nanocluster is highly sensitive toward hazardous mustard gas [18]. Beheshtian et al. demonstrated $\text{Be}_{12}\text{O}_{12}$ fullerene is a promising hydrogen storage material [19]. $\text{Mg}_{12}\text{O}_{12}$ nanocluster has the potential carriers for the delivery of hydroxyurea drug [20]. These results show that such nanoclusters are good candidate for drug delivery, hydrogen storage, and hazardous gas detection. The potential application of $\text{Be}_{12}\text{O}_{12}$, $\text{Mg}_{12}\text{O}_{12}$, and $\text{Ca}_{12}\text{O}_{12}$ as gas sensing material for TMA has not been reported yet. In addition, $\text{Be}_{12}\text{O}_{12}$, $\text{Mg}_{12}\text{O}_{12}$, and $\text{Ca}_{12}\text{O}_{12}$ as the Group-II alkaline earth oxide nanocages, the atomic electronegativity gradually decreases from Be (1.57), Mg (1.31) to Ca (1.00). However, few studies have focused on the effect of electronegativity of materials on gas adsorption and electron transfer. Therefore, it is meaningful to explore the relationship between the material components and the sensitive properties for developing high performance gas sensors.

Intrinsic sensing materials generally exhibit poor adsorption properties for most gases. Previous studies found that transition-metal (TM) doping is an effective strategy for improving adsorption and sensing properties [21–24]. Among the various TM elements, Cr, Fe and Zn have been widely used as dopants in various materials due to their special properties, and have achieved excellent results. For example, Chen et al. reported that Cr modification can significantly improve surface activity of the intrinsic GeSe monolayer, and exhibits better absorption and sensing performance to the SF_6 decomposed species [25]. Wang et al. found that Zn-doped MoO_3 hierarchical microflower exhibits the most amazing detection characteristics for 50 ppm CO, with a response of 31.23 [26]. Xiao et al. have shown that Fe doped MoS_2 greatly enhanced the adsorption stability of NO_2 gas due to strong mixture between TM-nd orbital and molecular orbital of NO_2 [27]. In addition, our previous research found that there is a positive potential near Cr, Fe and Zn atoms [28], which is beneficial for attracting gas molecules with negative potential, such as TMA. To the best of our knowledge, the potential application of Cr, Fe and Zn doped inorganic nanocage clusters for detecting TMA gas and corresponding sensitive mechanisms has not been reported from a theoretical perspective.

In this work, the adsorption properties of TMA on $\text{Be}_{12}\text{O}_{12}$, $\text{Mg}_{12}\text{O}_{12}$, and $\text{Ca}_{12}\text{O}_{12}$ nanocages were investigated using density functional theory (DFT) calculations. The results showed that $\text{Be}_{12}\text{O}_{12}$ with higher electronegativity alkaline metal atoms is more sensitive towards TMA gas. We also further explored the potential applications of doped $\text{Be}_{12}\text{O}_{12}$ with TM elements (Cr, Fe and Zn) in TMA gas sensors. Some important parameters, the adsorption energies, adsorption distances, energy gap, natural bond orbital charge, charge difference density etc. were calculated to reveal the sensing mechanism. In order to better comprehensively evaluate the performance of gas sensors, the effects of temperature, humidity, electric field, and interfering gases were also considered.

2. Computational methods

The geometric and electronic properties of isolated gas molecule and nanocages were optimized based on the DFT method with B3LYP [29, 30] functional. The 6-311G(d,p) [31] and LANL2TZ basis set [32] was utilized to depict nonmetal and TM atoms, respectively. The convergence criteria of geometry optimization are 45×10^{-5} hartree/bohr for max force, 3×10^{-4} hartree/radian for root-mean-square (rms) force, 18×10^{-4} bohr for max displacement and 12×10^{-4} radian for gradients of rms displacement. Previous reports have shown that this functional performs reasonably well to describe geometric and electronic properties of inorganic nanoclusters ($\text{Zn}_{12}\text{O}_{12}$, $\text{Zn}_{12}\text{S}_{12}$, $\text{B}_{12}\text{P}_{12}$, $\text{B}_{12}\text{N}_{12}$, $\text{Al}_{12}\text{P}_{12}$, $\text{Al}_{12}\text{N}_{12}$, $\text{Ga}_{12}\text{As}_{12}$, $\text{Ga}_{12}\text{N}_{12}$ etc.), and can provide reliable calculation results [33–36]. The frequency calculations have been

carried out at the same level to confirm that the all structures are local minima on the potential energy surfaces. The structural stability of nanocages can be estimated by calculating the cohesive energy, as following equation:

$$E_{\text{cho}} = [E_{\text{Nanocage}} - 12E_{\text{O}} - 12E_{\text{M}}] / 24 \quad (1)$$

where E_{Nanocage} , E_{M} and E_{O} are the total energies of nanocages, single metal atom and oxygen atom, respectively.

M06-2X as the hybrid meta functional exhibits a good reliability in calculating empirical dispersion energy correction [37–39]. Therefore, TMA adsorption properties with $\text{Be}_{12}\text{O}_{12}$, $\text{Mg}_{12}\text{O}_{12}$, $\text{Ca}_{12}\text{O}_{12}$ clusters were calculated at the M06-2X level of theory. The adsorption energy was calculated as follows:

$$E_{\text{ads}} = E_{\text{Complex}} - (E_{\text{Nanocage}} + E_{\text{TMA}}) + E_{\text{BSSE}} \quad (2)$$

where E_{Complex} , E_{Nanocage} and E_{TMA} are the total energy of adsorption system, nanocage and TMA molecule, respectively. E_{BSSE} is the basis set superposition error (BSSE) corrected [40], which can be obtained using the counterpoise method. Natural bond orbital (NBO) charge was implemented to analyze charge transfer quantity. The charge redistribution between the gas molecule and adsorbent material can be observed by the charge density difference (CDD), which was defined as:

$$\Delta\rho = \rho_{\text{Complex}} - \rho_{\text{Nanocage}} - \rho_{\text{TMA}} \quad (3)$$

where ρ_{Complex} , ρ_{Nanocage} and ρ_{TMA} represent the charge densities of the adsorption system, nanocage, and the TMA molecule, respectively. In addition, electrostatic potential (ESP) and independent gradient model based on Hirshfeld partition (IGMH) were calculated, and then processed using Multiwfn 3.8 code [41] and VMD 1.9.3 software [42]. All calculations were performed using Gaussian 16C program [43].

3. Results and discussion

3.1. Pristine nanocages ($\text{Be}_{12}\text{O}_{12}$, $\text{Mg}_{12}\text{O}_{12}$, $\text{Ca}_{12}\text{O}_{12}$) and their TMA gas adsorptions

The optimized structures and ESP surfaces of the pristine nanocages are depicted in Fig. 1, and the corresponding geometry parameters are listed in Table S1. Two bond lengths can be identified, where d_{46} is shared between hexagon ring and tetragon ring while d_{66} between two hexagon rings. As shown in Table S1, the lengths of d_{46} (d_{66}) are 1.58 (1.52), 1.95 (1.89), and 2.19 (2.15) Å for $\text{Be}_{12}\text{O}_{12}$, $\text{Mg}_{12}\text{O}_{12}$, $\text{Ca}_{12}\text{O}_{12}$ nanocages, respectively, which are in good agreement with previous calculations [44–46], indicating that our predicted values are reasonable. In addition, the cohesive energies (E_{cho}) are calculated to further evaluate the structural stability. The E_{cho} of $\text{Be}_{12}\text{O}_{12}$, $\text{Mg}_{12}\text{O}_{12}$, $\text{Ca}_{12}\text{O}_{12}$ nanocages are −6.89, −5.41 and −5.80 eV (Table S1), implying their high structural stability. The information of charge distribution can be obtained from the ESP, which is crucial for predicting the reaction sites of sensitive materials. Orange and cyan spheres represent the local maxima and minima of the ESP. As shown in Fig. 1, the global surface minimum is around the O atom of $\text{Be}_{12}\text{O}_{12}$, $\text{Mg}_{12}\text{O}_{12}$, and $\text{Ca}_{12}\text{O}_{12}$ with the values of about −0.86, −1.60, and −2.08 eV, respectively. The positive ESP belong to the Be, Mg, and Ca atom with the local surface maximum of about 3.16, 5.85, and 3.81 eV, respectively. For TMA, the most negative region is located near the N atom with a global surface minimum of −1.49 eV and a local surface maximum of 0.42 eV over the H atom.

After full relaxed geometries optimization, the most stable structures of TMA gas on pristine nanocages are depicted in Fig. 2. The distances between the Be, Mg, and Ca atom of nanocages and N atom of TMA gas molecule are 1.75, 2.13, and 2.54 Å, respectively. It is clear that the adsorption distance decreases gradually with the increase of alkaline earth electronegativity. $\text{Be}_{12}\text{O}_{12}$ exhibits shorter adsorption distance

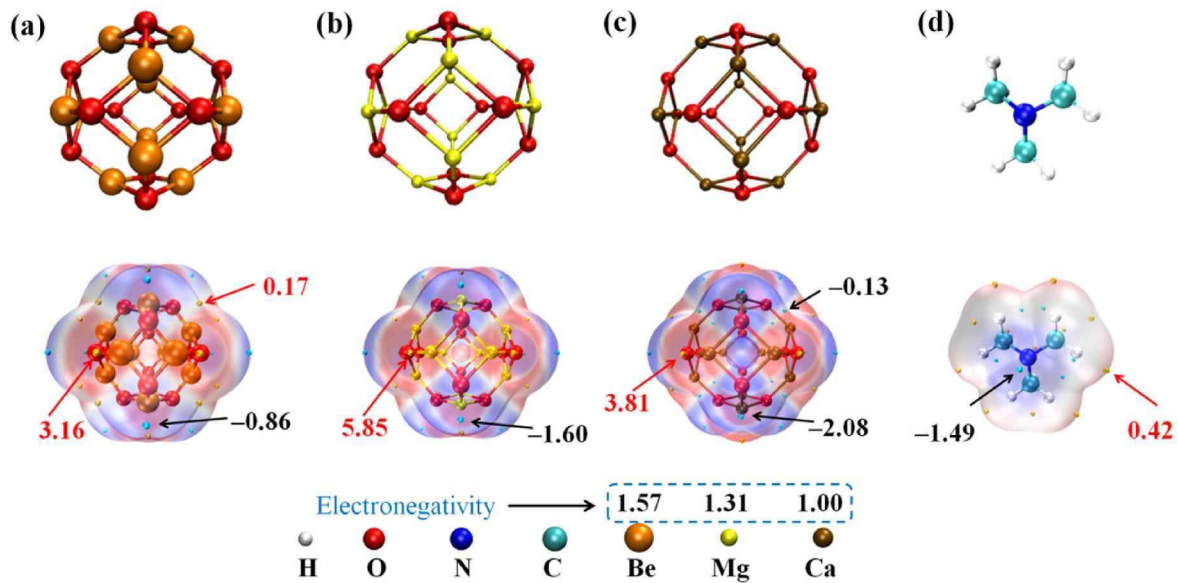


Fig. 1. The optimized structures (up) and electrostatic potential surfaces (down) of (a) $\text{Be}_{12}\text{O}_{12}$, (b) $\text{Mg}_{12}\text{O}_{12}$, and (c) $\text{Ca}_{12}\text{O}_{12}$, and (d) TMA. The unit of the electrostatic potential is eV.

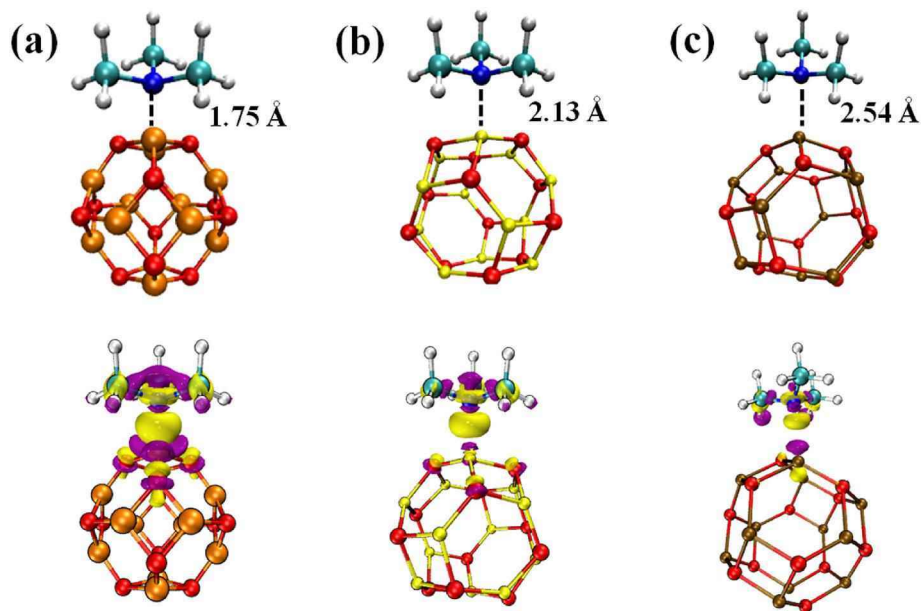


Fig. 2. The stable adsorption configuration (up) and charge density difference (down) of (a) $\text{Be}_{12}\text{O}_{12}/\text{TMA}$, (b) $\text{Mg}_{12}\text{O}_{12}/\text{TMA}$, and (c) $\text{Ca}_{12}\text{O}_{12}/\text{TMA}$. Yellow and magenta represent electron accumulation and depletion, respectively. The isosurface value was set to $0.003 \text{ e} \text{ \AA}^{-3}$. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

than that of MoO_3 [47], SnS/SnS_2 [48], and functionalized graphene [49] substrate (Table 1), which may be attributed to the strong

Table 1

The adsorption energy (E_{ads}) and interaction distance (D) of TMA molecules on different substrates.

| Substrate | E_{ads} (eV) | D (Å) | Reference |
|-------------------------------|-----------------------|--------------|-----------|
| MoO_3 | -1.37 | 2.55 | [47] |
| SnS/SnS_2 | -0.60 | 2.72 | [48] |
| Graphene/carbene | -0.01 to -0.51 | 1.56 to 3.12 | [49] |
| Metalloporphyrins | -0.16 to -0.73 | – | [53] |
| MoS_2/PANI | -0.75 | – | [54] |
| $\text{Be}_{12}\text{O}_{12}$ | -1.39 | 1.75 | This work |

interaction. The small distance allows favorable adsorption of TMA on $\text{Be}_{12}\text{O}_{12}$. Previous reports found that the obvious charge transfer and stronger adsorption energy may easily result in higher resistance changes and better gas sensing behavior [50–52]. To explore the electron transfer characteristics of the adsorption system, the charge density difference (CDD) is calculated and presented in Fig. 2. It can be seen from the CDDs that an obvious charge density redistribution was observed between $\text{Be}_{12}\text{O}_{12}$ and TMA, implying that there is a strong chemisorption. The electron sharing occurs between the $\text{Mg}_{12}\text{O}_{12}$, $\text{Ca}_{12}\text{O}_{12}$ and TMA gradually decreases, which indicates that their interaction is weak. The electron depletion is mainly localized at the TMA while the electron accumulation is mainly localized at the nanocages. Through NBO charge analysis (Fig. 3), 0.155, 0.069, and 0.029 e

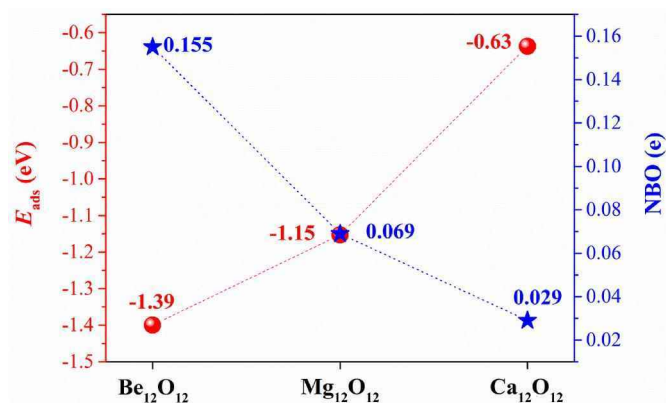


Fig. 3. The adsorption energy and natural bond orbital charge analysis of $\text{Be}_{12}\text{O}_{12}$, $\text{Mg}_{12}\text{O}_{12}$, and $\text{Ca}_{12}\text{O}_{12}$.

charge is transferred from TMA to $\text{Be}_{12}\text{O}_{12}$, $\text{Mg}_{12}\text{O}_{12}$, and $\text{Ca}_{12}\text{O}_{12}$, respectively. It is clear that nanocages with higher electronegativity alkaline metal atoms can capture more electrons, which is beneficial to causing a sharp improvement in the resistance of the gas sensor. More negative adsorption energy (E_{ads}) refer to strong interaction. As shown in Fig. 3, the order of adsorption strength of TMA gas on nanocages is: $\text{TMA}/\text{Be}_{12}\text{O}_{12}$ (-1.39 eV) > $\text{TMA}/\text{Mg}_{12}\text{O}_{12}$ (-1.15 eV) > $\text{TMA}/\text{Ca}_{12}\text{O}_{12}$ (-0.63 eV). In other word, TMA gas is more favorably adsorbed on nanocages with higher electronegativity alkaline metal atoms, which is consistent with the analysis of CDDs. The adsorption energy ($E_{\text{ads-1}}$) without BSSE corrected was also calculated, and the results are presented in Fig. S1. The values of the uncorrected ($E_{\text{ads-1}}$) are more negative than that of E_{ads} , accompanied by the BSSE corrected (E_{BSSE}) ranged from 0.14 to 0.23 eV. It is clear that the adsorption energy will be overestimated if the E_{BSSE} is ignored. In addition, we have also compared the adsorption strength with previously reported 2D materials in Table 1. It is clear that $\text{Be}_{12}\text{O}_{12}$ exhibits more negative adsorption energy compared with other reported 2D compounds [47–49,53,54]. In sum, the smaller adsorption distance, the larger adsorption energy and charge transfer amount indicated that $\text{Be}_{12}\text{O}_{12}$ exhibits excellent sensitivity to TMA gas. Therefore, nanocages with higher electronegativity alkaline metal atoms are more suitable for TMA gas sensor applications.

3.2. Doped $\text{Be}_{12}\text{O}_{12}$ nanocages as well as nanocage/TMA complexes

Heteroatom doping is an efficient approach to enhance the gas adsorption behavior of the substrate. In this work, Cr, Fe and Zn atoms with higher electronegativity as dopants, are introduced to improve the sensing performance of $\text{Be}_{12}\text{O}_{12}$ toward TMA gas. One of the main reasons due to these atoms show relatively positive potential and can form a better adsorption with TMA gas dominated by negative potential through electrostatic attraction. The optimized structures of Cr, Fe and Zn-doped $\text{Be}_{12}\text{O}_{12}$ nanocages are presented in Fig. S2. The bond lengths d_{46}/d_{66} of $\text{CrBe}_{11}\text{O}_{12}$, $\text{FeBe}_{11}\text{O}_{12}$, and $\text{ZnBe}_{11}\text{O}_{12}$ have been enlarged from 1.58/1.52 to 1.96/1.88, 1.90/1.84, and 1.99/1.91 Å (Table S1), respectively. These geometric change can be ascribed to the radius of Cr, Fe and Zn atoms is slightly larger than that of the Be atom. The E_{cho} of $\text{CrBe}_{11}\text{O}_{12}$, $\text{FeBe}_{11}\text{O}_{12}$, and $\text{ZnBe}_{11}\text{O}_{12}$ nanocages are -6.88 , -6.83 and -6.68 eV (Table S1), respectively, which are close to that of pristine $\text{Be}_{12}\text{O}_{12}$ (-6.89 eV), implying their high structural stability. Heteroatom doping can change the distribution of electrostatic potential of sensitive materials. As shown in Fig. S2, the most positive region is located near Cr, Fe, Zn atoms, and the global surface maximum is increased from 3.16 ($\text{Be}_{12}\text{O}_{12}$) to 7.24, 3.25, and 4.50 eV for $\text{CrBe}_{11}\text{O}_{12}$, $\text{FeBe}_{11}\text{O}_{12}$, and $\text{ZnBe}_{11}\text{O}_{12}$ nanocages, respectively. The global surface minimum of negative ESP has not changed significantly. Therefore, Cr, Fe and Zn-doped $\text{Be}_{12}\text{O}_{12}$ nanocages is more conducive to TMA gas adsorption.

The fully relaxed geometries of complexes were presented in Fig. 4. TMA molecule tends to be vertically adsorbed on the substrate, in which the N atom of TMA molecule is trapped by the dopants, with the bond lengths of N–Cr, N–Fe and N–Zn measured to be 2.17, 2.07, and 2.11 Å, respectively. The E_{ads} of TMA gas adsorbed to $\text{FeBe}_{11}\text{O}_{12}$ and $\text{ZnBe}_{11}\text{O}_{12}$ are -1.53 , and -1.52 eV, respectively, which are larger than that of $\text{Be}_{12}\text{O}_{12}$. Fe and Zn doping can effectively enhance the sensitivity to TMA gas. Cr doping has little effect on E_{ads} . The CDD in Fig. 4 showed a large accumulation of charges on $\text{CrBe}_{11}\text{O}_{12}$ and $\text{FeBe}_{11}\text{O}_{12}$ relative to the $\text{ZnBe}_{11}\text{O}_{12}$. TMA behaves as an electron-donor and transfers 0.172, 0.185, and 0.092 e to the $\text{CrBe}_{11}\text{O}_{12}$, $\text{FeBe}_{11}\text{O}_{12}$ and $\text{ZnBe}_{11}\text{O}_{12}$, respectively. It is clear that Cr and Fe dopants can capture more electrons from TMA molecule. The IGMH is a powerful technique for understanding the nature of the intermolecular interaction [55]. The IGMH method has more rigorous physics background and exhibits markedly better graphical effects compared to the IGM. The blue areas represent the strong attractions (such as H-bond, halogen bond ...), the green areas indicate van der Waals interactions. As shown in Fig. 5, the evident dark blue patch appears between the N atom and the Be, Cr, Fe and Zn atoms of the $\text{Be}_{12}\text{O}_{12}$, $\text{CrBe}_{11}\text{O}_{12}$, $\text{FeBe}_{11}\text{O}_{12}$, and $\text{ZnBe}_{11}\text{O}_{12}$, respectively, and corresponding the spike peak located around -0.06 a.u., indicating that their strong attractive interaction. For $\text{Be}_{12}\text{O}_{12}/\text{TMA}$ complex, there are also large areas of green between the O atom and six H atoms of $-\text{CH}_3$ group. For doped $\text{Be}_{12}\text{O}_{12}/\text{TMA}$ complex, TMA molecule has a slight inclination due to the dopants protruding slightly from the nanocages surface. Therefore, the number of green areas between O atom and H atoms of $-\text{CH}_3$ group decreases. Compared with Cr and Fe doped $\text{Be}_{12}\text{O}_{12}$, the electron density of green areas for $\text{ZnBe}_{11}\text{O}_{12}/\text{TMA}$ is very small with the spike peak located around 0 a.u. This phenomenon can be attributed to the long distance between H atoms of $-\text{CH}_3$ group and O atom.

The theory of atoms in molecules (AIM) can provide more information of chemical bond characteristics and electron density distribution. Topological properties such as electron density $\rho(r)$, Laplacian of electron density $\nabla^2\rho(r)$, Lagrangian kinetic energy $G(r)$, potential energy density $V(r)$, eigenvalues of Hessian λ_n , and bond ellipticity index ϵ are listed Table 2. The bond critical point (BCP) (3, -1) and bond paths are all shown in Fig. 5. The larger $\rho(r)$ values indicate higher interaction strength. The positive $\nabla^2\rho(r)$ can be described as highly polarized bond. The values of $G(r)/|V(r)|$ smaller than 0.5 (larger than 1) can be considered as purely covalent (noncovalent) interaction, respectively. As shown in Table 2, the larger $\rho(r)$, positive values of $\nabla^2\rho(r)$, and also $0.5 < G(r)/|V(r)| < 1$ reveal that N ... Be, N ... Cr, N ... Fe, and N ... Zn were detected as medium and partially covalent interaction. Lower ϵ values indicate the structural stability of the interaction. The more stability of $\text{Be}_{12}\text{O}_{12}/\text{TMA}$, $\text{FeBe}_{11}\text{O}_{12}/\text{TMA}$, and $\text{ZnBe}_{11}\text{O}_{12}/\text{TMA}$ complexes relative to $\text{CrBe}_{11}\text{O}_{12}/\text{TMA}$ case can also be inferred from smaller ϵ values.

The global reactivity descriptors such as chemical hardness (η), chemical potential (μ), and electrophilicity index (ω) can be used to study stability and reactivity of the system. Low η and high μ values correspond to high reactivity. As shown in Table 3, doping of Cr, Fe and Zn to $\text{Be}_{12}\text{O}_{12}$ have decreased the η of pristine $\text{Be}_{12}\text{O}_{12}$ from 3.92 to 1.62, 2.22, and 2.82 eV, respectively. The order of the μ is as follows: $\text{CrBe}_{11}\text{O}_{12} > \text{FeBe}_{11}\text{O}_{12} > \text{Be}_{12}\text{O}_{12} > \text{ZnBe}_{11}\text{O}_{12}$. These results suggest that Cr and Fe-doped $\text{Be}_{12}\text{O}_{12}$ can enhance the reactivity. The ω values of nanocages are reduced after the adsorption of TMA gas, and follow the trend for the complexes: $\text{FeBe}_{11}\text{O}_{12}/\text{TMA} > \text{ZnBe}_{11}\text{O}_{12}/\text{TMA} > \text{Be}_{12}\text{O}_{12}/\text{TMA} > \text{CrBe}_{11}\text{O}_{12}/\text{TMA}$. The results elucidate that $\text{ZnBe}_{11}\text{O}_{12}/\text{TMA}$ and $\text{FeBe}_{11}\text{O}_{12}/\text{TMA}$ composites are more stable in accepting electrons.

To gain more insight into the sensitivity of nanocages towards TMA molecule, some relevant electronic properties are calculated (Table 3). After Cr, Fe and Zn-doped $\text{Be}_{12}\text{O}_{12}$, the LUMO levels were shifted more negative, while the HOMO levels were shifted to the higher energy

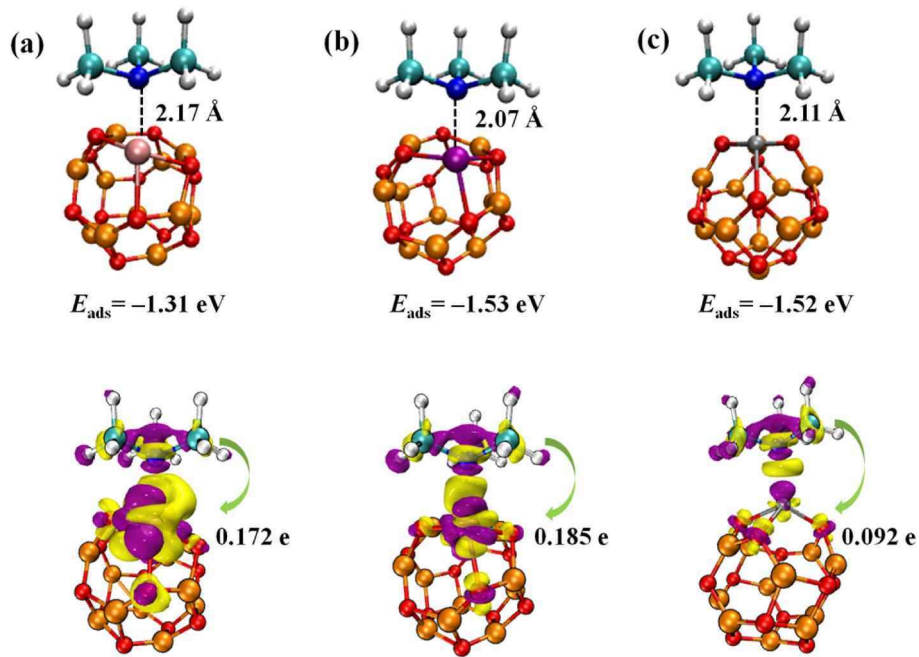


Fig. 4. The stable adsorption configuration (up) and charge density difference (down) of (a) $\text{CrBe}_{11}\text{O}_{12}/\text{TMA}$, (b) $\text{FeBe}_{11}\text{O}_{12}/\text{TMA}$, and (c) $\text{ZnBe}_{11}\text{O}_{12}/\text{TMA}$ complexes. The isosurface value was set to $0.003 \text{ e } \text{\AA}^{-3}$.

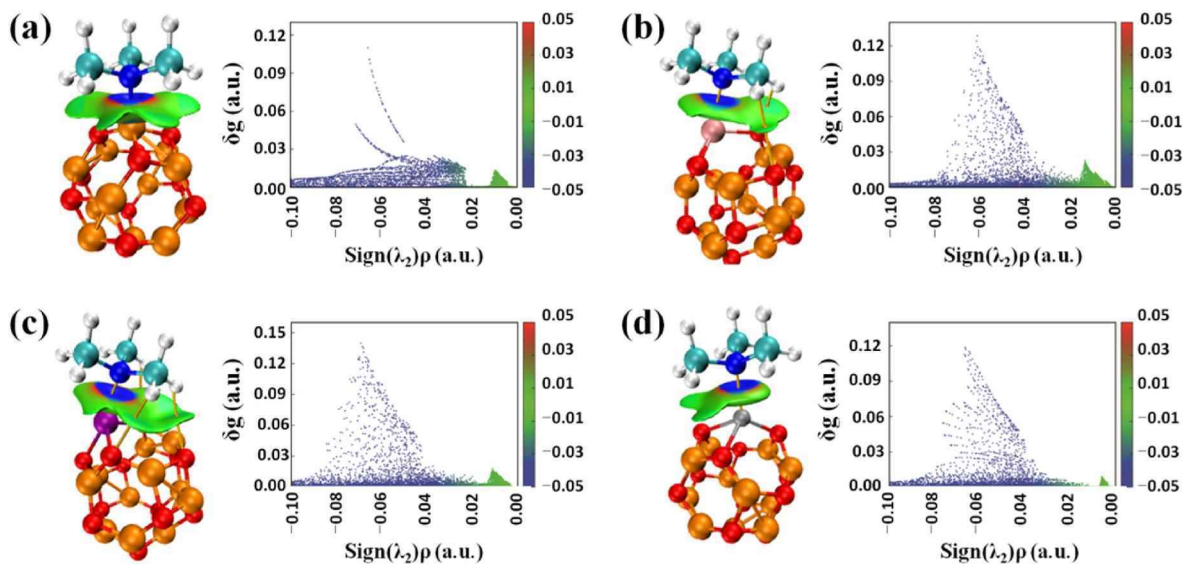


Fig. 5. Independent gradient model based on Hirshfeld partition isosurfaces (left) and scatter graphs (right) of (a) $\text{Be}_{12}\text{O}_{12}/\text{TMA}$, (b) $\text{CrBe}_{11}\text{O}_{12}/\text{TMA}$, (c) $\text{FeBe}_{11}\text{O}_{12}/\text{TMA}$, and (d) $\text{ZnBe}_{11}\text{O}_{12}/\text{TMA}$ complexes, respectively. Isosurfaces of $\delta g^{\text{inter}}(\rho) = 0.004 \text{ a.u.}$. Orange spheres and yellow lines represent BCPs and bond paths, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 2

Topological parameters of nanocages/TMA complexes at BCP in key residue pairs. The unit of the data is a.u.

| Substrate | Bond | $\rho(r)$ | $\nabla^2 \rho(r)$ | $G(r)$ | $V(r)$ | $G(r)/ V(r) $ | λ_1 | λ_2 | λ_3 | ϵ |
|---------------------------------|----------|-----------|--------------------|--------|---------|---------------|-------------|-------------|-------------|------------|
| $\text{Be}_{12}\text{O}_{12}$ | N ... Be | 0.0659 | 0.3627 | 0.0979 | -0.1051 | 0.9314 | -0.1292 | -0.1257 | 0.6177 | 0.0278 |
| $\text{CrBe}_{11}\text{O}_{12}$ | N ... Cr | 0.0611 | 0.2702 | 0.0777 | -0.0878 | 0.8849 | -0.1038 | -0.0220 | 0.3961 | 3.7181 |
| $\text{FeBe}_{11}\text{O}_{12}$ | N ... Fe | 0.0694 | 0.3729 | 0.1049 | -0.1167 | 0.8988 | -0.0727 | -0.0402 | 0.4859 | 0.8084 |
| $\text{ZnBe}_{11}\text{O}_{12}$ | N ... Zn | 0.0654 | 0.2810 | 0.0682 | -0.0829 | 0.8226 | -0.0809 | -0.0806 | 0.4426 | 0.0037 |

region, which leads to a decrease in the energy gap (E_g) of the pristine $\text{Be}_{12}\text{O}_{12}$ from 7.84 to 3.24, 4.45 and 5.64 eV for $\text{CrBe}_{11}\text{O}_{12}$, $\text{FeBe}_{11}\text{O}_{12}$ and $\text{ZnBe}_{11}\text{O}_{12}$, respectively. Smaller E_g means higher chemical

reactivity, which is easy to react with gas molecules [56–58]. By adsorbing the TMA gas, E_g of $\text{Be}_{12}\text{O}_{12}$ and $\text{FeBe}_{11}\text{O}_{12}$ systems are increased, while that of $\text{CrBe}_{11}\text{O}_{12}$ and $\text{ZnBe}_{11}\text{O}_{12}$ systems are increased.

Table 3

Sensitive response (S) and electronic properties including energy level, energy gap (E_g), chemical hardness (η), chemical potential (μ), electrophilicity index (ω) and work function (Φ). All electronic parameters are in eV.

| Systems | S | E_H | E_L | E_g | % E_g | η | μ | ω | Φ | % $\Delta\Phi$ |
|---|-----------------------|-------|-------|-------|---------|--------|-------|----------|--------|----------------|
| Be ₁₂ O ₁₂ | – | –8.63 | –0.79 | 7.84 | – | 3.92 | –4.71 | 2.82 | 4.71 | – |
| Be ₁₂ O ₁₂ /TMA | 5.41×10^4 | –8.00 | –0.44 | 7.56 | 3.57 | 3.78 | –4.22 | 2.35 | 4.22 | 10.40 |
| CrBe ₁₁ O ₁₂ | – | –5.33 | –2.09 | 3.24 | – | 1.62 | –3.71 | 4.24 | 3.71 | – |
| CrBe ₁₁ O ₁₂ /TMA | 6.18×10^8 | –4.35 | –0.59 | 3.76 | 16.04 | 1.88 | –2.47 | 1.62 | 2.47 | 33.42 |
| FeBe ₁₁ O ₁₂ | – | –6.56 | –2.11 | 4.45 | – | 2.22 | –4.33 | 4.22 | 4.34 | – |
| FeBe ₁₁ O ₁₂ /TMA | 1.62×10^3 | –5.61 | –1.35 | 4.26 | 4.26 | 2.13 | –3.48 | 2.84 | 3.48 | 19.81 |
| ZnBe ₁₁ O ₁₂ | – | –8.14 | –2.50 | 5.64 | – | 2.82 | –5.32 | 5.01 | 5.32 | – |
| ZnBe ₁₁ O ₁₂ /TMA | 1.31×10^{20} | –7.52 | –0.69 | 6.83 | 21.09 | 3.41 | –4.10 | 2.46 | 4.11 | 22.74 |

The variation of E_g for CrBe₁₁O₁₂ (16.04 %), FeBe₁₁O₁₂ (4.26 %) and ZnBe₁₁O₁₂ (21.09 %) are larger than that of the Be₁₂O₁₂ (3.57 %) by adsorption of TMA gas.

To obtain further insight into the influence of doping agents and gas adsorption effects on electronic characteristics, the density of states (DOS) and partial density of states (PDOS) along with HOMO and LUMO energy distributions are depicted in Fig. S3, and the corresponding values are listed in Table 3. Significant changes in DOS spectrum are observed after Cr, Fe and Zn-doped Be₁₂O₁₂, some new energy states are generated, which are in agreement with the HOMO and LUMO energy values in Table 3. For Be₁₂O₁₂, the HOMO density is delocalized on the surface, while the LUMO density is observed to be dominant in the ring cage. For doped Be₁₂O₁₂, HOMO density is primarily distributed on dopants and adjacent atoms, while LUMO density is majorly located on dopants. It is obvious that the chemical adsorption of TMA shows an impact on the DOS, making the whole DOS shift to the higher energy level relative to those of isolated nanocages. From the orbital PDOS, one can see that the nanocages contribute largely to the TDOS. In addition, the electron density maps reveal that there is an obvious charge transfer from TMA to nanocages, which is consistent with the CDD results.

To examine the feasibility of utilizing pristine and doped Be₁₂O₁₂ as an effective resistance-type sensor, the electrical conductivity (σ) and sensitive response (S) degree are analyzed by Eq. (4) and Eq. (5):

$$\sigma \propto \exp\left(\frac{-E_g}{2KT}\right) \quad (4)$$

$$S = \left(\frac{1}{\sigma_{\text{Complex}}} - \frac{1}{\sigma_{\text{Nanocage}}} \right) / \frac{1}{\sigma_{\text{Nanocage}}} \quad (5)$$

where K is the Boltzmann's constant, T is working temperature, σ_{Nanocage} and σ_{Complex} are electrical conductivity of the clean and TMA adsorbed substrate, respectively. As shown in Table 3, the ZnBe₁₁O₁₂ exhibits excellent sensitivity and discrimination detection ability for TMA molecule among all nanocages, with the S values of 1.31×10^{20} at room temperature. This sensitivity is significantly superior to that of Rh and Ru modified InSe monolayers [59].

The sensitivity of a sensor can be further assessed by work function (Φ), especially for the Φ -type sensor, which is defined as:

$$\Phi = E_{\text{vac}} - E_F \quad (6)$$

where E_{vac} and E_F are the energy of the vacuum level and the Fermi energy, respectively. The alteration of Φ related to the direction of charge transfer [60,61]. Table 3 shows that Φ of nanocages decrease after TMA gas adsorption, indicating TMA acts as electron donor in adsorption systems, which is consistent with the analysis of NBO and CDD. The change in the Φ of Be₁₂O₁₂, CrBe₁₁O₁₂, FeBe₁₁O₁₂ and ZnBe₁₁O₁₂ after gas adsorption reaches 10.40, 33.42, 19.81 and 22.74 %, respectively, which demonstrates that Cr, Fe and Zn-doped Be₁₂O₁₂ can be considered as an appropriate Φ -based sensor for TMA gas detection.

3.3. Effect of H₂O, N₂ and O₂ on sensitivity and selectivity

It is well known that the gas-sensing characteristics of the sensors are affected by humidity during actual operation. Clarifying the effect of humidity on the sensitive mechanism is highly required. From a theoretical perspective, it is difficult to determine the percentage of humidity in experiments. Therefore, the actual scenarios of relative humidity can be modelled by placing several H₂O molecules on the nanocages surface. This strategy is widely used in theoretical research [62–65]. The most stable adsorption configurations of H₂O onto Be₁₂O₁₂, CrBe₁₁O₁₂, FeBe₁₁O₁₂ and ZnBe₁₁O₁₂ are displayed in Fig. S4, and corresponding E_{ads} are –0.82, –1.21, –1.10 and –1.01 eV, respectively, which are smaller than that of the TMA adsorption system. It unravels that TMA is preferentially adsorbed on the nanocages when H₂O molecule exists at the same time. In other words, this unique nanocage sensor has good selectivity and can effectively detect TMA gas in a wet environment. Subsequently, gas adsorption ability of TMA gas on pristine and doped Be₁₂O₁₂ at different relative humidity is investigated. As shown in Fig. 6, E_{ads} of TMA adsorbed onto Be₁₂O₁₂, FeBe₁₁O₁₂, and ZnBe₁₁O₁₂ gradually decreases as the number of H₂O molecules increases, especially for Be₁₂O₁₂, which suggests that a humid environment has little effect on the FeBe₁₁O₁₂, and ZnBe₁₁O₁₂ adsorption systems compared to Be₁₂O₁₂. It is worth noting that E_{ads} of CrBe₁₁O₁₂ is more negative in the presence of humidity, especially when the number of H₂O molecules is 1, implying that gas selectivity and sensitivity can be improved when a certain of H₂O molecules are accompanied on the surface of CrBe₁₁O₁₂ nanocage.

N₂ and O₂ are one of the most abundant molecules in the atmosphere and can also come into contact with sensitive material surfaces, which may interfere with the detection of target gas. Therefore, it is necessary to study the adsorption ability of N₂ and O₂ on the surface of nanocages.

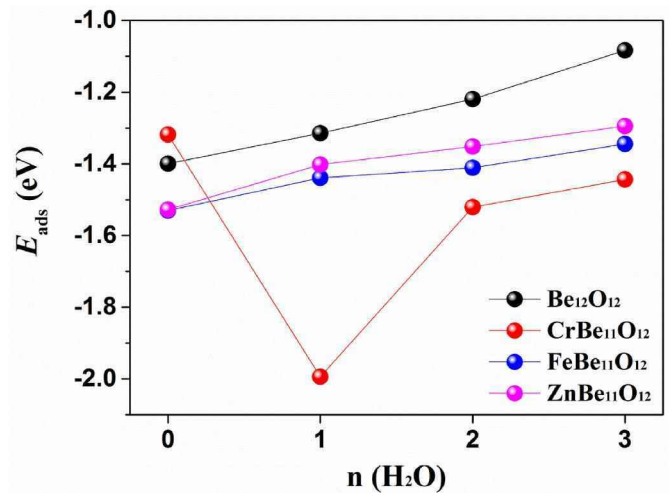


Fig. 6. Adsorption energies of TMA on the surface of Be₁₂O₁₂, CrBe₁₁O₁₂, FeBe₁₁O₁₂, and ZnBe₁₁O₁₂ accompanied by H₂O molecules.

As shown in Fig. S4, E_{ads} of N_2 on $\text{Be}_{12}\text{O}_{12}$, $\text{CrBe}_{11}\text{O}_{12}$, $\text{FeBe}_{11}\text{O}_{12}$ and $\text{ZnBe}_{11}\text{O}_{12}$ are -0.13 , -0.68 , -0.47 and -0.27 eV, respectively, there are lower than that of the TMA adsorption system, indicating that pristine and doped $\text{Be}_{12}\text{O}_{12}$ exhibit good selectivity for TMA gas in the presence of N_2 gas molecule. For O_2 -adsorbed system, it is important to note that O_2 is adsorbed onto $\text{CrBe}_{11}\text{O}_{12}$ with an E_{ads} of -2.97 eV, which is more negative than that of the TMA-adsorbed system. O_2 could adsorb on the $\text{CrBe}_{11}\text{O}_{12}$ rather than TMA, hindering the adsorption sites for adsorption/sensing of TMA. Although some target gases exhibit relatively weak adsorption properties in the presence of O_2 and N_2 from a theoretical perspective, in fact, sensitive materials have good sensitivity to target gases in experiments [66]. Therefore, the co-adsorption mechanism of O_2 and TMA on $\text{CrBe}_{11}\text{O}_{12}$ was been studied. As shown in Fig. 7 (a), TMA shows a favorable E_{ads} of -1.41 eV and short adsorption distance of 1.78 Å in the presence of O_2 . However, there is a small amount of charge transfer (0.052 e) between the sensitive material and the TMA gas (Fig. 7 (b)), which may have a negative impact on the resistance-type sensor.

3.4. Effect of electric field on adsorption process

Previous investigation has shown that electric field may induce the distortion of the adsorption configuration [62,67]. Therefore, the interaction distance between nanocage and TMA, as well as bond length between metal atom (Be, Fe and Zn) and oxygen atom (O) are selected to explore the deformation of adsorption structure under the applied electric field ranging from -0.008 to $+0.008$ a.u. The positive (negative) electric field was set to be perpendicular to the nanocage surface with an upward (downward) direction (the inset in Fig. 8(a)). The interaction distance decreases gradually as the electric field decreases from -0.008 to $+0.008$ a.u. (Fig. 8(a)), suggesting a strong interaction between nanocage and TMA. There is no significant change for bond length (Fig. 8(b)), implying higher the chemical stability of the sensitive materials under the electric fields. Applying an external electric field is an effective method to enhance gas-sensing performance by accelerating electron transfer and promoting gas adsorption [68,69]. In this study, the relationship between electric field with E_{ads} as well as NBO charge was systematically studied. As shown in Fig. 8(c), the applied electric fields had significant effects on the E_{ads} , the values are more negative gradually as the electric field decreases from -0.008 to $+0.008$ a.u. There is an obvious linear trend between the E_{ads} and the electric field, implying that the adsorption (desorption) process can be adjusted by applying a positive (negative) electric field. Fig. 8(d) shows the change of charges transfer under different electric fields. The results show that more charge moving from TMA to nanocages with the electric field changes from -0.008 to $+0.008$ a.u. In a word, applying negative electric fields is beneficial to enhance the sensing performance of nanocages for TMA detection.

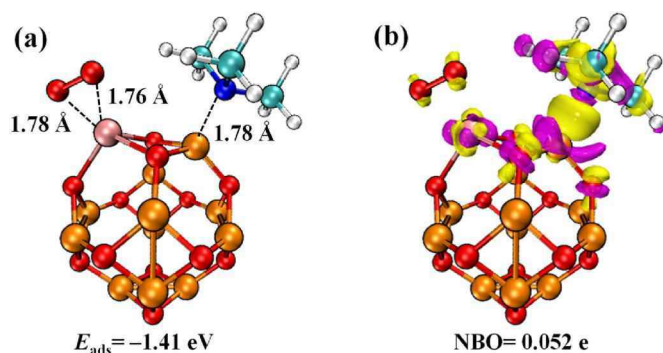


Fig. 7. (a) The stable co-adsorption configuration and (b) charge density difference of TMA on $\text{CrBe}_{11}\text{O}_{12}$ in the presence of O_2 .

3.5. Effect of temperature on recovery time

Recovery time (τ) refers to the time for gas desorption from the sensing material, which can be used to evaluate the reproducibility of the gas sensor. It could be calculated from van't-Hoff-Arrhenius expression in the transition state theory:

$$\tau = \nu_0^{-1} \exp\left(-\frac{E_{\text{ads}}}{KT}\right) \quad (7)$$

where ν_0 is attempt frequency (10^{12} s^{-1}). The high adsorption energy usually leads to a long recovery time. Previous experimental studies have reported that heating the sensor at high temperatures can promote gas desorption, resulting in an ideal recovery time [70–73]. Three temperatures (298 K, 398 K and 498 K) are considered to fully understand the desorption performance of nanocages, with calculated recovery time shown in Table 4. It is obvious that TMA is difficult to desorb from nanocages surface at room temperature due to strong chemical adsorption (-1.39 eV for $\text{Be}_{12}\text{O}_{12}$, -1.53 eV for $\text{FeBe}_{11}\text{O}_{12}$, -1.52 eV for $\text{ZnBe}_{11}\text{O}_{12}$). When the temperature reaches 498 K, TMA could be desorbed from $\text{FeBe}_{11}\text{O}_{12}$, $\text{ZnBe}_{11}\text{O}_{12}$ and $\text{Be}_{12}\text{O}_{12}$ within 16 min, 18 min and 45 s, respectively, which means that $\text{FeBe}_{11}\text{O}_{12}$, $\text{ZnBe}_{11}\text{O}_{12}$ and $\text{Be}_{12}\text{O}_{12}$ are reusable sensitive material for the detection of TMA through heating. In addition, molecular dynamic simulations are conducted to verify the thermal stability of $\text{FeBe}_{11}\text{O}_{12}$, $\text{ZnBe}_{11}\text{O}_{12}$ and $\text{Be}_{12}\text{O}_{12}$, at 498 K for 6 ps with interval of 2 fs, the potential energy curve of $\text{Be}_{12}\text{O}_{12}$, $\text{FeBe}_{11}\text{O}_{12}$, and $\text{ZnBe}_{11}\text{O}_{12}$ was depicted in Fig. S5. It is clear that there is a small change in the range of energy fluctuations, while their structures show no significant deformation. These results prove that $\text{Be}_{12}\text{O}_{12}$, $\text{FeBe}_{11}\text{O}_{12}$, and $\text{ZnBe}_{11}\text{O}_{12}$ are thermodynamically and dynamically stable at higher temperature. The dynamical stability of $\text{FeBe}_{11}\text{O}_{12}$, $\text{ZnBe}_{11}\text{O}_{12}$ and $\text{Be}_{12}\text{O}_{12}$ at 498 K is also verified by plotting IR spectra. Fig. S6 shows that there is no imaginary frequency, indicating their dynamical stability.

4. Conclusion

In this work, the adsorption behaviors of TMA gas on pristine and doped nanocages are systematically investigated from a theoretical perspective. In addition, the effects of temperature, humidity, electric field, and interfering gases on gas sensitivity performance are also considered. The significant conclusions can be summarized as follows.

- (i) Among pristine nanocages, $\text{Be}_{12}\text{O}_{12}$ with higher electronegativity alkaline metal atoms is more sensitive towards TMA gas due to its strong adsorption strength (-1.39 eV), shorter adsorption distance (1.75 Å) and significant charge transfer (0.155 e).
- (ii) The adsorption behavior can be enhanced by doping Fe (-1.53 eV) and Zn (-1.52 eV) atoms in $\text{Be}_{12}\text{O}_{12}$ substrate. The number of electrons transferred from TMA to Cr (0.172 e) and Fe-doped (0.185 e) $\text{Be}_{12}\text{O}_{12}$ is more than that from TMA to pristine $\text{Be}_{12}\text{O}_{12}$ (0.155 e).
- (iii) Fe and Zn-doped $\text{Be}_{12}\text{O}_{12}$ exhibit superior selectivity even in the presence of the humid (H_2O) and other interfere gas (N_2 and O_2). The adsorption strength and charges transfer of TMA on Fe and Zn-doped $\text{Be}_{12}\text{O}_{12}$ could be further enhanced by applying positive electric fields. In addition, TMA could be desorbed from Fe and Zn-doped $\text{Be}_{12}\text{O}_{12}$ through heating.
- (iv) The analyses of sensitive response and recovery time reveal the potentials of Zn-doped $\text{Be}_{12}\text{O}_{12}$ as reusability resistance-type gas sensors with comparable performances.

Overall, the calculated results show that Zn-doped $\text{Be}_{12}\text{O}_{12}$ has great potential in TMA sensing fields. We hope our studies can provide valuable information for understanding sensing mechanism and developing high-performance sensitive materials.

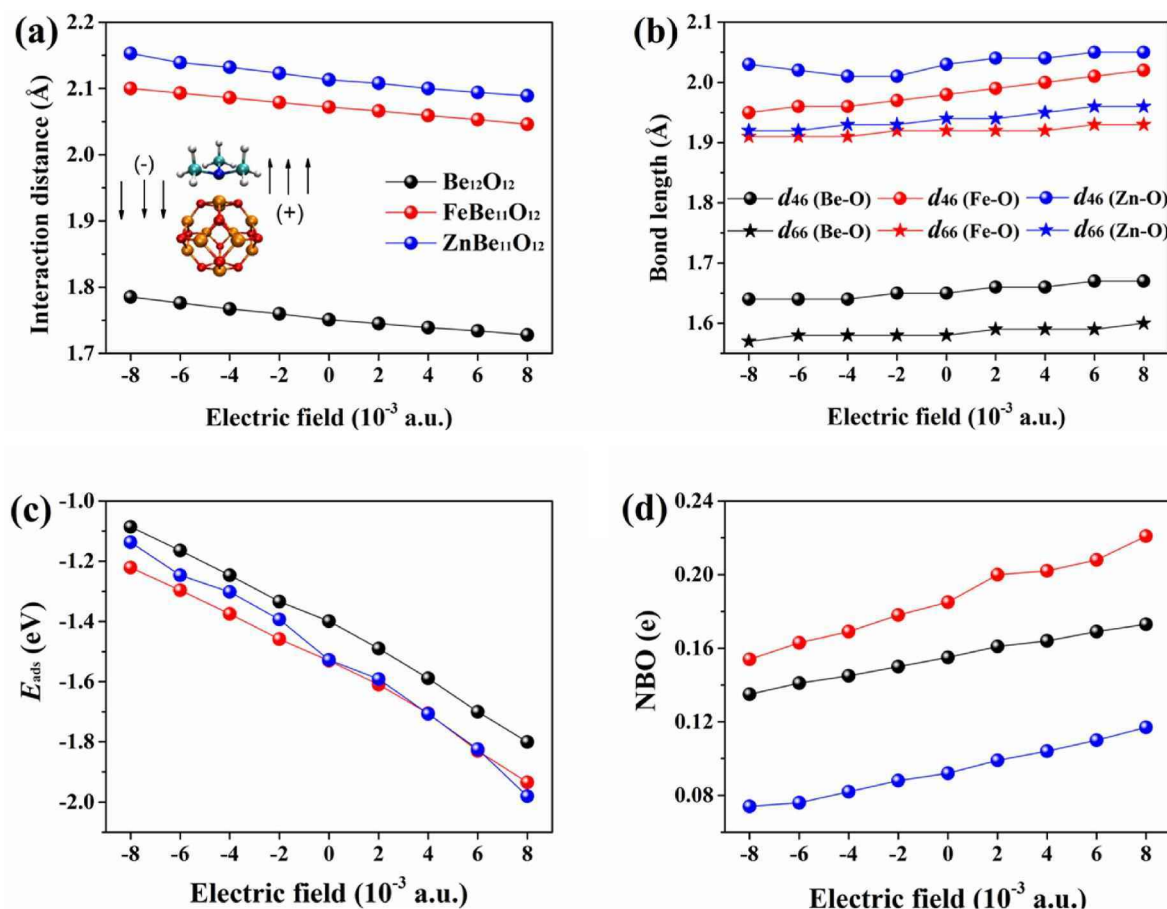


Fig. 8. (a) Interaction distance, (b) bond length, (c) adsorption energy and (d) natural bond orbital charge analysis of adsorption systems under the applied electric field.

Table 4

The recovery time (unit: s) of nanocage/TMA complexes at different temperatures.

| Substrate | 298 K | 398 K | 498 K |
|---------------------------------|----------------------|-------------------|-------------------|
| $\text{Be}_{12}\text{O}_{12}$ | 3.1×10^{11} | 2.2×10^5 | 45 |
| $\text{FeBe}_{11}\text{O}_{12}$ | 7.3×10^{13} | 9.7×10^6 | 9.7×10^2 |
| $\text{ZnBe}_{11}\text{O}_{12}$ | 4.9×10^{13} | 1.3×10^7 | 1.1×10^3 |

CRediT authorship contribution statement

Yuanchao Li: Writing - original draft, Methodology, Investigation. **Jing Sun:** Writing - review & editing, Visualization, Supervision, Formal analysis. **Cuijuan Jiang:** Formal analysis, Investigation. **Xiliang Yan:** Writing - review & editing, Visualization, Supervision, Software, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgements

This research was supported by the National Natural Science Foundation of China (22106025), the Basic and Applied Basic Research Foundation of Guangzhou, China (202201010541).

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.mtsust.2024.100668>.

References

- [1] S. Baachouai, S. Aldulajjan, F. Raouafi, R. Besbes, L. Sementa, A. Fortunelli, N. Raouafi, A. Dhouib, Pristine graphene covalent functionalization with aromatic aziridines and their application in the sensing of volatile amines—an ab initio investigation, *RSC Adv.* 11 (2021) 7070–7077.
- [2] X. Li, L. Jin, A. Ni, L. Zhang, L. He, H. Gao, P. Lin, K. Zhang, X. Chu, S. Wang, Tough and antifreezing MXene@Au hydrogel for low-temperature trimethylamine gas sensing, *ACS Appl. Mater. Interfac.* 14 (2022) 30182–30191.
- [3] E.-X. Chen, H.-R. Fu, R. Lin, Y.-X. Tan, J. Zhang, Highly selective and sensitive trimethylamine gas sensor based on cobalt imidazolate framework material, *ACS Appl. Mater. Interfac.* 6 (2014) 22871–22875.
- [4] Y.H. Cho, Y.C. Kang, J.-H. Lee, Highly selective and sensitive detection of trimethylamine using WO_3 hollow spheres prepared by ultrasonic spray pyrolysis, *Sensor. Actuator. B Chem.* 176 (2013) 971–977.
- [5] S. Zhai, X. Jiang, D. Wu, L. Chen, Y. Su, H. Cui, F. Wu, Single Rh atom decorated pristine and S-defected PdS_2 monolayer for sensing thermal runaway gases in a lithium-ion battery: a first-principles study, *Surface. Interfac.* 37 (2023) 102735.
- [6] M. Sun, A. Dong, Y. Gui, Gas-sensing properties of Pb, Pd modified C_3N_4 for SF_6 decomposition products detection: a DFT study, *Chem. Phys.* 570 (2023) 111898.
- [7] Y. Wang, Y. Gui, S. He, J. Yang, Hybrid nanogenerator driven self-powered SO_2/F_2 sensing system based on $\text{TiO}_2/\text{Ni}/\text{C}$ composites at room temperature, *Sensor. Actuator. B Chem.* 377 (2023) 133053.

- [8] H. Wu, Y. Xia, C. Zhang, S. Xie, S. Wu, H. Cui, Adsorptions of $C_5F_{10}O$ decomposed compounds on the Cu-decorated NiS_2 monolayer: a first-principles theory, *Mol. Phys.* 121 (2023) e2163715.
- [9] S. He, Y. Gui, Y. Wang, J. Yang, A self-powered β -Ni(OH) $_2$ /MXene based ethanol sensor driven by an enhanced triboelectric nanogenerator based on β -Ni(OH) $_2$ @PVDF at room temperature, *Nano Energy* 107 (2023) 108132.
- [10] A. Allangawi, M.A. Gilani, K. Ayub, T. Mahmood, First row transition metal doped $B_{12}P_{12}$ and $Al_{12}P_{12}$ nanocages as excellent single atom catalysts for the hydrogen evolution reaction, *Int. J. Hydrogen Energy* 48 (2023) 16663–16677.
- [11] E. Hosseinzadeh, A. Foroumadi, L. Firoozpour, A DFT study on the transition metal doped BN and AlN nanocages as a drug delivery vehicle for the cladribine drug, *J. Mol. Liq.* 374 (2023) 121262.
- [12] A. Wang, J. Cui, L. Zhang, L. Liang, Y. Cao, Q. Liu, Monitoring of COS, SO $_2$, H $_2$ S, and CS $_2$ gases by $Al_{24}P_{24}$ nanoclusters: a DFT inspection, *J. Mol. Model.* 29 (2023) 98.
- [13] S. Hussain, S.A. Shahid Chatha, A.I. Hussain, R. Hussain, M.Y. Mehboob, T. Gulzar, A. Mansha, N. Shahzad, K. Ayub, Designing novel Zn-decorated inorganic $B_{12}P_{12}$ nanoclusters with promising electronic properties: a step forward toward efficient CO $_2$ sensing materials, *ACS Omega* 5 (2020) 15547–15556.
- [14] L.-K. Li, Y.-Q. Ma, K.-N. Li, W.-L. Xie, B. Huang, Structural and electronic properties of H $_2$, CO, CH $_4$, NO, and NH $_3$ adsorbed onto $Al_{12}Si_{12}$ nanocages using density functional theory, *Front. Chem.* 11 (2023).
- [15] T. Condon-Baxendale, N. Ploysongsri, M. Petchmark, V. Ruangpornvisuti, Adsorption, sensing and catalytic properties of the pristine $C_{24}N_{24}$ nanocage to small gas molecules: a DFT-D3 investigation, *Vacuum* 209 (2023) 111798.
- [16] M. Mohammad Alizadeh, F. Salimi, G. Ebrahimzadeh-Rajaei, Sensing of sarin nerve agent by BN nanoclusters: DFT and TDDFT calculation, *Braz. J. Phys.* 52 (2022) 56.
- [17] S. Abdalkareem Jasim, F.H. Alsultany, M.Z. Mahmoud, D. Olegovich Bokov, W. Suksatan, Investigations of chemical sensing properties of $Al_{24}N_{24}$, $B_{24}N_{24}$, and $B_{24}P_{24}$ nanoclusters toward carbamazepine: a DFT study, *Inorg. Chem. Commun.* 142 (2022) 109644.
- [18] R. Kartika, F.H. Alsultany, A. Turki Jalil, M.Z. Mahmoud, M.N. Fenjan, H. Rajabzadeh, $Ca_{12}O_{12}$ nanocluster as highly sensitive material for the detection of hazardous mustard gas: density-functional theory, *Inorg. Chem. Commun.* 137 (2022) 109174.
- [19] J. Beheshtian, I. Ravaei, Hydrogen storage by BeO nano-cage: a DFT study, *Appl. Surf. Sci.* 368 (2016) 76–81.
- [20] X. Chen, Z. Sun, H. Zhang, S. Onori, Effect of metal atoms on the electronic properties of metal oxide nanoclusters for use in drug delivery applications: a density functional theory study, *Mol. Phys.* 118 (2020) e1692150.
- [21] L. Xu, Y. Gui, W. Li, Q. Li, X. Chen, Gas-sensing properties of Ptn-doped WSe_2 to SF $_6$ decomposition products, *J. Ind. Eng. Chem.* 97 (2021) 452–459.
- [22] H. Cui, C. Yan, P. Jia, W. Cao, Adsorption and sensing behaviors of SF $_6$ decomposed species on Ni-doped C3N monolayer: a first-principles study, *Appl. Surf. Sci.* 512 (2020) 145759.
- [23] J. Zhang, W. Feng, Y. Zhang, W. Zeng, Q. Zhou, Gas-sensing properties and first-principles comparative study of metal (Pd, Pt)-decorated $MoSe_2$ hierarchical nanoflowers for efficient SO $_2$ detection at room temperature, *J. Alloys Compd.* 968 (2023) 172006.
- [24] M. Wang, Q. Zhou, W. Zeng, Theoretical study on adsorption of SF $_6$ decomposition gas in GIS gas cell based on intrinsic and Ni-doped $MoTe_2$ monolayer, *Appl. Surf. Sci.* 591 (2022) 153167.
- [25] J. Chen, L. Jia, X. Cui, W. Zeng, Q. Zhou, Adsorption and gas-sensing properties of SF $_6$ decomposition components (SO $_2$, SOF $_2$ and SO $_2$ F $_2$) on Co or Cr modified GeSe monolayer: a DFT study, *Mater. Today Chem.* 28 (2023) 101382.
- [26] J. Wang, Q. Zhou, Z. Wei, L. Xu, W. Zeng, Experimental and theoretical studies of Zn-doped MoO_3 hierarchical microflower with excellent sensing performances to carbon monoxide, *Ceram. Int.* 46 (2020) 29222–29232.
- [27] Z. Xiao, W. Wu, X. Wu, Y. Zhang, Adsorption of NO $_2$ on monolayer MoS_2 doped with Fe, Co, and Ni, Cu: a computational investigation, *Chem. Phys. Lett.* 755 (2020) 137768.
- [28] Y. Li, X. Li, Y. Xu, The sensing mechanism of pristine and transition metals doped $Zn_{12}O_{12}$, $Sn_{12}O_{12}$ and $Ni_{12}O_{12}$ nanocages towards NH $_3$ and PH $_3$: a DFT study, *J. Mater. Chem. C* 9 (2021) 17382–17391.
- [29] C. Lee, W. Yang, R.G. Parr, Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density, *Phys. Rev. B* 37 (1988) 785–789.
- [30] A.D. Becke, A new mixing of Hartree-Fock and local density-functional theories, *J. Chem. Phys.* 98 (1993) 1372–1377.
- [31] A.D. McLean, G.S. Chandler, Contracted Gaussian basis sets for molecular calculations. I. Second row atoms, Z=11–18, *J. Chem. Phys.* 72 (2008) 5639–5648.
- [32] L.E. Roy, P.J. Hay, R.L. Martin, Revised basis sets for the LANL effective core potentials, *J. Chem. Theor. Comput.* 4 (2008) 1029–1031.
- [33] W. Pipornpong, S. Phunnarungsi, V. Ruangpornvisuti, DFT investigation on adsorption of di-, tri- and tetra-atomic gases on Sc-doped ZnO sodalite like cage for gas sensing purpose, *Mater. Chem. Phys.* 217 (2018) 63–73.
- [34] M.D. Mohammadi, H.Y. Abdullah, V. Kalamse, A. Chaudhari, Adsorption of alkali and alkaline earth ions on nanocages using density functional theory, *Comput. Theor. Chem.* 1204 (2021) 113391.
- [35] K. Chukwuemeka, H. Louis, I. Benjamin, P.A. Nyong, E.U. Ejiofor, E.A. Eno, A.-L. E. Manicun, Therapeutic potential of $B_{12}N_{12}$ -X (X = Au, Os, and Pt) nanostructured as effective fluorouracil (5Fu) drug delivery materials, *ACS Appl. Bio Mater.* 6 (2023) 1146–1160.
- [36] T. Afshari, M. Mohsenia, Transition metals doped ZnO nanocluster for ethylene oxide detection, A DFT study 42 (2019) 113–120.
- [37] M. Rouhani, Density functional theory study towards capability of Ga-doped boron nitride nanosheet as a nanocarrier for 3-allyl-2 selenohydantoin anticancer drug delivery, *Phys. E Low Dimens. Syst. Nanostruct.* 126 (2021) 114437.
- [38] A. Saberinasab, H. Raissi, H. Hashemzadeh, Molecular insight into the role of polyethylene glycol and cholesterol on the performance of graphene-based nanomaterials in blood-brain barrier delivery, *J. Mol. Liq.* 341 (2021) 117446.
- [39] N. Dastani, A. Arab, H. Raissi, DFT study of Ni-doped graphene nanosheet as a drug carrier for multiple sclerosis drugs, *Comput. Theor. Chem.* 1196 (2021) 113114.
- [40] S.F. Boys, F. Bernardi, The calculation of small molecular interactions by the differences of separate total energies. Some procedures with reduced errors, *Mol. Phys.* 19 (1970) 553–566.
- [41] T. Lu, F. Chen, Multiwfn: a multifunctional wavefunction analyzer, *J. Comput. Chem.* 33 (2012) 580–592.
- [42] W. Humphrey, A. Dalke, K. Schulten, VMD: visual molecular dynamics, *J. Mol. Graph.* 14 (1996) 33–38.
- [43] M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, G. Scalmani, V. Barone, G.A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B.G. Janesko, R. Gomperts, B. Mennucci, H.P. Hratchian, J. V. Ortiz, A.F. Izmaylov, J.L. Sonnenberg, Williams, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V.G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J.A. Montgomery Jr., J.E. Peralta, F. Ogliaro, M.J. Bearpark, J. J. Heyd, E.N. Brothers, K.N. Kudin, V.N. Staroverov, T.A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A.P. Rendell, J.C. Burant, S.S. Iyengar, J. Tomasi, M. Cossi, J.M. Millam, M. Klene, C. Adamo, R. Cammi, J.W. Ochterski, R.L. Martin, K. Morokuma, O. Farkas, J.B. Foresman, D.J. Fox, Gaussian 16, Rev. C.01, Wallingford, CT, 2016.
- [44] D. Zhao, Y. Li, M. Xu, Z. Li, H. Zhang, L. Yu, Identification of sulfur gases (environmental pollution) by BeO fullerenes: a DFT study, *J. Mol. Liq.* 343 (2021) 117528.
- [45] H.M. Badran, K.M. Eid, S. Baskoutas, H.Y. Ammar, $Mg_{12}O_{12}$ and $Be_{12}O_{12}$ Nanocages as Sorbents and Sensors for H $_2$ S and SO $_2$ Gases, A Theoretical Approach, *Nanomaterials*, 2022.
- [46] S. Ma, L. Yi, Z. Wu, Metal oxide (BeO-MgO-ZnO) nanoclusters as drug delivery systems for isoniazid anticancer drug: a DFT study, *Mol. Phys.* 120 (2022) e1986162.
- [47] T. Yang, S. Yang, W. Jin, Y. Zhang, N. Barsan, A. Hemeryck, S. Wageh, A.A. Al-Ghamdi, Y. Liu, J. Zhou, W. Chen, H. Zhang, Density functional investigation on α -MoO $_3$ (100): amines adsorption and surface chemistry, *ACS Sens.* 7 (2022) 1213–1221.
- [48] Q.a. Zhou, C. Zheng, L. Zhu, J. Wang, Tin sulfides heterostructure modified quartz crystal microbalance sensors with high sensitivity for hazardous trimethylamine gas, *Sens. Actuatur. B Chem.* 371 (2022) 132520.
- [49] S. Baachouai, L. Sementa, R. Hajlaoui, A. Fortunelli, A. Dhouib, N. Raouafi, Tailoring graphene functionalization with organic residues for selective sensing of nitrogenated compounds: structure and transport properties via QM simulations, *J. Phys. Chem. C* 127 (2023) 15474–15485.
- [50] E. Lee, Y.S. Yoon, D.-J. Kim, Two-dimensional transition metal dichalcogenides and metal oxide hybrids for gas sensing, *ACS Sens.* 3 (2018) 2045–2060.
- [51] J. Choi, Y.-J. Kim, S.-Y. Cho, K. Park, H. Kang, S.-J. Kim, H.-T. Jung, In situ formation of multiple Schottky barriers in a Ti $_3$ C $_2$ MXene film and its application in highly sensitive gas sensors, *Adv. Funct. Mater.* 30 (2020) 2003998.
- [52] X. Chang, X. Liu, W. Zheng, L. Zhou, J. Zhang, Monolayer fullerene network: a promising material for VOCs sensor, *Appl. Surf. Sci.* 637 (2023) 157909.
- [53] R. Lv, X. Huang, W. Ye, J.H. Aheto, H. Xu, C. Dai, X. Tian, Research on the reaction mechanism of colorimetric sensor array with characteristic volatile gases-TMA during fish storage, *J. Food Process. Eng.* 42 (2019) e12952.
- [54] X. Tian, X. Cui, Y. Xiao, T. Chen, X. Xiao, Y. Wang, Pt/MoS $_2$ /Polyaniline nanocomposite as a highly effective room temperature flexible gas sensor for ammonia detection, *ACS Appl. Mater. Interfaces* 15 (2023) 9604–9617.
- [55] T. Lu, Q. Chen, Independent gradient model based on Hirshfeld partition: a new method for visual study of interactions in chemical systems, *J. Comput. Chem.* 43 (2022) 539–555.
- [56] M. Luo, Z. Liang, S. Gouse Peera, M. Chen, C. Liu, H. Yang, J. Liu, U. Pramod Kumar, T. Liang, Theoretical study on the adsorption and predictive catalysis of MnN $_4$ embedded in carbon substrate for gas molecules, *Appl. Surf. Sci.* 525 (2020) 146480.
- [57] S. Demir, M.F. Fellah, A DFT study on Pt doped (4,0) SWCNT: CO adsorption and sensing, *Appl. Surf. Sci.* 504 (2020) 144141.
- [58] X. He, Y. Gui, J. Xie, X. Liu, Q. Wang, C. Tang, A DFT study of dissolved gas (C $_2$ H $_2$, H $_2$, CH $_4$) detection in oil on CuO-modified BNNT, *Appl. Surf. Sci.* 500 (2020) 144030.
- [59] D. Lu, L. Huang, J. Zhang, Y. Zhang, W. Feng, W. Zeng, Q. Zhou, Rh- and Ru-modified InSe monolayers for detection of NH $_3$, NO $_2$, and SO $_2$ in agricultural greenhouse: a DFT study, *ACS Appl. Nano Mater.* 6 (2023) 14447–14458.
- [60] S.U.D. Shamim, D. Roy, S. Alam, A.A. Piya, M.S. Rahman, M.K. Hossain, F. Ahmed, Doubly doped graphene as gas sensing materials for oxygen-containing gas molecules: a first-principles investigation, *Appl. Surf. Sci.* 596 (2022) 153603.
- [61] P.C.D. Mendes, V.K. Ocampo-Restrepo, J.L.F. Da Silva, Ab initio investigation of quantum size effects on the adsorption of CO $_2$, CO, H $_2$ O, and H $_2$ on transition-metal particles, *Phys. Chem. Chem. Phys.* 22 (2020) 8998–9008.
- [62] Y. Wu, X. Chen, K. Weng, Arramel, J. Jiang, W.-J. Ong, P. Zhang, X. Zhao, N. Li, Highly sensitive and selective gas sensor using heteroatom doping graphdiyne: a DFT study, *Adva Electron. Mat.* 7 (2021) 2001244.

- [63] M. Eslamian, A. Salehi, E. Nadimi, The role of oxygen vacancies on SnO₂ surface in reducing cross-sensitivity between ambient humidity and CO: a first principles investigation, *Surf. Sci.* 708 (2021) 121817.
- [64] Y. Li, Y. Meng, X. Li, J. Sun, X. Li, FeN₄-embedded warped nanographene as a potential candidate for scavenging and detecting sulfur-based gases: a DFT study, *J. Environ. Chem. Eng.* 11 (2023) 109705.
- [65] H. Xu, X. Tu, X. Wang, X. Liu, G. Fan, Theoretical study of the adsorption and sensing properties of pure and metal doped C₂₄N₂₄ fullerene for its potential application as high-performance gas sensor, *Mater. Sci. Semicond. Process.* 134 (2021) 106035.
- [66] D. Cortés-Arriagada, N. Villegas-Escobar, A DFT analysis of the adsorption of nitrogen oxides on Fe-doped graphene, and the electric field induced desorption, *Appl. Surf. Sci.* 420 (2017) 446–455.
- [67] S. Yang, G. Lei, H. Xu, B. Xu, H. Li, Z. Lan, Z. Wang, H. Gu, A DFT study of CO adsorption on the pristine, defective, In-doped and Sb-doped graphene and the effect of applied electric field, *Appl. Surf. Sci.* 480 (2019) 205–211.
- [68] J. Wang, X. Zhang, L. Liu, Z. Wang, Dissolved gas analysis in transformer oil using Ni-Doped GaN monolayer: a DFT study, *Superlattice. Microst.* 159 (2021) 107055.
- [69] H.-B. Li, Y.-T. Feng, Z.-G. Shao, C.-L. Wang, L. Yang, First-principles study of CO gas adsorption on pristine and Fe-doped H₄,4,4-graphyne, *Appl. Surf. Sci.* 586 (2022) 152749.
- [70] T. Gakhar, Y. Rosenwaks, A. Hazra, Fullerene (C₆₀) functionalized TiO₂ nanotubes for conductometric sensing of formaldehyde, *Sensor. Actuator. B Chem.* 364 (2022) 131892.
- [71] Z. Zhang, C. Yue, D. Dastan, D. Zhang, X. Zhang, X.-T. Yin, X. Ma, High response and selectivity of bimetallic MOFs-derived metal oxides Co₃O₄/In₂O₃ nanoparticles to TEA, *Sensor. Actuator. B Chem.* 398 (2024) 134727.
- [72] S. Mobtakeri, S. Habashyani, Ö. Çoban, H.F. Budak, A.E. Kasapoğlu, E. Gür, Effect of growth pressure on sulfur content of RF-magnetron sputtered WS₂ films and thermal oxidation properties of them toward using Pd decorated WO₃ based H₂ gas sensor, *Sensor. Actuator. B Chem.* 381 (2023) 133485.
- [73] H. Liu, Z. Wang, C. Sun, J. Shao, Z. Li, H. Zhang, M. Qiu, G. Pan, X. Yang, Construction of Co₃O₄/Fe₃O₄ heterojunctions from metal organic framework derivatives for high performance toluene sensor, *Sensor. Actuator. B Chem.* 375 (2023) 132863.



Reaching the Full Potential of Machine Learning in Mitigating Environmental Impacts of Functional Materials

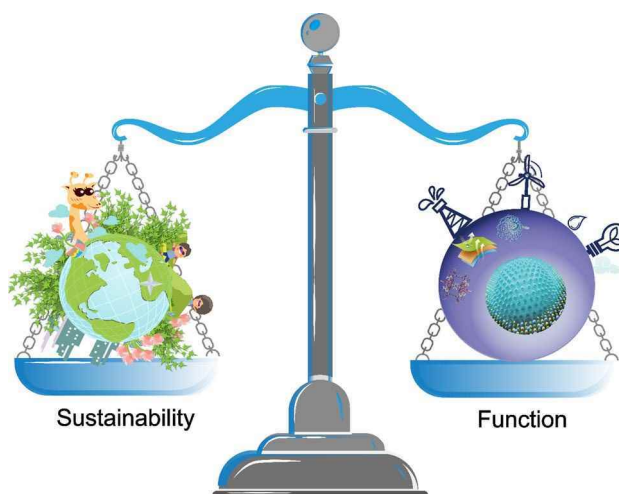
Ying He¹ · Guohong Liu^{1,2} · Chengjun Li^{1,2} · Xiliang Yan^{1,2} 

Received: 8 November 2022 / Accepted: 2 December 2022
© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

Abstract

In conventional ways of functional material design, the performance of synthesized materials is the focal point, whereas the toxicity of and environmental problems caused by synthesized materials are neglected to a large extent. Only with a balanced consideration of all the above-mentioned factors can we ensure the development of eco-friendly functional materials. In recent years, with big data generated by experiments and computing technology becoming increasingly accessible, data-driven solutions, especially machine learning methods have opened a new window for the discovery and rational design of eco-friendly functional materials. In this review, we first presented a brief introduction of functional materials, the most commonly used machine learning models and relevant processes. The applications of ML-based approaches and computational methods in functional prediction and material design were then summarized. More importantly, by combining machine learning methods with the toxicity prediction of functional materials, we proposed a framework for sustainable functional material design to achieve better functionality and eco-friendliness. Such a framework will ensure both the practicability and effectiveness of functional materials, balance their functionality and environmental sustainability, and eventually pave the path toward the Sustainable Development Goals set by the United Nations.

Graphical Abstract



Introduction

Functional materials are a group of materials with unique properties designed for specific purposes (Beaujuge and Fréchet 2011). The earliest concept of functional materials was proposed by Morton of the Bell Institute in 1965.

✉ Chengjun Li
cli@gzhu.edu.cn

✉ Xiliang Yan
yanxiliang1991@gzhu.edu.cn

Extended author information available on the last page of the article

Functional materials can be broadly divided into four categories, i.e., mechanical, chemical, physical, and biological materials, which include nine subcategories in a more sophisticated classification manner according to their properties (Fig. 1) (Jelliarko et al. 2010). Functional materials act as important solutions to many challenges in personalized healthcare, energy production, and storage sectors (Song et al. 2019). In recent years, new functional materials have been emerging endlessly with groundbreaking progress. These include superconducting materials (Polichetti et al. 2021), microelectronic materials (Yan et al. 2020a, b, c), photonic materials, information materials, energy conversion and energy storage materials, eco-environmental materials, and biomedical materials (Yu et al. 2012). Such emerging functional materials have a wide range of properties and promising applications, forming a broad market prospect and contributing to an extremely important strategic significance (Lin et al. 2020).

Despite significant economic and social benefits, functional materials and their development have caused environmental problems, leading to a shift in the focus of material design from performance to sustainability. With advances in chemistry, engineering, energy, materials science, and other related fields, a large number of new structural and functional materials have been developed. This depletes non-reusable natural resources and brings serious environmental pollution problems, threatening the sustainable development

of human society (Plata and Janković 2021). The sustainability of functional materials, therefore, is becoming an increasingly important factor involved in developing functional materials (Chen et al. 2022a, b). Such a change warrants higher standards for the sustainability and environmental safety of functional materials. This includes good performance, a high-resource utilization rate, and low/no negative impacts on the environment. In recent years, the research hotspots in functional materials for sustainable development mainly focus on environmental pollution mitigation, the reduction of the generation of harmful substances and waste, and the improvement of resource utilization. Replacing toxic materials with low-toxic or non-toxic materials for sustainable development is becoming a trending topic in the development of functional materials (Chen et al. 2022a, b).

Developing new functional materials is time-consuming and labor-intensive, which is further complicated by the requirements of sustainability. It usually takes several years to develop functional materials from design and testing to final applications, especially in demanding products. There are three major obstacles restricting rapid development of functional materials, i.e., complex structure, high level of required experience and knowledge, and sheer volume of involved trials. The structure of many materials is very complex, and the analysis of the relationship often requires a large number of tests and data support (Palmer et al. 2018). The development of materials also depends on researchers' experience and judgment (Wellmann 2021). Each process requires personal knowledge and understanding of design, implementation, analysis and test data interpretation, and decision-making (Duan et al. 2022). Due to our limited understanding of the structure and properties of materials, traditional development of functional materials adopts the trial-and-error method based on experience. It constantly strides forward to the correct goal through attempts and repeated experiments. Only preliminary trial production of materials can be completed in the laboratory. Repeated experiments are needed to accurately determine the process and operating parameters and obtain functional materials with stable performance.

Machine learning may revolutionize the way of functional material design. When machine learning is applied, researchers only need to propose material performance requirements, design and calculation schemes, and analyze calculation results. Machine learning can automatically generate a large number of candidate materials and compute tasks according to the demand and material database (Kim et al. 2020). It can also help reduce unnecessary experimentation (Wahl et al. 2021) and even predict material properties under extreme conditions that cannot be achieved by experiments (e.g., high temperature, high pressure, strong field, ultrafast, and radiation) (Wang and Ma 2013; Fang et al. 2022). More importantly, machine

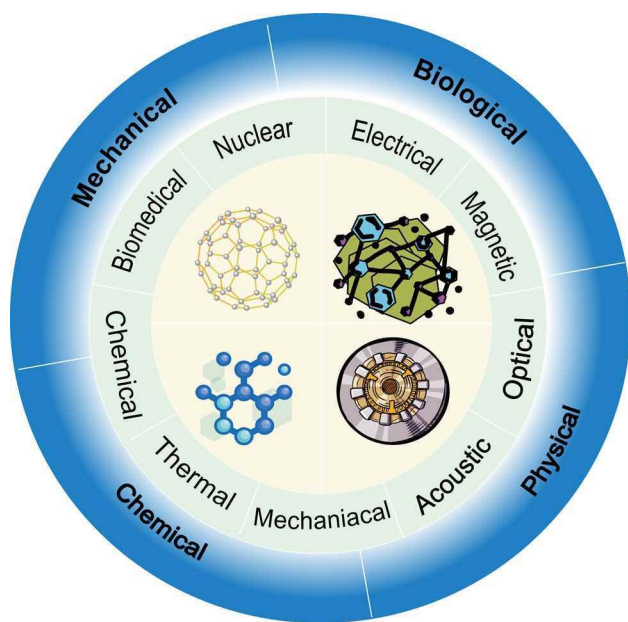


Fig. 1 Classification of functional materials and representative materials. Functional materials include mechanical, biological, chemical, and physical materials, which can be further divided into nine categories according to their specific properties. Carbon nanotubes, ion exchange columns, graphene, and biochips are displayed as representatives of four different functions

learning also has the potential to accelerate the process of functional material design with sustainability. Functional material design with sustainability requires systematic data research across at least two disciplines, one focusing on functional performance and the other on environmental performance. Conventional ways of functional material design can only focus on these two tasks separately, whereas machine learning can achieve both goals simultaneously.

In this study, we summarized the research progress of machine learning on functional material design, structure prediction, and toxicity prediction, emphasizing the importance of filling the research gaps in utilizing machine learning to mitigate the environmental impacts of functional materials and develop next-generation eco-friendly functional materials. To help achieve the above-mentioned goals, a framework has been proposed in this study as a promising standard operation procedure to consider both the performance of materials and the sustainable development of the environment by using machine learning. This framework can improve the efficiency of the development of sustainable functional materials, reduce their development costs, and most importantly help achieve the 2030 Agenda for Sustainable Development adopted by all United Nations Member States in 2015.

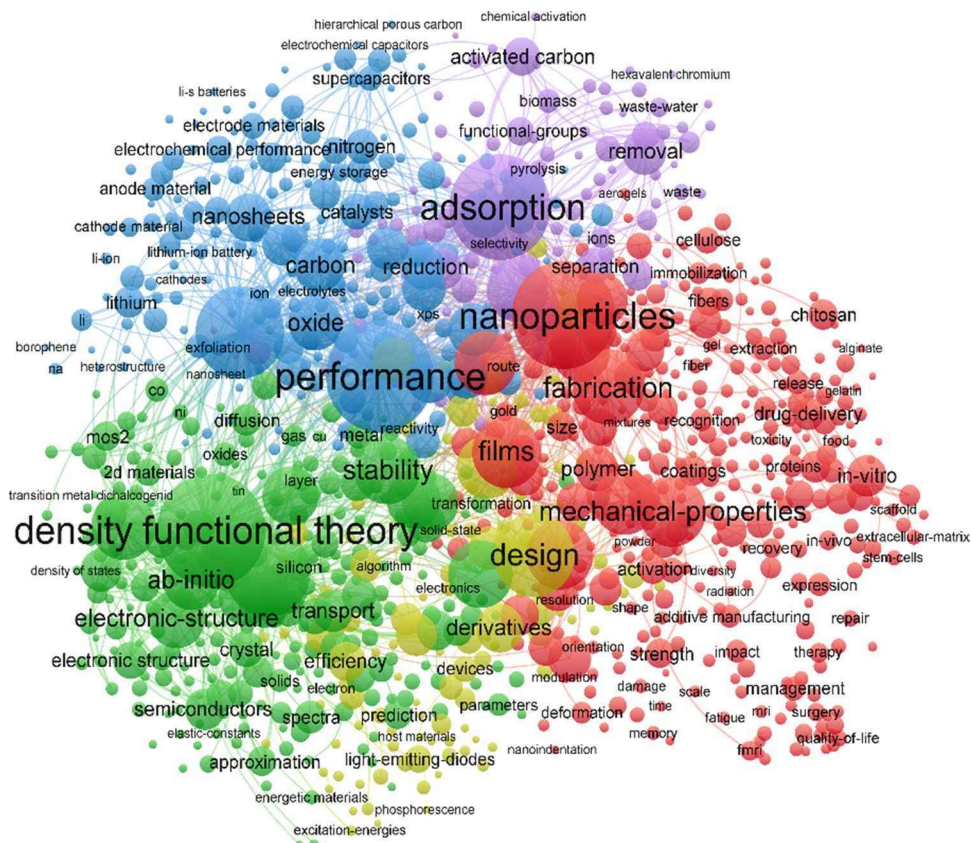
An Imbalance Between Performance and Environmental Risks

In the conventional process of developing functional materials, the priority is always to achieve the best performance, while not much attention has been paid to their potential environmental risks (Rydz et al. 2014). Due to the over-exploitation of limited natural resources and the deterioration of the environment, material scientists should actively explore ways to not only ensure the performance of materials, but also achieve environmental safety, and sustainable and healthy development (Kuznetsov and Edwards 2010).

The Status Quo of Research on Functional Materials

In the past four decades, more than 100,000 papers on functional materials have been published as shown in the Web of Science Core Collection Database, as of May 2022. Herein, keywords extracted from all published papers were analyzed and visualized using VOSviewer, through which research hotspots and emerging topics and their interactions can be identified (Eck and Waltman 2010). As shown in Fig. 2, the keywords of research hotspots in functional materials include adsorption, nanoparticles, mechanism research, performance, design, and density functional theory.

Fig. 2 The top keywords and their co-occurrence network. Top keywords were extracted from over 100,000 previously published articles (as of May 2022) related to the search topic “functional material” in the Web of Science Core Collection Database. Keywords are grouped into different color-coded categories using VOSviewer, depending on their intrinsic relationship. Dots of the same color form a cluster and five clusters are shown in the network. The font size of each keyword is in proportion to its frequency detected in the keyword and abstract sections of the previous studies. The size of the node indicates the frequency of occurrence. The curves between nodes indicate that they co-occurrence in the same publication. The shorter the distance between nodes, the more times the associated keywords co-occur



The blue cluster shows the hotspots related to the topic of performance, including electronic materials, biomaterials, catalysts, and others. A representative electronic material is NbTi/Nb₃Sn. It has been commercialized and applied in many fields, such as nuclear magnetic resonance human imaging, superconducting magnets, and large accelerator magnets (De Marzi et al. 2013). Biological functional materials are a group of natural or synthetic special functional materials developing rapidly in recent years. They are widely used in our daily life and can help diagnose, repair, replace, or induce the regeneration of human cells, tissues, and organs. Catalysts are used in many industrial processes which can change the reaction speed of a chemical reaction. For complex reactions, catalysts accelerate the rate of the main reaction, inhibit the side reaction, and improve the yield of the target product.

Hotspots in the red cluster are about the shape, size, and property of the material. The shape and size of the material microstructure are various and closely related to its surface morphology. Many factors affect the morphology of materials, including composition, preparation method, heat treatment, and heat processing technology. Material morphology is directly related to material properties.

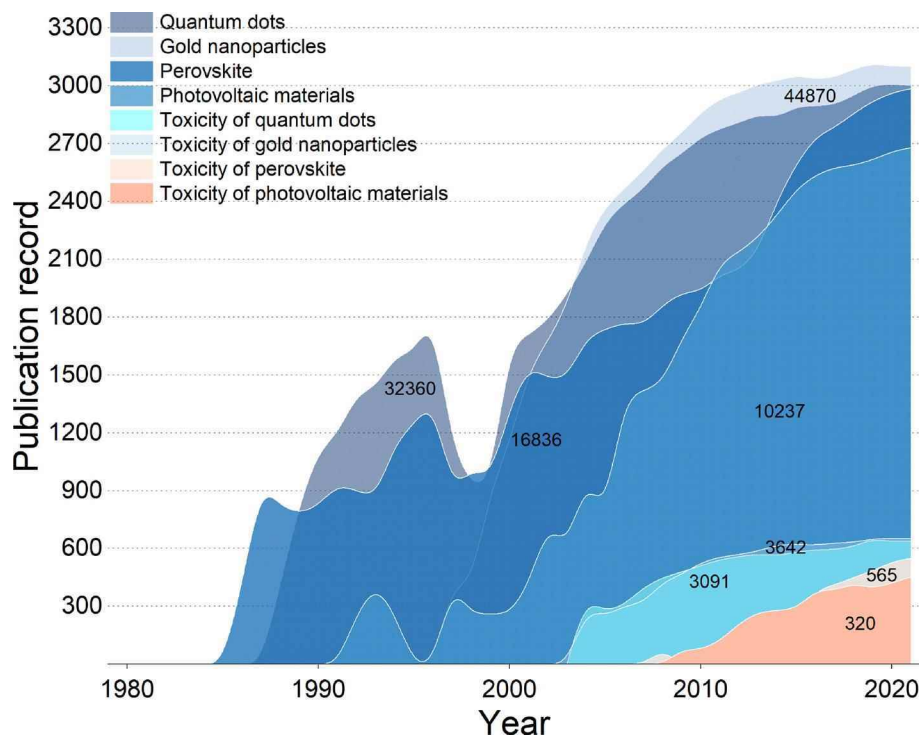
The green cluster is about molecular simulations and quantum chemical calculations. Due to the increasing popularity of computer applications, researchers use computer-aided methods to carry out theoretical calculations. Density functional theory (DFT) is a leading method for electronic structure calculation in many fields, which is often used to

study the properties of molecules and condensed matter (Tsuneda 2020). Ab-initio calculation is another method to solve the Schrodinger equation of ionic systems in computational materials. Computational theory programs are often used to model solid, interface, and surface properties for a variety of material systems, including ceramics, semiconductors, and metals (Lindan 2002).

However, research conducted on the environmental health and safety (EHS) of functional materials is often orders of magnitude less and decades behind (Plata and Janković 2021). As shown in Fig. 3, there were obvious disparities in terms of the number of articles between functional and toxicity research of commonly used functional materials, including Quantum dots (Qds), Gold nanoparticles (AuNPs), Perovskite, and Photovoltaic materials.

Take the research on Qds and gold nanoparticles as an example. In 1983, Bell Laboratory first reported that CdS has a size effect and other related properties, as the prelude to the research of Qds (Rossetti et al. 1983). In 1994, Rossetti et al. published an article on the construction of light-emitting diodes using CdSe, which promoted the application of quantum dots in the field of photoelectric conversion (Colvin et al. 1994). Then, the research of Qds in multi-color labeling, solar cells, fluorescence imaging, and other aspects has gradually increased (Giepmans et al. 2005; Robel et al. 2006; Smith et al. 2008). After 2004, Researchers gradually paid attention to green, low toxicity, strong compatibility, and higher luminous efficiency Qds, e.g., the development of CuInS, MnSe, InP, and other heavy metal-free Qds (Derfus

Fig. 3 Publication number of and disparities between functional and EHS research of functional materials in the past four decades. Disparities between functional and EHS research of functional materials as indicated by the number of publication records. Publications were categorized into eight groups, i.e., quantum dots, gold nanoparticles, perovskite, photovoltaic materials, and the toxicity research of the four materials, respectively. The numbers above each shaded area indicate the publication records of each publication group



et al. 2004). For AuNPs, at the beginning of the nineteenth century, Wallace et al. observed the tobacco mosaic virus labeled with AuNPs with an electron microscope, showing a high electron density and fine particles (Wallace et al. 1972). Now, AuNPs have been widely used in the fields of information, photonics, beauty, and skin care. However, it was not until 2006 that damage to algae and other organisms from AuNPs began to be noticed (Sani et al. 2021).

Functional Materials as Emerging Pollutants

Functional materials can become pollutants without proper EHS assessment, which has long been neglected. Previous functional material design barely considered the toxic effects of functional material on the environment and humans. As a result, some new materials have developed into emerging pollutants in the process of their synthesis, use, and disposal. Environmental pollution and the ecotoxicity of new pollutants have become major global environmental problems. In this section, we present the potential health and environmental toxicity of some functional materials after turning into new pollutants by introducing functional material-originated emerging pollutants such as nanomaterials, perfluorinated alkylated substances (PFAS), and ionic liquids (ILs).

Nanomaterials are defined as materials having a size of roughly 1 to 100 nm or at least one dimension in the nanometer range (Boyes and van Thriel 2020). Widespread applications of nanomaterials have led to the ubiquitous presence of nanoparticles, causing environmental pollution and adverse effects on human health (Pacurari et al. 2016). Nanoparticles (NPs) from engineered nanomaterials are released into the environment during the process of production, use, and disposal. NPs are absorbed by organisms through the respiratory system and attached to the alveoli and larger bronchi (Schneider and Lim 2019). Skin is an important pathway for NPs, which usually occurs in cosmetics containing TiO₂ and ZnO NPs (Xie et al. 2011). Administering and injecting drugs loaded with nanoparticles will also cause acute toxicity, bone marrow toxicity, cytotoxicity, and organotoxicity of organisms. Nanoplastics generated by the degradation of plastics are emerging pollutants. They can easily enter human and animal bodies via diet and respiration, and affect body functions. As an important defense line against alien antigens, immune cells are vulnerable to the attack of nanoplastics (Banerjee and Shelper 2021). Nanoplastics can stay in the soil for hundreds of years through interaction with soil organic debris, thus affecting the physical and chemical properties of the soil and causing groundwater pollution (Allouzi et al. 2021).

PFAS are a group of the most important chemical products in the twentieth century. It is composed of completely fluorinated anions in perfluorinated acid sulfate (Uwayezu et al. 2022). PFAS have the characteristics of oil and water

drainage at the same time, and is widely used in the production of surface antifouling agents such as textiles, leather products, furniture, and carpets (Herzke et al. 2012). Nowadays, they are considered environmental pollutants with systemic multiple-organ toxicity, such as genetic toxicity (Logeshwaran et al. 2021), male reproductive toxicity (Qiu et al. 2013), neurotoxicity (Huang et al. 2021), developmental toxicity (Yilmaz et al. 2020), and endocrine interference (Belchior et al. 2019).

In recent years, ILs, as a new type of green solvent, have been rapidly applied in organic synthesis, electrochemistry, chemistry, pharmaceutical, and biomedical fields due to their high extraction rate, strong solubility, and recyclability (Belchior et al. 2019; Greer et al. 2020; Tiago et al. 2020; Simões et al. 2021). However, researchers found that ILs are highly persistent in aquatic and terrestrial environments due to their stability which threatens the safety of eco-environments and human health (Wei et al. 2021). To sum up, for functional materials, functional performance often comes first in their design while sustainable development is seldomly considered, which will cause environmental problems and threaten human health.

Why We Need Machine Learning in Functional Material Design

With in-depth studies of functional materials, their EHS including environmental and biological effects should not be ignored. EHS is an important prerequisite for the healthy and sustainable development of the functional material industry. In order to achieve this goal, researchers need to constantly innovate research methods. As a data-driven research paradigm, artificial intelligence (AI) is increasingly favored by researchers for its advantages in data analysis (Gupta et al. 2021).

AI is a new cutting-edge science to develop theories, methods, technologies, and application systems for simulating, extending, and expanding human intelligence (Ligeza 1995). Computational problems such as face recognition require a multi-level feature structure, and different features have increasing complexity at a deeper level. At each level, simple data are converted through non-linear functions, and the generated information is further transmitted to display complex output. Such processing procedures of AI resemble the way how a child's brain operates (Fig. 4). Machine learning is a subset of AI that allows "the brain" to master cognitive ability and guide computers to learn from data, and then use the experience to improve performance. Deep learning is a new research direction of machine learning methods, acting as an efficient learning system in processing new types of data such as images, audio, videos, and natural languages (Donahue et al. 2013; Socher 2014; Srivastava et al. 2014).

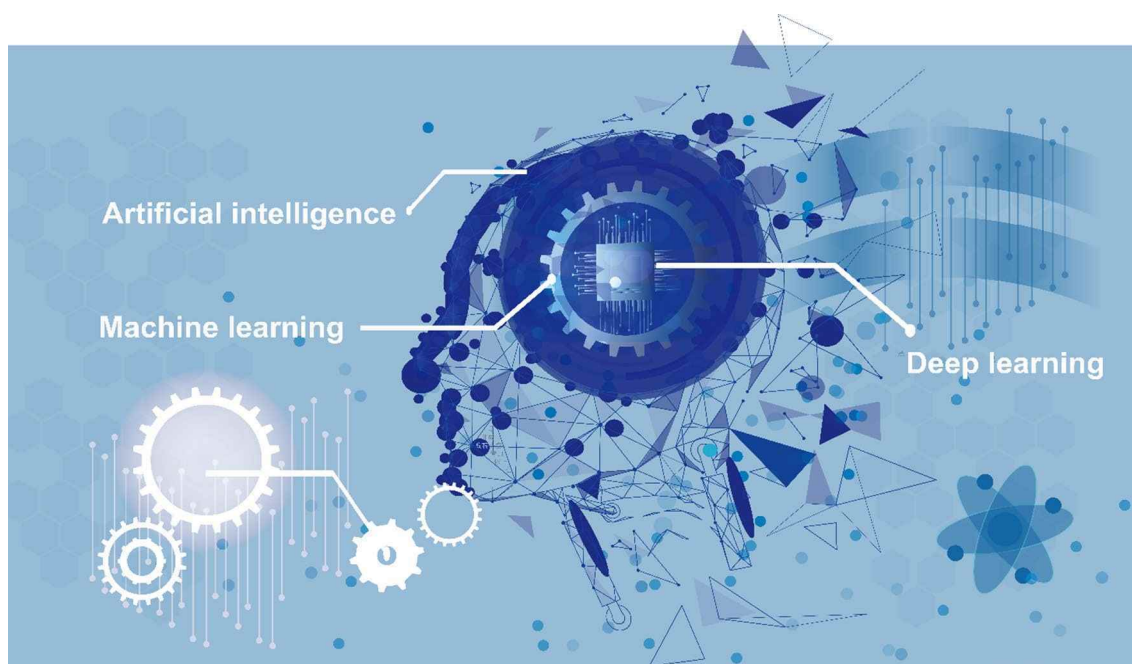


Fig. 4 A concept scheme illustrating the relationship between AI, ML, and deep learning. AI, including ML and deep learning, resembles the way how a child's brain operates. ML, as a subset of AI, enables computers to master cognitive ability and learn from data, and

then use the experience to improve performance. Deep learning is an emerging subset of machine learning that can accurately recognize, classify, and describe objects within the data

The differences between deep learning and machine learning can be concluded mainly from datasets, feature engineering, and manual intervention. Deep learning is applicable to large amounts of data (Miotto et al. 2018). On the other hand, machine learning is suitable for small structured or tokenized datasets. In addition, due to a large amount of data and multiple computing levels, deep learning requires a lot of training time. The training time of machine learning depends on the size of the dataset and the nature of the features considered. Under proper feature engineering, the prediction result of the machine learning model is better. The advantage of deep learning is that, as it continues to penetrate the network, it will automatically identify more advanced basic features. Machine learning decomposes the problem into several parts, and then summarizes all the results into the final output to get the final output. Deep learning solves end-to-end problems (Lloyd et al. 2013). Human intervention in machine learning is limited to providing appropriate data and adjusting models to fit the correct data. On the contrary, the deep learning network does not need such supervision. Convolutional neural network (CNN) is an architecture based on an image deep learning framework and realized by computational mathematics to extract image features. Because of its ability to capture different levels of features, CNN is suitable for describing some complex material properties. In a recent study, CNN was used to predict physicochemical properties (i.e., logP and zeta

potential) and biological activities (i.e., cellular uptake and protein adsorption) of 147 unique nanoparticles (Yan et al. 2020a, b, c). This method can directly learn nanostructure features from the nanoparticle images. No other external information or characteristics of the nanoparticles need to be calculated. This is difficult to achieve with traditional machine learning methods.

A Brief Introduction to Machine Learning Methods

The connotation of machine learning is to guide computers to learn from previous data and experience to improve their performance. In machine learning, the algorithm will be trained many times to find patterns and correlations from large datasets, and then make the best decisions and predictions based on the results of data analysis. Machine learning involves three types of learning: supervised learning (parametric/non-parametric algorithms, support vector machines, kernels, neural networks), unsupervised learning (clustering, dimensionality reduction, recommender systems, deep learning), and reinforcement learning (deep reinforcement learning; deep RL, model-free, model-based) (Lloyd et al. 2013; Anandakumar and Arulmurugan 2019). Both techniques are widely used in different scenarios and with different datasets. Supervised learning includes regression models and classification models. In the design of functional materials, supervised learning can build a new model,

which can predict the properties of new materials according to the functional activity and properties of known materials. If the functional properties of materials are continuous quantities (e.g., the conversion efficiency of organic photovoltaic materials), the process is called regression. If the predicted functional activity is a discrete target (e.g., whether toxic or not), the process is called classification. Figure 5 shows the general classification of machine learning algorithms and typical algorithms within each classification.

Nuts and Bolts of Machine Learning in Designing Functional Materials

The general process of material discovery and activity prediction based on machine learning is mainly divided into four steps as shown in Fig. 6, i.e., (i) data preparation, (ii) descriptor generation and screening, (iii) model construction

and verification, and (iv) material prediction and experimental verification.

Firstly, relevant performance data of the functional materials should be prepared, respectively. Existing databases collect a lot of material information obtained from calculations and experiments, which is widely used in computational materials science. Table 1 lists major publicly available databases containing a large number of material structures and properties. We provide a concise description of the characteristics of each database. In the process of data collection, data from multiple databases are usually selected.

Then, model-related descriptors should be computed and filtered. Material descriptors can be categorized into experimental descriptors and theoretical descriptors (Table 2). The experimental descriptors are all experimental measurements, such as octanol–water partition coefficient, molar refractive index, polarization, and any general physical and chemical

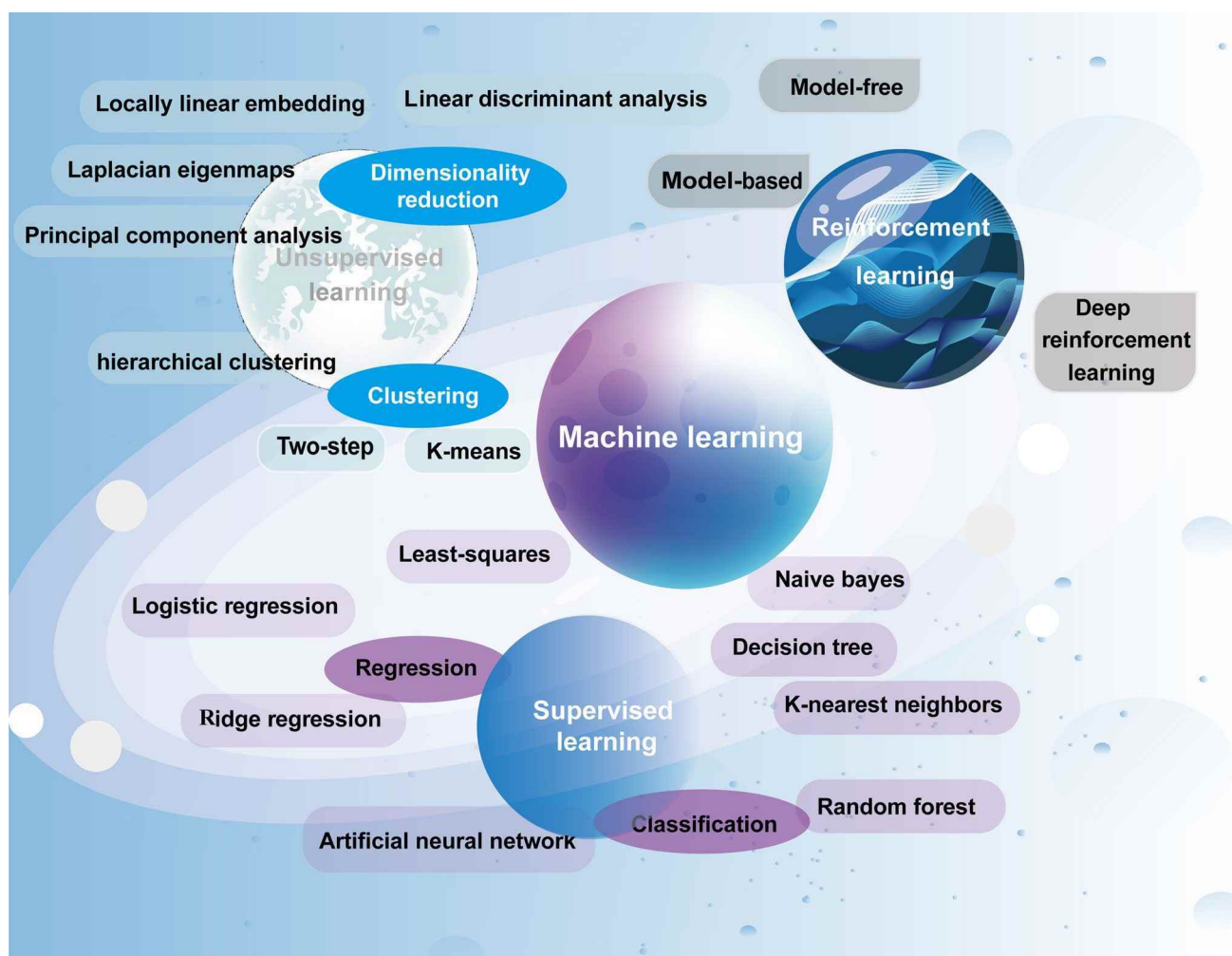


Fig. 5 Classification of important machine learning methods. Machine learning involves three types of learning: supervised learning, unsupervised learning, and reinforcement learning. Unsupervised learning included dimensionality reduction and clustering.

Regression and classification models belong to supervised learning. The main methods of reinforcement learning are model-free, model-based, and deep reinforcement learning. Only commonly used methods for different learning types are shown

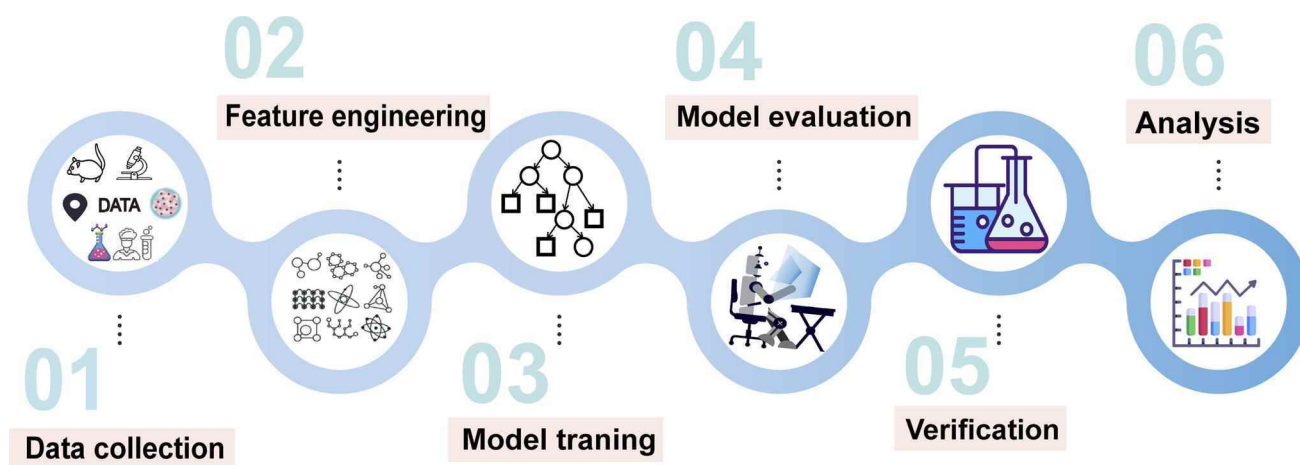


Fig. 6 Generic workflow for functional materials discovery and design based on Machine Learning. The general machine learning process includes data preparation (collection and processing), feature

engineering (descriptor generation and screening), model construction (model training and evaluation), and experimental verification (result analysis)

properties, which are obtained by applying the specified experimental procedures (Cui et al. 2021). By contrast, theoretical molecular descriptors are those descriptors obtained by a well-specified chemoinformatic algorithm applied to an unequivocal molecular representation such as atomic number, molecular weight, and some energy parameters based on molecular dynamics simulations (Adhikari and Mishra 2018). Experimental descriptors are directly measured through characterization experiments under specific conditions, which can more accurately describe the corresponding properties of functional materials (Sahoo et al. 2016). However, the measurement of experimental descriptors is time-consuming and laborious with poor repeatability. Theoretical descriptors are easier to obtain but depend on computational methods (Liang et al. 2017). There are many commercial and freely available softwares for computing theoretical descriptors, some of which have been developed specifically for the computation of molecular descriptors (Table 2).

Finally, after building the model, we need to find the materials with the required properties according to the model and test all possible candidate materials one by one. The best material can be synthesized and its actual performance will be verified by experiments. If the experimental results are consistent with the predicted results, it proves that the model is suitable for the screening of this material.

The Role of Machine Learning in Functional Material Design

Machine learning can facilitate the discovery of new materials (Butler et al. 2018). Traditional methods for material performance prediction are often limited to the

structure and parameters, which restricts the improvement of prediction efficiency (Zheng 2018). Through machine learning, a large number of structure and performance relationships can be analyzed, and relevant models can be obtained to quickly screen unknown materials (Menon and Ranganathan 2022). In recent years, utilizing machine learning to predict the performance of new materials has been increasingly favored by researchers, and a series of achievements have been made in superconductors, magnets, catalysts, alloys, and other materials (Zahrt et al. 2019; Machak et al. 2021).

Unlike the traditional material design that needs a lot of calculation simulation and experimental verification, machine learning can accelerate the calculation without experiments. It can quickly optimize the composition and microstructure of materials, reduce the research and development time from months to a few weeks, and significantly improve the comprehensive performance of materials (Liu et al. 2015). At the same time, with the continuous development of functional material theory and experimental research, the data generated in the experiment and calculation simulation, including failed data, are collected to form large-scale databases. Machine learning can select the optimal chemical reaction route of the optimal synthetic materials through big data learning (Juan et al. 2020). Using machine learning technology can accurately calculate the parameters of each stage of the preparation process, and improve the synthesis speed and accuracy. In general, the main role of machine learning in functional material design is to predict the properties of materials, conduct virtual screening from a large number of candidate materials, and give a reasonable explanation of the model (Fig. 7).

Table 1 Publicly accessible structure and property databases for functional materials

| Source | Database | Description |
|--------------|---|---|
| Experimental | Cambridge structural database (http://webcsd.ccdc.cam.ac.uk/) | Contains organic and metal–organic crystal structures (Groom and Allen 2014) |
| | ChemSpider (http://www.chemspider.com/Default.aspx) | Provides millions of chemical formulas and integrates many online services (Little et al. 2012) |
| | CoRE MOF (https://zenodo.org/record/3370144#.Y1JCvxByUk) | CoRE MOF Datasets are derived from Cambridge Structural Database (CSD) and also from the World Wide Web. It includes the solvent-free atomic coordinates and pore characteristics of metal–organic materials (Nazarian et al. 2016) |
| | Inorganic crystal structure database (https://icsd.products.fiz-karlsruhe.de/) | Includes detailed information on more than 210,000 experimentally characterized inorganic crystal structures published since 1913 (Zagorac et al. 2019) |
| | MatWeb (https://www.matweb.com/) | Includes data on thermoplastic and thermosetting polymers |
| | Pauling file (http://www.paulingfile.com/) | Belongs to the inorganic crystal material database, which contains the phase diagram and physical properties of materials (Nazarian et al. 2016) |
| | Pubchem (https://pubchem.ncbi.nlm.nih.gov/) | Contains a large number of physical and chemical properties of organic molecular materials, and the application of each molecule can be traced back to relatively comprehensive literature reports (Curtarolo et al. 2012) |
| Theoretical | AFLOWlib (http://www.afowlib.org/) | Stores more than 3.56 million material structures and 700 million first-principle calculations including inorganic compounds, binary alloys, and multi-alloys (Wang et al. 2022) |
| | Atomly (https://www.atomly.net/) | Contains 180,000 calculated inorganic crystal structures, detailed electronic structure information, and thermodynamic phase diagram from the ICSD database (Hachmann et al. 2011) |
| | Harvard clean energy project | Includes 2.3 million candidate materials, and provides data support for research on solar cell materials (Jain et al. 2013) |
| | Materials project (https://legacy.materialsproject.org/) | Focuses on supporting the development of advanced materials such as fuel cells, photovoltaics, and thermoelectricity (Upadhyay et al. 2020) |
| | JARVIS-DFT (https://jarvis.nist.gov/) | Focuses DFT prediction of material properties, especially crystalline materials |
| | Computational 2D materials database (https://cmrdb.fysik.dtu.dk/c2db/) | Contains the structural, thermodynamic, elastic, electronic, magnetic, and optical properties of about 4000 two-dimensional (2D) materials (Shen et al. 2021) |
| | The open quantum materials database (https://oqmd.org/materials/) | Includes calculated thermodynamic and structural properties of 1.02 million materials by DFT, among which perovskite data are the most important (Kirklin et al. 2015) |

What Machine Learning Can Do in Functional Material Design

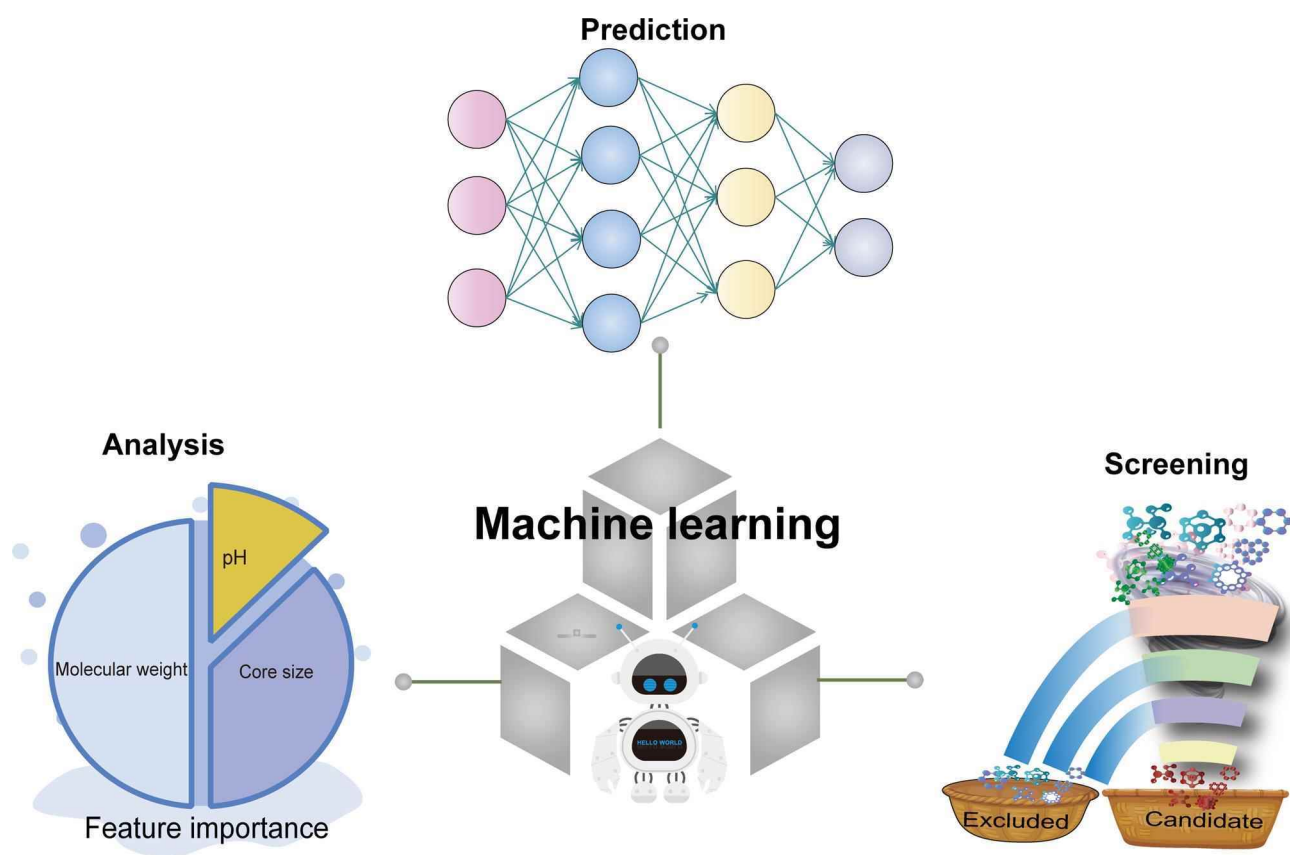
Material science has been a research hotspot in recent years with a large amount of data information obtained through experimental means and computational simulation methods (Pei et al. 2021). With the help of machine learning, effective information can be mined from numerous material data, which plays an important role in the design, synthesis, characterization, and optimization of materials (Jiang et al. 2020).

Guiding Chemical Synthesis

Traditional methods of synthesizing new functional materials often need to design synthesis routes, and thousands of combination schemes may occur in each synthesis step. It is often accompanied by extreme conditions that are hard to achieve. Many factors, including time, cost, and toxicity, need to be considered when designing the synthesis scheme, which is difficult to meet by using traditional methods. Researchers often can only constantly make trial-and-error reactions based on fuzzy theoretical knowledge.

Table 2 A list of classification of descriptors and commonly used generation software

| Classification | Properties | Generation method/software |
|--------------------------|----------------------------------|---|
| Experimental descriptors | Microstructure | Scanning electron microscope |
| | | Liquid chromatography (Zhang et al. 2021) |
| | | Terahertz time-domain spectroscopy |
| | | Phase analysis of X-ray diffraction |
| | Phase structure analysis | Raman spectroscopy |
| Theoretical descriptors | Composition analysis | Nuclear magnetic resonance |
| | | Atomic absorption spectroscopy |
| | | Mass spectrum |
| | | X-ray fluorescence |
| | | DRAGON (Mauri et al. 2006) |
| | Physical and chemical properties | RDKit (Lovri et al. 2019) |
| | | MOE (Qiao and Guo 2005) |
| | | Gaussian (Press 2007) |
| | Quantum chemical properties | VASP (Bear et al. 2002) |
| | | Molpro (Papad and Schaefer 2006) |
| | Molecular dynamics properties | GROMACS (Pronk et al. 2013) |
| | | LAMMPS (Plimpton et al. 2011) |

**Fig. 7** The role of machine learning in functional material mainly includes virtual screening, prediction properties, and model interpretation. Virtual screening removes a large number of candidate struc-

tures. Furthermore, the properties of the material can be predicted by constructing the model and the model can be explained by analyzing the importance of features

Machine learning was first applied by organic chemists in the field of chemical synthesis and has become increasingly popular in recent years. In 1969, Corey and Wipke developed the ‘organic simulation synthesis program’ and tried to automate the synthesis of materials by relying on computers (Corey and Wipke 1969). This was the very first time that the concept and potential of machine learning application in material synthesis were appreciated. In recent years, a large number of studies have shown that the material synthesis route designed by using machine learning can completely replace the traditional empirical method when the conditions and synthesis rules are determined (Pillong et al. 2017; Xu et al. 2018; Qian et al. 2019).

Take the development of metal–organic frameworks (MOFs) as an example. Researchers have been looking for clean energy, and MOFs are expected to be alternative medium materials to convert CO₂. MOFs can be used to capture and convert carbon dioxide in the short term and help to produce and store hydrogen in the long term, and use this material as a tool to eventually form a carbon–neutral energy cycle (Gulati et al. 2023). However, the synthesis of MOFs needs to consider metal oxidation state, potential, ion radius, and other factors (Zhang and Fei 2019). Constant adjustments of temperature, pH, reaction concentration, and other parameters are needed in the synthesis. To tackle this problem, researchers used the random forests or neural networks algorithm to analyze the material morphology under different synthesis schemes and synthesized MOFs in batches (Luo et al. 2022). It was found that the important factors affecting the morphology of MOFs were the concentration of water and formic acid. The correlation function between morphology and experimental conditions was determined by using a random forest algorithm. The optimal conditions for preparing MOFs, therefore, were determined, and complex synthesis steps were decomposed to obtain more complex MOFs.

The reaction conditions obtained through experiments are usually saved in texts by different laboratories, so it is difficult to use machine learning for training. Luo has built an open reaction database (ORD), which comprehensively covers the key experimental details of repeatability, and captures the most important information in a structured format which is shown in Fig. 8 (Kearnes et al. 2021). ORD can generate structured datasets and increase the probability of successful modeling. The synthesis path, reaction products, and yield according to machine learning can be analyzed. In addition, the catalyst selection and material optimization can be done based on the selection of low toxicity, emission reduction, and sustainable development (Jin et al. 2014; Bagheri et al. 2022). Machine learning starts with a large number of experimental data, explores the synthesis conditions through learning, and reduces the use and production of toxic and side effects. Through continuous learning of

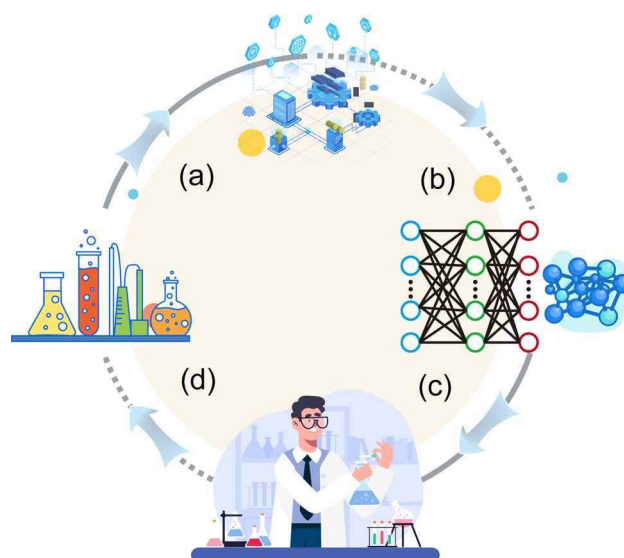


Fig. 8 A scheme explaining the computer-aided chemical discovery cycle. (a) The Open Reaction Database; (b) machine learning and cheminformatics; (c) human or automated interpretation and material design; (d) manual or robotic chemical synthesis

experience, the best way to influence the synthesis process of materials is obtained. This can greatly save synthesis time, and reduce experimental loss and toxic substance emission, which has gradually become an important means of new material synthesis.

Predicting the Physical and Chemical Properties of Functional Materials

Machine learning can also predict the performance of new functional materials by establishing a structure–activity relationship model (Takeichi et al. 2007). The input variables of the model are the structural characteristics of the material, and the output variables are the property we need to predict. Through machine learning, the complex relationship between input and output is learned to obtain the known learning rules (Saito et al. 2019). In the research and development of materials with specific functions (e.g., adsorption materials, photovoltaic materials, superconductors, catalysts, etc.), it is very important to find the corresponding relationship between structure and properties (Chan et al. 2020). In many cases, the relationships between the structure and properties of materials are not simply linear. When the existing theoretical knowledge cannot explain the relationship between structure and properties, machine learning is used for training and learning to predict the structure and properties of materials. Mining the corresponding relationship is also a feasible method. For example, in order to design effective materials for organic photovoltaic (OPV), it is necessary

to determine the most relevant attribute parameters and use these parameters (i.e., descriptors) to build models for power conversion efficiency (PCE). Recently, Sahu et al. established a model to predict PCE by building a dataset of 280 small molecule OPV systems and using 13 important microscopic characteristics of organic materials as descriptors to build a model (Sahu et al. 2018). The gradient enhancement model can be applied to the high-throughput virtual screening of potential new donor molecules, which is crucial for the efficient screening of OPV.

Predicting the Biological Effect of Functional Materials

Functional materials have developed rapidly in many fields, such as electronic machinery, medical and chemical industry, energy conversion and storage, and may enter environmental media in their life cycle. Therefore, it is imperative to evaluate the environmental risk of functional materials. The impacts of materials on the environment include (i) intrinsic toxic effects of certain materials, such as nanomaterials, micro/nanoplastics, and gasoline additives (Henry 1998); (ii) damage to organisms induced by toxic by-products released into the environment during production and use of functional materials; and (iii) toxic effects during the degradation of functional materials, such as the damage to the environment caused by unreasonable recycling of functional materials containing heavy metals.

Due to the rapid development of nanotechnology and the wide application of products (Alaa et al. 2015), research scientists began to pay attention to the potential adverse effects of nanomaterials on human health and the environment (Xu et al. 2020). Conventional safety testing of nanomaterials mainly relies on animal studies. Although animal experiments are still essential for mechanical and chronic toxicology studies, they are not suitable for pre-test in the design and safe production of new materials (Puzyn et al. 2011). Machine learning can replace animal testing and can conduct high-throughput testing (Nel et al. 2013). For example, researchers studied the cytotoxicity of 17 metal oxide nanoparticles on *Escherichia coli*, and based on the toxicity data and the calculated structural descriptors, they established a model to predict the cytotoxicity of other nanomaterials (Bin Hafeez et al. 2021). Ten nanomaterials were used as the training set and the remaining as the test set. Twelve descriptors were calculated for each nanomaterial to account for the reactivity and electronic properties of the nanoparticles. The correlation between the toxicological characteristics of nanoparticles and their descriptors was obtained via the multiple regression method, which was further optimized to find the best parameters.

How to Reach the Full Potential of Machine Learning: A New Paradigm for Sustainable Functional Materials Design

Machine learning can play a key role not only in predicting the toxicity and performance of functional materials but also in their sustainable design and development (Hamilton et al. 2009). In the process of material design and synthesis, machine learning can minimize the use and production of toxic and side effects, and reduce the development cost of materials.

Design of Functional Materials Through Eco-friendly Synthetic Pathways

Machine learning approaches can guide the design of eco-friendly functional materials by optimizing synthetic pathways such as, but not limited to, the use of green catalysts and selecting the most suitable conditions to reduce the generation of toxic by-products (Fig. 9). For the sustainable development of materials, machine learning can avoid using certain heavy metals and toxic organic compounds as the cores and ligands of materials at the beginning of material design, for example, toxic elements such as lead and cadmium in QDs solar cells can be replaced by copper, indium, and selenium (Tagit et al. 2017). Researchers combined machine learning and DFT to design a framework for efficiently searching for stable lead-free organic-inorganic hybrid perovskites (HOIPs) with appropriate bandgaps (Capecci et al. 2021). Combining 11 kinds of non-toxic organic small molecules and 32 types of non-toxic divalent metal atoms, 5152 untapped potential HOIPs were generated. Further, the structure-property relationship was mined through machine learning, and high-performance non-toxic HOIPs were screened. In a recent study presented by Sushil et al., they proposed a machine learning method to identify efficient green solvents in the synthesis of covalent organic frameworks (COFs) (Sushil et al. 2021). Their results demonstrated that the formation of crystalline or amorphous COFs can be quantitatively related to the properties of solvent media and amine precursors. Together with computational models trained by Monte Carlo tree search and reinforcement learning, the valid synthesis pathways with shorter routes and greener solvents were successfully identified for fine chemicals or pharmaceuticals (Wang et al. 2020). Besides, researchers have proposed a new method to generate synthetic routes conditional on a Markov decision process (MDP) of target molecules (Ferraiolo et al. 2001). It only suggests molecular structures that can be synthesized. This method ensures that molecules consist of materials that can be purchased and that the chemical reactions that take place between these materials obey the laws of chemistry. However, this method only considers the construction of

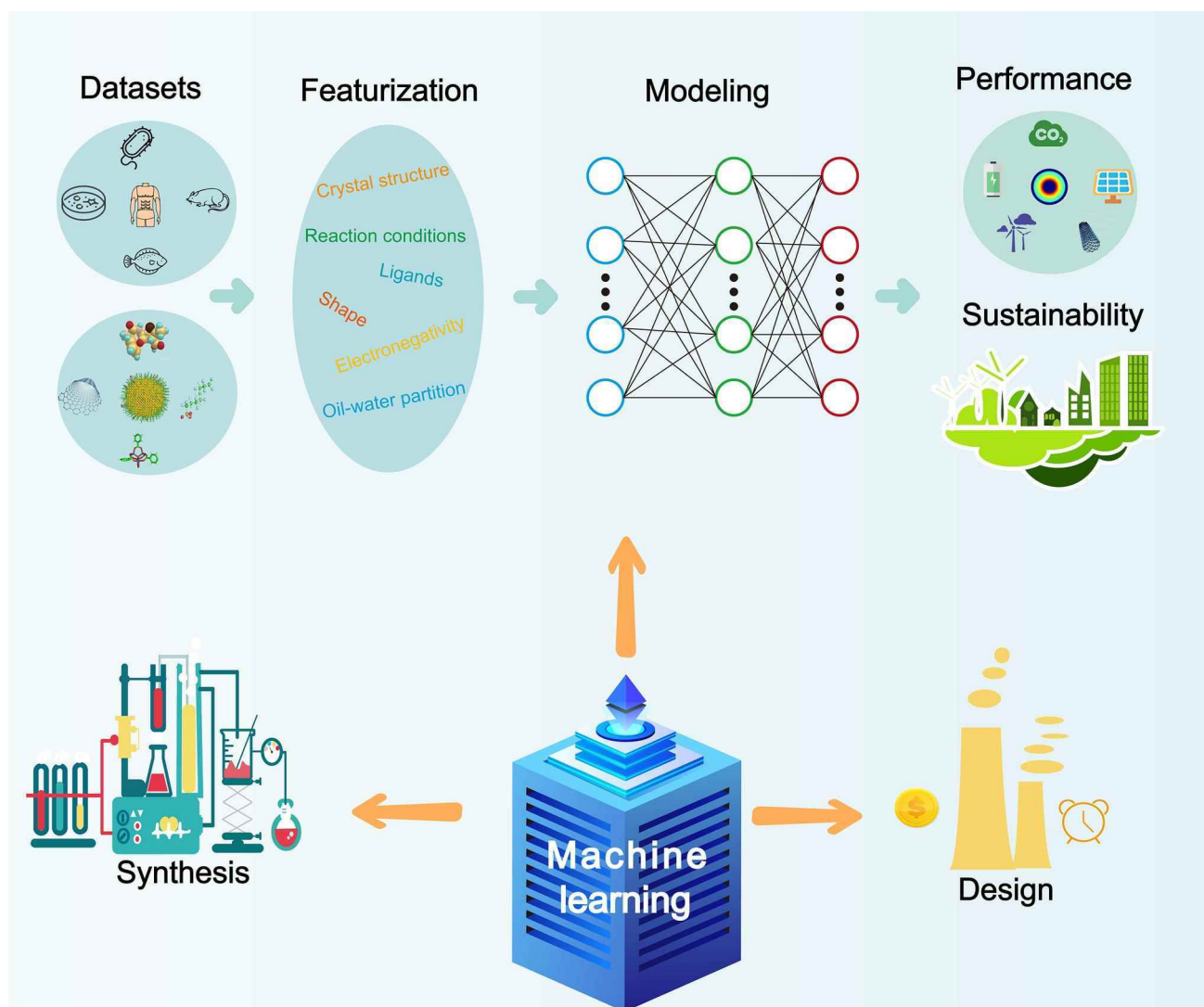


Fig. 9 The screening processes of functional materials with sustainability. In the machine learning model, performance and sustainability are used as double objectives to establish the model and predict. In addition, machine learning plays an important role in the synthesis and design of sustainable functional materials. In material

design, non-toxic and sustainable functional materials can be rapidly designed through machine learning. By predicting the conditions for synthesizing functional materials, the efficiency is accelerated and the cost is reduced

new molecular building blocks and the construction of new molecules through reactions. In the design process, we can add the evaluation module of the use of green catalysts and the generation of toxic by-products and design the synthesis route of new molecules for sustainable development.

In addition, the life cycle and recycling of materials are also key to sustainable development. Cost control and the life cycle of candidate functional materials in the process of successful commercialization and sustainable development also deserve attention. For example, high cost and safety problems limit the wide application of lithium-ion batteries (LIBs) (Hachmann et al. 2011). Improving the quality and reducing the manufacturing cost in battery production is a

key challenge because LIBs involve a complex electrochemical system. Traditional quality control measures (such as aging) are time-consuming and costly. In recent years, several data-driven methods have been proposed to analyze the status and quality of LIBs using various analysis methods. In order to predict the life or remaining useful life (RUL), various methods in the fields of stochastic process, screening and artificial intelligence methods are applied. The prediction accuracy of the linear regression model and artificial neural network (ANN) was compared by using the online measurement data of 29 NMC111 electrode/graphite soft-coated batteries. A total of 24 features were extracted from the comprehensive electrochemical impedance spectroscopy

(EIS) and cyclic datasets. The optimal artificial neural network achieved a test error of 10.1% in 2 days. During wetting, reliable classification of high-life batteries can be achieved using EIS measurements alone.

Screening of Eco-friendly Functional Materials Through Simultaneous Prediction of Toxicity and Functionality

In traditional machine learning, the optimization of specific indicators, such as the performance of materials, is the top priority. In order to achieve relevant goals, researchers usually train a model or a group of models to perform the tasks, and then fining tuning these models is performed until the target parameters are no longer improved. However, when screening functional materials with sustainability, we must focus on multiple tasks to consider both the performance and biocompatibility of materials at the same time by sharing related tasks.

Herein, we propose a new framework for the screening process of sustainable functional materials (Fig. 9), in which an additional toxicity target is added. In our proposed framework, two datasets for different model training should be first prepared (Tables 1 and 3). The next steps are consistent with the traditional machine learning process to screen functional materials with good performance and low toxicity for further synthesis and validation. When assessing the environmental exposure and health risk of functional materials, it is necessary to comprehensively consider their transformation, dose, and interaction with pollutants in the environment. After ingestion, it is also necessary to consider whether it will cause some problems such as inflammation, hemolysis, and gastrointestinal injury (Ferraiolo et al. 2001).

Take the design of antimicrobial peptides as an example. Antimicrobial peptides are small molecules that have

biological activities against bacteria, fungi, and viruses (Bin et al. 2021). However, they are not widely used because their hemolysis can selectively act on red blood cells, change the biochemical properties of red blood cells at a certain concentration, and enhance the fluidity of their membranes or cause red blood cell rupture (Wei et al. 2020). The design of antimicrobial peptides by machine learning can consider both activity and hemolysis. For example, Capecchi et al. trained a combination of recurrent neural networks (RNN) for generation and prediction using the sequence information, antibacterial activity, and hemolysis data from the Database of Antimicrobial Activity and Structure of Peptides (DBAASP) (Capecchi et al. 2021). In order to screen non-hemolytic antimicrobial peptides in the generated sequence, two RNN models were used to predict antimicrobial activity and hemolysis. The combination of antibacterial activity and non-hemolysis was taken as the positive group. A total of 148 and 160 peptides were filtered out from the activity model and the hemolytic model, respectively. Finally, 28 screened sequences were synthesized and tested, and 12 new antimicrobial peptides were produced, eight of which were non-hemolytic.

As a class of widely used functional materials, biocompatible nanomaterials and their design have also drawn a great amount of attention from researchers during the past decade. For example, the quantitative structure–activity relationship (QSAR) models were built to screen surface-modified carbon nanotubes (CNTs) from a combinatorial virtual library of 240,000 ligands (Fourches et al. 2016). The constructed QSAR models finally identified multiple CNTs with desired protein-binding activity and low cytotoxicity. Similarly, in a more recent study, seven gold nanoparticles with desired cellular uptake and oxidative stress were successfully identified using machine learning models (Wang et al. 2017). The above results demonstrate that machine

Table 3 A list of molecular toxicity prediction websites or software

| Databases | Description |
|---|--|
| ChEMBL https://www.ebi.ac.uk/chembl/ | Collected biological activity data of various targets and compounds (Michal et al. 2017) |
| CTD http://ctdbase.org/ | Linked toxicological information on chemicals, genes, phenotypes, diseases, and exposures (Mattingly et al. 2010) |
| DrugBank https://www.drugbank.com/ | Combined bioinformatics and chemical information resources related to small molecule (Wishart et al. 2008) |
| PubChem https://pubchem.ncbi.nlm.nih.gov/ | Contains information on the chemical and physical properties, biological activities, safety, and toxicity of a large number of molecular materials (Pouliot et al. 2011) |
| RepDose https://repdose.item.fraunhofer.de/ | Comprises a subset of about 200 high-quality subacute studies with oral exposure (Bitsch et al. 2006) |
| TOXNET https://www.toxnet.net/ | Includes information on drug toxicology, hazardous chemicals, and other related fields (Wexler 2001) |
| TOPKAT http://www.star.bris.ac.uk/~mbt/topcat/ | Collected data from 16 animal toxicity models of the compounds (Cariello 2002) |

learning methods can maximize the functionality of materials and minimize their toxicities during the process of functional material design.

In summary, two major approaches are involved in our proposed framework for eco-friendly functional material design, i.e., through eco-friendly synthetic pathways, and simultaneous prediction of toxicity and functionality. To achieve this purpose, some important parameters should be considered in performing these approaches. One of the key principles of eco-friendly synthesis is to reduce or eliminate the use and generation of hazardous substances. Additionally, scientists should consider the use of more economical raw materials and improve their utilization efficiency to achieve an eco-friendly synthesis and other similar optimized synthetic routes such as multicomponent reactions and one-pot cascade synthesis. Furthermore, multi-task deep learning holds great promise for simultaneously predicting the toxicity and functionality of materials. At its core, multi-task deep learning aims to learn generalized representations that can be shared across different tasks. Therefore, multi-task deep learning models can fully leverage useful hidden information in multiple related tasks to improve the performance of all tasks, i.e., maximum functionality and minimum toxicity.

Challenges and Perspectives

Since the discovery and development of new materials are time-consuming processes with high investment and risks, it is extremely important to determine the toxicity of materials and their interaction with the environment at the early stage of material design for sustainable development. Computational models have been widely used to predict the toxicity of materials because of their advantages in speed, cost, and accuracy. However, large-scale development of machine learning still faces challenges, in which the low-quality and small-scale datasets of materials have become the main obstacles. Therefore, there is an urgent need to develop efficient synthesis and characterization methods, e.g., the combinatorial chemistry approach and other high-throughput screening methods, to generate high-quality and large-scale experimental data. The combinatorial chemistry approach has proven to be capable of quickly synthesizing a large number of molecules and materials in a single process, through systematic, repetitive, and covalent linkage of various “building blocks.” In the past decade, researchers have applied high-throughput screening techniques to test thousands to millions of chemicals for physicochemical properties and biological activities. Additionally, the data quality should also be strictly controlled by performing uniform experimental standards.

Apart from data generation, great efforts should be devoted to relevant data collection. At present, the priority should be the collection and sharing of toxicity data of materials (Yan et al. 2020a, b, c). In the process of data collection, not only should the data format be highly unified, but also the quality of the data needs to be reviewed. The construction of material function and toxicity databases involves long-term efforts. Researchers need to give priority to the establishment of thematic material databases of some popular material systems, so that the construction and use of the database can be carried out simultaneously.

Moreover, in machine learning, feature engineering is often important. The selection of descriptors plays an important role in the prediction performance of machine learning models. As long as descriptors closely related to the properties of the prediction target are selected as the input of the model, reasonable prediction performance and beyond can be guaranteed regardless of the types of chosen machine learning models. Future studies need to focus on developing more advanced descriptors of some complex materials, such as nanomaterials (Yan et al. 2019). Such advanced descriptors that can be used for machine learning should comprehensively cover the structure, physical, and chemical properties of materials, and have the characteristics of structural interpretation, good correlation with the prediction target and convenient calculation. In addition, to predict the biological effect or toxicity of materials, descriptors should also consider their transformation after release into the environment and ingestion by organisms, such as degradation, oxidation, and adsorption.

Another challenge comes from the interpretation of machine learning models. The ultimate goal of machine learning is to extract interpretable knowledge from data and emphasize its interpretability while pursuing algorithm accuracy. However, the complexity and diversity of functional materials lead to intricate relationships between influencing factors (e.g., structural features) and targets (e.g., physicochemical properties). An interpretable machine learning model allows us to understand how the model makes predictions and helps us design functional materials with targeted properties. Therefore, it is recommended to select machine learning algorithms that are inherently interpretable such as the decision tree, random forest, and linear regression. Besides, external interpretation can be introduced to input different feature combinations into the trained model to find the relationship between the input features and the model decision results, so as to find the decision rules of the model to improve the interpretability of the model.

Last but not least, advanced machine learning models for designing functional materials are urgently needed. Such models will help to narrow the gap between experiment and calculation, and between material designers and environmentalists, to meet the urgent needs of sustainable

development. For instance, scientists have proposed an inverse design to construct the property–structure relationship and realize the design of related high-performance molecules (Yao et al. 2021). Functional materials can be designed directly according to the required performance and the intrinsic properties of the raw materials. Moreover, a standard functional material sustainable system should be established to validate and reliably assess machine learning approaches, which will promote the development and application of new functional materials. By doing so, we believe sustainable functional material science will eventually develop in the near future (Li et al. 2022).

Acknowledgements This study was supported by the National Natural Science Foundation of China (Grant Nos. 22106025, 22006025).

Author Contributions XY contributed to conceiving the idea, writing, reviewing, and editing of the manuscript. YH contributed to the writing of the original draft, and writing, reviewing, and editing of the manuscript. CL contributed to the writing of the original draft, and writing, reviewing, and editing of the manuscript. GL contributed to data collection, analysis, reviewing, and editing of the manuscript. All authors read and approved the final manuscript.

Data Availability All data generated or analyzed during this study are included in this published article.

Declarations

Conflict of interest The authors declare that they have no potential conflict of interest with respect to the research, authorship or publication of this article.

References

- Adhikari C, Mishra K (2018) Quantitative structure-activity relationships of aquatic narcosis: a review. *Curr Comput Aided Drug Des* 14:7–28
- Alaa T, Tarek G, Mohamed M, Fouad V, Snasel A (2015) Towards an automated zebrafish-based toxicity test model using machine learning. *ScienceDirect* 65:643–651
- Allouzi M, Tang D, Chew KW, Rinklebe J, Bolan N, Allouzi S, Show PL (2021) Micro (nano) plastic pollution: the ecological influence on soil-plant system and human health. *Sci Total Environ* 788:147815
- Anandakumar H, Arulmurugan R (2019) Supervised, unsupervised and reinforcement learning-A detailed perspective. *J Dyn Control Syst* 11:429–433
- Bagheri S, Esfandiary N, Yliniemi J (2022) Porous SB-CuI two-dimensional metal-organic framework: the green catalyst towards C N bond-forming reactions. *Colloid Surface* 637:128202–128208
- Banerjee A, Shelper WL (2021) Micro- and nanoplastic induced cellular toxicity in mammals: a review. *Sci Total Environ* 755:142518
- Bear JE, Svitkina TM, Krause M, Schafer DA, Loureiro J, Strasser G, Maly IV, Chaga O, Cooper J, Borisy G (2002) Antagonism between Ena/VASP proteins and actin filament capping regulates fibroblast motility. *Cell* 109:509–521
- Beaujuge PM, Fréchet JM (2011) Molecular design and ordering effects in π -functional materials for transistor and solar cell applications. *J Am Chem Soc* 133:20009–20029
- Belchior D, Duarte IF, Freire G (2019) Ionic liquids in bioseparation processes. *Adv Biochem Eng Biotechnol* 168:1–29
- Bin A, Jiang X, Bergen J, Zhu Y (2021) Antimicrobial peptides: an update on classifications and databases. *Int J Mol Sci* 22:11691–11695
- Bin Hafeez A, Jiang X, Bergen J, Zhu Y (2021) Antimicrobial peptides: an update on classifications and databases. *Int J Mol Sci* 22:21
- Bitsch A, Jacobi S, Melber C, Wahnschaffe U, Simetska N, Mangelsdorf I (2006) REPDOSE: a database on repeated dose toxicity studies of commercial chemicals—a multifunctional tool. *Regul Toxicol Pharmacol* 46:202–210
- Boyes WK, van Thriel C (2020) Neurotoxicology of nanomaterials. *Chem Res Toxicol* 33:1121–1144
- Butler Keith T, Davies DW, Cartwright H, Isayev O (2018) Machine learning for molecular and materials science. *Nature* 559:547–555
- Capecci A, Cai X, Personne H, Khler T, Delden V, Reymond L (2021) Machine learning designs non-hemolytic antimicrobial peptides. *Chem Sci* 12:9221–9232
- Cariello N (2002) Comparison of the computer programs DEREK and TOPKAT to predict bacterial mutagenicity. *Mutagenesis* 17:321–329
- Chan H, Cherukara M, Loeffler D, Narayanan B, Sankaranarayanan S (2020) Machine learning enabled autonomous microstructural characterization in 3D samples. *NPJ Comput Mater* 6:1–9
- Chen H, Gao Y, Li J, Fang Z, Bolan N, Bhatnagar A, Gao B, Hou D, Wang S, Song H, Yang X, Shaheen SM, Meng J, Chen W, Rinklebe J, Wang H (2022a) Engineered biochar for environmental decontamination in aquatic and soil systems: a review. *Carbon Res* 1:1–25
- Chen Y, Sun K, Wang Z, Zhang E, Yang Y, Xing B (2022b) Analytical methods, molecular structures and biogeochemical behaviors of dissolved black carbon. *Carbon Res* 1:1–19
- Colvin VL, Schlamp MC, Alivisatos AP (1994) Light-emitting diodes made from cadmium selenide nanocrystals and a semiconducting polymer. *Nature* 370:354–357
- Corey J, Wipke W (1969) Computer-assisted design of complex organic syntheses. *Science* 166:178–192
- Cui X, Yang R, Li S, Liu J, Wu Q, Li X (2021) Modeling and insights into molecular basis of low molecular weight respiratory sensitizers. *Mol Divers* 25:847–859
- Curtarolo S, Setyawan W, Wang S, Xue J, Yang K, Taylor H, Nelson J, Hart W, Sanvito S, Buongiorno-Nardelli M, Mingo N, Levy O (2012) AFLOWLIB.ORG: a distributed materials properties repository from high-throughput ab initio calculations. *Comp Mater Sci* 58:227–235
- De Marzi G, Morici L, Muzzi L, Della CA, Nardelli MB (2013) Strain sensitivity and superconducting properties of Nb_3Sn from first principles calculations. *J Phys Condens Matter* 25:135702
- Derfus AM, Chan W, Bhatia S (2004) Probing the cytotoxicity of semiconductor quantum dots. *Nano Lett* 4:11–18
- Donahue J, Jia Y, Vinyals O, Hoffman J, Darrell T (2013) DeCAF: a deep convolutional activation feature for generic visual recognition. *JMLR* 32:647–655
- Duan C, Nandy A, Kulik HJ (2022) Machine learning for the discovery, design, and engineering of materials. *Annu Rev Chem Biomol* 13:405–429
- Eck N, Waltman L (2010) Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics* 84:523–538
- Fang L, Cheng L, Glerum JA, Bennett J, Cao J, Wagner GJ (2022) Data-driven analysis of process, structure, and properties of additively manufactured Inconel 718 thin walls. *NPJ Comput Mater*. <https://doi.org/10.1038/s41524-022-00808-5>
- Ferraiolo F, Sandhu R, Gavrilu S, Kuhn R, Chandramouli R (2001) Proposed NIST standard for role-based access control. *AMIA Annu Symp Proc* 4:224–274

- Fourches D, Pu D, Li L, Zhou H, Mu Q, Su G, Yan B, Tropsha A (2016) Computer-aided design of carbon nanotubes with the desired bioactivity and safety profiles. *Nanotoxicology* 10:374–383
- Giepmans B, Deerinck TJ, Smarr BL, Jones YZ, Ellisman MH (2005) Correlated light and electron microscopic imaging of multiple endogenous proteins using Quantum dots. *Nat Methods* 2:743–749
- Greer AJ, Jacquemin J, Hardacre C (2020) Industrial applications of ionic liquids. *Molecules* 25:5207
- Groom CR, Allen FH (2014) The Cambridge structural database in retrospect and prospect. *Angew Chem Int Ed Engl* 53:662–671
- Gulati S, Vijayan S, Mansi KS, Harikumar B, Trivedi M VS (2023) Recent advances in the application of metal-organic frameworks (MOFs)-based nanocatalysts for direct conversion of carbon dioxide (CO₂) to value-added chemicals. *Coord Chem Rev* 474:214853–214860
- Gupta R, Srivastava D, Sahu M, Tiwari S, Ambasta R, Kumar P (2021) Artificial intelligence to deep learning: machine intelligence approach for drug discovery. *Mol Divers* 25:1315–1360
- Hachmann J, Olivares-Amaya R, Atahan-Evrenk S, Amador-Bedolla C, Sánchez-Carrera S, Gold-Parker A, Vogt L, Brockway M, Aspuru-Guzik A (2011) The Harvard clean energy project: large-scale computational screening and design of organic photovoltaics on the world community grid. *J Phys Chem Lett* 2:2241–2251
- Hamilton R, Wu N, Porter D, Buford M, Holian A (2009) Particle length-dependent titanium dioxide nanomaterials toxicity and bioactivity. *Part Fibre Toxicol* 6:35
- Henry A (1998) Composition and toxicity of petroleum products and their additives. *Hum Exp Toxicol* 17:111–123
- Herzke D, Olsson E, Posner S (2012) Perfluoroalkyl and polyfluoroalkyl substances (PFASs) in consumer products in Norway—a pilot study. *Chemosphere* 88:980–987
- Huang J, Sun L, Mennigen JA, Liu Y, Liu S, Zhang M, Wang Q, Tu W (2021) Developmental toxicity of the novel PFOS alternative OBS in developing zebrafish: an emphasis on cilia disruption. *J Hazard Mat* 409:124491
- Jain A, Ping S, Geoffroy O, Hautier G, Chen W, Richards D, Dacek S, Cholia S, Gunter D, Skinner D, Ceder G, Persson A (2013) Commentary: the materials project: a materials genome approach to accelerating materials innovation. *APL Mater* 1:011002-011002-011011
- Jelliarko P, Sik H, Jung K, Lee M (2010) Ionic liquids for acetylene and ethylene separation: material selection and solubility investigation. *Chem Eng Process* 49:192–198
- Jiang J, Chen M, Fan A (2020) Deep neural networks for the evaluation and design of photonic devices. *Nat Rev Mater* 6:679–700
- Jin X, Jin M, Sheng L (2014) Three dimensional quantitative structure-toxicity relationship modeling and prediction of acute toxicity for organic contaminants to algae. *Comput Biol Med* 51:205–213
- Juan Y, Dai Y, Yang Y, Zhang J (2020) Accelerating materials discovery using machine learning. *J Mater Sci Technol* 20:178–190
- Kearnes S, Maser R, Wlekliński M, Kast A, Doyle G, Dreher D, Hawkins M, Jensen F, Cole W (2021) The open reaction database. *ACS* 143:18820–18826
- Kim S, Noh J, Gu GH, Aspuru-Guzik A, Jung Y (2020) Generative adversarial networks for crystal structure prediction. *ACS* 6:1412–1420
- Kirklin S, Saal E, Meredig Thompson A, Doak W, Aykol M, Rühl S, Wolverton C (2015) The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *Comp Mater Sci* 171:1–28
- Kuznetsov VL, Edwards P (2010) Functional materials for sustainable energy technologies: four case studies. *Condensed Matter* 3:44–58
- Li C, Jiang G, Ren H (2022) The common vision toward one health. *Eco Environ Health* 1:1–2
- Liang C, Qiao Q, Lian Z (2017) Determination of reversed-phase high performance liquid chromatography based octanol-water partition coefficients for neutral and ionizable compounds: methodology evaluation. *J Chromatogr A* 1528:25–34
- Ligeza A (1995) Artificial intelligence: a modern approach. *Neurocomputing* 9:215–218
- Lin F, Jia M, Sun Z, Fu Z (2020) Highly sensitive self-referencing thermometry probe and advanced anti-counterfeiting based on the CDs/YVO₄:Eu³⁺ composite materials. *Scripta Mater* 186:298–303
- Lindan P (2002) First-principles simulation: ideas, illustrations and the CASTEP code. *J Phys Condens Mat* 14:2717
- Little JL, Williams AJ, Tkachenko PV (2012) Identification of “Known Unknowns” utilizing accurate mass data and ChemSpider. *J Am Soc Mass Spectrom* 23:179–185
- Liu R, Kumar A, Chen Z, Agrawal A, Sundararaghavan V, Choudhary A (2015) A predictive machine learning approach for microstructure optimization and materials design. *Sci Rep* 5:11551–11555
- Lloyd S, Mohseni M, Rebentrost P (2013) Quantum algorithms for supervised and unsupervised machine learning. *Quantum Phys* 470:457–461
- Logeshwaran P, Sivaram AK, Surapaneni A, Kannan K, Megharaj M (2021) Exposure to perfluorooctanesulfonate (PFOS) but not perfluorooctanoic acid (PFOA) at ppb concentration induces chronic toxicity in *Daphnia carinata*. *Sci Total Environ* 769:144577
- Lovri M, Molero JM, Kern R (2019) PySpark and RDKit: moving towards big data in cheminformatics. *QSAR Comb Sci* 38:1–4
- Luo Y, Bag S, Zaremba O, Cierpka A, Andreo J, Wuttke S, Friederich P, Tsotsalas M (2022) MOF synthesis prediction enabled by automatic data mining and machine learning. *Angew Chem Int Ed Engl* 61:e202200242
- Machak R, Motsi T, Raganya M, Radingoana M, Chikocha S (2021) Machine learning-based prediction of phases in high-entropy alloys: a data article. *Data Brief* 38:107346
- Mattingly CJ, Rosenstein M, Colby G, Forrest Boyer JJ (2010) The Comparative Toxicogenomics Database (CTD): a resource for comparative toxicological studies. *J Exp Zool A Comp Exp Biol* 305:689–692
- Mauri A, Consonni V, Pavan M, Todeschini R (2006) DRAGON software: an easy approach to molecular descriptor calculations. *Match Commun Math CO* 56:237–248
- Menon D, Ranganathan R (2022) A generative approach to materials discovery, design, and optimization. *ACS Omega* 7:25958–25973
- Michał M, Gaulton A, Mendez D, Bento AP, Leach A (2017) Using ChEMBL web services for building applications and data processing workflows relevant to drug discovery. *Expert Opin Drug Discov* 12:1–11
- Miotto R, Wang F, Wang S, Jiang X, Dudley T (2018) Deep learning for healthcare: review, opportunities and challenges. *Brief Bioinform* 19:1236–1246
- Nazarian D, Camp JS, Sholl S (2016) A comprehensive set of high-quality point charges for simulations of metal-organic frameworks. *Chem Mater* 28:785–793
- Nel A, Zhao Y, Mdlor L (2013) Environmental health and safety considerations for nanotechnology. *Accounts Chem Res* 46:605–610
- Pacurari M, Lowe K, Tchounwou PB, Kafoury R (2016) A review on the respiratory system toxicity of carbon nanoparticles. *Int J Env Res Pub He* 13:325
- Palmer BA, Gur D, Weiner S, Addadi L, Oron D (2018) The organic crystalline materials of vision: structure-function considerations from the nanometer to the millimeter scale. *Adv Mater* 30:e1800006
- Papas B, Schaefer H (2006) Concerning the precision of standard density functional programs: Gaussian, molpro, nwchem, Q-chem, and gamess. *J Mol Struct (theochem)* 768:175–218

- Pei Z, Rozman K, Doan M, Wen Y, Gao N, Holm A, Hawk A, Alman E, Gao C (2021) Machine-learning microstructure for inverse material design. *Adv Sci (weinh)* 8:e2101207
- Pillong M, Marx C, Piechon P, Wicker P, Cooper I, Wagner T (2017) A publicly available crystallisation data set and its application in machine learning. *Cryst Eng Comm* 19:3737–3745
- Plata DL, Janković NZ (2021) Achieving sustainable nanomaterial design through strategic cultivation of big data. *Nat Nanotechnol* 16:612–614
- Plimpton S, Thompson A, Crozier P (2011) Molecular dynamics simulations from SNL'S large-scale atomic/molecular massively parallel simulator (LAMMPS). *Philos Trans A Math Phys Eng Sci* 362:1373–1386
- Polichetti M, Galluzzi A, Buchko VK, Tomov V, Pace S (2021) A precursor mechanism triggering the second magnetization peak phenomenon in superconducting materials. *Sci Rep* 11:7247
- Pouliot Y, Chiang A, Butte AJ (2011) Predicting adverse drug reactions using publicly available PubChem BioAssay data. *Clin Pharmacol Ther* 90:90–99
- Press M (2007) Approximation methods for Gaussian process regression. *Mit Press* 14:333–350
- Pronk S, Páll S, Schulz R, Larsson P, Lindahl E (2013) GROMACS 4.5. *Bioinformatics* 29:845–854
- Puzyn T, Rasulev B, Gajewicz A, Hu X, Dasari P, Michalkova A, Hwang M, Toropov A, Leszczynska D, Leszczynski J (2011) Using nano-QSAR to predict the cytotoxicity of metal oxide nanoparticles. *Nat Nanotechnol* 6:175–178
- Qian Y, Nie S, Yi C, Kon L, Fang C, Qian T, Ding H, Shi Y, Wang Z, Weng H (2019) Topological electronic states in HfRuP family superconductors. *NPJ Comput Mater* 5:121–126
- Qiao Y, Guo S (2005) Concise applications of molecular modeling software-MOE. *Comput Appl Chem* 2:157–160
- Qiu L, Zhang X, Zhang X, Zhang Y, Gu J, Chen M, Zhang Z, Wang X, Wang SL (2013) Sertoli cell is a potential target for perfluorooctane sulfonate-induced reproductive dysfunction in male mice. *Toxicol Sci* 135:229–240
- Robel I, Subramanian V, Kuno M, Kamat PV (2006) Quantum dot solar cells. Harvesting light energy with CdSe nanocrystals molecularly linked to mesoscopic TiO₂ films. *ACS* 128:2385–2393
- Rossetti R, Nakahara S, Brus LE (1983) Quantum size effects in the redox potentials, resonance Raman spectra, and electronic spectra of CdS crystallites in aqueous solution. *JCP* 79:1086–1088
- Rydz J, Sikorska W, Kyulavska M, Christova D (2014) Polyester-based (bio)degradable polymers as environmentally friendly materials for sustainable development. *Int J Mol Sci* 16:564–596
- Sahoo S, Adhikari C, Kuamar M, Mishra K (2016) A short review of the generation of molecular descriptors and their applications in quantitative structure property/activity relationships. *Curr Comput Aided Drug Des* 12:181–205
- Sahu H, Rao W, Troisi A, Ma H (2018) Toward predicting efficiency of organic solar cells via machine learning and improved descriptors. *Adv Energy Mater* 8:1801032.1–1801032.9
- Saito Y, Shin K, Terayama K, Desai S, Onga M, Nakagawa M, Itahashi Y, Iwasa Y, Yamada M, Tsuda K (2019) Deep-learning-based quality filtering of mechanically exfoliated 2D crystals. *NPJ Comput Mater* 5:1–6
- Sani A, Cao C, Cui D (2021) Toxicity of gold nanoparticles (AuNPs): a review. *Biochem Biophys Rep* 26:100991
- Schneider SL, Lim HW (2019) A review of inorganic UV filters zinc oxide and titanium dioxide. *Photodermatol Photoimmunol Photomed* 35:442–446
- Shen L, Zhou J, Yang T, Yang M, Feng YP (2021) High-throughput discovery and intelligent design of 2D functional materials for various applications. *arXiv e-prints* 1:1–29
- Simões M, Pereira AR, Simões LC, Cagide F, Borges F (2021) Biofilm control by ionic liquids. *Drug Discov Today* 26:1340–1346
- Smith AM, Duan H, Mohs AM, Nie S (2008) Bioconjugated quantum dots for in vivo molecular and cellular imaging. *Adv Drug Deliver Rev* 60:1226–1240
- Socher R (2014) Recursive deep learning for natural language processing and computer vision. *ACM Comput Surv* 52:1474
- Song MM, Wang YM, Liang XY, Zhang XQ, Zhang S, Li BJ (2019) Functional materials with self-healing properties: a review. *Soft Matter* 15:6615–6625
- Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res* 15:1929–1958
- Sushil K, Gergo I, Gyorgy S (2021) Synthesis of covalent organic frameworks using sustainable solvents and machine learning. *Green Chem* 23:D1100–D1107
- Tagit O, Ruiter M, Brasch M, Ma Y, Cornelissen J (2017) Quantum dot encapsulation in virus-like particles with tuneable structural properties and low toxicity. *RSC Adv* 7:38110–38118
- Takeichi N, Tanaka K, Tanaka H, Ueda T, Kamiya Y, Tsukahara M, Miyamura H, Kikuchi S (2007) Hydrogen storage properties of Mg/Cu and Mg/Pd laminate composites and metallographic structure. *J Alloys Compd* 446:543–548
- Tiago G, Matias I, Ribeiro A, Martins L (2020) Application of ionic liquids in electrochemistry-recent advances. *Molecules* 25:5812
- Tsunedo T (2020) Density functional theory as a data science. *Chem Rec* 20:618–639
- Upadhyay R, Kosuri S, Tamasi M, Meyer TA, Atta S (2020) Automation and data-driven design of polymer therapeutics. *Adv Drug Deliv Rev* 33:1–15
- Uwayezu N, Yeung L, Bckstrm M (2022) Sorption of Perfluorooctane sulfonate (PFOS) including its isomers on hydrargillite as a function of pH, humic substances and Na₂SO₄. *J Environ Sci* 111:263–272
- Wahl CB, Aykol M, Swisher JH, Montoya JH, Suram SK, Mirkin CA (2021) Machine learning-accelerated design and synthesis of polyelemental heterostructures. *Sci Adv* 7:eabj5505
- Wallace A, Abou-Zamzam AM, Mueller RT (1972) Transport of sodium into the xylem exudate of tobacco. *Plant Physiol* 50:388–390
- Wang Y, Ma Y (2013) Perspective: crystal structure prediction at high pressures. *JCP* 140:631–2378
- Wang W, Sedykh A, Sun H, Zhao L, Russo D, Zhou H, Yan B, Zhu H (2017) Predicting nano-bio interactions by integrating nanoparticle libraries and quantitative nanostructure activity relationship modeling. *ACS Nano* 11:12641–12649
- Wang X, Qian Y, Gao H, Colry W, Mo Y, Barzilay R, Jensen K (2020) Towards efficient discovery of green synthetic pathways with Monte Carlo tree search and reinforcement learning. *Chem Sci* 11:10959–10972
- Wang Y, Liang Y, Bo T, Meng S, Liu M (2022) Orbital dependence in single-atom electrocatalytic reactions. *J Phys Chem Lett* 13:5969–5976
- Wei H, Xie Z, Tan X, Guo R, Zhang Y (2020) Temporin-like peptides show antimicrobial and anti-biofilm activities against *Streptococcus mutans* with reduced hemolysis. *Molecules* 25:5724
- Wei P, Pan X, Chen CY, Li Y, Yan X, Li C, Chu Y, Yan B (2021) Emerging impacts of ionic liquids on eco-environmental safety and human health. *Chem Soc Rev* 50:13609–13627
- Wellmann PJ (2021) The search for new materials and the role of novel processing routes. *Discov Med* 1:14
- Wexler P (2001) TOXNET: an evolving web resource for toxicology and environmental health information. *Toxicology* 157:3–10
- Wishart D, Craig K, Guo A, Cheng D, Savota S, Dan T, Bijaya G, Murtaza H (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res* 36:D901–D906

- Xie S, Wang F, Wang Y, Zhu L, Dong Z, Wang X, Li X, Zhou W (2011) Acute toxicity study of tilmicosin-loaded hydrogenated castor oil-solid lipid nanoparticles. *Part Fibre Toxicol* 8:33
- Xu Q, Li Z, Liu M, Yin J (2018) Rationalizing perovskite data for machine learning and materials design. *J Phys Chem Lett* 9:6948–6954
- Xu T, Ngan K, Ye L, Xia M, Xie Q, Zhao B, Simeonov A, Huang R (2020) Predictive models for human organ toxicity based on in vitro bioactivity data and chemical structure. *Chem Res Toxicol* 33:731–741
- Yan X, Sedykh A, Wang W, Zhao X, Yan B, Zhu H (2019) In silico profiling nanoparticles: predictive nanomodeling using universal nanodescriptors and various machine learning approaches. *Nanoscale* 11:8352–8362
- Yan X, Sedykh A, Wang W, Yan B, Zhu H (2020a) Construction of a web-based nanomaterial database by big data curation and modeling friendly nanostructure annotations. *Nat Commun* 11:2519
- Yan X, Zhang J, Daniel P, Zhu H, Yan B (2020b) Prediction of nanobio interactions through convolutional neural network analysis of nanostructure images. *ACS Sustain Chem Eng* 8:19096–19104
- Yan X, Zheng M, Gao X, Zhu M, Hou Y (2020c) Giant current performance in lead-free piezoelectrics stem from local structural heterogeneity. *Acta Mater* 187:29–40
- Yao Z, Sánchez-Lengeling B, Bobbitt S, Bucior J, Aspuru-Guzik A (2021) Inverse design of nanoporous crystalline reticular materials with deep generative models. *Nat Machine Intell* 3:76–86
- Yilmaz B, Terekeci H, Sandal S, Kelestimur F (2020) Endocrine disrupting chemicals: exposure, effects on human health, mechanism of action, models for testing and strategies for prevention. *Rev Endocr Metab Dis* 21:127–147
- Yu L, Shin M, Lee H, Jun I, Kang K, Park C, Shin H (2012) Polydopamine-mediated immobilization of multiple bioactive molecules for the development of functional vascular graft materials. *Biomaterials* 33:8343–8352
- Zagorac D, Müller H, Ruehl S, Zagorac J, Rehme S (2019) Recent developments in the inorganic crystal structure database: theoretical crystal structure data and related features. *J Appl Crystallogr* 52:918–925
- Zahrt F, Henle J, Rose T, Wang Y, Darrow T, Denmark S (2019) Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning. *Science* 363:6424
- Zhang G, Fei H (2019) Synthesis and applications of porous organosulfonate-based metal-organic frameworks. *Top Curr Chem* 32:2364–8961
- Zhang T, Zhang Z, Arnold A (2021) Crystal structure-free method for dielectric and polarizability characterization of crystalline materials at Terahertz frequencies. *Appl Spectrosc* 75:647–653
- Zheng-Dong A (2018) Macro-architected cellular materials: properties, characteristic modes, and prediction methods. *Front Mech Eng* 13:442–459
- Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Authors and Affiliations

Ying He¹ · Guohong Liu^{1,2} · Chengjun Li^{1,2} · Xiliang Yan^{1,2} 

¹ Key Laboratory for Water Quality and Conservation of the Pearl River Delta, Ministry of Education, Institute of Environmental Research at Greater Bay Area, Guangzhou University, Guangzhou 510006, China

² School of Agriculture and Biological Sciences, Qiannan Normal University for Nationalities, Duyun 558000, China

Comprehensive Interrogation on Acetylcholinesterase Inhibition by Ionic Liquids Using Machine Learning and Molecular Modeling

Jiachen Yan, Xiliang Yan,* Song Hu, Hao Zhu, and Bing Yan*



Cite This: *Environ. Sci. Technol.* 2021, 55, 14720–14731



Read Online

ACCESS |



Metrics & More



Article Recommendations



Supporting Information

ABSTRACT: Quantitative structure–activity relationship (QSAR) modeling can be used to predict the toxicity of ionic liquids (ILs), but most QSAR models have been constructed by arbitrarily selecting one machine learning method and ignored the overall interactions between ILs and biological systems, such as proteins. In order to obtain more reliable and interpretable QSAR models and reveal the related molecular mechanism, we performed a systematic analysis of acetylcholinesterase (AChE) inhibition by 153 ILs using machine learning and molecular modeling. Our results showed that more reliable and stable QSAR models ($R^2 > 0.85$ for both cross-validation and external validation) were obtained by combining the results from multiple machine learning approaches. In addition, molecular docking results revealed that the cations and organic anions of ILs bound to specific amino acid residues of AChE through noncovalent interactions such as π interactions and hydrogen bonds. The calculation results of binding free energy showed that an electrostatic interaction ($\Delta E_{\text{ele}} < -285$ kJ/mol) was the main driving force for the binding of ILs to AChE. The overall findings from this investigation demonstrate that a systematic approach is much more convincing. Future research in this direction will help design the next generation of biosafe ILs.

KEYWORDS: artificial intelligence, molecular docking, molecular dynamic simulation, toxicity of ionic liquids, emerging pollutants, design of green chemicals



1. INTRODUCTION

As organic salts mostly composed of organic cations and organic/inorganic anions, ionic liquids (ILs) have been widely used in the fields of chemical synthesis,¹ biomedicine,² and catalysis³ due to their special physical and chemical properties, such as negligible vapor pressure, high conductivity, and high solubility in both water and lipids. According to reports from Global Market Insights (www.gminsights.com), the IL market size exceeded a value of USD 530 million in 2014, and the consumption is expected to reach over 60 kilotons by 2022. Due to their increased production and use, ILs will inevitably enter the environment, mainly including soil and contamination of groundwater and even drinking water, thereby causing great human health risks.^{4,5} Furthermore, their high solubility and slow degradation make ILs potential persistent aquatic pollutants. As with many chemicals of future, current, or past use, ILs cannot be discriminately used before in-depth toxicity analyses are performed and pertinent toxicity mechanisms are fully understood.^{6,7}

The toxicity of ILs has been evaluated by resorting to different biological models, including bacteria⁸ and fungi,⁹ invertebrates,¹⁰ fish,¹¹ algae,¹² plants,¹³ and mammalian cell lines.¹⁴ Recently,¹⁵ a high level of an IL (i.e., 1-octyl-3-methylimidazolium, also referred to as M8OI) was detected in soils around a landfill waste site in the northeast of England, and the IL was found to have the potential to trigger primary

biliary cholangitis. According to statistics,^{6,16} the possible synthesis of ILs can reach up to a quintillion (i.e., 10^{18}), and even more if binary and ternary systems are considered. It is impossible to experimentally evaluate the toxicity of such a large number of ILs by *in vitro* and/or *in vivo* toxicology tests. Therefore, a computational approach is promising for this purpose.¹⁷

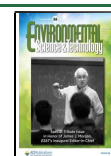
Previously, the QSAR (quantitative structure–activity relationship) modeling approach has been applied to predict the toxicity of ILs.^{18–23} However, most QSAR models have been constructed by using one machine learning method. Therefore, the resulting models were limited by the power of individual approaches. In addition, although the use of some new types of descriptors has improved model prediction accuracy, it has also led to the lack of interpretability of the modeling results. More importantly, the QSAR modeling only considered the structure of the IL itself, while ignoring the overall interactions between ILs and biological systems, such as

Received: May 7, 2021

Revised: August 5, 2021

Accepted: September 28, 2021

Published: October 12, 2021



ACS Publications

© 2021 American Chemical Society

14720

<https://doi.org/10.1021/acs.est.1c02960>
Environ. Sci. Technol. 2021, 55, 14720–14731

proteins. Therefore, it is difficult to understand the toxicity mechanisms of ILs by QSAR alone.

In order to obtain more reliable and interpretable QSAR models and reveal the related toxicity mechanisms, we performed a systematic modeling of ILs to AChE enzyme inhibition. As a key enzyme in biological nerve conduction, AChE can catalyze the breakdown of acetylcholine and of some other choline esters that function as neurotransmitters, thereby preventing the excitatory effect of a neurotransmitter on the postsynaptic membrane and ensuring the normal delivery of nerve signals in an organism. Previous studies^{24,25} have proved that some ILs can affect the activities of AChE enzyme, and the inhibition of AChE enzyme will lead to a number of biomedical problems, such as agitation, miosis, and even severe neuromuscular disorder. Hence, designing ILs with low AChE enzyme toxicity is conducive to the sustainability of the IL industry.²⁶ To this end, in combination with MOE (molecular operating environment)²⁷ and Dragon²⁸ descriptors, *k*NN (*k*-nearest neighbor), RF (random forest), XGBoost (extreme gradient boosting), and ANN (artificial neural network) algorithms were first applied to construct QSAR models. Molecular docking and molecular dynamics simulations (MD) were then used to understand the molecular mechanisms of interactions between the ILs and AChE enzyme. We believe that the constructed predictive models and in-depth mechanism analysis can help in the design of environmentally benign ILs in the future.

2. MATERIALS AND METHODS

2.1. Data Set. Herein, the data set used was about the toxicity of 153 ILs toward AChE enzyme, which was collected and curated from the literature.²⁹ The enzyme toxicity data set was originally generated from the UFT (center for environmental research and sustainable technology) at the University of Bremen,²⁶ and the uniform experimental standards in single laboratory tests can ensure the quality of these toxicity data. In addition, the data set has been rigorously checked and widely used for QSAR modeling in previous studies.^{29–31} Hence, the high-quality data can show promise for the following machine learning and molecular modeling. In the data set, the AChE enzyme inhibition of an IL was experimentally determined by the half maximum effective (EC₅₀) concentration (μM), which were converted into a logarithmic form. The initial data set was randomly divided into a training set (80% of the whole set) and test set (20% of the whole set). Detailed information about the IL data set can be found in Table S1.

2.2. Descriptor Calculation and Machine Learning Approaches. On the basis of the canonical SMILES (simplified molecular input line entry specification) representations of the corresponding cations and anions, the molecular descriptors of the considered ILs were computed using the MOE²⁷ and Dragon²⁸ software. Here, a total of 206 MOE descriptors and 3150 Dragon descriptors was generated for each cation or anion. The calculated descriptors covered the physicochemical, constitutional, geometrical, topological, and spatial properties of the ILs. In the present study, three traditional machine learning approaches (i.e., random forest, *k*-nearest neighbor, and extreme gradient boosting) and one deep learning approach (i.e., artificial neural network) were applied to construct the predictive QSAR models. The RF predictor³² consisted of a large number of decision trees and combined the outputs from the individual tree to generate the final predictions. The key point of the *k*NN method³³ was to

define a number of training samples closest in distance to the new sample, and it used the weighted average of the defined nearest neighbors as its prediction. XGBoost³⁴ is an optimized gradient boosting method and used a more regularized model formation to control overfitting. As implied by its name, the artificial neural network³⁵ was designed to simulate the structure and functionalities of biological neural networks. The general structure of an artificial neural network mainly contained one input layer, one or more hidden layers, and one output layer. The machine learning library in Python, scikit-learn v0.19.2, was used to construct the RF and XGBoost regression model. The *k*NN method was implemented by an in-house program, and the deep learning model was constructed by the Keras v2.2.0 library using TensorFlow v1.14.0 as the back end. On the basis of the predictions generated from four individual models, we also constructed an extra consensus model to avoid the instability of predictions by randomly selecting a certain machine learning method.

All generated models were validated using 5-fold cross-validation and external validation. The model predictivity was accessed by the determination coefficient (*R*²) (eq 1) and the root-mean-square error (RMSE) (eq 2). As the machine learning models have already obtained high predictive ability under default parameters, we did not perform hyperparameter tuning, to avoid wasting computational resources. In the training of RF, XGBoost, and *k*NN models, default parameters of the scikit-learn package or our in-house program were used are given in Table S2. In this study, the DNN models were implemented with three hidden layers containing 512, 128, and 64 neurons, sequentially. Other parameters are shown in Table S2. Model training was done for 300 epochs when loss values no longer significantly decreased (Figure S1). The training loss and validation loss did not show much difference, indicating that the DNN model was not overfitting. The source codes of all machine learning models can be found at <https://github.com/yanxiliang1991/ILs>.

$$R^2 = \frac{\sum_{i=1}^n (y_i^{\text{obs}} - y_i^{\text{pred}})^2}{\sum_{i=1}^n (y_i^{\text{obs}} - \bar{y}_i^{\text{obs}})^2} \quad (1)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^n (y_i^{\text{obs}} - y_i^{\text{pred}})^2}{n}} \quad (2)$$

y_i^{pred} is the predicted value for each particular IL, y_i^{obs} is the observed value for each particular IL; \bar{y}_i^{obs} is the mean value over all ILs, and *n* is the number of ILs.

2.3. Molecular Docking. The AChE protein structure was obtained from the RSCB Protein Data Bank (PDB ID: 4BDT). The crystal structure of AChE formed a complex with the original reference ligands huprine W and tacrine, respectively. Molecular docking was carried out using MOE software. The canonical SMILES of cations and organic anions were first loaded into MOE and converted to energy-minimized three-dimensional structures with the MMFF94x force field (merck molecular force field 94x). After water molecules and one excess ligand (i.e., tacrine) were removed, the ionization state and implicit hydrogens were then assigned to the processed protein structures using the protonate 3D procedure of MOE.

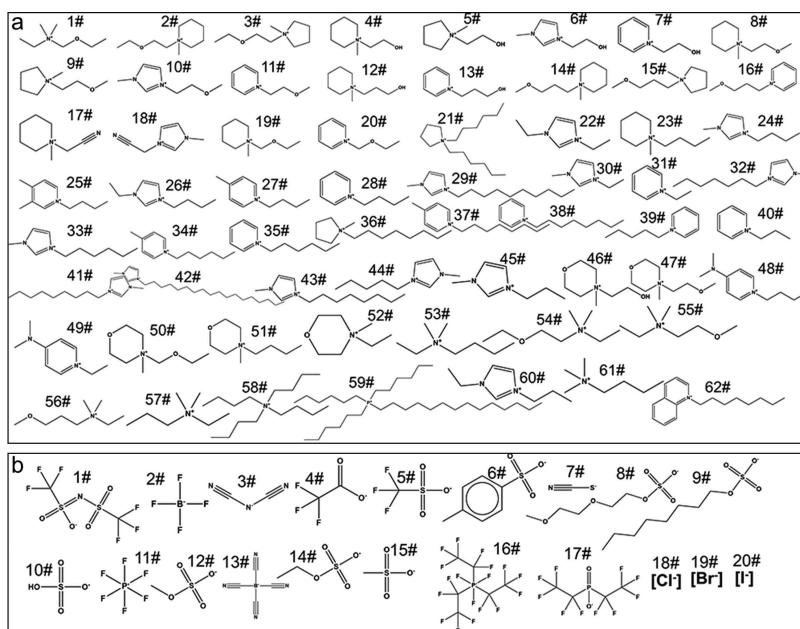


Figure 1. Visualization of the cation and anion structures. Kekulé representations of (a) 62 cations and (b) 20 anions. These structures were then converted to SMILES format for descriptor calculations and optimized for molecular docking and molecular dynamics simulation.

The coordinates of the binding site were defined on the basis of the cocrystallized ligand: i.e., huprine W. Ligand binding poses were generated by docking cations/organic anions to the binding site. The triangle matcher method was used to place the ligand in the binding site, and 30 initial ligand placements were first evaluated by the ASE scoring function. Next, the initial placements were further refined to 5 poses by the affinity dG scoring function. From these 5 refined poses, we selected the best docking structures (i.e., the highest docking score) for a docking analysis and MD simulation.

2.4. Molecular Dynamics Simulation. GROMACS v2020.4³⁶ was used to perform the MD simulation of ILs binding with AChE enzyme. Here, two representative ILs, i.e., IL152 and IL153, were selected. The best docking structure of the IL-AChE complex derived from the molecular docking results was used as the initial conformation of the MD simulation. The IL-AChE complex was first placed at the geometric center of a cubic box with a length of 8 nm. Water molecules simulated with the SPC/E (extended simple point charge) model were added to the box for IL-AChE complex solvation. On consideration of the negative charge on the surface of the protein, Na⁺ was randomly introduced into the water box to ensure the absolute electrical neutrality of a simulation system. The OPLS-AA (all-atom optimized potentials for liquid simulations) force field³⁷ was used to develop the topology files of ILs and AChE enzyme. To remove the bad initial contacts that may affect subsequent equilibrium simulations, energy minimization of the system was first performed using the steepest descent method. The energy-minimized system was then respectively equilibrated in two phases: i.e., the NVT (moles, volume, and temperature are conserved) and NPT (moles, pressure, and temperature are conserved) ensembles. On equilibration, the position restraints were released and the unrestrained production MD simulation was performed. Berendsen coupling was used for both temperature and pressure to quickly equilibrate. The temperature was fixed at 310 K with a coupling constant of 0.1 ps, and

the pressure was kept at 1 bar with a coupling constant of 2 ps. The protein and nonproteins were coupled as separate groups. All nonbonded interactions were truncated at a cutoff of 1.4 nm, and the particle-mesh-Ewald algorithm was used to calculate the long-range electrostatic interactions. The covalent bonds were constrained using the Lincs algorithm. Each simulation was run for 50 ns with a time step of 2 fs. The g_mmpbsa tool,³⁸ implementing the MM/PBSA (molecular mechanics/Poisson Boltzmann surface area) approach, was applied to estimate the binding free energies between the ILs and AChE enzyme.

2.5. DFT Calculation for Structure Optimization and Atomic Charge Generation. The molecular structures in the ground state of all cations and anions were fully optimized using density functional theory (DFT) calculations. In this study, B3LYP (Becke-three-parameter–Lee–Yang–Parr) hybrid functional in combination with the 6-31+G(d,p) basis set was employed. The CHELPG (charges from electrostatic potentials using a grid-based method) scheme at B3LYP/6-31+ G(d,p) was applied to fit atomic charges in order to reproduce the electrostatic potential at the surface of the molecule.³⁹ Here, the CHELPG charges were obtained from a single-point Gaussian run of the optimized structures. All DFT calculations were implemented with the Gaussian 09 package and Gauss-View visualization program. As shown in Figure S2, the optimized structures were then used for descriptor calculations and atomic charge generation.

3. RESULTS AND DISCUSSION

3.1. Diversity of the Structure and Toxicity Data Set of ILs. The data set contained a total of 153 ILs that was composed of 62 cations and 20 anions. The detailed structures of the cations and anions are shown in Figure 1. According to the structures of 62 cations, the 153 ILs can be divided into quaternary amine ILs (e.g., cation1), piperidine ILs (e.g., cation2), pyrrolidine ILs (e.g., cation3), imidazole ILs (e.g., cation6), pyridine ILs (e.g., cation7), morpholine ILs (e.g.,

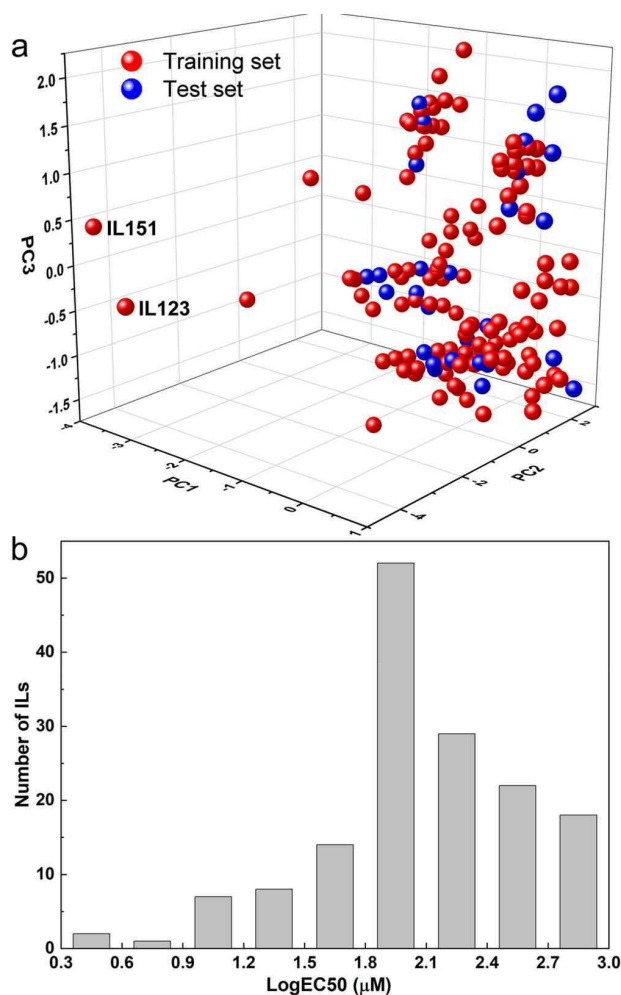


Figure 2. Analysis of IL structure diversity and AChE enzyme inhibition values. (a) Principal component analysis of 122 training set ILs (red) and 31 test set ILs (blue) based on the calculated MOE descriptors. (b) Histogram of experimental toxicity values. The diversity of IL structures and wide distribution of toxicity values were conducive to machine learning.

Table 1. Coefficient of Determination (R^2) and Root Mean Square Error (RMSE) (in Parentheses) from k NN, RF, XGBoost, ANN, and Consensus Modeling Results

| machine learning method | MOE descriptor | | Dragon descriptor | |
|-------------------------|-------------------------|---------------------|-------------------------|---------------------|
| | 5-fold cross validation | external validation | 5-fold cross validation | external validation |
| RF | 0.77 (0.25) | 0.95 (0.11) | 0.79 (0.23) | 0.94 (0.11) |
| k NN | 0.87 (0.18) | 0.91 (0.15) | 0.84 (0.20) | 0.94 (0.12) |
| XGBoost | 0.85 (0.20) | 0.94 (0.12) | 0.82 (0.22) | 0.92 (0.14) |
| ANN | 0.73 (0.27) | 0.78 (0.23) | 0.84 (0.21) | 0.77 (0.23) |
| consensus | 0.87 (0.18) | 0.93 (0.13) | 0.88 (0.17) | 0.94 (0.12) |

cation46), quaternary phosphorus ILs (e.g., cation59), and quinoline ILs (e.g., cation62). Although most of the anions were organics, a total of 58 ILs contained one of three halogen anions, which were chloride, bromide, and iodide. These structures were then converted into SMILES format for descriptor calculations and optimized for molecular docking

and molecular dynamics simulation. To visualize the chemical space of 153 ILs, we performed a principal component analysis (PCA) on the basis of the calculated descriptors. Here, the calculated MOE descriptors were used for each IL. The top three principal components, which accounted for 49% of the total descriptor variance, were used to show the distribution of 153 ILs in a 3D chemical space.

As shown in Figure 2a, in a view as the chemical space of ILs, all ILs were structurally different due to various cations and anions. Furthermore, there were also two structural outliers (i.e., IL123 and IL151) in the training set. As shown in Table S1, IL123 and IL151 were the only two quaternary phosphorus ILs that contained the same cation (i.e., cation59) but different anions. The central phosphorus atom and four long side chains made cation59 different from other cations (Figure 1). In the data set, the toxicity values ($\log EC_{50}$) also exhibited a wide distribution, ranging from 0.3 to 2.98 (Figure 2b). IL153 composed of cation62 and anion2 caused the highest enzyme activity inhibition, while IL152 composed of cation47 and anion18 was the least toxic. The diversity of IL structures and wide distribution of toxicity values were conducive to the subsequent computational modeling, especially for machine learning.

3.2. Development of Reliable QSAR Models by Combining Various Machine Learning Approaches. In order to design efficient and safe ILs, we set out to develop predictive QSAR models. Using calculated MOE and Dragon descriptors and four machine learning approaches (i.e., RF, k NN, XGBoost, and ANN), various QSAR models were constructed to predict the toxicological activity of ILs in AChE enzyme inhibition. In order to eliminate the instability of constructed models caused by random selection of machine learning methods, an extra consensus model was generated by averaging the predicted values of the RF, k NN, XGBoost, and ANN models. The model performance was evaluated by the R^2 and RMSE values of 5-fold cross-validation (R^2_{SCV} and $RMSE_{SCV}$) and external validation (R^2_{val} and $RMSE_{val}$) procedures and is shown in Table 1. Overall, all machine learning models showed good predictive ability with $R^2 > 0.73$ and $RMSE < 0.27$. To avoid correlation by chance of the machine learning models, we additionally performed Y-scrambling permutation tests. Briefly, we constructed 100 random machine learning models, where molecular features remained the same but toxicity values underwent different permutations. As shown in Figure S3, $RMSE_{SCV}$ and $RMSE_{val}$ for QSAR models were at least 2 times lower than those for randomly obtained models, indicating that all machine learning models were not obtained by chance.

In comparison with RF and ANN models, XGBoost and k NN models exhibited better predictive ability. Specially, we found that the deep learning (i.e., ANN) model did not show better predictive ability in comparison to models constructed by traditional machine learning methods (i.e., RF, k NN, and XGBoost) in the present study. Although deep learning methods have made many remarkable achievements in the fields of drug discovery,⁴⁰ chemical synthesis,⁴¹ and image recognition,⁴² the higher predictive ability of this method was driven by big data. For such a small data set, the superiority of deep learning cannot be reflected.^{43,44} In comparison to individual models, consensus models were normally advantageous due to the avoidance of potential overfitting and arbitrary selection of modeling approaches on the basis of our previous studies.^{44,45} Expectedly, as shown in Table 1, the

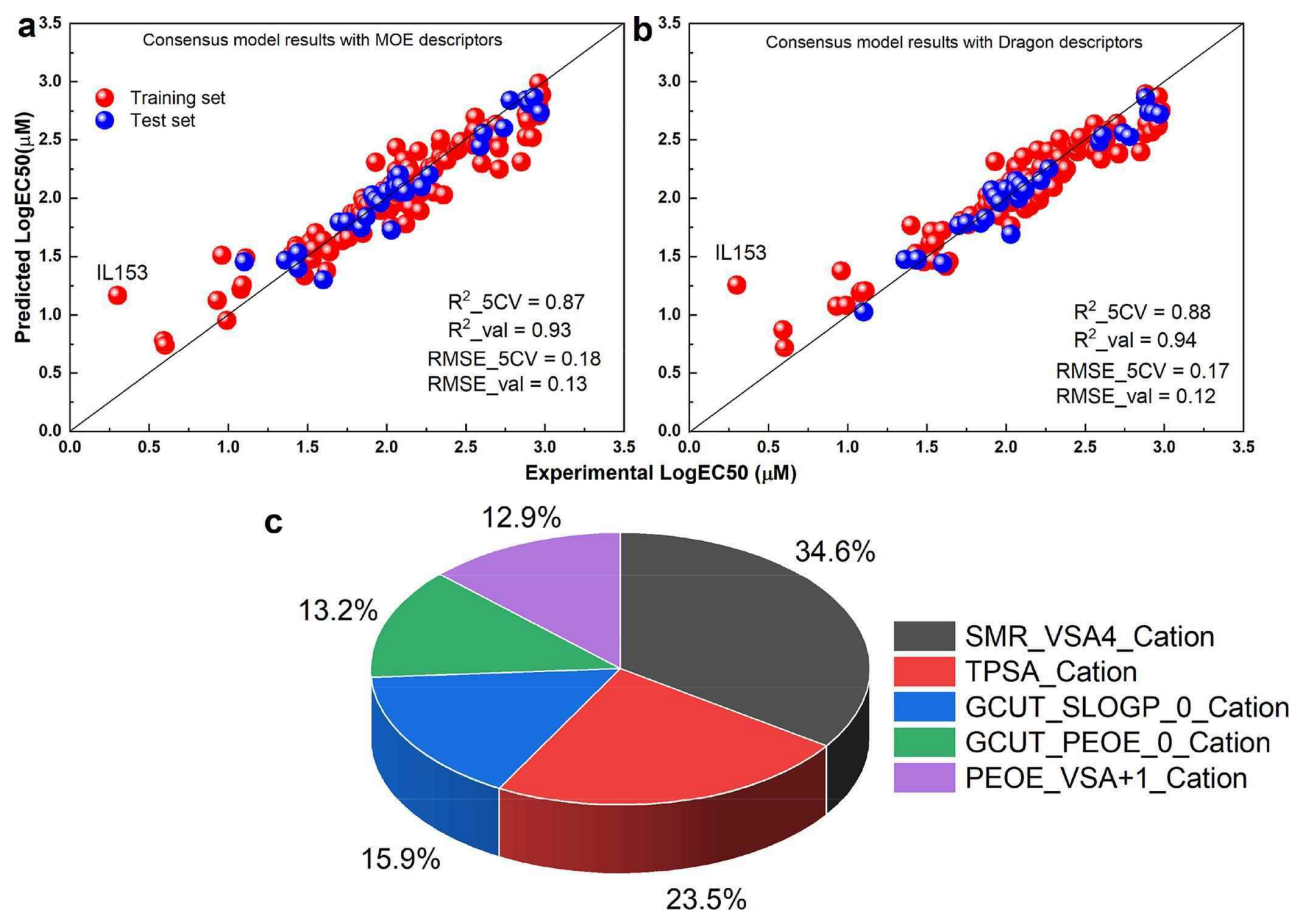


Figure 3. Performance and interpretation of the machine learning models. Correlations between experimental values and consensus model predictions using (a) MOE descriptors and (b) Dragon descriptors. Red dots are ILs in the training set, and blue dots are those in the test set. The coefficient of determination (R^2) and root-mean-square error (RMSE) from consensus modeling results are also shown. (c) Contributions of the top five MOE descriptors from kNN modeling results. The percentage contributions of descriptors were harmonized.

consensus models achieved almost the same predictive ability as the optimal performance of all individual models for both cross-validation and external validation. Therefore, the outputs of consensus models were selected as the final prediction results. Figure 3a,b shows the correlations between the experimental values of log EC₅₀ and the predictions obtained from consensus models. The high R^2 and low RMSE values indicated that the predictions were accurate for both training and external validation purposes. However, prediction outliers were also noticeable. For example, the experimental toxicity value of IL153 was 0.3 μM, while the consensus predictions were 1.17 μM (MOE descriptors) and 1.26 μM (Dragon descriptors). In the IL data set, the closest structural analogue of IL153 was IL141, which had a relatively higher activity value of 1.48 μM. IL153 and IL141 had the same anions (i.e., anion2) but different cations (i.e., cation62 and cation34). At present, neither MOE nor Dragon descriptors can discriminate these two cations. This issue suggested a direction for the future development of more advanced descriptors. In addition, we also explored the effect of molecular descriptors calculated from different input structures on model performance. As described above, we have generated 206 MOE and 3150 Dragon descriptors using SMILES representations. Similarly, we calculated same MOE and Dragon descriptors using optimized structures from DFT calculations. These new

generated descriptors were then used to build machine learning models with the same parameters shown in Table S2. The model performances (R^2) did not show much differences from those of machine learning models constructed with SMILES-based descriptors (Figure S4). The reason was that these descriptors were two-dimensional descriptors (e.g., molecular weight, number of atoms, and pharmacophore features) that only used the atoms and the connection information of the molecule for the calculation. 3D coordinates and individual conformations were not considered when these two-dimensional descriptors were applied.

As one of the OECD (organization for economic cooperation and development) principles for QSAR, model interpretation can help us identify a number of structural features (i.e., molecular descriptors) responsible for the corresponding toxicity. These could be used to elucidate potential toxicity mechanisms and thus provide guidance for green IL design. On the basis of the use frequency of the descriptors among the kNN model, we obtained the top five ranking descriptors, as shown in Figure 3c. The high frequency of a descriptor use indicated its critical contribution to the final models and also its important role in the corresponding toxicity. Our results showed that the cations played the most important role in the AChE enzyme inhibition, which was consistent with the experimental results. For example, IL1 and

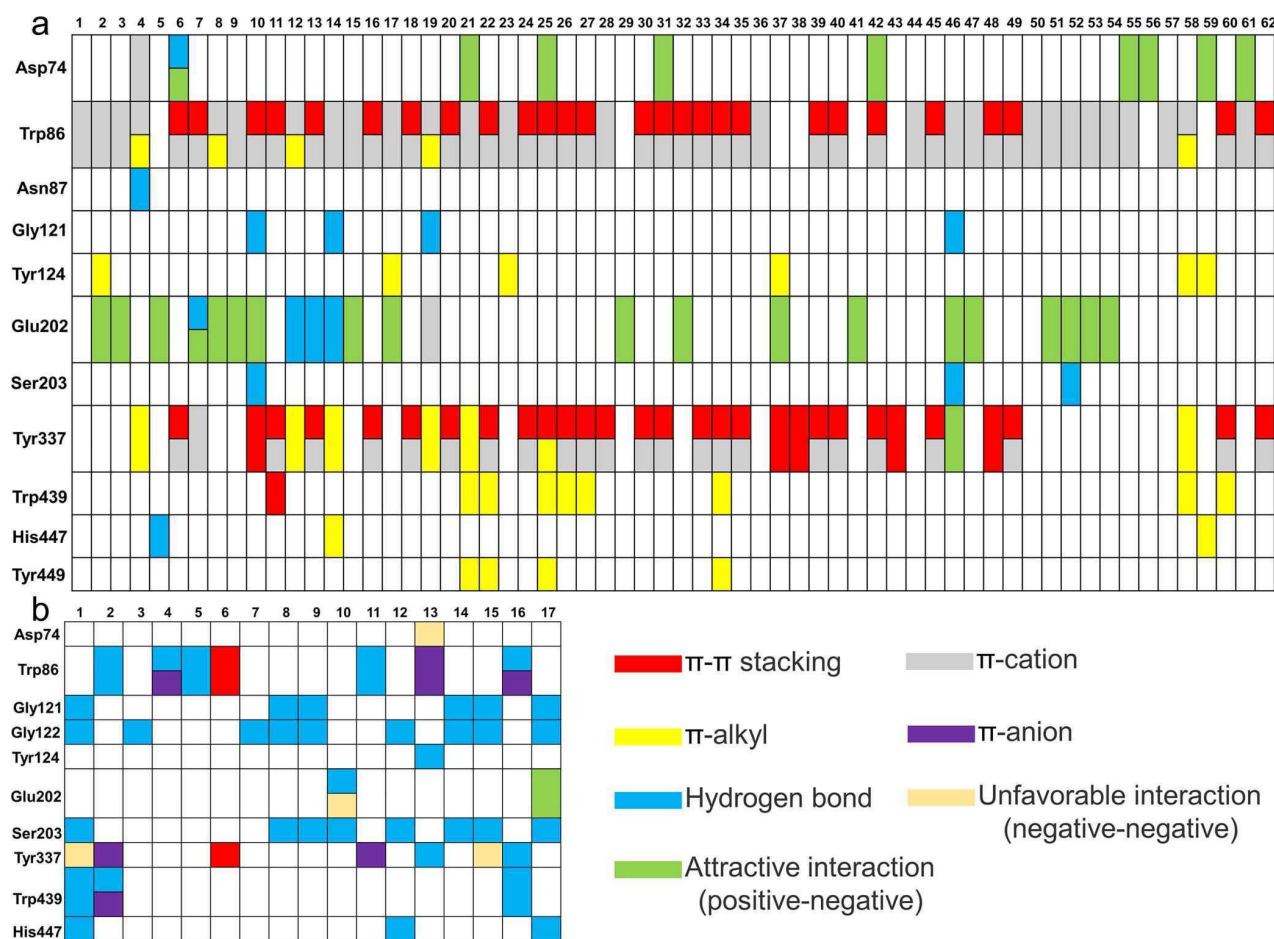


Figure 4. Summary of the main noncovalent interactions between the ligands and AChE enzyme from molecular docking results. Interactions between amino acid residues and (a) 62 cations and (b) 17 organic anions were respectively generated from their binding pose with the highest docking score. The horizontal numbers represent different cations and anions, and vertical labels represent different amino acid residues.

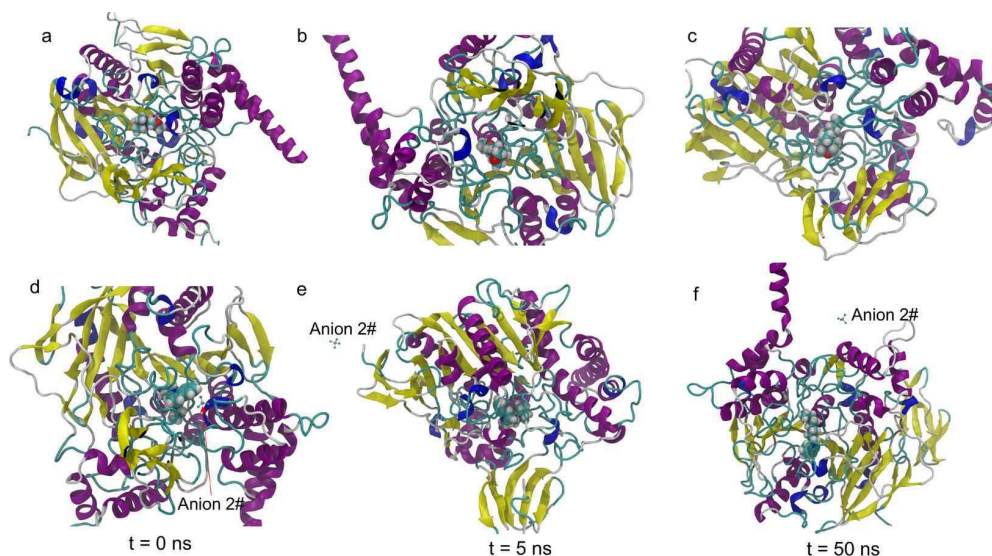


Figure 5. Snapshots of the molecular dynamics trajectory of an IL-protein complex using OPLS-AA parameters. Representative snapshots from the molecular dynamics trajectory of (a–c) the IL152-AChE complex and (d–f) the IL153-AChE complex at 0, 5, and 50 ns. The protein is shown in the NewCartoon drawing method, cation47 and cation62 are drawn in the VDW drawing method, and anion2 is displayed in the CPK drawing method. Water and ions are not shown for clarity. All images were rendered using the VMD software.

Table 2. Cation47/Cation62-AChE Binding Free Energy

| | interaction energy with AChE (kJ/mol) | |
|--|---------------------------------------|----------------------|
| | cation47 (IL152) | cation62 (IL153) |
| van der Waals energy (ΔE_{vdw}) | -94.787 ± 0.805 | -132.685 ± 1.000 |
| electrostatic energy (ΔE_{ele}) | -285.669 ± 2.060 | -406.670 ± 1.715 |
| polar solvation energy (ΔG_{polar}) | 175.372 ± 1.413 | 202.334 ± 2.105 |
| nonpolar solvation energy ($\Delta G_{\text{nonpolar}}$) | -12.719 ± 0.080 | -19.157 ± 0.124 |
| binding energy ($\Delta G_{\text{binding}}$) | -217.803 ± 1.895 | -356.179 ± 2.066 |

IL2 had the same cations (i.e., cation1) but different anions (i.e., anion18 and anion1), and their toxicity values were similar (2.36 and 2.3 μM). We also found that the van der Waals surface area, log P , and atomic partial charge contributed significantly to the AChE enzyme inhibition potential of ILs (Figure 3c). Previous studies have also shown that the chain length,⁴⁶ hydrophobicity,⁴⁷ and charge properties²¹ of ILs had a significant effect on its toxicity. For example, a longer alkyl chain IL usually has a van der Waals surface area, resulting in stronger inhibition of algal growth.⁴⁶ Detailed information about the top five ranked MOE descriptors can be seen in Table S3.

Furthermore, we applied this modified QSAR methodology covering multiple descriptors and machine learning methods to predict other ligand–protein interactions. Herein, three different data sets were used: (1) external data set1 contained 66 ligands interacting with human hepatic cytochrome P450 3A4 (CYP3A4),⁴⁸ (2) external data set2 contained 113 ligands interacting with the main protease of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2 M^{pro}),⁴⁹ and (3) external data set3 contained 43 ligands interacting with kappa-type opioid receptor (KOR) (PubChem Assay ID 1344503). External data set1 and data set2 were used for binary classification modeling, and external data set3 was used for regression modeling. Detailed information about the three external data sets can be found in Tables S4–S6. The predictivity of classification models was accessed by accuracy, and the predictivity of regression models was accessed by R^2 . All machine learning models were evaluated using 5-fold cross-validation and external validation. As shown in Table S7, all machine learning models can accurately predict the ligand–enzyme interactions. In comparison with individual models, the consensus predictions were normally close to the top performance of all individual models, which was consistent with the above analysis of IL toxicity prediction. These results

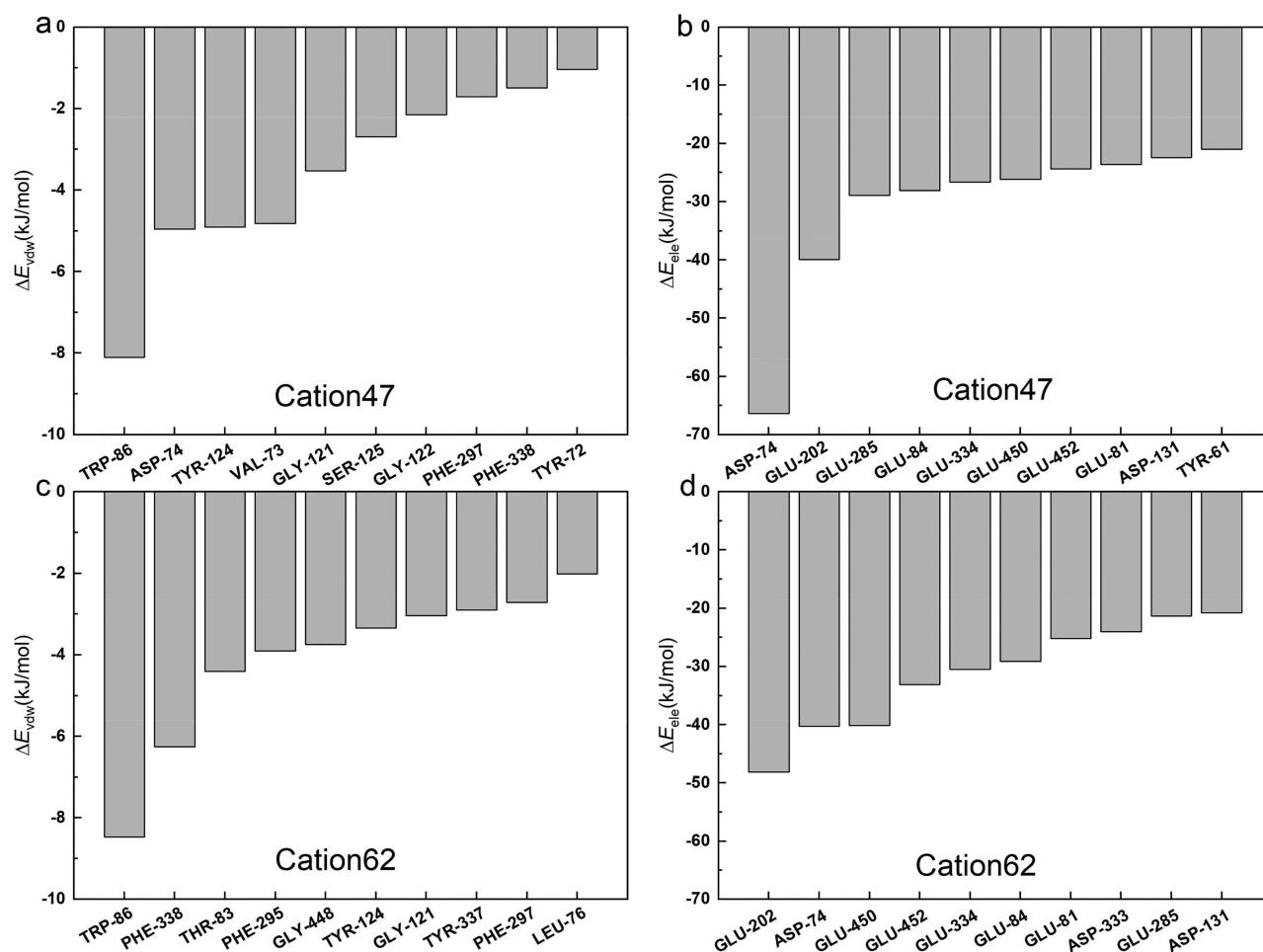


Figure 6. Analysis of residue-specific energy decomposition. The electrostatic energies (ΔE_{ele}) and van der Waals energies (ΔE_{vdw}) were quantitatively decomposed into residue-specific contributions for cation47–AChE binding (a, b) and cation62–AChE binding (c, d). The top ten amino acid residues with the most favorable contributions are presented.

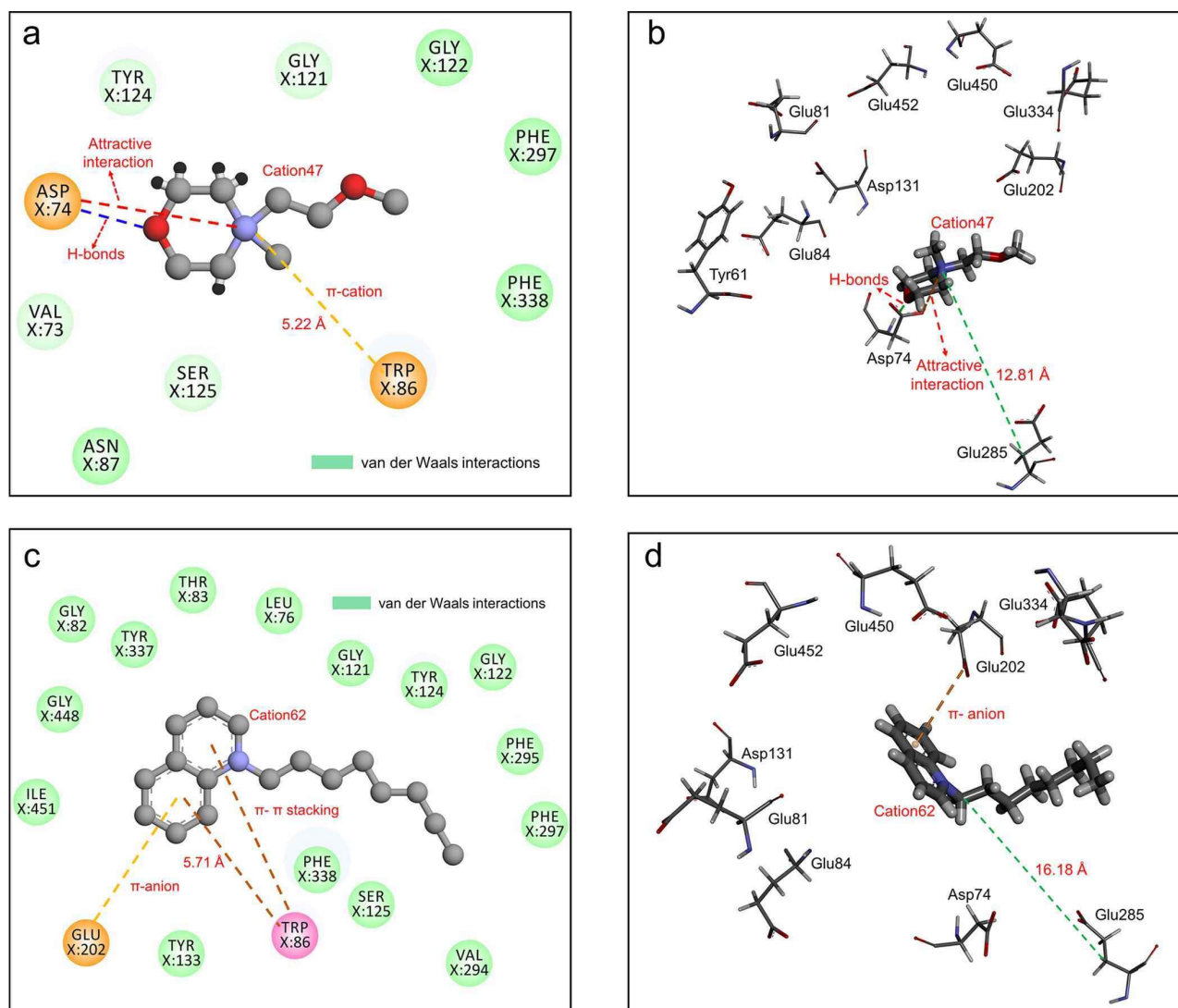


Figure 7. Binding modes of cation47/cation62 in the area of van der Waals interactions and electrostatic interactions. Key residues for van der Waals interactions between cation47/cation62 and AChE are shown in spheres (a, c), and these key residues for electrostatic interactions between cation47/cation62 and AChE are shown in sticks (b, d). The residue number and some noncovalent interactions (e.g., H bonds, π interactions) are also shown. The results were obtained from a frame of the last 5 ns molecular dynamics simulation.

show that QSAR models combined with multiple descriptors and machine learning methods can obtain more reliable predictions.

3.3. Identification of IL–Protein Binding Mode by Molecular Docking. In order to better understand the interactions between the ILs and AChE enzyme and provide an in-depth molecular mechanism analysis, molecular docking and molecular dynamics simulation were performed in the present study. Molecular docking can provide preliminary insights into the interaction between ILs and protein binding sites, while a molecular dynamics simulation can provide a more detailed analysis of the dynamic process of ILs binding with AChE enzyme. The binding poses of 62 cations and 17 organic anions that had the highest affinity to AChE LBD (ligand binding domain), as determined by the affinity dG score, were analyzed. The pocket site of AChE enzyme was surrounded by Asp74, Thr83, Trp86, Asn87, Gly120, Gly121, Gly122, Tyr124, Glu202, Ser203, Phe297, Tyr337, Phe338,

Tyr341, Trp439, Pro446, His447, and Tyr449 (Figure S5a). As shown in Figure 4, the cations and organic anions mainly interacted with amino acids at the binding site. Furthermore, we also found that the cations and organic anions bound to specific amino acid residues mainly through π -interactions (e.g., π -cation, π - π stacking, and π -alkyl interactions) and hydrogen bonds. The binding modes of some representative cations and organic anions can be seen in Figure S5b–e.

As noncovalent interactions, the π -interactions can be formed between the electron-rich π system and a cation (i.e., π -cation interactions), an anion (i.e., π -anion interactions), another molecule (e.g., π -alkyl interactions), and even another π system (i.e., π - π interactions). As shown in Figure 4a, π -interactions mainly existed between the cations and two key amino acid residues (i.e., Trp86 and Tyr337) containing an electron-rich π system. Similarly, the hydrogen bond was also a noncovalent interaction and mainly existed between a partially positively charged hydrogen atom attached to a highly

electronegative atom and another nearby electronegative atom. The electronegative atoms were particularly the second-row elements: e.g., nitrogen (N), oxygen (O), and fluorine (F). The N, O, and F atoms in the amino acid residues, as well as in the structure of the ILs, could form hydrogen bonds with each other. It was observed that the organic anions could form hydrogen bonds with several key amino acid residues, such as Gly121, Gly122, and Ser203 (Figure 4b). In addition, we also observed that there were halogen bonds and chalcogen bonds between certain anions and amino acid residues (Figure S5e). Similar to hydrogen bonds, a halogen bond or chalcogen bond occurs when there is a net attractive interaction between an electrophilic region associated with a halogen atom or chalcogen atom in a molecular entity and a nucleophilic region in another. Both halogen bonds and chalcogen bonds are beneficial to the binding of anions with AChE enzyme. However, the docking scores of anions were lower than those of the cations, indicating that the binding of anions to pocket sites was unstable. The main reason was that there were unfavorable repulsive interactions between the organic anions and negatively charged amino acid residues (e.g., Asp74). On the other hand, these negatively charged amino acid residues can strongly attract the cations through electrostatic interaction. A detailed discussion about the electrostatic interactions can be seen in the following analysis of MD simulation. The results further verified that the enzyme activity inhibition was mainly caused by the cations. The binding modes predicted by docking simulations, for the cations with higher docking scores, suggested them as potential candidates for AChE inhibitors.

3.4. Enzyme Binding Energy Is Correlated with the Toxicity of ILs. To explore the dynamic process and binding free energy of the interaction between ILs and the AChE enzyme, molecular dynamics simulations for two representative ILs (IL152 and IL153 with highest and lowest toxicity values, respectively) were carried out using the GROMACS software. The dynamic process of ILs binding an enzyme can be observed intuitively by visualizing snapshots of the dynamic trajectory at different time points. Figure 5 shows several representative snapshots of IL–protein complexes at 0, 5, and 50 ns of a molecular dynamics simulation; it is apparent the cations and anions behaved quite differently in the biological systems. The two cations (i.e., cation47 and cation62) were tightly bound to the protein during the whole process of molecular dynamics simulation, while the organic anion (i.e., anion2) broke away from the protein and entered the aqueous solution at 5 ns. The MD simulation results were consistent with the molecular docking score, showing that organic anions mainly acted as counterions. In addition, as an important property for molecular modeling, the effect of atomic charges on molecular docking and MD simulation were also explored. Herein, the atomic charges of ILs were further generated from DFT calculations using the CHELPG charge calculation scheme. Figure S6 shows the best docking poses of IL153 (i.e., cation62 and anion2) generated using the atomic charges assigned by MOE and the CHELPG scheme. It can be seen that these docking poses almost overlapped with each other and obtained similar docking scores. Similarly, the MD simulation results also did not show much difference when OPLS-AA parameters (Figure Sd–f) and CHELPG charges were used (Figure S7a–c). The anion2 broke away from the protein after 8 ns, while the cation62 can remain near the binding site during the 50 ns simulation (Figure S7a–c). As

shown in Table S8, most atomic charges of IL153 were similar when CHELPG charges and OPLS-AA parameters were used (i.e., 1.14*CM1A atomic charges), indicating that our docking and MD simulation results can be trusted.

In order to quantitatively explore the interaction between the ligand and the receptor, the binding free energies of cation47/cation62 and AChE enzyme were calculated using the MM/PBSA method. As shown in eq 3, the basic principle of the MM/PBSA method is to calculate the difference between the free energy of two solvated molecules in the bound and unbound states or to compare the free energies of different solvation conformations of the same molecule.³⁸

$$\Delta G_{\text{bind}} = G_{\text{complex}} - (G_{\text{receptor}} + G_{\text{ligand}}) \quad (3)$$

Many studies^{50,51} have demonstrated the success of the MM/PBSA method for estimating the binding free energies of small molecules to macromolecules. Before calculating the binding free energies, we first analyzed the dynamic stability of the two IL/AChE complexes through an RMSD (root-mean-square deviation) analysis. Herein, the time-dependent RMSD values of the complex backbone atoms were calculated with the program gmx rms by least-squares fitting of the structure to the reference structure. As shown in Figure S8, the conformations of the AChE/IL153 complex achieved equilibrium at around 20 ns while the equilibrium time for AChE/IL152 complex was around 30 ns, indicating that the AChE/IL153 complex was more stable than the AChE/IL152 complex. Overall, RMSD values were relatively small (<0.5 nm) for both complexes, and the fluctuation of RMSD values was within 0.1 nm after 30 ns. Therefore, it is rational to do the binding free energy calculation and free energy decomposition on the basis of the trajectories extracted from the last 5 ns MD simulations.

According to the calculation principle of the MM/PBSA method,³⁸ the binding free energy can be further decomposed into the vacuum potential energy (mainly including electrostatic interactions and van der Waals interactions) and the free energy of solvation (i.e., polar solvation and nonpolar solvation interactions). The values of four aforementioned energies, shown in Table 2, can be used to investigate the main driving force of IL–AChE enzyme binding. Overall, the binding free energies (i.e., -217.803 ± 1.895 and -356.179 ± 2.066 kJ/mol) indicated that both cation47 and cation62 could bind tightly to AChE enzyme. Obviously, cation62 with a lower $\Delta G_{\text{binding}}$ value can cause more severe enzyme activity inhibition. In comparison with the other binding energies, ΔE_{ele} (-285.669 ± 2.060 and -406.670 ± 1.715 kJ/mol, respectively) was relatively large due to the surface charge in the two cations and the amino acid residues, indicating that electrostatic interactions played the most important role in the binding of ILs with AChE enzyme. Furthermore, the van der Waals energy (ΔE_{vdw}), characterizing the strength of hydrophobic interactions, also provided a beneficial contribution to the binding process of the ILs with AChE enzyme. On the other hand, due to the large volume (Figure S5a) of the AChE binding pocket exposed to solvent, the free energy of solvation ΔG_{polar} in the two systems produced relatively large positive values, which was not beneficial to the binding process of the ILs with AChE enzyme. In summary, the electrostatic interactions between the cations and negatively charged amino acid residues can attract ILs toward the protein, after which the van der Waals interactions between the alkyl chain of cations and protein backbone/side chain also provide an important driving force for IL binding with the AChE enzyme.

In addition to the calculation of the total interaction energy of IL–AChE binding, the MM/PBSA method can provide a better description of the noncovalent interactions in terms of per-residue interaction energies. In this approach, the total interaction energy can be partitioned into the contributions from individual amino acid residues, allowing the quantification of their energy contributions. This information is essential in the design of green ILs since it helps to identify crucial amino acid residues for IL–AChE binding and also allows each interaction to be tuned. In this study, we decomposed ΔE_{ele} and ΔE_{vdw} into residue-specific contributions to analyze the contributions of different noncovalent interactions to the van der Waals and electrostatic components. Figure 6 illustrates the residue-specific interaction energy partitioning for the cation47-AChE and cation62-AChE complexes; it can be seen that the energy contribution profiles are similar for the two complexes. As shown in Figure 7, the contributions of van der Waals energies mainly come from the van der Waals interactions between the cations and the amino acid residues near the pocket site (e.g., Trp86, Asp74, and Phe338), while the contributions of electrostatic energies mainly come from the long-range electrostatic interactions between the cations and the negatively charged amino acid residues (e.g., Asp74, Glu202, and Glu285). In addition, some other noncovalent interactions, such as H bonds, π – π stacking, and π –cation, can also make beneficial contributions to the van der Waals energies and electrostatic energies (Figure 7). These analyses can help us better understand the toxicity mechanisms of ILs to AChE enzyme and design environmentally friendly ILs in the future.

4. ENVIRONMENTAL SIGNIFICANCE

Are ILs safe? The present data do not justify ILs to be classified as environmentally safe chemicals. Although ILs are safe to the atmosphere due to their low vapor pressure, their environmental accumulations, excellent miscibility with most media, and high stability have made them a persistent threat to environmental waters and soils. Actually, a large number of studies have shown that ILs can cause toxicity to various organisms, such as algae,¹² bacteria,⁸ fish,¹¹ plants,¹³ and even mammalian cell lines.¹⁴ Therefore, the toxicity risk of ILs and the underlying mechanisms must be elucidated to ensure their large-scale safe applications. On the basis of the experimental data of 153 ILs to AChE inhibition, we constructed predictive QSAR models using various machine learning methods and systematically explored the molecular mechanisms using molecular docking and molecular dynamics simulation. We found that more reliable and stable QSAR models can be constructed by combining the results from multiple machine learning approaches. Molecular docking results revealed that the cations and organic anions bound to specific amino acid residues through π interactions and hydrogen bonds. The binding energy calculations from molecular dynamics simulations showed that the electrostatic interaction energy was the main driving force for IL binding with an enzyme. Findings from this study provided a complementary approach to assess the toxicity of ILs and also contributed to a better understanding of the molecular mechanisms. Applications of artificial intelligence and molecular simulation provided a new paradigm for toxicological research to supplement time-consuming and laborious experimental approaches and solve unclear toxicity mechanisms from *in vitro* or *in vivo* experiments.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.est.1c02960>.

Training loss and validation loss against epochs, representative optimized structures generated from DFT calculations, model performances of Y-scrambling permutation tests, comparison of model performances using SMILES-based descriptors and optimized structure descriptors, the pocket site and binding modes of representative ionic liquids, comparison of docking results obtained using different atomic charges, analysis of the molecular dynamics trajectory of the IL153-protein complex using CHELPG charges, and backbone RMSD analysis during the MD simulations (PDF)

Inhibition log EC₅₀ values of 153 ILs toward AChE, critical parameters used in the machine learning models, detailed descriptions of the top five ranked MOE descriptors from the kNN model, external data set1 about 66 ligands interacting with human hepatic CYP3A4, external data set2 about 113 ligands interacting with SARS-CoV-2 M^{pro}, external data set3 about 43 ligands interacting with KOR, accuracy (external data set1 and external data set2) and determination coefficients (external data set3) from kNN, RF, XGBoost, ANN, and consensus modeling results, and comparison of atomic charges obtained from the CHELPG scheme and OPLS-AA parameters (XLSX)

■ AUTHOR INFORMATION

Corresponding Authors

Xiliang Yan – Institute of Environmental Research at Greater Bay Area, Key Laboratory for Water Quality and Conservation of the Pearl River Delta, Ministry of Education, Guangzhou University, Guangzhou 510006, People's Republic of China; orcid.org/0000-0003-4173-6228; Email: yanxiliang1991@gzhu.edu.cn

Bing Yan – Institute of Environmental Research at Greater Bay Area, Key Laboratory for Water Quality and Conservation of the Pearl River Delta, Ministry of Education, Guangzhou University, Guangzhou 510006, People's Republic of China; School of Environmental Science and Engineering, Shandong University, Qingdao 266237, People's Republic of China; orcid.org/0000-0002-7970-6764; Email: drbingyan@yahoo.com

Authors

Jiachen Yan – Institute of Environmental Research at Greater Bay Area, Key Laboratory for Water Quality and Conservation of the Pearl River Delta, Ministry of Education, Guangzhou University, Guangzhou 510006, People's Republic of China

Song Hu – School of Environmental Science and Engineering, Shandong University, Qingdao 266237, People's Republic of China

Hao Zhu – The Rutgers Center for Computational and Integrative Biology, Camden, New Jersey 08102, United States; orcid.org/0000-0002-3559-6129

Complete contact information is available at: <https://pubs.acs.org/doi/10.1021/acs.est.1c02960>

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (22036002), the National Key R&D Program of China (2016YFA0203103), and the introduced innovative R&D team project under the “The Pearl River Talent Recruitment Program” of Guangdong province (2019ZT08L387).

■ REFERENCES

- (1) Duan, C. W.; Hu, L. X.; Ma, J. L. Ionic Liquids as an Efficient Medium for the Mechanochemical Synthesis of α -AlH₃ Nano-Composites. *J. Mater. Chem. A* **2018**, *6* (15), 6309–6318.
- (2) Egorova, K. S.; Gordeev, E. G.; Ananikov, V. P. Biological Activity of Ionic Liquids and Their Application in Pharmaceuticals and Medicine. *Chem. Rev.* **2017**, *117* (10), 7132–7189.
- (3) Werner, S.; Haumann, M.; Wasserscheid, P. Ionic Liquids in Chemical Engineering. *Annu. Rev. Chem. Biomol. Eng.* **2010**, *1*, 203–230.
- (4) Costa, S. P. F.; Azevedo, A. M. O.; Pinto, P. C. A. G.; Saraiva, M. L. M. F. S. Environmental Impact of Ionic Liquids: Recent Advances in (Eco)Toxicology and (Bio)Degradability. *ChemSusChem* **2017**, *10* (11), 2321–2347.
- (5) Amde, M.; Liu, J. F.; Pang, L. Environmental Application, Fate, Effects, and Concerns of Ionic Liquids: A Review. *Environ. Sci. Technol.* **2015**, *49* (21), 12611–12627.
- (6) Chatel, G.; Naffrechoux, E.; Draye, M. Avoid the PCB Mistakes: A More Sustainable Future for Ionic Liquids. *J. Hazard. Mater.* **2017**, *324*, 773–780.
- (7) Oskarsson, A.; Wright, M. C. Ionic Liquids: New Emerging Pollutants, Similarities with Perfluorinated Alkyl Substances (PFASs). *Environ. Sci. Technol.* **2019**, *53* (18), 10539–10541.
- (8) Ghanem, O. B.; Mutalib, M. I. A.; El-Harbawi, M.; Gonfa, G.; Kait, C. F.; Alitheen, N. B. M.; Leveque, J.-M. Effect of Imidazolium-Based Ionic Liquids on Bacterial Growth Inhibition Investigated via Experimental and QSAR Modelling Studies. *J. Hazard. Mater.* **2015**, *297*, 198–206.
- (9) Petkovic, M.; Ferguson, J.; Bohn, A.; Trindade, J.; Martins, I.; Carvalho, M. B.; Leitão, M. C.; Rodrigues, C.; Garcia, H.; Ferreira, R.; Seddon, K. R.; Rebelo, L. P. N.; Silva Pereira, C. Exploring Fungal Activity in the Presence of Ionic Liquids. *Green Chem.* **2009**, *11* (6), 889–89.
- (10) Luo, Y. R.; San-Hu, W.; Li, X. Y.; Yun, M. X.; Wang, J. J.; Sun, Z. J. Toxicity of Ionic Liquids on the Growth, Reproductive Ability, and ATPase Activity of Earthworm. *Ecotoxicol. Environ. Saf.* **2010**, *73* (5), 1046–1050.
- (11) Ruokonen, S. K.; Sanwald, C.; Sundvik, M.; Polnick, S.; Vyavaharkar, K.; Duša, F.; Holding, A. J.; King, A. W. T.; Kilpeläinen, I.; Lämmerhofer, M.; Panula, P.; Wiedmer, S. K. Effect of Ionic Liquids on Zebrafish (*Danio Rerio*) Viability, Behavior, and Histology; Correlation between Toxicity and Ionic Liquid Aggregation. *Environ. Sci. Technol.* **2016**, *50* (13), 7116–7125.
- (12) Ventura, S. P. M.; Gurbisz, M.; Ghavre, M.; Ferreira, F. M. M.; Gonçalves, F.; Beadham, I.; Quilty, B.; Coutinho, J. A. P.; Gathergood, N. Imidazolium and Pyridinium Ionic Liquids from Mandelic Acid Derivatives: Synthesis and Bacteria and Algae Toxicity Evaluation. *ACS Sustainable Chem. Eng.* **2013**, *1* (4), 393–402.
- (13) Pawłowska, B.; Telesiński, A.; Platkowski, M.; Stręk, M.; Śnioszek, M.; Biczak, R. Reaction of Spring Barley and Common Radish on the Introduction of Ionic Liquids Containing Asymmetric Cations to the Soil. *J. Agric. Food Chem.* **2017**, *65* (23), 4562–4571.
- (14) Mikkola, S. K.; Robciuc, A.; Lokajová, J.; Holding, A. J.; Lämmerhofer, M.; Kilpeläinen, I.; Holopainen, J. M.; King, A. W. T.; Wiedmer, S. K. Impact of Amphiphilic Biomass-Dissolving Ionic Liquids on Biological Cells and Liposomes. *Environ. Sci. Technol.* **2015**, *49* (3), 1870–1878.
- (15) Probert, P. M.; Leitch, A. C.; Dunn, M. P.; Meyer, S. K.; Palmer, J. M.; Abdelghany, T. M.; Lakey, A. F.; Cooke, M. P.; Talbot, H.; Wills, C.; McFarlane, W.; Blake, L. I.; Rosenmai, A. K.; Oskarsson, A.; Figueiredo, R.; Wilson, C.; Kass, G. E.; Jones, D. E.; Blain, P. G.; Wright, M. C. Identification of a Xenobiotic as a Potential Environmental Trigger in Primary Biliary Cholangitis. *J. Hepatol.* **2018**, *69* (5), 1123–1135.
- (16) Mell, A.; Kragl, U. Ionic Liquids. *Clean Prod. Process.* **1999**, *1* (4), 223–236.
- (17) Cao, L.; Zhu, P.; Zhao, Y.; Zhao, J. Using Machine Learning and Quantum Chemistry Descriptors to Predict the Toxicity of Ionic Liquids. *J. Hazard. Mater.* **2018**, *352*, 17–26.
- (18) Zhao, Y.; Zhao, J.; Huang, Y.; Zhou, Q.; Zhang, X.; Zhang, S. Toxicity of Ionic Liquids: Database and Prediction via Quantitative Structure-Activity Relationship Method. *J. Hazard. Mater.* **2014**, *278*, 320–329.
- (19) Cho, C. W.; Yun, Y. S. Correlating Toxicological Effects of Ionic Liquids on *Daphnia Magna* with in Silico Calculated Linear Free Energy Relationship Descriptors. *Chemosphere* **2016**, *152*, 207–213.
- (20) Melo, E. B. De. A Structure-Activity Relationship Study of the Toxicity of Ionic Liquids Using an Adapted Ferreira-Kirali Hydrophobicity Parameter. *Phys. Chem. Chem. Phys.* **2015**, *17* (6), 4516–4523.
- (21) Kang, X.; Chen, Z.; Zhao, Y. Assessing the Ecotoxicity of Ionic Liquids on *Vibrio Fischeri* Using Electrostatic Potential Descriptors. *J. Hazard. Mater.* **2020**, *397*, 122761.
- (22) Cruz-Monteagudo, M.; Ancede-Gallardo, E.; Jorge, M.; Cordeiro, M. N. D. S. Chemoinformatics Profiling of Ionic Liquids-Automatic and Chemically Interpretable Cytotoxicity Profiling, Virtual Screening, and Cytotoxicophore Identification. *Toxicol. Sci.* **2013**, *136* (2), 548–565.
- (23) Jafari, M.; Keshavarz, M. H.; Salek, H. A Simple Method for Assessing Chemical Toxicity of Ionic Liquids on *Vibrio Fischeri* through the Structure of Cations with Specific Anions. *Ecotoxicol. Environ. Saf.* **2019**, *182*, 109429.
- (24) Stock, F.; Hoffmann, J.; Ranke, J.; Ondruschka, B.; Jastorff, B. Effects of Ionic Liquids on the Acetylcholinesterase-a Structure-Activity Relationship Consideration. *Green Chem.* **2004**, *6* (6), 286–290.
- (25) Thuy Pham, T. P.; Cho, C.-W.; Yun, Y.-S. Environmental Fate and Toxicity of Ionic Liquids: A Review. *Water Res.* **2010**, *44* (2), 352–372.
- (26) Ranke, J.; Stolte, S.; Sto, R.; Arning, J.; Jastorff, B. Design of Sustainable Chemical Products s The Example of Ionic Liquids. *Chem. Rev.* **2007**, *107* (6), 2183–2206.
- (27) Vilar, S.; Cozza, G.; Moro, S. Medicinal Chemistry and the Molecular Operating Environment (MOE): Application of QSAR and Molecular Docking to Drug Discovery. *Curr. Top. Med. Chem.* **2008**, *8* (18), 1555–1572.
- (28) Mauri, A.; Consonni, V.; Pavan, M.; Todeschini, R. Dragon Software: An Easy Approach to Molecular Descriptor Calculations. *Match* **2006**, *56* (2), 237–248.
- (29) Zhu, P.; Kang, X.; Zhao, Y.; Latif, U.; Zhang, H. Predicting the Toxicity of Ionic Liquids toward Acetylcholinesterase Enzymes Using Novel QSAR Models. *Int. J. Mol. Sci.* **2019**, *20* (9), 2186.
- (30) Basant, N.; Gupta, S.; Singh, K. P. Predicting Acetyl Cholinesterase Enzyme Inhibition Potential of Ionic Liquids Using Machine Learning Approaches: An Aid to Green Chemicals Designing. *J. Mol. Liq.* **2015**, *209*, 404–412.
- (31) Cho, C.; Yun, Y. Interpretation of Toxicological Activity of Ionic Liquids to Acetylcholinesterase Inhibition via in Silico Modelling. *Chemosphere* **2016**, *159*, 178–183.
- (32) Svetnik, V.; Liaw, A.; Tong, C.; Christopher Culberson, J.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (6), 1947–1958.
- (33) Zhang, Z. Introduction to Machine Learning: K-Nearest Neighbors. *Ann. Transl. Med.* **2016**, *4* (11), 218–218.

- (34) Chen, T.; He, T.; Benesty, M.; Khotilovich, V.; Tang, Y.; Cho, H. Xgboost: Extreme Gradient Boosting. *R Packag. version* **2015**, *4* (2), 1–4.
- (35) Krogh, A. What Are Artificial Neural Networks? *Nat. Biotechnol.* **2008**, *26* (2), 195–197.
- (36) Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. GROMACS: Fast, Flexible, and Free. *J. Comput. Chem.* **2005**, *26* (16), 1701–1718.
- (37) Dodda, L. S.; De Vaca, I. C.; Tirado-Rives, J.; Jorgensen, W. L. LigParGen Web Server: An Automatic OPLS-AA Parameter Generator for Organic Ligands. *Nucleic Acids Res.* **2017**, *45* (W1), W331–W336.
- (38) Kumari, R.; Kumar, R.; Lynn, A. G_mmpbsa-A GROMACS Tool for High-Throughput MM-PBSA Calculations. *J. Chem. Inf. Model.* **2014**, *54* (7), 1951–1962.
- (39) Breneman, C. M.; Wiberg, K. B. Determining Atom-Centered Monopoles from Molecular Electrostatic Potentials. The Need for High Sampling Density in Formamide Conformational Analysis. *J. Comput. Chem.* **1990**, *11* (3), 361–373.
- (40) Zhavoronkov, A.; Ivanenkov, Y. A.; Aliper, A.; Veselov, M. S.; Aladinskiy, V. A.; Aladinskaya, A. V.; Terentiev, V. A.; Polykovskiy, D. A.; Kuznetsov, M. D.; Asadulaev, A.; Volkov, Y.; Zholus, A.; Shayakhmetov, R. R.; Zhebrak, A.; Minaeva, L. I.; Zagribelnyy, B. A.; Lee, L. H.; Soll, R.; Madge, D.; Xing, L.; Guo, T.; Aspuru-Guzik, A. Deep Learning Enables Rapid Identification of Potent DDR1 Kinase Inhibitors. *Nat. Biotechnol.* **2019**, *37* (9), 1038–1040.
- (41) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. *Nature* **2018**, *555* (7698), 604–610.
- (42) Ouyang, W.; Winsnes, C. F.; Hjelmare, M.; Cesnik, A. J.; Åkesson, L.; Xu, H.; Sullivan, D. P.; Dai, S.; Lan, J.; Jinmo, P.; Galib, S. M.; Henkel, C.; Hwang, K.; Poplavskiy, D.; Tunguz, B.; Wolfinger, R. D.; Gu, Y.; Li, C.; Xie, J.; Buslov, D.; Fironov, S.; Kiselev, A.; Panchenko, D.; Cao, X.; Wei, R.; Wu, Y.; Zhu, X.; Tseng, K. L.; Gao, Z.; Ju, C.; Yi, X.; Zheng, H.; Kappel, C.; Lundberg, E. Analysis of the Human Protein Atlas Image Classification Competition. *Nat. Methods* **2019**, *16* (12), 1254–1261.
- (43) Rodríguez-Pérez, R.; Miyao, T.; Jasial, S.; Vogt, M.; Bajorath, J. Prediction of Compound Profiling Matrices Using Machine Learning. *ACS Omega* **2018**, *3* (4), 4713–4723.
- (44) Yan, X.; Sedykh, A.; Wang, W.; Yan, B.; Zhu, H. Construction of a Web-Based Nanomaterial Database by Big Data Curation and Modeling Friendly Nanostructure Annotations. *Nat. Commun.* **2020**, *11* (1), 1–10.
- (45) Yan, X.; Sedykh, A.; Wang, W.; Zhao, X.; Yan, B.; Zhu, H. *In Silico* Profiling Nanoparticles: Predictive Nanomodeling Using Universal Nanodescriptors and Various Machine Learning Approaches. *Nanoscale* **2019**, *11* (17), 8352–8362.
- (46) Cho, C. W.; Pham, T. P. T.; Jeon, Y. C.; Vijayaraghavan, K.; Choe, W. S.; Yun, Y. S. Toxicity of Imidazolium Salt with Anion Bromide to a Phytoplankton *Selenastrum Capricornutum*: Effect of Alkyl-Chain Length. *Chemosphere* **2007**, *69* (6), 1003–1007.
- (47) Das, R. N.; Roy, K. Predictive Modeling Studies for the Ecotoxicity of Ionic Liquids towards the Green Algae *Scenedesmus Vacuolatus*. *Chemosphere* **2014**, *104*, 170–176.
- (48) Zhang, Y.; Wang, Y.; Liu, A.; Xu, S. L.; Zhao, B.; Zhang, Y.; Zou, H.; Wang, W.; Zhu, H.; Yan, B. Modulation of Carbon Nanotubes' Perturbation to the Metabolic Activity of CYP3A4 in the Liver. *Adv. Funct. Mater.* **2016**, *26* (6), 841–850.
- (49) Alves, V. M.; Bobrowski, T.; Melo-filho, C. C.; Korn, D.; Auerbach, S.; et al. QSAR Modeling of SARS-CoV M^{pro} Inhibitors Identifies Sufugolix, Cenicriviroc, Proglumetacin, and Other Drugs as Candidates for Repurposing against SARS-CoV-2. *Mol. Inf.* **2021**, *40* (1), 2000113.
- (50) Lai, T. T.; Eken, Y.; Wilson, A. K. Binding of Per- And Polyfluoroalkyl Substances to the Human Pregnane X Receptor. *Environ. Sci. Technol.* **2020**, *54* (24), 15986–15995.
- (51) Lin, W.; Yan, Y.; Ping, S.; Li, P.; Li, D.; Hu, J.; Liu, W.; Wen, X.; Ren, Y. Metformin-Induced Epigenetic Toxicity in Zebrafish: Experimental and Molecular Dynamics Simulation Studies. *Environ. Sci. Technol.* **2021**, *55* (3), 1672–1681.



Predicting cytotoxicity of binary pollutants towards a human cell panel in environmental water by experimentation and deep learning methods

Jiahui Wang^a, Gaoxing Su^{b,*}, Xiliang Yan^{c,**}, Wei Zhang^c, Jianbo Jia^c, Bing Yan^{c,***}

^a School of Chemistry and Chemical Engineering, Shandong University, Jinan, 250100, China

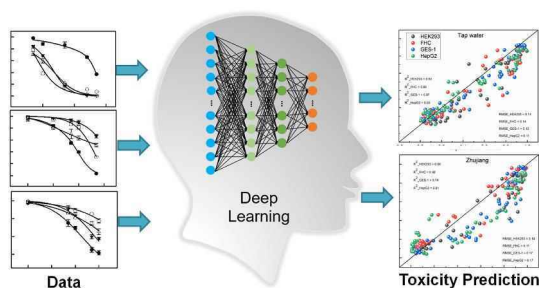
^b School of Pharmacy, Nantong University, Nantong, 226001, China

^c Key Laboratory for Water Quality and Conservation of the Pearl River Delta, Ministry of Education, Institute of Environmental Research at Greater Bay, Guangzhou University, Guangzhou, 510006, China

HIGHLIGHTS

- A human cell panel was used to evaluate cytotoxicity of environmental water.
- Deep learning could predict toxicity of pollutant mixtures in environmental water.
- Pollutant mixtures exhibited synergistic or antagonistic effects in pure water.
- Improved deep learning models showed better predict ability ($R^2 > 0.74$, RMSE < 0.17).

GRAPHICAL ABSTRACT



ARTICLE INFO

Handling Editor: Jian-Ying Hu

Keywords:

Water quality
Deep learning
Human health risk
Cell panel
Prediction

ABSTRACT

Biological assays are useful in water quality evaluation by providing the overall toxicity of chemical mixtures in environmental waters. However, it is impossible to elucidate the source of toxicity and some lethal combination of pollutants simply using biological assays. As facile and cost-effective methods, computation model-based toxicity assessments are complementary technologies. Herein, we predicted the human health risk of binary pollutant mixtures (i.e., binary combinations of As(III), Cd(II), Cr(VI), Pb(II) and F(I)) in water using in vitro biological assays and deep learning methods. By employing a human cell panel containing human stomach, colon, liver, and kidney cell lines, we assessed the human health risk mimicking cellular responses after oral exposures of environmental water containing pollutants. Based on the experimental cytotoxicity data in pure water, multi-task deep learning was applied to predict cellular response of binary pollutant mixtures in environmental water. Using additive descriptors and single pollutant toxicity data in pure water, the established deep learning model could predict the toxicity of most binary mixtures in environmental water, with coefficient of determination (R^2) > 0.65 and root mean squared error (RMSE) < 0.22 . Further combining the experimental data on synergistic and antagonistic effects of pollutant mixtures, deep learning helped improve the predictive ability of the model ($R^2 > 0.74$ and RMSE < 0.17). Moreover, predictive models allowed us identify a number of

* Corresponding author.

** Corresponding author.

*** Corresponding author.

E-mail addresses: sugaoxing@ntu.edu.cn (G. Su), yanxiliang1991@163.com (X. Yan), drbingyan@yahoo.com (B. Yan).

<https://doi.org/10.1016/j.chemosphere.2021.132324>

Received 8 June 2021; Received in revised form 12 September 2021; Accepted 20 September 2021

Available online 21 September 2021

0045-6535/© 2021 Elsevier Ltd. All rights reserved.

toxicity source-related physiochemical properties. This study illustrates the combination of experimental findings and deep learning methods in the water quality evaluation.

1. Introduction

Complex chemical mixtures as well as numerous transformation products in surface waters cause huge potential risk to human health and ecosystems and raise the concerns of water quality all over the world (He and Li, 2020; Li et al., 2017; Zhang et al., 2021). The environmental water quality is currently assessed by analyzing a short list of so-called “priority substances” according to concentration criteria. The recommended safety concentration limit for each pollutant is derived from toxicity tests of very limited individual chemicals. However, numerous reports have indicated that the conventional water quality standards are not sufficient to estimate the toxicity of complex chemical contaminations to human health and ecosystems (Deville et al., 2020; Naidu et al., 2016; Xu et al., 2020). This is because there are hundreds and thousands of pollutants (not to mention numerous emerging pollutants) in the environmental waters, but only tens of them are “listed” in the national standards (Geissen et al., 2015). Besides, the coexisting pollutants in the water are quite likely to have synergistic or antagonistic effects, which are not considered in the standards (Yang et al., 2017). Therefore, new approaches to assess and predict water quality are on highly demand.

In recent years, to offset the shortfalls of conventional water quality evaluation methods, biological assays using whole organisms or human cells have been increasingly developed for water safety evaluation (Altenburger et al., 2015; Di Paolo et al., 2016; Neale et al., 2017; Xu et al., 2020). Such assays are powerful approaches to evaluate the safety risk of all pollutants and their joint toxicity in water. Standard protocols were developed for effect-based biological assays to estimate the biological activity of environmental samples, and effect-based values were calculated to access the safety risk of chemical mixtures. Because single biological assay is limited to one particular effect, panels of biological assays are often employed to comprehensively evaluate water quality (Di Paolo et al., 2016). To mimic human oral exposures, we have used a human cell panel containing four human cell lines to evaluate the safety risk of ZrO_2 nanoparticles and their pollutant adducts in a water remediation process (Wang et al., 2018b). Although biological assays are feasible tools for estimating the potential health risks of pollutants in environmental water, the chemical mixtures in real environmental water are considerable complexity, with the characteristics of diversity of pollutants, indefinitely mixing ratios, and diverse concentrations. It is practically impossible to experimentally measure the toxicity of all chemical mixtures. Therefore, it is necessary to develop experiment-based computation models to predict chemical mixture toxicity for water quality evaluation.

The computational approaches offer possibilities for establishing quantitative relationships between any chemical mixture and their toxicity endpoints (Kar and Leszczynski, 2019; Raies and Bajic, 2016). In previous studies, concentration addition (CA) model and independent action (IA) are two mostly widely used methods for toxicity prediction of chemical mixtures (Nys et al., 2017; Ukic et al., 2019). However, the CA or IA models are based on the idea that the components in the mixture have similar or dissimilar mode of action, which greatly limits the universal of these methods. More importantly, the CA or IA models only consider the effect of concentration on toxicity, while ignoring other factors, such as the physical and chemical properties of the mixtures. As a result, the CA or IA models cannot be used to predict the toxicity of unknown mixtures. Machine learning, a subdomain of artificial intelligence, broadly refers to computer algorithms that can automatically learn from big data. Recent breakthroughs in artificial intelligence have led to the fast development of powerful deep learning algorithms that can extract hierarchical features from data, with better predictive performance and less human intervention. Deep learning techniques have

exhibited great potentials in data-driven environmental toxicity evaluation. These technological advances may greatly benefit environmental water quality evaluation and prediction (Asadollah et al., 2021; Baek et al., 2021; Bui et al., 2020; Chen et al., 2020; Kang et al., 2017; Najah Ahmed et al., 2019). Until now, multiple machine learning models and deep learning models, including artificial neural network models, radial basis function network models, random forest models, decision tree models, and support vector models have been developed for water quality prediction (Bui et al., 2020; Kang et al., 2017; Lu and Ma, 2020). Since application of deep learning does not require a prior knowledge of the underlying processes and the recognition of all the complex relationships between chemical mixtures and organisms or cells, deep learning-based toxicity prediction of chemical mixtures is highly desirable.

Herein, we applied experimental and multi-task deep learning methods to predict human health risk of binary mixtures of some very toxic and persistent pollutants in environmental water (Fig. 1A). In this study, the Pear River was selected as the study area since it is the most complex water system in South China. As the most developed regions in China, the Pearl River Delta is highly polluted due to the rapid increase in population and industrial development in recent decades. Therefore, the Pear River can be used as a representative water system to explore the water quality evaluation in South China. The training set was obtained by measuring the cytotoxicity of individuals and combinations of pollutants in a human cell panel mimicking oral exposures. Furthermore, the synergistic or antagonistic effects of binary mixtures were obtained and employed in the deep learning process to further improve the prediction power of the model.

2. Materials and methods

2.1. Materials

As(III) ($NaAsO_2$), Cd(II) ($CdCl_2$) and Cr(VI) ($K_2Cr_2O_7$) were purchased from Sigma-Aldrich. Pb(II) ($Pb(CH_3COOH)_2 \cdot 3H_2O$) was purchased from Macklin Biochemical Co., Ltd. (Shanghai, China). F(I) (NaF) was purchased from Sinopharm chemical reagent Co., Ltd. (Shanghai, China). The pollutants were dissolved to 2 mg/mL with ultrapure water, Pearl River water or tap water and stored at 4 °C before use.

2.2. Cell cultures

The human cell panel consists of four cell lines. They are normal epithelial cell lines of human gastric mucosa (GES-1), normal epithelial cell lines of the human colon (FHC), human hepatoma cells (HepG2) and human embryonic kidney cells (HEK293). All four cell lines were purchased from ATCC (Manassas, VA). They were cultured at 37 °C in the incubator (humidified atmosphere, 5% CO_2). GES-1 and FHC cells were grown in RPMI 1640 medium (basic (1x), Gibco) supplemented with 10% fetal bovine serum, 100 μ g/mL penicillin and 100 U/mL streptomycin. HepG2 and HEK293 cells were grown in Dulbecco's modified eagle's medium (DMEM, basic (1x), Gibco) with the same supplements as RPMI 1640.

2.3. Water samples

The Pearl River water was collected from Pearl River in Guangdong, China. The Pearl River Basin is located between latitude 21°31'–26°49' N, and longitude 102°14'–115°53' E. The sampling site was located in the south of Guangdong, at longitude 113°16' E, and latitude 23°3' N. In order to avoid sample contamination, the sampling container was rinsed

with water to be sampled 5 times before sampling. The procedures of quality assurance and quality control (QA/QC) referred to previous studies (He and Li, 2020; Zhang et al., 2021), and “Technical Specifications for Surface Water and Wastewater Monitoring” (HJ/T 91–2002) (Ministry of Environmental Protection of the People’s Republic of China, 2002). Water samples were filtered through 0.45 μm glass membranes and stored at 4 $^{\circ}\text{C}$ in dark conditions, and then send it to the laboratory as soon as possible to analyze the parameters (Table S1). Throughout the analysis process, experimenters always wear gloves, lab coats and gloves made of 100% cotton (Zhou et al., 2021). The tap water was acquired in Qingdao, China. Before preparing the medium, we boiled the tap water for 15 min to sterilize and eliminate the cytotoxic interference of chlorine.

2.4. Cytotoxicity assays

The HEK293, HepG2, FHC, and GES-1 cells were seeded in 96-well plates at a density of 5000 cells/mL. After incubation for 24 h, cells were exposed to single pollutants or binary mixtures at designed concentrations for 48 h. The control groups were cultured in growth media. For evaluating the toxicity of pollutants in different water samples, RPMI 1640 or DMEM powder (Gibco, USA) and sodium bicarbonate were dissolved with Pearl River water, tap water and the ultrapure water under the same conditions (Ren et al., 2017). The concentrations of pollutants used in this study were listed in Table S2. The pH of medium was adjusted to 7.4 using HCl or NaOH. In order to remove bacteria, prepared cell culture media were passed through Millipore Express membranes with 0.22 μm Filter unit (Merck Millipore Ltd.).

We measured the cell viability based on the level of adenosine triphosphate (ATP), which signals the presence of metabolically active cells. We applied the CellTiter-Glo[®] luminescent cell viability assay kit (Promega Corporation, Madison, USA) to quantify ATP levels (Yang et al., 2015). The cell viability obtained from the experimental groups divided by the respective control groups. We conducted three replications every test to obtain the mean value of cell viability.

2.5. TU values calculation

The toxic unit (TU) method has been successfully used as a tool to assess combined toxicity. We applied the TU approach to assess joint effects of mixtures in this study. The TU_i was calculated from the concentrations of individual pollutants divided by its EC_{50} . The equation of sum of TU was as below:

$$TU = \sum TU_i = \sum_{i=1}^n \frac{M_i^{r+}}{EC_{50i}} \quad (1)$$

where, i is the identity of the As(III), Cd(II), Cr(VI), F(I) and Pb(II). The n is the number of pollutants. EC_{50} is the concentration of pollutants that results in 50% cell viability. The EC_{50} values were calculated by the SigmaPlot 12.5 based on the single toxicity data. Based on this method, the combined effects are divided into three categories. The toxic unit (TU) < 1.0 indicates synergism, $TU = 1.0$ indicates additivity, and $TU > 1.0$ is interpreted as antagonism.

2.6. Descriptors calculation and multi-task learning with deep neural networks

To obtain the descriptors of the mixture, the descriptors of each component (*i.e.*, CdCl_2 , $\text{K}_2\text{Cr}_2\text{O}_7$, $\text{Pb}(\text{CH}_3\text{COOH})_2$, NaF, and NaAsO_2) in the mixture was first obtained. All components were uniformly converted to SMILES (simplified molecular input line entry specification) format before descriptors calculation. The Molecular Operating Environment (MOE) software was used to calculate the physical and chemical properties (*e.g.*, molecular weight, van der Waals volume, atom counts, and bond counts) of each component. A total of 206 theoretical descriptors were generated for each component. Based on previous studies (Mikolajczyk et al., 2019), the additive descriptor was used for each mixture, which is represented as formula 2.

$$D_{\text{mix}} = \sum_{i=1}^n D_i x_i \quad (2)$$

D_{mix} represents the mixture descriptor, D_i represents the descriptor of component i , x_i represents the concentration of component i . Combined with the respective concentration values of the five components, a total of 211 variables were used to represent each mixture.

Once a series of descriptors for the mixture were generated, we built multi-task deep learning models. These models were built using deep neural networks (DNN). Multi-task learning aims to learn multiple different tasks simultaneously while maximizing performance on one or all of the tasks through model parameters sharing. A deep neural network is an artificial neural network with multiple hidden layers between the input (*i.e.*, the descriptors) and output (*i.e.*, the predicted values) layers. As shown in Fig. 1B, the deep neural network contained three hidden layers with 128, 64, and 32 nodes respectively. To prevent overfitting, a dropout layer with dropout rate of 0.3 is added before the output layer. The RMSprop (Root Mean Square prop) and MSE (mean squared error) were used as optimizer and loss function to compile the

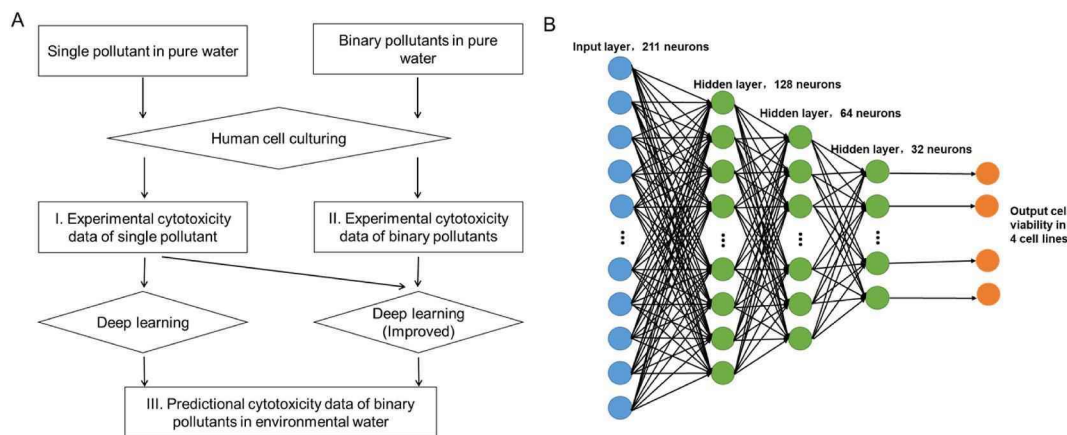


Fig. 1. Workflow for prediction of cytotoxicity of binary pollutants using experimentation and deep learning methods. (A) Essential steps in cytotoxicity assays and model building. (B) The deep neural network architecture used in the present study. The input layer contained 211 neurons. Three hidden layers with 128, 64, and 32 neurons were set between input layers and output layers.

DNN model in this study. The learning rate was set as the default value of the RMSprop optimizer. Each DNN model was trained for 500 epochs when no significant changes of MSE were observed. The deep learning model was implemented with *TensorFlow* 1.14.0 and *Keras* 2.2.5. The model performance was accessed by the determination coefficient (R^2) and the root mean square error (RMSE), which were represented as formula 3 and 4.

$$R^2 = \frac{\sum_{i=1}^n (y_i^{obs} - y_i^{pred})^2}{\sum_{i=1}^n (y_i^{obs} - \bar{y}^{obs})^2} \quad (3)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i^{obs} - y_i^{pred})^2}{n}} \quad (4)$$

3. Results and discussions

3.1. Cytotoxicity evaluation of single pollutant using a human cell panel

During oral exposure to water, pollutants in such water enter gastric system first and, after absorption, will accumulate in liver and kidney. Therefore, a human cell panel consisting of four human cell lines derived from human stomach, intestine, liver, and kidney was employed in this investigation to evaluate the toxicity of pollutants. The four cell lines are normal gastric mucosa cells (GES-1), normal human colorectal mucosa cells (FHC), liver hepatocellular carcinoma cells (HepG2), and human embryonic kidney cells (HEK293). To prioritize pollutants evaluation, we selected As(III), Cd(II), Cr(VI), F(I) and Pb(II) as model pollutants since they are the most toxic and persistent pollutants in environmental waters.

The dose-dependent cell viability of individual pollutant in pure water was examined first (Fig. 2), and EC_{20} and EC_{50} values of the five pollutants in four cell lines were calculated to evaluate their toxicity (Table 1). Among all pollutants, it was found As(III) showed severe cytotoxicity for the human cell panel, especially for GES-1 and FHC cells. The EC_{50} values are less than 0.7 $\mu\text{g/mL}$ for all cell lines, and are 0.15, and 0.23 $\mu\text{g/mL}$ for GES-1 and FHC cells, respectively. According to previous study, As(III) can induce serious cytotoxicity, due to As(III) has

the capacity to induce apoptotic and autophagic cell death (Chiu et al., 2010; Hettick et al., 2015; Meister et al., 2016; Sanjay Kumar, 2014). Cd (II) showed much higher cytotoxicity to HepG2 and HEK293 cells ($EC_{50} = 0.99$, and 0.65 $\mu\text{g/mL}$, respectively) than that of GES-1 and FHC cells ($EC_{50} = 2.76$, and 9.53 $\mu\text{g/mL}$, respectively). Cr(VI) exhibited high cytotoxicity to three of four cell lines, with the EC_{50} values of 0.10, 0.17, and 0.15 $\mu\text{g/mL}$ for FHC, HepG2 and HEK293 cells, respectively, while less toxic to GES-1 cells ($EC_{50} = 3.66$ $\mu\text{g/mL}$). Earlier studies have reported that the Cr(VI) is carcinogenic and mutagenic to living organisms and cause serious damages to bacteria, plants and animals (Guo et al., 2020; Kamarudheen et al., 2020; Kim et al., 2015; Marikkani et al., 2019), because of its oxidizing properties as well as a tetrahedral arrangement of oxygen groups, making it structurally similar to sulfate and phosphate (Wadhawan et al., 2013; Zhitkovich, 2005). Compared to above three pollutants, F(I) and Pb(II) were relatively less toxic, the EC_{50} values are higher than 4 $\mu\text{g/mL}$ for all cell lines, while both of them showed highest cytotoxicity to GES-1 cells. These results revealed that different pollutants showed different toxicity in different cell lines and there was no single cell lines could evaluate cytotoxicity for all pollutants. By employing a human cell panel, the human health risk of multiple pollutants can be evaluated fully and, at the same time, enormous cellular data can be collected for deep learning model building.

EC_{20} is the concentration of the pollutant that results in 20% cell death. In general, pollutant concentrations above the EC_{20} values are considered as low toxic (Iwasawa et al., 2013; Song et al., 2010). Compared with the current national water standard for reclaimed water (GB8978-1996) (Table S2), the EC_{20} values are even lower for As(III) in all four cell lines, Cr(VI) in three cell lines, F(I) in two cell lines, and Pb (II) in one cell line. For example, the national standard for Cr(VI) is 0.5 $\mu\text{g/mL}$, the EC_{20} values are 0.045, 0.066, and 0.037 $\mu\text{g/mL}$ for FHC, HepG2, and HEK293 cells, respectively. These measurements were performed only for single pollutant in water. In case of mixtures, the toxicity will be aggravated due to the joint effects. Therefore, the concentrations in current national standard do not help much in health risk assessment.

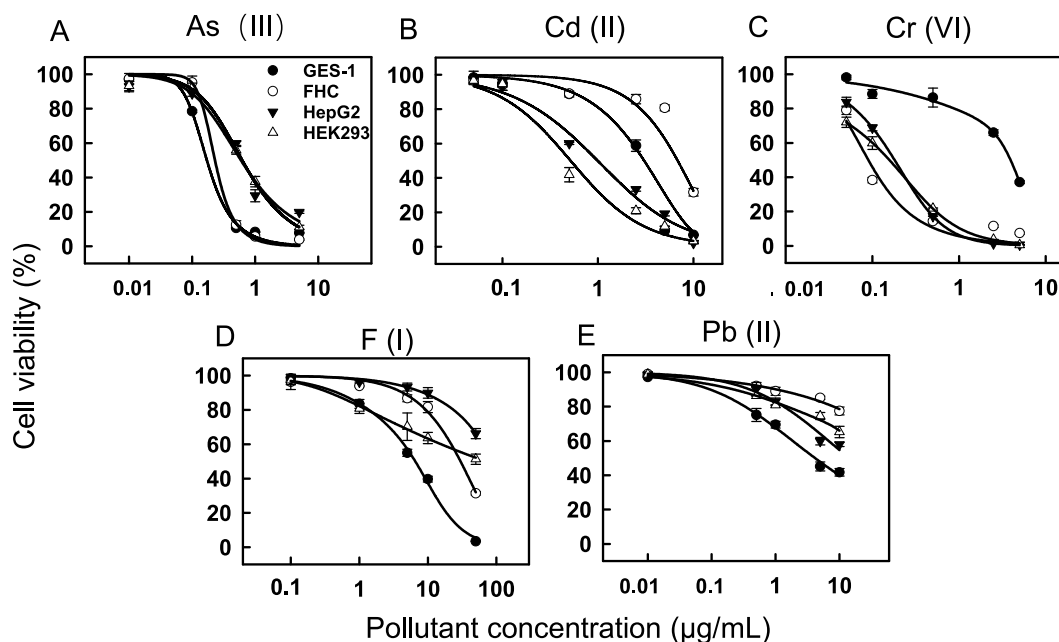


Fig. 2. Dose-dependent toxicity of individual pollutant for the human cell panel in pure water. (A) As(III); (B) Cd(II); (C) Cr(VI); (D) F(I); (E) Pb(II). Four human cell lines were used: GES-1, FHC, HepG2, and HEK293. CellTiter-Glo luminescent cell viability assays were employed to measure the cell viability. Incubation time: 48 h. Data represent mean \pm standard deviation ($n = 3$).

Table 1

The EC₂₀ and EC₅₀ values of As(III), Cd(II), Cr(VI), F(I) and Pb(II) for the human cell panel in pure water. (n = 3).

| | EC ₂₀ (μg/mL) | | | | EC ₅₀ (μg/mL) | | | |
|----|--------------------------|----------------|----------------|-----------------|--------------------------|---------------|---------------|--------------|
| | GES-1 | FHC | HepG2 | HEK293 | GES-1 | FHC | HepG2 | HEK293 |
| As | 0.096 ± 0.0021 | 0.20 ± 0.081 | 0.30 ± 0.017 | 0.19 ± 0.021 | 0.15 ± 0.23 | 0.23 ± 0.16 | 0.57 ± 0.17 | 0.62 ± 0.063 |
| Cd | 1.17 ± 0.088 | 4.13 ± 0.50 | 0.17 ± 0.014 | 0.18 ± 0.029 | 2.76 ± 0.46 | 9.53 ± 0.67 | 0.99 ± 0.25 | 0.65 ± 0.35 |
| Cr | 1.13 ± 0.31 | 0.045 ± 0.0022 | 0.066 ± 0.0088 | 0.037 ± 0.00027 | 3.66 ± 0.83 | 0.10 ± 0.020 | 0.17 ± 0.0037 | 0.15 ± 0.015 |
| F | 1.28 ± 0.26 | 10.48 ± 1.14 | 30.03 ± 5.10 | 1.52 ± 0.49 | 6.42 ± 0.30 | 29.99 ± 3.16 | 63.54 ± 5.70 | 27.12 ± 5.47 |
| Pb | 0.25 ± 0.037 | 11.88 ± 0.072 | 5.74 ± 0.41 | 5.15 ± 0.46 | 4.36 ± 0.78 | 155.54 ± 6.16 | 29.01 ± 5.30 | 13.02 ± 0.77 |

3.2. Initial toxicity prediction for binary mixtures using single pollutant cytotoxicity data

The ultimate goal of this research is to build a deep learning model that can be used to accurately predict the toxicity of binary mixtures in different water environment. Tap water is directly drunk in most places. The Pearl River has the second largest flow in China. Therefore, Tap water and Pearl River water were chosen as the representative water environment. To this end, we designed two deep learning models: 1) the toxicity data of single pollutant in pure water was used as the training set to build an initial model, and 2) both the toxicity data of single pollutant and binary mixture in pure water were selected as training data to build an advanced model. The initial model was expected to be able to make accurate predictions of the toxicity endpoint produced by the additive mode with the additive descriptors. The advanced model was expected to learn synergistic and antagonistic action modes from the training set including the toxicity data of binary mixtures in pure water.

In the present study, the constructed deep learning model was a classic feedforward neural network that was trained by the backpropagation algorithm (also sometimes called reverse mode differentiation). In simple terms, after each forward passing through the network,

backpropagation will use the chain rule to perform backward propagation, and at the same time adjust the parameters (weights and biases) of the model to achieve the minimum error. After a series of backpropagation calculations from the top layers to the bottom layers, the deep neural network can correctly map the inputs (*i.e.*, the descriptors) to outputs (*i.e.*, the cell viability values). Generally speaking, a deep neural network with two hidden layers is sufficient to solve general problems. However, considering the high dimensionality of the input variables (211 descriptors), three hidden layers were used to get an optimum prediction, and also did not significantly increase the complexity of the deep learning models. In fact, the loss curves of the training sets and the test sets also showed that the deep learning models were not overfitting (Fig. 3A and C). As shown in Formulas (5)–(8), after multiple linear and non-linear transformations of the input descriptors (*x*), the deep learning models can finally output the predicted values (*y*) of cell viability.

$$h1_j = f_{h1} \left(\sum_{i=1}^{211} w_{ij} \cdot x_i + b_j \right) (j = 1, 2, \dots, 128) \quad (5)$$

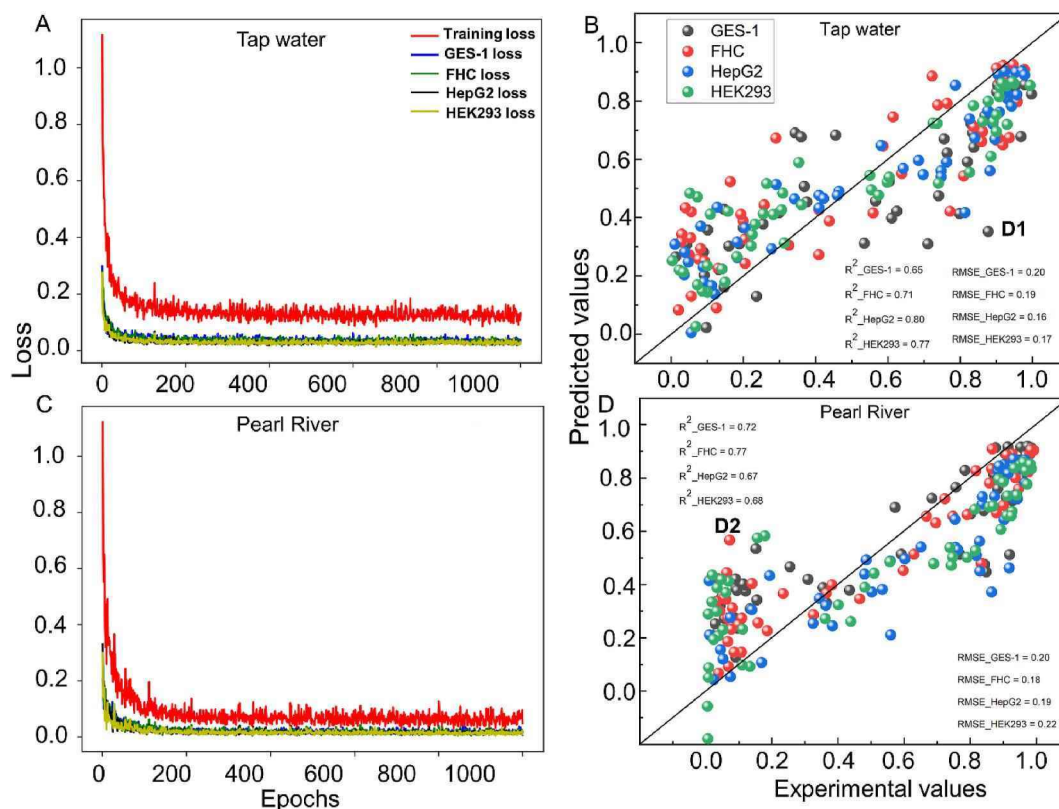


Fig. 3. Deep learning models for predicting binary toxicity using single pollutants toxicity data as training set. (A, C) Validation loss and training loss of the deep learning models for cytotoxicity of binary pollutants in tap water (A) and Pearl River water (C). (B, D) Correlations between experimental and predicted cytotoxicity of binary pollutants in tap water (B) and Pearl River water (D). D1 and D2 data points represent two predicted outliers.

$$h2_k = f_{h2} \left(\sum_{i=1}^{128} w_{ij} \cdot x_i + b_k \right) (k=1, 2, \dots, 64) \quad (6)$$

$$h3_l = f_{h3} \left(\sum_{i=1}^{64} w_{ij} \cdot x_i + b_l \right) (l=1, 2, \dots, 32) \quad (7)$$

$$y = f_y \left(\sum_{i=1}^{32} w_{ij} \cdot x_i + b_m \right) (m=1) \quad (8)$$

where, w is the weights of the connection of different layers, and b is the bias value of three hidden layers and the output layer. In addition, the multi-task deep learning could improve the generalization ability of the models through parameters sharing between different tasks (Liao et al., 2019). Fig. 3B and D showed the correlations between the predictions and experimental values when the toxicity data of single pollutant in pure water was selected as the training set. It can be seen that the resultant models had high predictability with high R^2 values ($R^2 > 0.65$) and low RMSE values (RMSE < 0.22), indicating the predict models are highly reliable that is constructed through single pollutant toxicity data and multi-task deep learning. The prediction results showed that most of the binary mixtures exhibited additive interactions. However, it should be noted that some outliers are also noticeable. For example, the data point D1 (Fig. 3B) generated by mixing Cr(VI) and Pb(II) in GES-1 cells, the experimental value is 0.88 while the predicted value is 0.35. Similarly, the data point D2 (Fig. 3D) generated by mixing As(III) and F(I) in FHC cells, the experimental value is 0.07 while the predicted value is 0.57. These outliers may be caused by only considering the additive effects while synergistic and antagonistic effects also existed, so we designed advanced deep learning models to see if these outliers can be eliminated.

3.3. Synergistic or antagonistic effects of binary mixtures for the human cell panel

To evaluate the joint effects of mixtures, the cytotoxicity of the 10 binary mixtures for the human cell panel were measured (Table S3). Five concentrations for each pollutant were used here, which were set according to the cytotoxicity assays of single pollutant. For majority

pollutants, the five concentrations were the same as the single pollutant cytotoxicity assays. Concentrations of each pollutant were listed in Table S3. The toxic unit (TU) values of individual pollutant and binary mixtures for the four cells lines were calculated according to Formula (1) (Wang et al., 2017, 2020). The dose-dependent relationships between TU values and cell viability were plotted in Fig. 4 and Fig. S1. It was observed that the dose-dependent curves of binary mixtures were not overlapped with individual pollutants, indicating the binary mixtures have certain of synergistic (left shift) or antagonistic (right shift) effects.

To further verify the combined effects of the binary mixtures, the TU_{50} values (TU values at the 50% inhibition of the cell growth in the mixture) for all cell lines were calculated according to the dose-dependent curves. According to TU value theory, when the TU_{50} value is equal to 1.0, the combined toxic effect is additive, while the TU_{50} values are greater or less than 1.0, are defined as antagonistic or synergistic effects, respectively (Wang et al., 2018a, 2020). As listed in Table 2, TU_{50} values of 12 binary mixtures are in the range of 0.8–1.2, indicating those mixture showed weak synergistic or antagonistic effects. TU_{50} values of 13 binary mixtures are in the range of 0.6–0.8, and 8 mixtures in the range of 1.2–1.4, indicating medium synergistic and antagonistic effects existed in these mixtures. Besides, five mixtures exhibited strong synergistic effects and two mixtures exhibited strong antagonistic effects, with TU_{50} values smaller than 0.6 or greater than 1.4, respectively. Same binary mixtures, like As–Cd mixtures, exhibited totally different combined effects between different cell lines, i.e.

Table 2
 TU_{50} values of binary pollutants in the human cell panel ($n = 3$).

| | GES-1 | FHC | HepG2 | HEK 293 |
|-------|---------------|---------------|---------------|---------------|
| As–Cd | 1.049 ± 0.192 | 1.276 ± 0.387 | 0.936 ± 0.209 | 0.666 ± 0.369 |
| As–Cr | 0.862 ± 0.113 | 0.947 ± 0.050 | 1.358 ± 0.029 | 0.862 ± 0.043 |
| As–F | 0.540 ± 0.046 | 1.359 ± 0.209 | 1.176 ± 0.054 | 1.185 ± 0.050 |
| As–Pb | 0.716 ± 0.095 | 0.919 ± 1.329 | 0.917 ± 0.085 | 0.582 ± 0.089 |
| Cd–Cr | 0.756 ± 0.138 | 0.741 ± 0.086 | 0.781 ± 0.004 | 0.763 ± 0.002 |
| Cd–F | 1.674 ± 0.218 | 1.208 ± 0.213 | 1.325 ± 0.403 | 0.701 ± 0.098 |
| Cd–Pb | 0.953 ± 0.092 | 0.427 ± 0.066 | 0.833 ± 0.192 | 0.374 ± 0.244 |
| Cr–F | 1.211 ± 0.072 | 0.689 ± 0.091 | 0.615 ± 0.008 | 0.670 ± 0.023 |
| Cr–Pb | 0.560 ± 0.080 | 0.885 ± 0.315 | 0.761 ± 0.013 | 0.617 ± 0.043 |
| F–Pb | 0.827 ± 0.162 | 1.376 ± 0.982 | 1.311 ± 0.368 | 1.561 ± 0.197 |

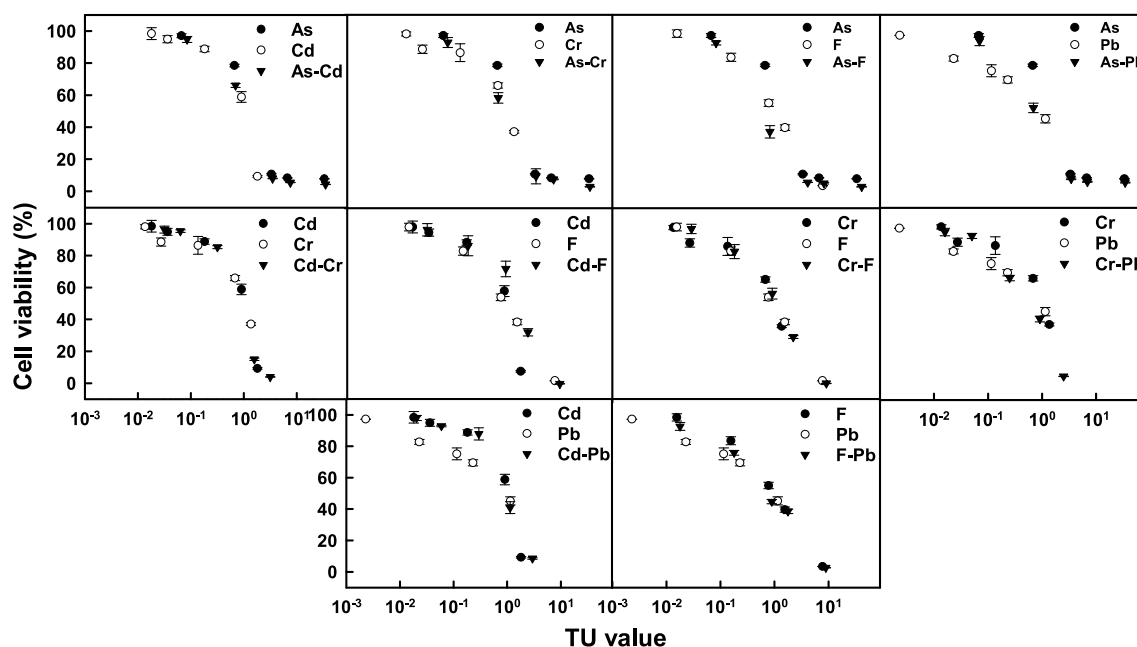


Fig. 4. Relationships between TU values and cell viability of GES-1. TU values of individual pollutant and binary mixtures were obtained from pollutant concentrations and EC_{50} values according to equation (1). Data were mean ± standard deviation from three independent measurements ($n = 3$).

additive effects in GES-1 and HepG2 cells, antagonistic effect in FHC cells, and synergistic effect in HEK293 cells, which might be the reason that the modes of actions of the mixtures were different in different cell lines. Also, same cell lines have different combined effects for different mixtures. The combined effects are very complicated and it is might impossible to employ traditional modeling methods to predict the toxicity of mixtures only using toxicity data of single pollutant as training set. If the computer could learn the synergistic and antagonistic effects of mixtures, the deep learning model might be improved.

3.4. Model predictability improvements

In order to build a deep learning model that can better predict the toxicity of binary mixtures in tap water and Pearl River water, the toxicity data of binary mixtures in pure water were added to the training set. Fig. 5A–D showed the learning curves of the deep learning models and correlations between the predictions and experimental values when the toxicity data of single pollutants and binary mixtures in pure water were selected as the training set. There was no significant difference between training loss and validation loss (Fig. 5A and C), indicating that the model results were not overfitting after dropout regularization. It can be seen that the predictive ability of the model is improved (the average R^2 increased from 0.73 to 0.86, and from 0.71 to 0.83 in tap water and Pearl River water, respectively). More importantly, we found that some outliers were eliminated that were generated by the initial deep learning models. For example, the former data points D1 and D2 were outliers (Fig. 3B and D), but now the predicted values of D1 and D2 were close to the experimental values (Fig. 5B and D). As discussed in Table 2, the binary mixture (D1 data point) of Cr(VI) and Pb(II) in GES-1 cell showed antagonistic interactions, and the binary mixture (D2 data point) of As(III) and F(I) in FHC cell showed synergistic interactions. The predicted results indicated that the deep learning models could learn the

synergistic and antagonistic effects from the toxicity data of binary mixtures in pure water. In addition, the increased volume of data has been proven to be helpful for improving the prediction accuracy of deep learning models. As a result, adding the toxicity data of binary mixtures to the training set can improve the predictive ability of the model. Furthermore, the model's predictive performance can be affected by the molar fractions and composition of mixtures. Therefore, the model performance should be further investigated for any compositions of mixture from the training set and mixtures formed by novel pure compound absent in the training set in the future.

Analysis of the resultant predictive models allowed us to identify a number of physicochemical properties (i.e., input descriptors) responsible for the cytotoxicity, which can be used to elucidate potential toxicity mechanisms. Although deep neural network was once criticized as a “black box” (Poon and Sung, 2021), now some advanced methods could allow us to visualize the prediction process of the deep learning models (Gunning et al., 2019; Tang et al., 2019). For example, the Grad-CAM (Gradient-weighted Class Activation Mapping) method can make the convolutional neural network (CNN)-based models more transparent by visualizing the regions of input that are important for predictions from these models (Selvaraju et al., 2020). Here, a method, named VarImpVIANN (de Sá, 2019), was applied to measure the relative importance of features in artificial neural networks (ANN) models. Its underlying principle assumes that the more important a feature is, the more the weights, connected to the respective input neuron, will change during the training of the model. In the present study, we found that the top four most important descriptors were the concentrations of As(III) and Cr(VI), the molecular refractivity (the BCUT_SM_R_1 descriptor), and the total positive partial charge (the PEOE_PC + descriptors), which were considered to be responsible for the cytotoxicity. Compared with other heavy metals, the As(III) and Cr(VI) induced the most severe cytotoxicity with lowest EC₅₀ values (Table 1). The molecular

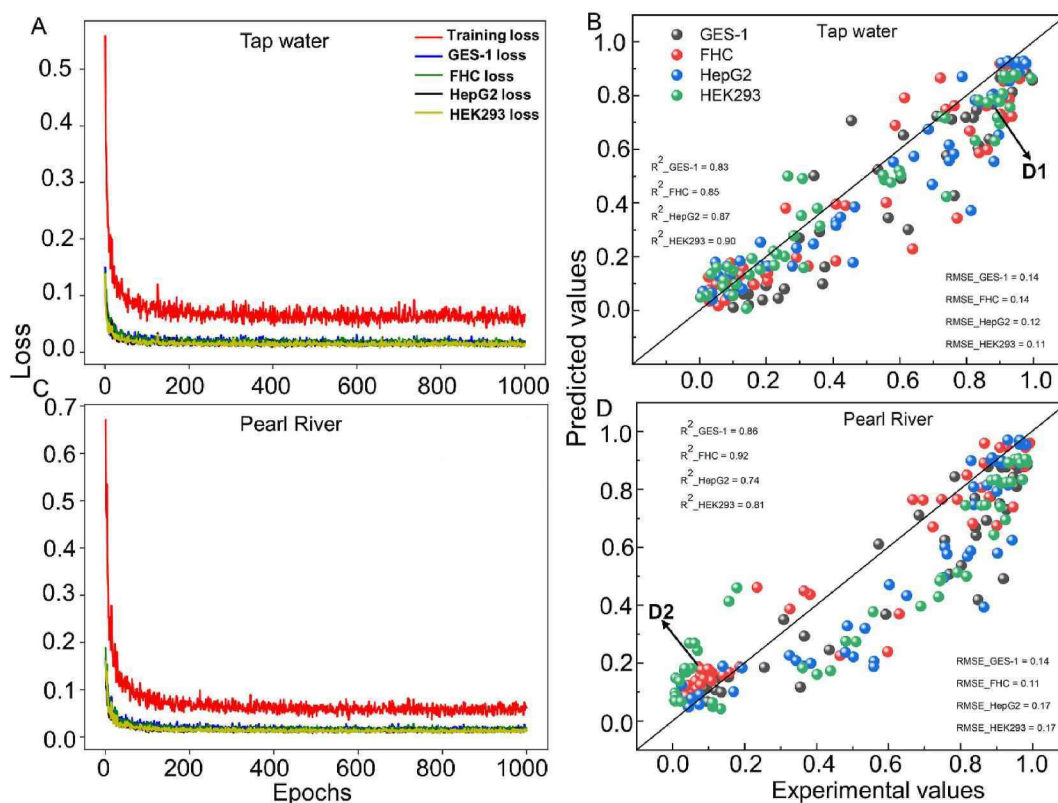


Fig. 5. Learning curves and performances of the improved deep learning models. (A, C) Validation loss and training loss of the deep learning models for cytotoxicity of binary pollutants in tap water (A) and Pearl River water (C). (B, D) Correlations between experimental and predicted cytotoxicity of binary pollutants in tap water (B) and Pearl River water (D). D1 and D2 points in Fig. 5B and D were the same D1 and D2 points in Fig. 3B and D.

refractivity and partial charge were properties related to electron density that has been proven to be closely related to the toxicity of heavy metals (Jin et al., 2014; Su et al., 2010).

In the present study, using the toxicity data in pure water as training set, the deep learning model can accurately predict the toxicity of the binary mixtures in tap water and Pearl River water. However, some drawbacks and prospects should also be pointed out. The high prediction accuracy is partially benefited from the similar physical and chemical properties between tap water, Pearl River water and pure water. Considering the complexity of actual water environment, it is not enough to build reliable deep learning models just using the properties of pollutants as descriptors. In the future, the differences between the physical and chemical properties of different water samples also need to be taken into account, such as the pH, conductivity, and total organic carbon (TOC). As alternatives for animal studies, the in vitro methods including high-throughput cell assays and computational approaches provide a new paradigm for in vivo toxicity assessment. In previous studies, several machine learning models have been constructed to establish the relationships between in vitro activity profiles and in vivo toxicity (Huang et al., 2016; Kim et al., 2016). Therefore, it is an ultimate goal to predict the human health risk using in vitro toxicity data and deep learning models.

It should also be noted that every coin has two sides, and deep learning is no exception. Undoubtedly, deep learning significantly outperforms traditional machine learning methods in multiple domains such as speech recognition and computer vision. Also, like what we do in this study, the deep learning models can simultaneously predict multiple targets. However, deep learning requires very large amount of data and computational resources in order to perform better than traditional machine learning, and the model interpretability is also a big challenge for deep learning. In addition, there is still a lot that needs to improve models for chemical mixtures' toxicity prediction: 1) More experimental data should be generated, 2) appropriate descriptors describing the properties of mixtures should be developed, and 3) more rigorous external validation should be implemented using mixtures formed by novel pure compounds absent in the training set.

4. Conclusions

To mimic the human health risk after oral exposure of environmental water containing pollutants, a human cell panel included human stomach, colon, liver, and kidney cells was employed to assess the water quality experimentally. Using collected toxicity data as training set, multi-task deep learning models were successfully constructed to predict the toxic effects of binary pollutant mixtures (i.e., binary combinations of As(III), Cd(II), Cr(VI), Pb(II) and F(I)). The main findings are as follows. (1) A cell panel is necessary for evaluation of pollutants' cytotoxicity, since different cell lines showed very different sensitivity to the same pollutant. (2) Cytotoxicity of binary pollutants in environmental water could be predicted by deep learning methods using single pollutant toxicity data ($R^2 > 0.65$ and RMSE < 0.22). (3) Synergistic and antagonistic effects of pollutant mixtures were existing and quite distinct between different cell lines. (4) After considering the synergistic and antagonistic effects, the predictive abilities of the deep learning models were further improved (R^2 improved to > 0.74 and RMSE improved to < 0.17). The average R^2 of tap water and Pearl River water has increased from 0.73 to 0.86, and from 0.71 to 0.83, respectively. More importantly, we found that some outliers generated by the initial deep learning model were eliminated. The high prediction accuracy of the deep learning models in both tap water and Pearl River water has proved the effectiveness and universality of our methods. Combined with more experimental data generated in the future, the deep learning methods can also be easily extended to explore the chemical mixtures' toxicity in other water systems.

Credit author statement

Jiahui Wang: Methodology, Investigation, Visualization, Validation. **Gaoxing Su:** Conceptualization, Methodology, Investigation, Writing – original draft. **Xiliang Yan:** Formal analysis, Methodology, Software, Writing – review & editing. **Wei Zhang:** Resources, Writing – review & editing. **Jianbo Jia:** Resources, Writing – review & editing. **Bing Yan:** Conceptualization, Resources, Writing – review & editing, Supervision

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was funded by the National Key R&D Program of China (2016YFA0203103), the National Natural Science Foundation of China (22036002, 22076085), and the introduced innovative R&D team project under the “The Pearl River Talent Recruitment Program” of Guangdong Province, China (2019ZT08L387).

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.chemosphere.2021.132324>.

References

- Altenburger, R., Ait-Aissa, S., Antczak, P., Backhaus, T., Barceló, D., Seiler, T.-B., Brion, F., Busch, W., Chipman, K., de Alda, M.L., de Aragão Umbuzeiro, G., Escher, B.L., Falciani, F., Faust, M., Focks, A., Hilscherova, K., Hollender, J., Hollert, H., Jäger, F., Jahnke, A., Kortenkamp, A., Krauss, M., Lemkine, G.F., Munthe, J., Neumann, S., Schymanski, E.L., Scrimshaw, M., Segner, H., Slobodnik, J., Smedes, F., Kughathas, S., Teodorovic, I., Tindall, A.J., Tollefsen, K.E., Walz, K.-H., Williams, T.D., Van den Brink, P.J., van Gils, J., Vrana, B., Zhang, X., Brack, W., 2015. Future water quality monitoring — adapting tools to deal with mixtures of pollutants in water resource management. *Sci. Total Environ.* 512–513, 540–551. <https://doi.org/10.1016/j.scitotenv.2014.12.057>.
- Asadollah, S.B.H.S., Sharafati, A., Motta, D., Yaseen, Z.M., 2021. River water quality index prediction and uncertainty analysis: a comparative study of machine learning models. *J. Environ. Chem. Eng.* 9 (1), 104599. <https://doi.org/10.1016/j.jece.2020.104599>.
- Baek, S.-S., Choi, Y., Jeon, J., Pyo, J., Park, J., Cho, K.H., 2021. Replacing the internal standard to estimate micropollutants using deep and machine learning. *Water Res.* 188, 116535. <https://doi.org/10.1016/j.watres.2020.116535>.
- Bui, D.T., Khosravi, K., Tiefenbacher, J., Nguyen, H., Kazakis, N., 2020. Improving prediction of water quality indices using novel hybrid machine-learning algorithms. *Sci. Total Environ.* 721, 137612. <https://doi.org/10.1016/j.scitotenv.2020.137612>.
- Chen, K., Chen, H., Zhou, C., Huang, Y., Qi, X., Shen, R., Liu, F., Zuo, M., Zou, X., Wang, J., Zhang, Y., Chen, D., Chen, X., Deng, Y., Ren, H., 2020. Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. *Water Res.* 171, 115454. <https://doi.org/10.1016/j.watres.2019.115454>.
- Chiu, H.W., Lin, J.H., Chen, Y.A., Ho, S.Y., Wang, Y.J., 2010. Combination treatment with arsenic trioxide and irradiation enhances cell-killing effects in human fibrosarcoma cells in vitro and in vivo through induction of both autophagy and apoptosis. *Autophagy* 6 (3), 353–365. <https://doi.org/10.4161/auto.6.3.11229>.
- de Sá, C.R., 2019. Variance-based feature importance in neural networks. In: Kralj Novak, P., Šmuc, T., Džeroski, S. (Eds.), *Discovery Science*. Springer International Publishing, Cham, pp. 306–315. https://doi.org/10.1007/978-3-030-33778-0_24.
- Deviller, G., Lundy, L., Fatta-Kassinos, D., 2020. Recommendations to derive quality standards for chemical pollutants in reclaimed water intended for reuse in agricultural irrigation. *Chemosphere* 240. <https://doi.org/10.1016/j.chemosphere.2019.124911>.
- Di Paolo, C., Ottermanns, R., Keiter, S., Ait-Aissa, S., Bluhm, K., Brack, W., Breitholtz, M., Buchinger, S., Carere, M., Chalon, C., Cousin, X., Dulio, V., Escher, B.L., Hamers, T., Hilscherová, K., Jarque, S., Jonas, A., Maillot-Marchal, E., Marneffe, Y., Nguyen, M. T., Pandard, P., Schifferli, A., Schulze, T., Seidensticker, S., Seiler, T.-B., Tang, J., van der Oost, R., Vermeirssen, E., Zouneková, R., Zwart, N., Hollert, H., 2016. Bioassay battery interlaboratory investigation of emerging contaminants in spiked water extracts – towards the implementation of bioanalytical monitoring tools in water quality assessment and monitoring. *Water Res.* 104, 473–484. <https://doi.org/10.1016/j.watres.2016.08.018>.

- Geissen, V., Mol, H., Klumpp, E., Umlauf, G., Nadal, M., van der Ploeg, M., van de Zee, S., E.A.T.M., Ritsema, C.J., 2015. Emerging pollutants in the environment: a challenge for water resource management. *Int. Soil Water Conse.* 3 (1), 57–65. <https://doi.org/10.1016/j.iswcr.2015.03.002>.
- Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., Yang, G.-Z., 2019. XAI-Explainable artificial intelligence. *Sci. Robot.* 4 (37) <https://doi.org/10.1126/scirobotics.aay7120>.
- Guo, H., Chen, Y., Hu, H., Zhao, K., Li, H., Yan, S., Xiu, W., Coyte, R.M., Vengosh, A., 2020. High hexavalent chromium concentration in groundwater from a deep aquifer in the baiyangdian basin of the North China plain. *Environ. Sci. Technol.* 54 (16), 10068–10077. <https://doi.org/10.1021/acs.est.0c02357>.
- He, X., Li, P., 2020. Surface water pollution in the middle Chinese loess plateau with special focus on hexavalent chromium (Cr⁶⁺): occurrence, sources and health risks. *Expos. Health* 12 (3), 385–401. <https://doi.org/10.1007/s12403-020-00344-x>.
- Hettick, B.E., Canas-Carrell, J.E., French, A.D., Klein, D.M., 2015. Arsenic: a review of the element's toxicity, plant interactions, and potential methods of remediation. *J. Agric. Food Chem.* 63 (32), 7097–7107. <https://doi.org/10.1021/acs.jafc.5b02487>.
- Huang, R., Xia, M., Sakamuru, S., Zhao, J., Shahane, S.A., Attene-Ramos, M., Zhao, T., Austin, C.P., Simeonov, A., 2016. Modelling the Tox21 10 K chemical profiles for in vivo toxicity prediction and mechanism characterization. *Nat. Commun.* 7 <https://doi.org/10.1038/ncomms10425>.
- Iwasawa, A., Ayaki, M., Niwano, Y., 2013. Cell viability score (CVS) as a good indicator of critical concentration of benzalkonium chloride for toxicity in cultured ocular surface cell lines. *Regul. Toxicol. Pharmacol.* 66 (2), 177–183. <https://doi.org/10.1016/j.yrtph.2013.03.014>.
- Jin, H., Wang, C., Shi, J., Chen, L., 2014. Evaluation on joint toxicity of chlorinated anilines and cadmium to Photobacterium phosphoreum and QSAR analysis. *J. Hazard Mater.* 279, 156–162. <https://doi.org/10.1016/j.jhazmat.2014.06.068>.
- Kamarudheen, N., Chacko, S.P., George, C.A., Chettiparambil Somachandran, R., Rao, K. V.B., 2020. An ex-situ and in vitro approach towards the bioremediation of carcinogenic hexavalent chromium. *Prep. Biochem. Biotechnol.* 50 (8), 842–848. <https://doi.org/10.1080/10826068.2020.1755868>.
- Kang, G., Gao, J.Z., Xie, G., 2017. Data-driven water quality analysis and prediction: a survey. In: 2017 IEEE Third International Conference on Big Data Computing Service and Applications (BigDataService), pp. 224–232. <https://doi.org/10.1109/BigDataService.2017.40>.
- Kar, S., Leszczynski, J., 2019. Exploration of computational approaches to predict the toxicity of chemical mixtures. *Toxics* 7 (1), 15. <https://doi.org/10.3390/toxics7010015>.
- Kim, M.T., Huang, R., Sedykh, A., Wang, W., Xia, M., Zhu, H., 2016. Mechanism profiling of hepatotoxicity caused by oxidative stress using antioxidant response element reporter gene assay models and big data. *Environ. Health Perspect.* 124 (5), 634–641. <https://doi.org/10.1289/ehp.1509763>.
- Kim, Y.H., Wyrzykowska-Ceradini, B., Touati, A., Krantz, Q.T., Dye, J.A., Linak, W.P., Gullett, B., Gilmour, M.I., 2015. Characterization of size-fractionated airborne particles inside an electronic waste recycling facility and acute toxicity testing in mice. *Environ. Sci. Technol.* 49 (19), 11543–11550. <https://doi.org/10.1021/acs.est.5b03263>.
- Li, P., Feng, W., Xue, C., Tian, R., Wang, S., 2017. Spatiotemporal variability of contaminants in lake water and their risks to human health: a case study of the shahu lake tourist area, Northwest China. *Expos. Health* 9 (3), 213–225. <https://doi.org/10.1007/s12403-016-0237-3>.
- Liao, Q., Ding, Y., Jiang, Z.L., Wang, X., Zhang, C., Zhang, Q., 2019. Multi-task deep convolutional neural network for cancer diagnosis. *Neurocomputing* 348, 66–73. <https://doi.org/10.1016/j.neucom.2018.06.084>.
- Lu, H., Ma, X., 2020. Hybrid decision tree-based machine learning models for short-term water quality prediction. *Chemosphere* 249, 126169. <https://doi.org/10.1016/j.chemosphere.2020.126169>.
- Marikkani, S., Kumar, J.V., Muthuraj, V., 2019. Design of novel solar-light driven sponge-like Fe₂V₄O₁₃ photocatalyst: a unique platform for the photoreduction of carcinogenic hexavalent chromium. *Sol. Energy* 188, 849–856. <https://doi.org/10.1016/j.solener.2019.06.075>.
- Meister, M.T., Boedicker, C., Graab, U., Hugle, M., Hahn, H., Klingebiel, T., Fulda, S., 2016. Arsenic trioxide induces Noxa-dependent apoptosis in rhabdomyosarcoma cells and synergizes with antimicrotubule drugs. *Canc. Lett.* 381 (2), 287–295. <https://doi.org/10.1016/j.canlet.2016.07.007>.
- Mikolajczyk, A., Sizochenko, N., Mulkiewicz, E., Malankowska, A., Rasulev, B., Puzyn, T., 2019. A chemoinformatics approach for the characterization of hybrid nanomaterials: safer and efficient design perspective. *Nanoscale* 11 (24), 11808–11818. <https://doi.org/10.1039/c9nr01162e>.
- Ministry of Environmental Protection of the People's Republic of China, 2002. Technical specifications for surface water and sewage monitoring (HJ/T 91–2002). China Environmental Science Press, Beijing, China.
- Naidu, R., Espana, V.A.A., Liu, Y., Jit, J., 2016. Emerging contaminants in the environment: risk-based analysis for better management. *Chemosphere* 154, 350–357. <https://doi.org/10.1016/j.chemosphere.2016.03.068>.
- Najah Ahmed, A., Binti Othman, F., Abdulmohsin Afan, H., Khaleel Ibrahim, R., Ming Fai, C., Shabbir Hossain, M., Ehteram, M., Elshafie, A., 2019. Machine learning methods for better water quality prediction. *J. Hydrol.* 578, 124084. <https://doi.org/10.1016/j.jhydrol.2019.124084>.
- Neale, P.A., Altenburger, R., Ait-Aissa, S., Brion, F., Busch, W., de Aragão Umbuzeiro, G., Denison, M.S., Du Pasquier, D., Hilscherová, K., Hollert, H., Morales, D.A., Novák, J., Schlichting, R., Seiler, T.-B., Serra, H., Shao, Y., Tindall, A.J., Tollefsen, K.E., Williams, T.D., Escher, B.L., 2017. Development of a bioanalytical test battery for water quality monitoring: fingerprinting identified micropollutants and their contribution to effects in surface water. *Water Res.* 123, 734–750. <https://doi.org/10.1016/j.watres.2017.07.016>.
- Nys, C., Versieren, L., Cordery, K.I., Blust, R., Smolders, E., De Schamphelaere, K.A.C., 2017. Systematic evaluation of chronic metal-mixture toxicity to three species and implications for risk assessment. *Environ. Sci. Technol.* 51 (8), 4615–4623. <https://doi.org/10.1021/acs.est.6b05688>.
- Poon, A.I.F., Sung, J.J.Y., 2021. Opening the black box of AI-Medicine. *J. Gastroenterol. Hepatol.* 36 (3), 581–584. <https://doi.org/10.1111/jgh.15384>.
- Raies, A.B., Bajic, V.B., 2016. In silico toxicology: computational methods for the prediction of chemical toxicity. *WIREs. Compu. Mol. Sci.* 6 (2), 147–172. <https://doi.org/10.1002/wcms.1240>.
- Ren, X., Kou, Y.Y., Kim, T., Chae, K.J., Ng, H.Y., 2017. Toxicity study of reclaimed water on human embryonic kidney cells. *Chemosphere* 189, 390–398. <https://doi.org/10.1016/j.chemosphere.2017.08.134>.
- Sanjay Kumar, C G Y a P B T, 2014. Arsenic trioxide induces oxidative stress, DNA damage, and mitochondrial pathway of apoptosis in human leukemia (HL-60) cells. *J. Exp. Clin. Oncol. Res.* 33 <https://doi.org/10.1186/1756-9966-33-42>.
- Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2020. Grad-CAM: visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vis.* 128 (2), 336–359. <https://doi.org/10.1007/s11263-019-01228-7>.
- Song, M.-M., Song, W.-J., Bi, H., Wang, J., Wu, W.-L., Sun, J., Yu, M., 2010. Cytotoxicity and cellular uptake of iron nanowires. *Biomaterials* 31 (7), 1509–1517. <https://doi.org/10.1016/j.biomaterials.2009.11.034>.
- Su, L.M., Zhao, Y.H., Yuan, X., Mu, C.F., Wang, N., Yan, J.C., 2010. Evaluation of combined toxicity of phenols and lead to photobacterium phosphoreum and quantitative structure-activity relationships. *Bull. Environ. Contam. Toxicol.* 84 (3), 311–314. <https://doi.org/10.1007/s00128-009-9665-0>.
- Tang, Z., Chuang, K.V., DeCarli, C., Jin, L.-W., Beckett, L., Keiser, M.J., Dugger, B.N., 2019. Interpretable classification of Alzheimer's disease pathologies with a convolutional neural network pipeline. *Nat. Commun.* 10 <https://doi.org/10.1038/s41467-019-10212-1>.
- Ukic, S., Sigurnjak, M., Cvetnic, M., Markic, M., Stankov, M.N., Rogosic, M., Rasulev, B., Bozic, A.L., Kusic, H., Bolanca, T., 2019. Toxicity of pharmaceuticals in binary mixtures: assessment by additive and non-additive toxicity models. *Ecotoxicol. Environ. Saf.* 185 <https://doi.org/10.1016/j.ecoenv.2019.109696>.
- Wadhawan, A.R., Stone, A.T., Bouwer, E.J., 2013. Biogeochemical controls on hexavalent chromium formation in estuarine sediments. *Environ. Sci. Technol.* 47 (15), 8220–8228. <https://doi.org/10.1021/es401159b>.
- Wang, X., Ji, D., Chen, X., Ma, Y., Yang, J., Ma, J., Li, X., 2017. Extended biotic ligand model for predicting combined Cu-Zn toxicity to wheat (*Triticum aestivum* L.): incorporating the effects of concentration ratio, major cations and pH. *Environ. Pollut.* 230, 210–217. <https://doi.org/10.1016/j.envpol.2017.06.037>.
- Wang, X., Luo, X., Wang, Q., Liu, Y., Naidu, R., 2020. Predicting the combined toxicity of binary metal mixtures (Cu-Ni and Zn-Ni) to wheat. *Ecotoxicol. Environ. Saf.* 205, 111334. <https://doi.org/10.1016/j.ecoenv.2020.111334>.
- Wang, X., Meng, X., Ma, Y., Pu, X., Zhong, X., 2018a. The prediction of combined toxicity of Cu-Ni for barley using an extended concentration addition model. *Environ. Pollut.* 242 (Pt A), 136–142. <https://doi.org/10.1016/j.envpol.2018.06.070>.
- Wang, Y., Zhou, L., Wang, X., Liu, X., Jiang, L., Wang, J., Sun, H., Jiang, C., Xing, X., Zhang, Y., Pan, B., Yan, B., 2018b. A human cell panel for evaluating safe application of nano-ZrO₂/polymer composite in water remediation. *Ecotoxicol. Environ. Saf.* 166, 474–481. <https://doi.org/10.1016/j.ecoenv.2018.09.098>.
- Xu, J., Wei, D., Wang, F., Bai, C., Du, Y., 2020. Bioassay: a useful tool for evaluating reclaimed water safety. *J. Environ. Sci.* 88, 165–176. <https://doi.org/10.1016/j.jes.2019.08.014>.
- Yang, G., Chen, C., Wang, Y., Peng, Q., Zhao, H., Guo, D., Wang, Q., Qian, Y., 2017. Mixture toxicity of four commonly used pesticides at different effect levels to the epigeic earthworm. *Eisenia fetida*. *Ecotox. Environ. Safe.* 142, 29–39. <https://doi.org/10.1016/j.ecoenv.2017.03.037>.
- Yang, Y., Lu, Y., Wu, Q.Y., Hu, H.Y., Chen, Y.H., Liu, W.L., 2015. Evidence of ATP assay as an appropriate alternative of MTT assay for cytotoxicity of secondary effluents from WWTPs. *Ecotoxicol. Environ. Saf.* 122, 490–496. <https://doi.org/10.1016/j.ecoenv.2015.09.006>.
- Zhang, Z., Guo, Y., Wu, J., Su, F., 2021. Surface Water Quality and Health Risk Assessment in Taizhou City, Zhejiang Province (China). *Expos. Health*. <https://doi.org/10.1007/s12403-021-00408-6>.
- Zhitkovich, A., 2005. Importance of chromium-DNA adducts in mutagenicity and toxicity of chromium(VI). *Chem. Res. Toxicol.* 18 (1), 3–11. <https://doi.org/10.1021/tx049774+>.
- Zhou, X.X., He, S., Gao, Y., Li, Z.C., Chi, H.Y., Li, C.J., Wang, D.J., Yan, B., 2021. Protein corona-mediated extraction for quantitative analysis of nanoplastics in environmental waters by pyrolysis gas chromatography/mass spectrometry. *Anal. Chem.* 93 (17), 6698–6705. <https://doi.org/10.1021/acs.analchem.1c00156>.



Developmental toxicity of fenbuconazole in zebrafish: Effects on mitochondrial respiration and locomotor behavior

Yingju Qin^{a,b}, Xiaohong Wang^{a,b,*}, Xiliang Yan^{a,b,*}, Di Zhu^c, Jia Wang^c, Siying Chen^{a,b}, Shuo Wang^c, Yang Wen^d, Christopher J. Martyniuk^e, Yuanhui Zhao^{c,**}

^a Key Laboratory for Water Quality and Conservation of the Pearl River Delta, Ministry of Education, Guangzhou University, Guangzhou 510006, China

^b Institute of Environmental Research at Greater Bay Area, Ministry of Education, Guangzhou University, Guangzhou 510006, China

^c State Environmental Protection Key Laboratory of Wetland Ecology and Vegetation Restoration, School of Environment, Northeast Normal University, Changchun, Jilin 130117, China

^d Key Laboratory of Environmental Materials and Pollution Control, The Education Department of Jilin Province, School of Environmental Science and Engineering, Jilin Normal University, Siping, Jilin 136000, China

^e Center for Environmental and Human Toxicology, Department of Physiological Sciences, College of Veterinary Medicine, UF Genetics Institute, Interdisciplinary Program in Biomedical Sciences in Neuroscience, University of Florida, Gainesville, Florida, 32611, USA

ARTICLE INFO

Keywords:

Fenbuconazole
Mitochondrial function
Locomotion
Zebrafish
Aquatic toxicology

ABSTRACT

Triazole fungicides are used to control the disease of cereal crops but may also cause adverse effects on non-target organisms. There is a lack of toxicity data for some triazoles such as fenbuconazole in aquatic organisms. This research was conducted to evaluate the toxicity of fenbuconazole at environmentally relevant concentrations with attention on the mitochondria, antioxidant system, and locomotor activity in zebrafish. Zebrafish were exposed to one concentration of 5, 50, 200 or 500 ng/L fenbuconazole for 96 h. There was no effect on survival nor percentage of fish hatched, but exposure to 200 and 500 ng/L fenbuconazole resulted in malformation and hypoactivity in zebrafish. Oxygen consumption rates (OCR) of embryos were measured to determine if the fungicide impaired mitochondrial respiration. Exposure to 500 ng/L fenbuconazole reduced basal OCR and oligomycin-induced ATP linked respiration in exposed fish. Fenbuconazole reduced mitochondrial membrane potential and reduced the activities of mitochondrial Complex II and III. Transcript levels of both *sdhc* and *cyc1*, each related to Complex II and III, were also altered in expression by fenbuconazole exposure, consistent with mitochondrial dysfunction in embryos. Fenbuconazole activated the antioxidant system, based upon both transcriptional and enzymatic data in zebrafish. Consistent with mitochondrial impairment, molecular docking confirmed a strong binding capacity of the fungicide at the Q_i site of Complex III, revealing this complex is susceptible to fenbuconazole. This study reveals potential toxicity pathways related to fenbuconazole exposure in aquatic organisms; such data can improve risk assessments for triazole fungicides.

1. Introduction

Fungal pathogens are ubiquitous and problematic in agriculture, and many fungal species are demonstrating increased resistance to pesticides. As a result, fungicides must be applied more abundantly and frequently to improve crop yields and maintain the quality of agricultural products. Within the global market, there are many different classes of fungicide, and each class is characterized by unique chemical structures. Triazoles comprise the second largest proportion of the

global market for fungicides, and have been used in wood preservatives, paint production and textiles (WBISS Consulting Co, Ltd, 2016). Triazoles are chemically and photochemically stable, showing low propensity for degradation in the environment, leading to a relatively long half-life (Aladaghlo et al., 2019). Through surface runoff and soil leaching, in addition to spray residue in the air, water or soil, triazoles can enter aquatic areas and pose an exposure risk to organisms. Several studies report that triazoles are present in aquatic or soil environments around the world (Table S1). Triazole fungicides are frequently detected

* Corresponding authors at: Key Laboratory for Water Quality and Conservation of the Pearl River Delta, Ministry of Education, Guangzhou University, Guangzhou 510006, China.

** Corresponding author.

E-mail addresses: wangxh@gzhu.edu.cn (X. Wang), yanxiliang1991@163.com (X. Yan), zhaoyh@nenu.edu.cn (Y. Zhao).

<https://doi.org/10.1016/j.tox.2022.153137>

Received 1 November 2021; Received in revised form 17 February 2022; Accepted 22 February 2022

Available online 24 February 2022

0300-483X/© 2022 Elsevier B.V. All rights reserved.

in aquatic ecosystems at microgram per liter levels, suggesting that aquatic wildlife may be susceptible to exposures leading to toxicity.

Fenbuconazole is a triazole fungicides that is widely used to control black star disease, brown rot, and leaf spot in fruits (Mondal et al., 2005; Xu et al., 2010). It has a long half-life (61 days) for biodegradation in the field soil (Pesticide Properties DataBase, 1992). Compared to other triazoles, fenbuconazole has been given less attention regarding its potential for toxicity to aquatic organisms. Studies have shown that fenbuconazole has been detected in waterways in the Chinese city Xiamen, with concentrations ranging 0.22–6.98 ng/L (Wu et al., 2018). Fenbuconazole enantiomers were also detected in soil samples collected from Langfang, China, with residual concentrations ranging 10.49–23.54 µg/kg (Li et al., 2012). In addition, fenbuconazole residues have been detected in pears, peaches, grapes, and oranges at concentrations ranging 0.11–0.34 mg/kg (European Food Safety Authority, 2015). Human exposure risks are therefore a concern due to evidence of toxicity in rodents. For example, fenbuconazole exposure increased liver weight of mice in a dose-dependent manner, and this triggered hepatic histological injury (Juberg et al., 2006). As such, agricultural residues may pose hazards to human health.

Studies investigating potential toxicity of fenbuconazole to aquatic organisms such as fish are currently lacking. Triazole fungicides can exert toxicity to early staged zebrafish, including developmental toxicity, apoptosis, endocrine disruption, cardiovascular toxicity, adverse motor response, and disruptions in energy metabolism, specifically lipid and amino acid metabolism (Hermesen et al., 2012; Teng et al., 2018a,b; Wang et al., 2017; Icoglu Aksakal and Ciltas, 2018; Souders II et al., 2019; Sun et al., 2020; Shen et al., 2021). For instance, the 96 h- LC_{50} values for triazoles to *Danio rerio* typically range 1.70–26.4 mg/L (Mu et al., 2013; Cao et al., 2019; Sun et al., 2020; Wang et al., 2020; Weng et al., 2021). These data indicate that triazoles can be classified as moderately toxic to zebrafish, based upon 96 h- LC_{50} values ranging 0.1–100 mg/L (Zubrod et al., 2019). However, the toxicity of fenbuconazole to aquatic organisms such as fish has yet to be adequately addressed, despite growing literature reporting adverse effects related to other triazole fungicides.

Previous studies investigating triazoles point to developmental toxicity, metabolic, and bioenergetic alterations in zebrafish. Thus, we focused on these endpoints in early staged zebrafish following exposure to fenbuconazole. Mitochondria are responsible for oxidative phosphorylation of cells and dysfunction in ATP production can lead to several terminal effects in fish, such as delayed development and aberrant behavior. Mitochondrial dysfunction is also highly related to increased reactive oxygen species (ROS) and antioxidant system reduction or depletion (Mailloux et al., 2013; Luo et al., 2017; Bailey et al., 2018). Several agricultural chemicals have been confirmed to cause mitochondrial damage and oxidative stress in different fish species (Jin et al., 2010; Mu et al., 2015; Yang et al., 2016; Yang et al., 2020; Zhang et al., 2020; Park et al., 2021). However, the outcomes linked with oxidative stress response and mitochondria remain to be determined with fenbuconazole.

In this study, we assessed if fenbuconazole negatively affected mitochondria in early staged zebrafish at environmentally relevant concentrations. To achieve this, a mitochondrial stress test was conducted to test oxygen consumption rates in *Danio rerio* embryos. In addition, mitochondrial membrane potential, activities of mitochondrial complex I – V, and molecular docking was measured to evaluate more comprehensively mitochondrial responses in zebrafish. The mitochondrial respiratory chain is deemed the primary site for ROS production (Rigoulet et al., 2011; Nickel et al., 2014; Zhao et al., 2019), and we thus determined the concentrations of ROS, MDA and GSH, total antioxidant capacity (T-AOC), activities of SOD and CAT following exposure, as well as transcript levels of oxidative stress-related genes. Lastly, locomotor activity of zebrafish larvae was quantified, as locomotor activity can be directly related to the availability of ATP. This study reveals potential toxicity pathways related to fenbuconazole exposure in aquatic

organisms; such data can improve risk assessments for triazole fungicides.

2. Materials and methods

2.1. Zebrafish protocols, developmental exposure and harvesting

Parent zebrafish (wildtype AB-strain, *Danio rerio*, 5–6-month-old) were purchased from Nanjing YiShu LiHua Biotechnology Co., Ltd. (China). The females and males were separately housed in tanks at 27 ± 1 °C with a light/dark cycle of 14 h: 10 h and fed with brine shrimp three times every day. The dissolved oxygen concentration remained $\geq 80\%$ of air saturation. For fish breeding, adult fish were randomly caught and put together at a ratio of 2:2 (male/female) the previous night before embryos collection. The next morning, the fertilized eggs were collected and rinsed with embryo rearing medium (ERM) for exposure experiments.

Fenbuconazole (CAS: 119611–00–6; purity $\geq 99.1\%$) was purchased from Anpu Experimental Technology Co., Ltd. (Shanghai, China). Analytical acetone (0.001%) was used as a cosolvent to prepare the exposure solutions with ERM. The embryos were randomly placed in the ERM only group, solvent control group, or in one concentration of 5, 50, 200, or 500 ng/L fenbuconazole. In our experiments, glass petri dishes were used to accommodate 50 zebrafish embryos with the 30 mL exposure solution. We prepared three dishes for each treatment and place the dishes into an incubator at a controlled temperature of 27 ± 1 °C, with a light/dark cycle of 14 h:10 h. The exposure solutions were newly prepared every 24 h with a 50% water change. We observed the development of fish using a stereoscope (Leica, S9i, Wetzlar, Germany) each day, and we removed dead embryos or larvae until 96 h.

To assess mitochondrial bioenergetics of embryos, 48 hpf zebrafish per group were collected into a 24-well Islet Capture Microplate for detection of oxygen consumption rates. For endpoints associated with oxidative stress response, mitochondrial membrane potential, and mitochondrial complexes activities, 30 fish from each Petri dish were collected into one tube. For gene expression analysis, 10 fish were collected from each Petri dish as a biological replicate. For behavioral endpoints, fish were exposed to fenbuconazole for up to 6 dpf, and 5–6 fish from each dish were placed into a 96-well plate for activity assessment.

2.2. Mitochondrial stress test

The oxygen consumption rate (OCR) of whole embryos following 48 h exposure was determined using the Seahorse XFe24 Analyzer (Agilent Technologies Inc.). We used a 24-well Islet Capture Microplate to hold embryos, and each well contained one fish with 0.001% acetone or one concentration of 5, 50, 200 or 500 ng/L fenbuconazole ($N = 4/\text{group}$). The basal OCR was initially determined for 12 cycles. Next, we used mitochondrial inhibitors to probe mitochondrial respiration, assessing the respiratory sources. Based on our previously published methods (Wang et al., 2020), oligomycin, carbonyl cyanide 4-(trifluoromethoxy) phenylhydrazone (FCCP) and sodium azide were added into each well in sequence, with the final concentrations of 9.4 µM, 6 µM and 20 mM, respectively. Oligomycin is an inhibitor of ATP production, thus one can evaluate the ATP-linked OCR following injection. Here, 18 measurement cycles were conducted during this period. FCCP is a mitochondrial uncoupler, as such it was injected to maximize the OCR and 18 cycles were conducted following injection. Using this response, the maximal OCR and spare respiratory capacity of fish embryos can be calculated. Lastly, sodium azide was used to thoroughly inhibit mitochondrial respiration. There were 65 cycles performed in this final stage.

2.3. JC-1 measurement

The mitochondrial membrane potential (MMP) of zebrafish was

determined by the JC-1 kit purchased from Beyotime Institute of Biotechnology. Embryos following exposure were collected from the petri dishes for measurement ($N = 3/\text{group}$). The mitochondria of fish embryos were isolated by using the Mitochondria Isolation Kit bought from Shanghai Beyotime Institute of Biotechnology according to the protocols provided by manufacturer. The isolated mitochondria were mixed with JC-1 solution for 20 min at 37 °C. The red and green fluorescence values of JC-1 were measured at excitation/emission wavelengths of 530/590 nm and 485/528 nm, respectively, using a microplate reader (BioTek Instruments, Vermont, USA). The result is presented as the red/green fluorescence ratio, as an indicator of MMP.

2.4. Activities of mitochondrial respiratory chain complexes

At 96 hpf, zebrafish were collected to determine the activities of mitochondrial respiratory chain complexes. Three replicates were prepared for each group. Zebrafish were washed with cold normal saline and transferred into each tube with a mass to volume ratio of 1: 9. The zebrafish were homogenized using an ultrasonic grinder with an amplitude of 14 m for 30 s, and centrifuged for 10 min at a low-speed of 1000 rpm/min. Supernatants were then collected and centrifuged at a high speed of 12,000 g for 15 min at 4 °C. The precipitates were obtained as isolated mitochondria and added to cold homogenizing medium for subsequent experiments. The activities of Complex I-V were determined using the Mitochondrial Respiratory Chain Complex Assay Kit, based on the manufacturer's instructions (Shanghai Meilian Biotechnology Co., Ltd.). The activities of mitochondrial respiratory chain complexes were determined using the double antibody sandwich method. The isolated mitochondria were added into enzyme wells pre-coated with the corresponding antibodies. Then the horseradish peroxidase-labelled recognition antigens were added into each well for a 30 min incubation at 37 °C. Finally, each well was added with tetramethyl benzidine and then measured at 450 nm using a microplate reader.

2.5. Molecular docking

To better understand the molecular mechanism of the interaction between fenbuconazole and the mitochondrial respiratory complex enzyme, we performed *in silico* modeling based on molecular docking. An autodock approach aims to model the interaction between a protein and a small molecule and can improve understanding of *in vitro* or *in vivo* toxicity induced by fenbuconazole. Herein, the human mitochondrial respiratory complex III was selected as the protein receptor, and the corresponding crystal structure (PDB ID: 5XTE) was obtained from the Protein Data Bank (PDB) structure database (<https://www.rcsb.org/>). Fenbuconazole structure (Compound CID: 86138) used for molecular docking was obtained from the PubChem database (<https://pubchem.ncbi.nlm.nih.gov/>). AutoDock Vina (Oleg Trott, 2010) was used to perform molecular docking of fenbuconazole to the human mitochondrial respiratory complex III. Ligand binding poses were evaluated by the ΔG_{bind} scoring function (Oleg Trott, 2010) (Eq. 1), and the pose with the highest ΔG_{bind} score was selected for further analysis.

$$\Delta G_{\text{bind}} = \Delta G_{\text{vdw}} + \Delta G_{\text{H-bond}} + \Delta G_{\text{ele}} + \Delta G_{\text{int}} \quad (1)$$

Where, ΔG_{vdw} , $\Delta G_{\text{H-bond}}$, ΔG_{ele} , and ΔG_{int} represents van der Waals interaction, hydrogen bond interaction, electrostatic interaction, and intramolecular interaction, respectively.

2.6. Oxidative stress response

Following 96 h exposure, zebrafish samples were fully washed and homogenized in cold PBS buffer, then moved to centrifuge at a speed of 12,000 g at a low temperature, to obtain the supernatants for determining concentrations of ROS, malonaldehyde (MDA), and glutathione (GSH), total antioxidant capacity (T-AOC), and activities of superoxide

dismutase (SOD) and catalase (CAT) ($N = 3/\text{group}$). The experimental protocols were conducted according to manufacturer's instructions of commercial kits bought from Jiancheng Bioengineering Institute, Nanjing, China. For each sample, protein contents were determined using a BCA Protein Assay Kit bought from Beyotime Institute of Biotechnology.

2.7. Gene expression analysis

After exposure, approximately 10 zebrafish were collected as one replicate for transcript analysis ($N = 3/\text{group}$). Firstly, total RNA from samples was extracted using RNAPure Pure Tissue Kit purchased from Beijing Tiangen Biotech. The quantity and quality of RNA were confirmed by NanoDrop One (ThermoFisher, USA) using the absorbance at 260/280 nm. Then 1.5 μg purified RNA from each sample was used for cDNA synthesis using Takara PrimeScript™ RT Master Mix after the manufacturer's protocols.

Quantification of transcription was conducted using a Roche Light-Cycler® 96 Instrument with Takara TB Green™ Premix Ex Taq™ II. We prepared three samples with no reverse transcriptase during cDNA synthesis and three samples with no cDNA template during real-time PCR. Three biological replicates measured in triplicate were run for each group. The genes measured covered superoxide dismutase (*sod1*, *sod2*), catalase (*cat*), succinic dehydrogenase (*sdha*, *sdhb*, *sdhc*), mitochondrial cytochrome b (*cytb*), cytochrome c1 (*cyc1*) and ubiquinol-cytochrome c reductase core protein II b (*uqcrc2b*). Target genes were normalized to the housekeeping gene (ribosomal subunit 18, *rps18*), using the method of relative $\Delta\Delta C_q$ based on Pfaffl (2001). Primer sets are listed in Table S2.

2.8. Behavioral response test

We conducted a visual motor response test for zebrafish exposed up until 5 dpf. The animal procedure was carried out in accordance with the Guidelines for Care and Use of Laboratory Animals of Guangzhou University and approved by the Animal Ethics Committee of Guangzhou University. Following exposure, fish were transferred into a 96-well plate ($N = 15/\text{group}$) and this plate was placed into an observation chamber (Noldus Information Technology) in the darkness to acclimate for 24 h. In the next day, 6 dpf fish were individually and simultaneously tracked by an infrared analog camera in the observation chamber. The first acclimation was initially recorded for 10 min, then the Noldus White Routine began as alternating 10 min light and dark with a 50 min tracking in total. Data from each period were independently collected, and behavioral data for zebrafish from one group were binned into an averaged value for each minute. Also, fish total distance travelled, and velocity were tracked, however we show data for zebrafish activity as data were similar across endpoints.

2.9. Data processing

Statistical analyses were conducted by a statistical software GraphPad Prism 8.0. Significant differences between groups were analyzed by one-way ANOVA followed by a Dunnett's post-hoc test. The F ratio when reported is accompanied with degree of freedom for the numerator (DFn), and denominator (DFd) or F (DFn, DFd). Data are presented as a mean value \pm standard deviation (SD). Significant difference was considered $p < 0.05$. Different letters indicate a significant difference between groups. All figures were generated in GraphPad Prism 8.0 software.

3. Results

3.1. Developmental toxicity induced by fenbuconazole

We first evaluated the acute toxicity of fenbuconazole at

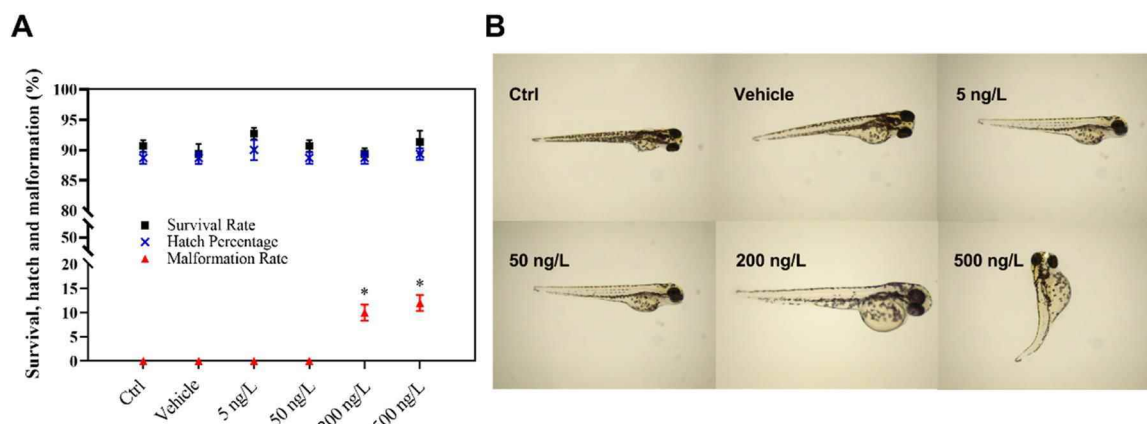


Fig. 1. Developmental toxicity induced by fenbuconazole. (A) Survival rate, hatch percentage and malformation rate of zebrafish embryos following exposure to vehicle, 5, 50, 200 and 500 ng/L fenbuconazole. (B) Morphological changes of fish embryos following exposure to vehicle, 5, 50, 200 and 500 ng/L fenbuconazole. N = 3/group. Asterisks (*) indicate a significant difference between the treatment and the control at $p < 0.05$.

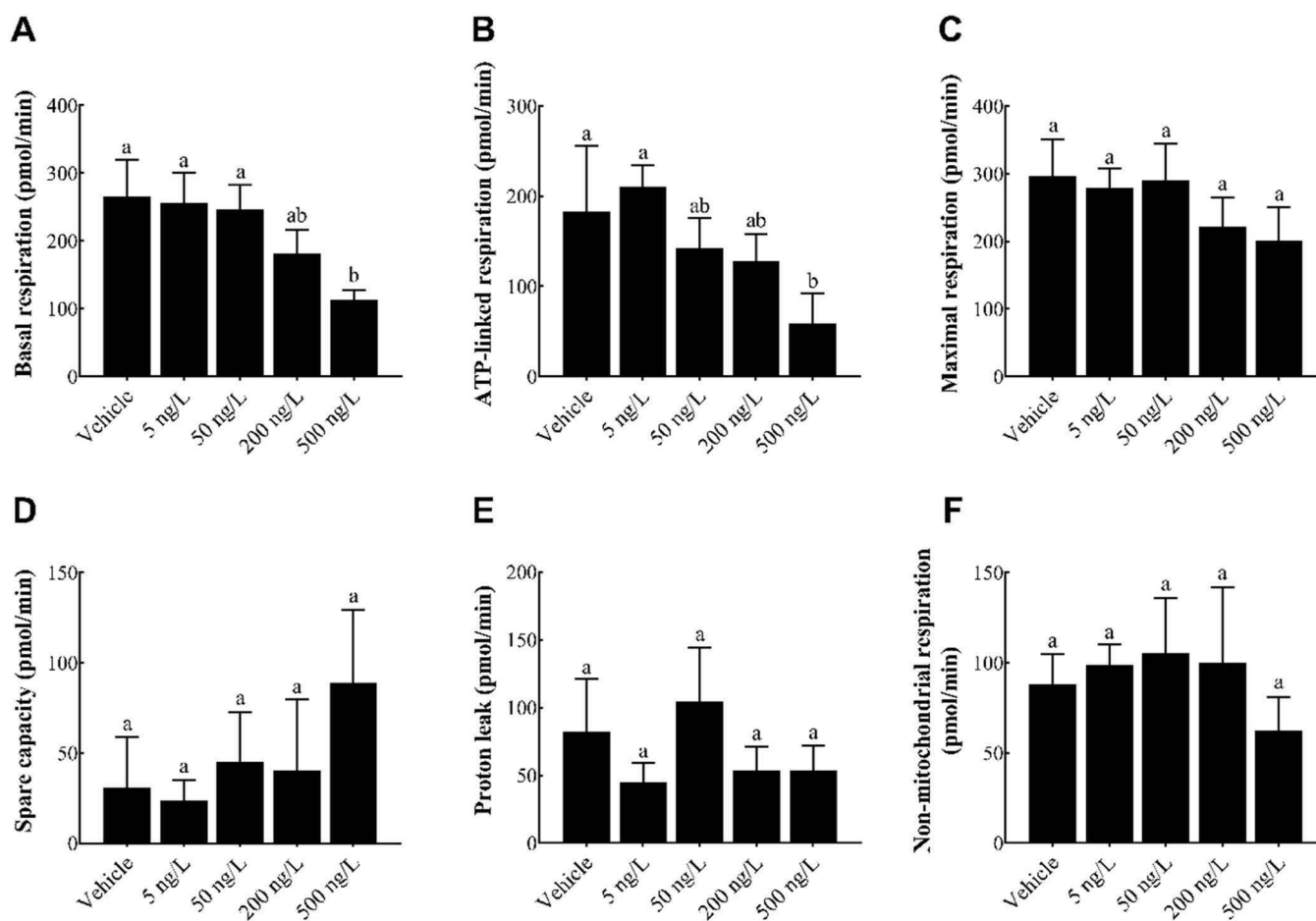


Fig. 2. Mitochondrial oxygen consumption rate (OCR) of zebrafish embryos following a 48-hour exposure to fenbuconazole. (A) Basal OCR, (B) Oligomycin induced ATP-linked production, (C) FCCP-induced maximal OCR, (D) Spare respiratory capacity, (E) Proton leak, and (F) Non-mitochondrial OCR. Results are shown as average value \pm standard deviation. N = 4/group. Different letters denote a significant difference between groups at $p < 0.05$.

concentrations ranging from 5 to 500 ng/L, to capture potential effects at environmental levels observed for several triazoles. After acute exposure, zebrafish developing in either vehicle solution or fenbuconazole at 5, 50, 200 and 500 ng/L showed no significant differences in survival rate and hatch percent ($F_{(5, 12)} = 2.945$; $p = 0.0581$; $F_{(5, 12)} = 0.7875$; $p = 0.5782$). However, zebrafish exposed to either 200 or 500 ng/L presented with malformations, with rates of 10% and 12%,

respectively ($F_{(5, 12)} = 110.2$; $p < 0.0001$) (Fig. 1B).

3.2. Mitochondrial bioenergetic responses induced by fenbuconazole

Using a mitochondrial stress test, our results showed that basal OCR ($F_{(4, 15)} = 10.74$; $p = 0.0003$) and oligomycin-induced ATP ($F_{(4, 15)} = 7.522$; $p = 0.0016$) were reduced by 500 ng/L fungicide (Fig. 2A-B).

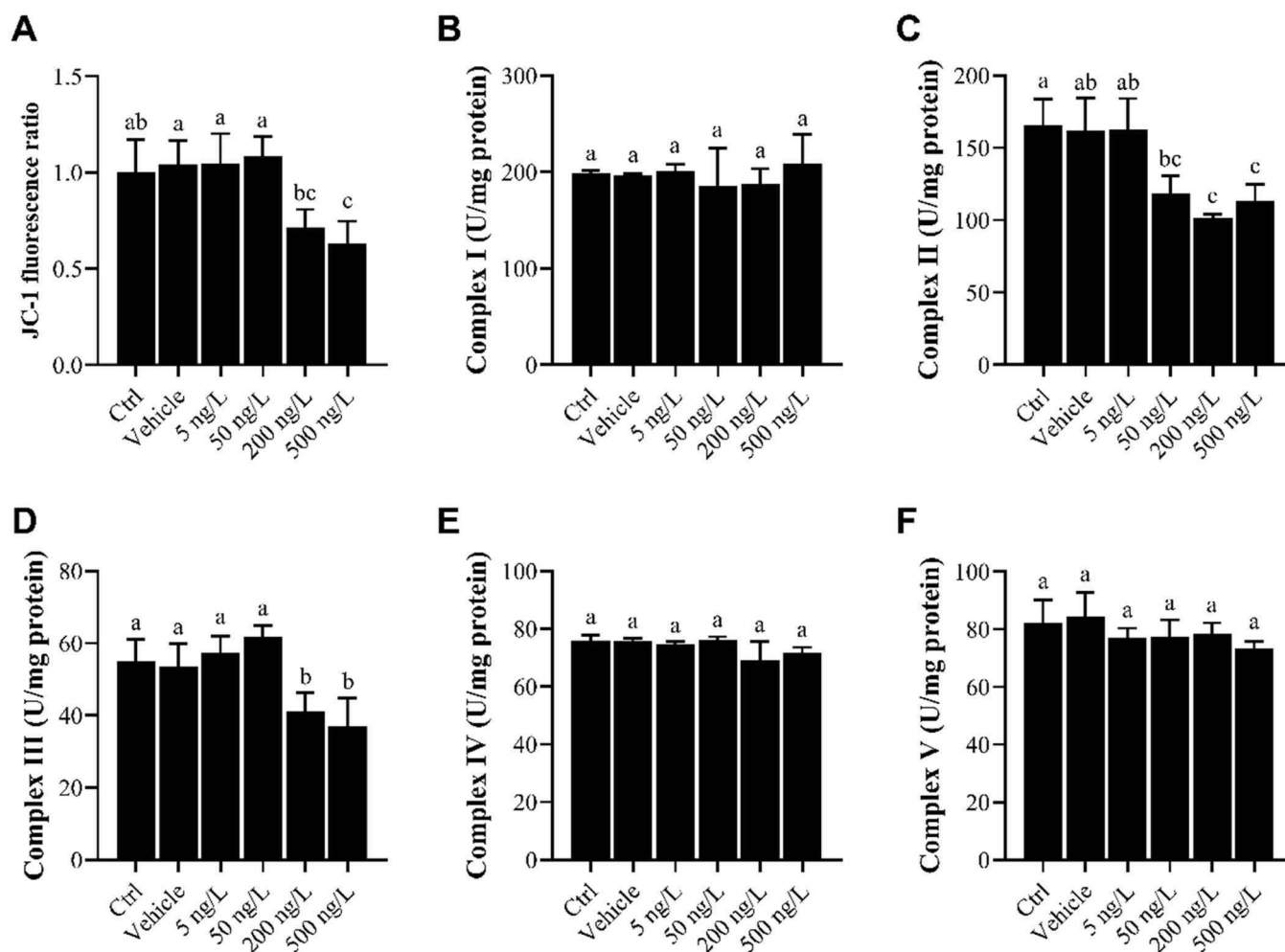


Fig. 3. Mitochondrial electron transfer chain activity of zebrafish following 96 h exposure to fenbuconazole. (A) Mitochondrial membrane potential, (B) Activity of Complex I, (C) Activity of Complex II, (D) Activity of Complex III, (E) Activity of Complex IV, (F) Activity of Complex V. Results are showed as average value \pm standard deviation. $N = 3/\text{group}$. Different letters denote a significant difference between groups at $p < 0.05$.

The maximal OCR, proton leak, spare capacity and non-mitochondrial respiration were not affected by the fungicide at the tested concentrations (Fig. 2C-F).

We next tested the MMP and activities of Complex I-V, components of the ETC. The MMP of zebrafish was decreased by 200 and 500 ng/L fungicide ($F_{(5, 18)} = 9.029$; $p = 0.0002$) (Fig. 3A). The activities of Complex II and III in zebrafish were also remarkably reduced with exposure to 200 and 500 ng/L fungicide ($F_{(5, 12)} = 9.457$; $p = 0.0008$; $F_{(5, 12)} = 8.634$; $p = 0.0011$) (Fig. 3C-D).

We further measured several transcriptional responses related to mitochondrial Complex II and III of the ETC in zebrafish, including *sdha*, *sdhb*, *sdhc*, *uqcrc2b*, *cytb* and *cyc1*. The transcription levels of *sdhc* and *cyc1*, each a component of Complex II and III, were significantly altered in zebrafish following exposure to the fungicide ($F_{(4, 10)} = 75.71$; $p < 0.0001$; $F_{(4, 10)} = 11.98$; $p = 0.0008$) (Fig. 4C, F). Transcripts of *sdhc* were upregulated in fish following exposure to 5 and 50 ng/L fungicide, while the transcriptions of *sdhc* were downregulated following exposure to 200 and 500 ng/L fungicide. The expression pattern of *cyc1* also corresponded to patterns observed with *sdhc*, and *cyc1* transcripts were increased in abundance in fish following exposure to 5 ng/L but were decreased in abundance following exposure to 500 ng/L fenbuconazole.

3.3. Binding mode of fenbuconazole to cytochrome b subunit

As mitochondrial Complex III was significantly affected by

fenbuconazole, we conducted molecular in silico docking between fenbuconazole and the cytochrome *b* subunit. In Fig. 5A, the human mitochondrial respiratory complex III can be divided into three main regions: intermembrane space, transmembrane space, and the mitochondrial matrix. In Fig. 5B, these chains include the cytochrome *b* subunit, where two hemes (b_L and b_H) and two ubiquinone binding sites (Q_0 and Q_i) are located. Two well-known inhibitors (azoxystrobin (Rodrigues et al., 2013) and antimycin (Lai et al., 2005) were regarded as positive controls to evaluate the binding ability of fenbuconazole with Q_0 and Q_i sites. The docking results revealed that fenbuconazole showed higher binding energy at Q_i site (Table 1). Fenbuconazole (-9.0 kcal/mol) exhibited higher binding energy than antimycin (-8.3 kcal/mol), showing a stronger binding capacity at Q_i site than classical inhibitors. Based on the mode analysis from molecular docking, fenbuconazole bound to the Q_0 site or the Q_i site through π interactions (Fig. 5C, D). At the Q_0 site, the 1,2,4-triazole can form π -cation interaction with Arg358, and the benzene ring can form π -H interaction with Gly251. At the Q_i site, the 1,2,4-triazole can form π -H interaction with Tyr358, and the benzene ring can form π - π interaction with Phe95.

3.4. Oxidative stress responses induced by fenbuconazole

We measured several endpoints related to oxidative stress responses in fish. In fish, MDA levels at 500 ng/L fenbuconazole and ROS levels at 50, 200 or 500 ng/L fenbuconazole were increased relative to control

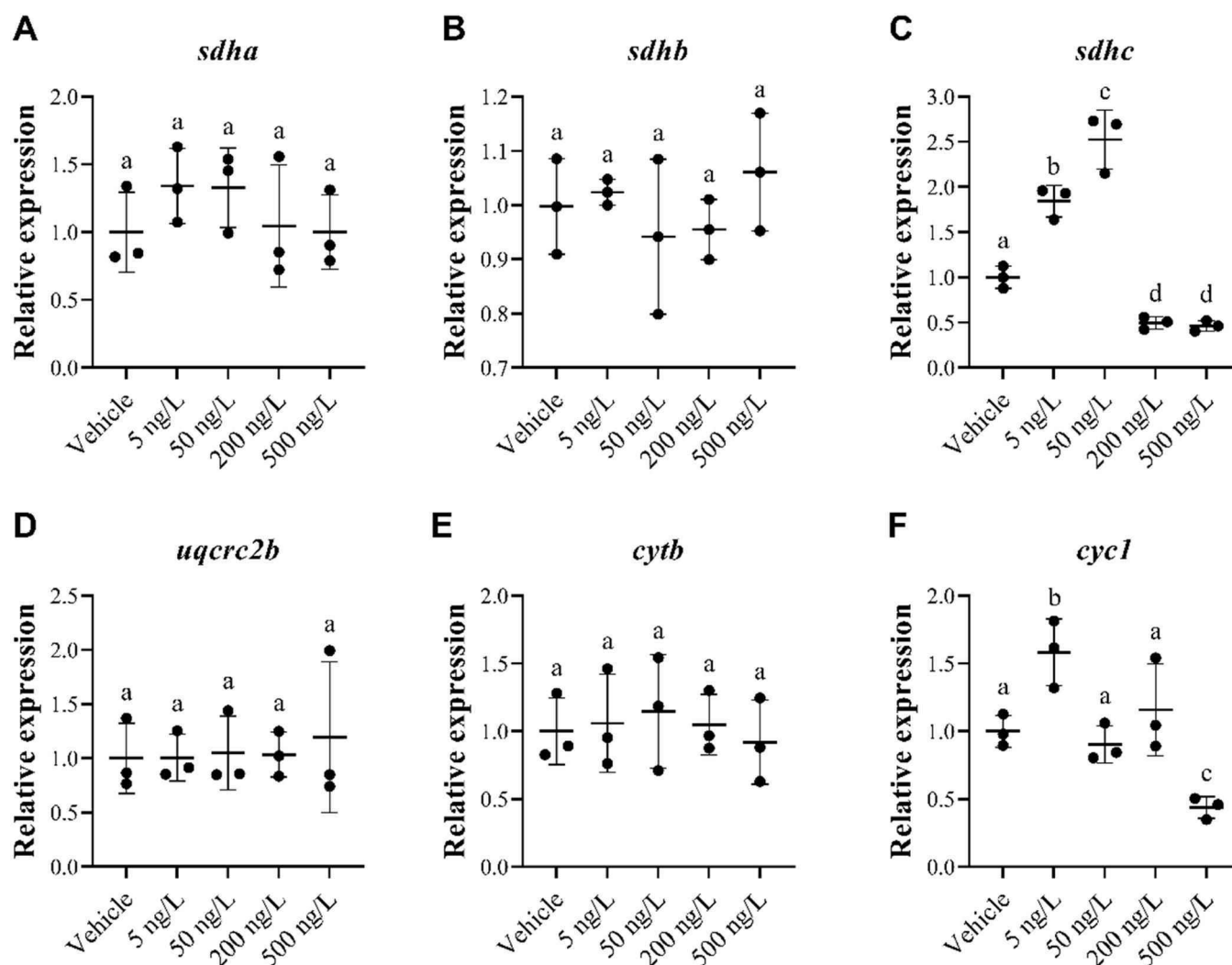


Fig. 4. The transcript levels of (A) *sdha*, (B) *sdhb*, (C) *sdhc*, (D) *uqcrc2b*, (E) *cytb*, (F) *cycl* genes related to mitochondrial electron transfer chain in zebrafish embryos following 96 h exposure to fenbuconazole. Results are showed as average value \pm standard deviation. N = 3/group. Different letters denote a significant difference between groups at $p < 0.05$.

fish ($F_{(5, 12)} = 24.35$; $p < 0.0001$; $F_{(5, 12)} = 242.5$; $p < 0.0001$; $F_{(5, 12)} = 242.5$; $p < 0.0001$) (Fig. 6A, B).

The total antioxidant capacity (T-AOC) was decreased in zebrafish by the fungicide at all tested concentrations, and the inhibition was concentration-dependent ($F_{(5, 12)} = 448$; $p < 0.0001$) (Fig. 6C). GSH levels were increased in fish exposed to 50, 200 and 500 ng/L fenbuconazole ($F_{(5, 12)} = 40.86$; $p < 0.0001$) (Fig. 6D). Additionally, SOD activity was induced by 5 and 50 ng/L fenbuconazole, while it was unaltered in fish exposed to 200 and 500 ng/L fenbuconazole ($F_{(5, 12)} = 76.52$; $p < 0.0001$) (Fig. 6E). CAT activity was reduced in fish by 50, 200 and 500 ng/L fenbuconazole ($F_{(5, 12)} = 319.2$; $p < 0.0001$) (Fig. 6F).

Moreover, the expression levels of *sod1* and *sod2* in fish were different from those of the control following fenbuconazole exposure ($F_{(4, 10)} = 38.68$; $p < 0.0001$; $F_{(5, 12)} = 42.6$; $p < 0.0001$; respectively) (Fig. 7A, B). The transcript levels of both genes were upregulated by lower concentrations of fenbuconazole; however, these transcripts were unaltered when fish were exposed to 200 and 500 ng/L fenbuconazole. The expression levels of *cat* were not changed by fenbuconazole at any concentration.

3.5. Changes in locomotor behaviors

We measured zebrafish swimming activity following fenbuconazole treatment. Fenbuconazole at 5 ng/L induced hyperactivity in zebrafish in the third period of the dark cycle ($F_{(4, 45)} = 22.96$; $p < 0.0001$). At higher concentrations in the dark period, hypoactivity was observed in zebrafish (at 500 ng/L) (the first period of the dark: $F_{(4, 45)} = 12.39$, $p < 0.0001$; the second period of the dark: $F_{(4, 45)} = 12.36$; $p < 0.0001$). No differences in zebrafish activity were noted in both light periods (Fig. 8).

4. Discussion

Based on literature, fenbuconazole appears to induce significantly higher toxicity compared to other triazoles, despite many being labelled as moderately toxic (Zubrod et al., 2019). The acute toxicity of fenbuconazole to zebrafish embryos was 1.68 mg/L (LC_{50}), much lower than that observed for several triazoles (~ 2 –40 mg/L) (Sanches et al., 2017; Sanches et al., 2018; Teng et al., 2019; Sun et al., 2020; Pang et al., 2020; Wang et al., 2020; Zhang et al., 2020). Additionally, morphological defects were noted with exposure to 200 ng/L, also much lower than that of other triazoles (~ 2000 – 7×10^6 ng/L) (Souders II et al., 2019; Teng et al., 2019; Tian et al., 2019; Cao et al., 2019; Cao et al., 2016; Wu

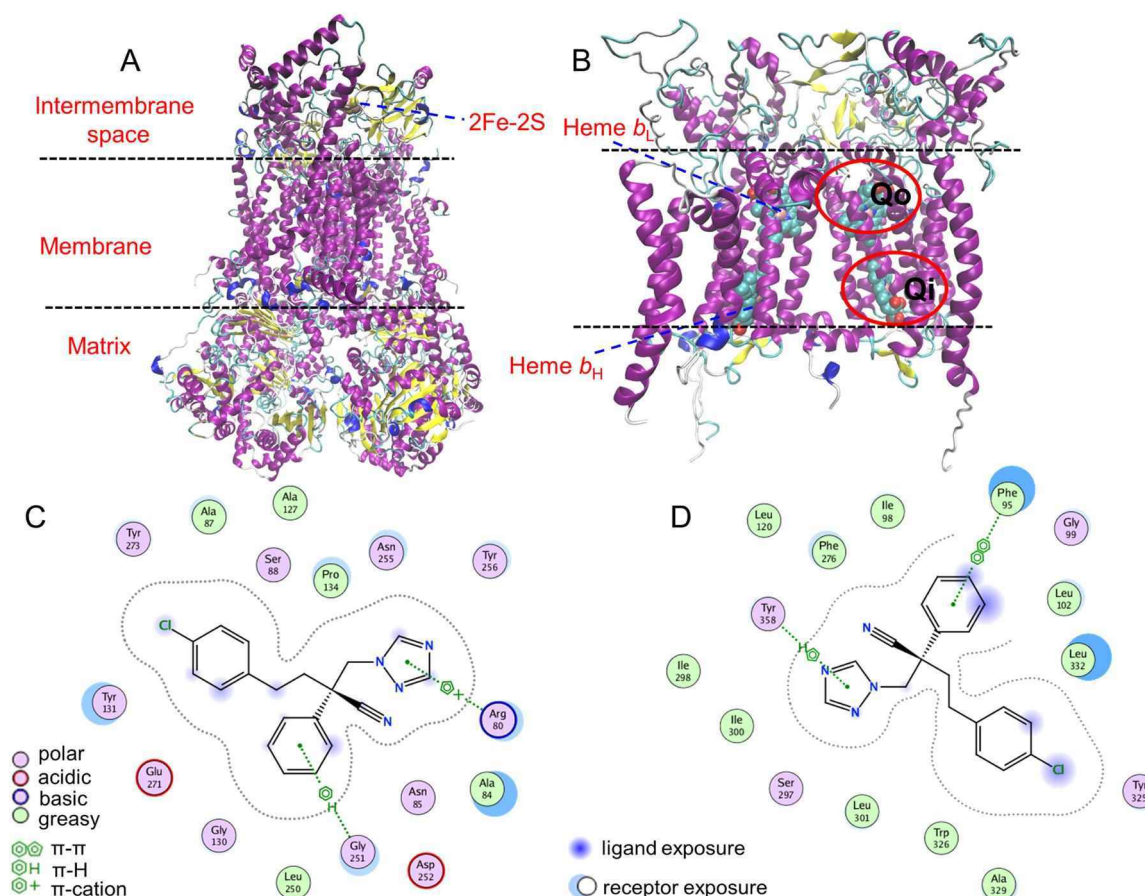


Fig. 5. Binding mode analysis of fenbuconazole at Qo and Qi sites. (A) 3D schematic representation of the full x-ray crystallographic molecular structure model from human mitochondrial respiratory complex III (PDB ID: 5XTE). The different regions (i.e., intermembrane and membrane space and mitochondrial matrix) were also labelled in the figure. (B) 3D schematic representation of the transmembrane region where Qo and Qi sites were located. (C) Binding mode of Fenbuconazole at Qo site. (D) Binding mode of Fenbuconazole at Qi site.

et al., 2018; Zhu et al., 2014; Liu et al., 2016; Mu et al., 2013; Mu et al., 2015; Mu et al., 2016; Toni et al., 2011; Aksakal and Ciltas, 2018; Zoupa and Machera, 2017). Thus, developmental toxicity of fenbuconazole to aquatic life should be ranked as relatively high in relation to other triazole fungicides.

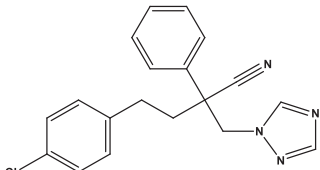
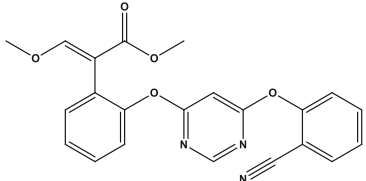
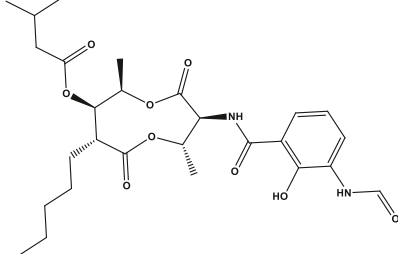
Triazoles have been reported to exert toxicity through multiple modes of action (MOAs) in fish, including oxidative stress, apoptotic signaling, endocrine disruption, metabolic disruption, and mitochondrial dysfunction (Jia et al., 2019; Mu et al., 2016; Souders II et al., 2019; Zhang et al., 2020; Zhao et al., 2020). The mitochondria are the major producers of ATP, and mitochondrial damage can adversely affect developmental trajectories. This is demonstrated in aquatic organisms during exposure to chemicals, as demonstrated by adverse effects on mitochondrial DNA of Atlantic killifish (*Fundulus heteroclitus*) exposed to benzo[a]pyrene, mitochondrial membrane potential of mussel (*Mytilus galloprovincialis*) eggs exposed to nano-sized Ag, and mitochondrial OCR of *Danio rerio* embryos exposed to pyraclostrobin (Jung et al., 2009; Auguste et al., 2018; Kumar et al., 2020). Nevertheless, few studies have been conducted with fenbuconazole in fish species, especially investigations at environmentally relevant levels. As such, we assessed the mitochondrial integrity and function in zebrafish embryos exposed to the fungicide.

The basal OCR and oligomycin-induced ATP were reduced by 500 ng/L fungicide, thus fenbuconazole negatively affects functional aspects of mitochondrial bioenergetics in fish. This supports the hypothesis that the fungicide impairs the mitochondrial respiratory chain, or electron transport chain (ETC) of zebrafish embryos, which can lead to compromised ATP production and oxidative respiration (Schwarz

et al., 2014; Wang et al., 2020). We next measured the ETC function, and according to the decreased MMP and activities of Complex II and III, we propose that reduced mitochondrial Complex II and III activities subsequently impair the electron transfer process and alters the transmembrane potential of the mitochondria (Zhao et al., 2019; Binukumar et al., 2010). These events could directly or indirectly lead to a decrease in OCR and ATP levels in developing fish embryos. Additionally, due to the same expression pattern of *cyc1* and *sdhc*, we propose that the ETC may be upregulated to cope with low exposure concentrations of the fungicide, but it may become impaired with higher exposure concentrations. However, mitochondrial biogenesis is regulated by nuclear-encoded proteins, transcription factors and co-activators, also maintains a degree of autonomy (Cagin and Enriquez, 2015). For mammalian, mitochondrial DNA encodes for only 13 proteins, all of which are subunits of OXPHOS complexes, but other OXPHOS protein subunits are encoded by nuclear DNA, such as SDH and CYC1. (Ghezzi and Zeviani, 2012; Lapuente-Brun et al., 2013). Thus, both *sdhc* and *cyc1*, each related to Complex II and III, were altered in expression, indicating that fenbuconazole exposure may affect mitochondrial ETC through disrupting genomic expression. Taken together, these data suggest that fenbuconazole at environmentally relevant concentrations affect the mitochondria and ETC in zebrafish at the transcriptional level, leading to functional impairments of mitochondria in early staged zebrafish. Molecular docking also offers novel insight into how fenbuconazole interacts with the human mitochondrial respiratory complex III at the atomic level.

According to an adverse outcome pathway proposed by a published literature (Souders II et al., 2018), chemicals adversely affect the

Table 1Comparison of binding energy between fenbuconazole and two inhibitors at Q_o site or Q_i site.

| | Molecule | Structure | ΔG_{bind} with Q _o site (kcal/mol) | ΔG_{bind} with Q _i site (kcal/mol) |
|---|---------------|---|--|--|
| 1 | Fenbuconazole |  | -7.6 | -9.0 |
| 2 | Azoxystrobin |  | -10.6 | — |
| 3 | Antimycin |  | — | -8.3 |

mitochondrial respiratory chain, or mitochondrial membrane proteins, may also reduce mitochondrial membrane potential, leading to reduced oxidative phosphorylation. This can further cause oxidative stress responses and even cell death. Here, fenbuconazole at 500 ng/L impaired the ETC, and this could prompt zebrafish to up-regulate antioxidant systems to cope with the overproduction of MDA and ROS. Increased MDA and ROS suggest that the fenbuconazole exposure caused oxidative damage in fish. In response to oxidative stress, nonenzymatic and enzymatic antioxidants in fish may function in the detoxification of the fungicide. GSH level, an indicator of the antioxidant level (Cheng et al., 2017), was increased in zebrafish, further confirming oxidative stress as a mechanism of toxicity. Additionally, two prominent antioxidant enzymes, SOD and CAT were assessed for activity and the transcript levels of *sod1* and *sod2* were also assessed in zebrafish following exposure to fenbuconazole (Khan et al., 2013; Ighodaro and Akinloye, 2018). The activity of SOD and the transcript levels of *sod1* and *sod2* were up-regulated by lower concentrations of fenbuconazole; however, the activity and those transcripts were unaltered by 200 and 500 ng/L fenbuconazole. Taken together, both transcript abundance and enzymatic activity levels were sensitive to fenbuconazole exposure. Other triazoles can alter activities of antioxidant enzymes and the transcripts of relevant genes in zebrafish. For example, after an 8-day incubation, difenconazole at 1.0 mg/L decreased the activity of SOD in zebrafish embryos and zebrafish livers (Mu et al., 2015). In addition, adult zebrafish suffered from a 14-day exposure of 0.5 mg/L prothioconazole, showed reduced antioxidant capacity, with marked decrease in the activities of SOD and GST (Zhang et al., 2020). Lastly, the expression levels of *sod1* were significantly downregulated to 83%, 46% and 55% of the control, respectively, in zebrafish subjected to 0.8, 1.6 and 2.4 mg/L penconazole (Aksakal and Ciltas, 2018). Taken together, there is compelling evidence that exposures to triazole fungicides affect the anti-oxidant system of zebrafish.

There have been several investigations providing convincing evidence that triazoles can affect fish activity. Studies indicate that the dose of fenbuconazole causing zebrafish hypoactivity was much lower than that of other triazoles causing zebrafish hypoactivity (~3–35 mg/L) (Table S3). Zebrafish locomotion in the second dark period seemed to

reflect response patterns noted with the transcript abundance of *sdhc*, *cyc1*, *sod1* and *sod2*, as well as the SOD activity. Locomotor activity in the first dark period also showed the same response pattern as the activities of antioxidant enzyme CAT, mitochondrial Complex II, III, total antioxidant capacity, and mitochondrial bioenergetics. There have been several studies also pointing to a correlation between mitochondrial dysfunction and locomotor behaviors in zebrafish (Li et al., 2019; Cao et al., 2019). Other classes of agrochemicals related to mitochondrial damage and loss of ATP production have previously been associated with hypoactive behavior in developmental *Danio rerio*, for instances, fluazinam and strobilurins (Wang et al., 2018b; Li et al., 2021). Taken together, we suggest that the compromised mitochondrial respiration and antioxidant system are related, and directly or indirectly contribute to hypoactivity of early developmental zebrafish.

Altered locomotor activity can also act as an indicator for neurotoxicity. Several pesticides exert neurotoxicity and alter fish behavioral responses, for examples, paraquat and pyraclostrobin (Wang et al., 2018a; Li et al., 2019). Mitochondrial dysfunction is often associated with neurodegeneration and impaired synaptic transmission, and is linked to complex neurodegenerative disorders, such as Alzheimer disease and Parkinson disease (Burté et al., 2015; Calkins et al., 2011; Liu et al., 2012), thus fenbuconazole may be a potential risk factor for neurotoxicity in aquatic organisms.

5. Conclusion

In summary, fenbuconazole showed higher toxicity to *Danio rerio* embryos than most previously studied triazoles. The present working model is that fenbuconazole can inhibit the activities of mitochondrial complex II and III and affect the transcript levels of *sdhc* and *cyc1*; thus, fenbuconazole impairs mitochondrial oxidative respiration of early developmental zebrafish at concentrations approaching 500 ng/L. Molecular docking showed that there is a stronger binding capacity of fenbuconazole at the Q_i site of Complex III compared to other mitochondrial toxicants such as antimycin. Impaired mitochondrial bioenergetics can result in oxidative stress, which can lead to changes in antioxidant capacity and the activities of CAT and SOD. In addition,

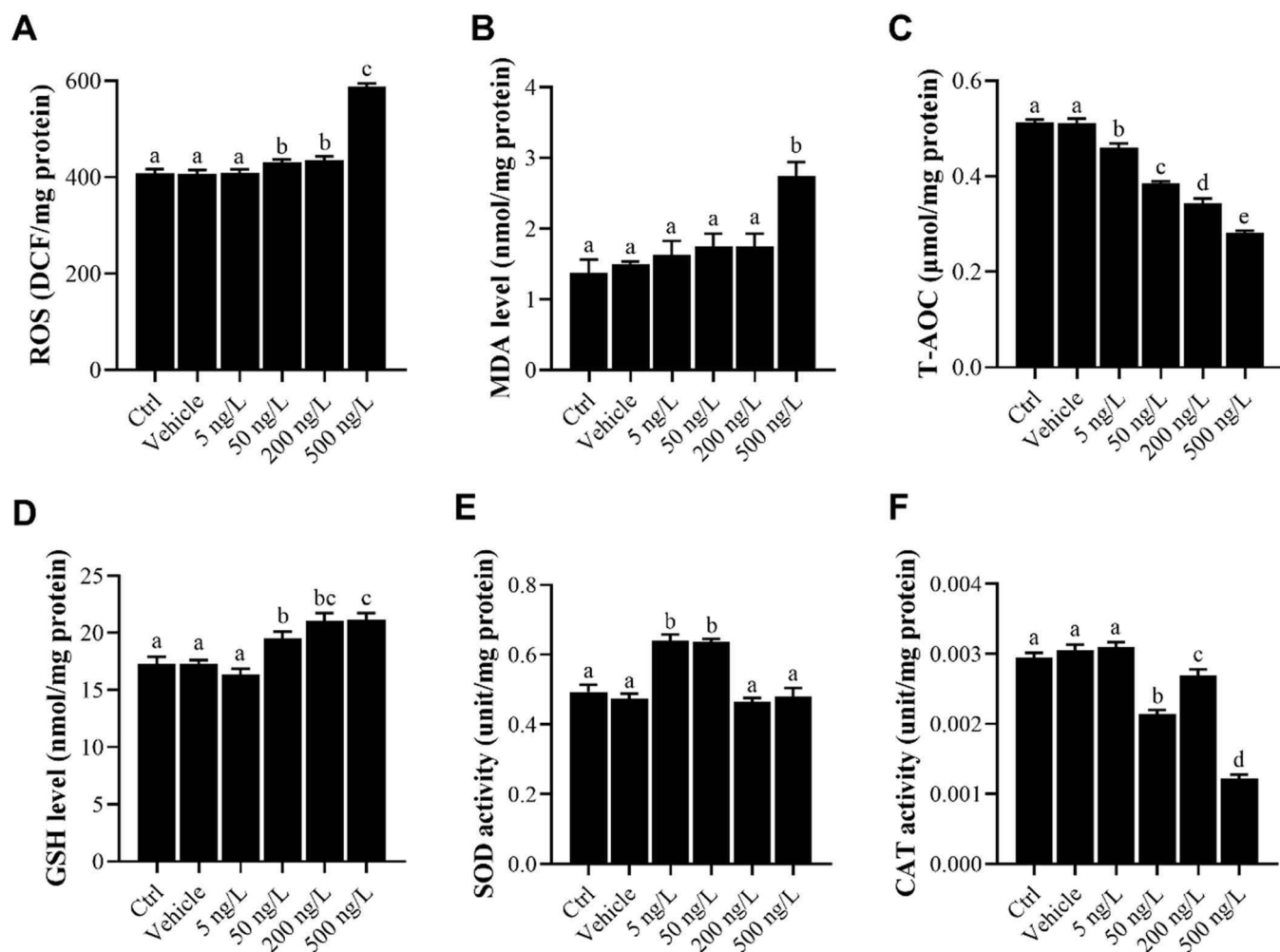


Fig. 6. Oxidative stress responses in zebrafish embryos following 96 h exposure to fenbucarb. (A) ROS level, (B) MDA level, (C) T-AOC, (D) GSH level, (E) Activity of SOD, (F) Activity of CAT. Results are shown as average value \pm standard deviation. N = 3/group. Different letters denote a significant difference between groups at $p < 0.05$.

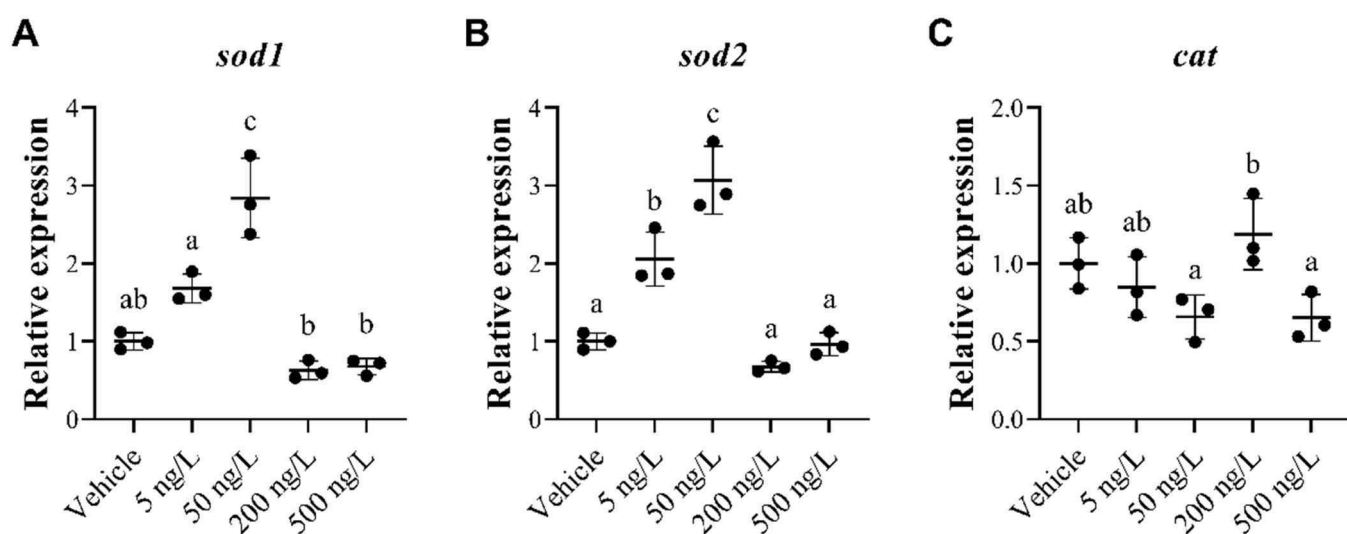


Fig. 7. The transcript levels of (A) *sod1*, (B) *sod2*, (C) *cat* genes associated with oxidative stress in zebrafish embryos following 96 h exposure to fenbucarb. Results are presented as average value \pm standard deviation. N = 3/group. Different letters denote a significant difference between groups at $p < 0.05$.

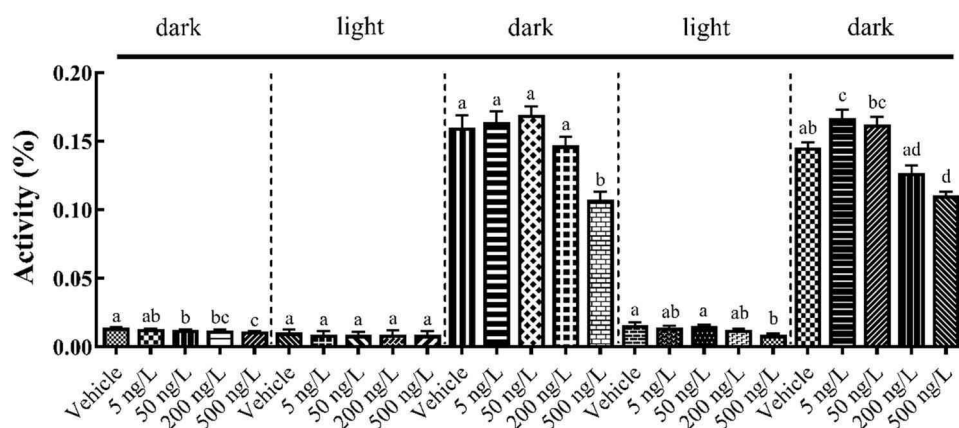


Fig. 8. Behavioral activity of zebrafish at 6 dpf following exposure to fenbuconazole. Results from each period were analyzed individually. $N = 15/\text{group}$. Different letters denote a significant difference between groups at $p < 0.05$.

exposure to 5 or 50 ng/L fenbuconazole may lead to a compensatory response for transcripts associated with the ETC and oxidative damage response, based upon the induction of *sdhc*, *cyc1*, *sod1* and *sod2*. Under this circumstance, the transcript responses may be sufficient to maintain oxidative phosphorylation and redox balance in zebrafish. However, for higher exposures (200 or 500 ng/L), the ETC and antioxidant system in zebrafish were either impaired or less active, as determined by enzyme activity and transcript data. This cascade may subsequently lead to adverse effects in developmental morphology and locomotor activities. Studies such as this are necessary for assessing the sublethal effect of fenbuconazole, as the fungicide is present in aquatic environments and can cause exposure risks to aquatic organisms.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The authors declare they have no potential or actual competing financial interests. This work is supported by the National Natural Science Foundation of China (21976026, 21777022), the Natural Science Foundation of Guangdong Province, China (2021A1515012319), and the Science and Technology Program of Guangzhou (202102020326).

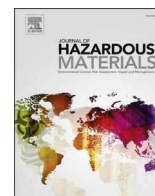
Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at [doi:10.1016/j.tox.2022.153137](https://doi.org/10.1016/j.tox.2022.153137).

References

- Aksakal, F.I., Ciltas, A., 2018. Developmental toxicity of penconazole in zebrafish (*Danio rerio*) embryos. *Chemosphere* 200, 8–15.
- Aladaghlo, Z., Fakhari, A.R., Alavioon, S.I., Dabiri, M., 2019. Ultrasound assisted dispersive solid phase extraction of triazole fungicides by using an N-heterocyclic carbene copper complex supported on ionic liquid-modified graphene oxide as a sorbent. *Microchim. Acta* 186, 1–8.
- Auguste, M., Ciacci, C., Balbi, T., Brunelli, A., Caratto, V., Marcomini, A., Cuppini, R., Canesi, L., 2018. Effects of nanosilver on *Mytilus galloprovincialis* hemocytes and early embryo development. *Aquat. Toxicol.* 203, 107–116.
- Bailey, D.C., Todt, C.E., Burchfield, S.L., Pressley, A.S., Denney, R.D., Snapp, I.B., Negga, R., Traynor, W.L., Fitsanakis, V.A., 2018. Chronic exposure to a glyphosate-containing pesticide leads to mitochondrial dysfunction and increased reactive oxygen species production in *Caenorhabditis elegans*. *Environ. Toxicol. Pharmacol.* 57, 46–52.
- Binukumar, B.K., Bal, A., Sunkaria, A., Gill, K.D., 2010. Mitochondrial energy metabolism impairment and liver dysfunction following chronic exposure to dichlorvos. *Toxicology* 270, 77–84.
- Burté, F., Carelli, V., Chinnery, P.F., Yu-Wai-Man, P., 2015. Disturbed mitochondrial dynamics and neurodegenerative disorders. *Nat. Rev. Neurol.* 11, 11–24.
- Cagin, U., Enriquez, J.A., 2015. The complex crosstalk between mitochondria and the nucleus: What goes in between? *Int. J. Biochem. Cell Biol.* 63, 10–15.
- Calkins, M.J., Manczak, M., Mao, P., Shirendeb, U., Reddy, P.H., 2011. Impaired mitochondrial biogenesis, defective axonal transport of mitochondria, abnormal mitochondrial dynamics and synaptic degeneration in a mouse model of Alzheimer's disease. *Hum. Mol. Genet.* 20, 4515–4529.
- Cao, C., Wang, Q., Jiao, F., Zhu, G., 2016. Impact of co-exposure with butachlor and triadimefon on thyroid endocrine system in larval zebrafish. *Exp. Toxicol. Pathol.* 68 (8), 463–469.
- Cao, F., Souders II, C.L., Li, P., Pang, S., Qiu, L., Martyniuk, C.J., 2019. Developmental toxicity of the triazole fungicide cyproconazole in embryo-larval stages of zebrafish (*Danio rerio*). *Environ. Sci. Pollut. Res.* 26, 4913–4923.
- Cheng, S.B., Liu, H.T., Chen, S.Y., Lin, P.T., Lai, C.Y., Huang, Y.C., 2017. Changes of oxidative stress, glutathione, and its dependent antioxidant enzyme activities in patients with hepatocellular carcinoma before and after tumor resection. *PLoS One* 12, e0170016.
- European Food Safety Authority, 2015. The 2013 European Union report on pesticide residues in food. *EFSA J.* 13, 4038.
- Ghezzi, D., Zeviani, M., 2012. Assembly factors of human mitochondrial respiratory chain complexes: Physiology and pathophysiology. *Adv. Exp. Med. Biol.* 748, 65–106.
- Hermesen, S.A., Pronk, T.E., van den Brandhof, E.J., van der Ven, L.T., Piersma, A.H., 2012. Triazole-induced gene expression changes in the zebrafish embryo. *Reprod. Toxicol.* 34, 216–224.
- Ighodaro, O.M., Akinloye, O.A., 2018. First line defence antioxidants-superoxide dismutase (SOD), catalase (CAT) and glutathione peroxidase (GPX): their fundamental role in the entire antioxidant defence grid. *Alex. J. Med.* 54, 287–293.
- Jia, M., Wang, Y., Wang, D., Teng, M., Yan, J., Yan, S., Meng, Z., Li, R., Zhou, Z., Zhu, W., 2019. The effects of hexaconazole and epoxiconazole enantiomers on metabolic profile following exposure to zebrafish (*Danio rerio*) as well as the histopathological changes. *Chemosphere* 226, 520–533.
- Jin, Y., Zhang, X., Shu, L., Chen, L., Sun, L., Qian, H., Liu, W., Fu, Z., 2010. Oxidative stress response and gene expression with atrazine exposure in adult female zebrafish (*Danio rerio*). *Chemosphere* 78, 846–852.
- Juberg, D.R., Mudra, D.R., Hazelton, G.A., Parkinson, A., 2006. The effect of fenbuconazole on cell proliferation and enzyme induction in the liver of female CD1 mice. *Toxicol. Appl. Pharmacol.* 214, 178–187.
- Jung, D., Cho, Y., Collins, L.B., Swenberg, J.A., Di Giulio, R.T., 2009. Effects of benzo[a]pyrene on mitochondrial and nuclear DNA damage in Atlantic killifish (*Fundulus heteroclitus*) from a creosote-contaminated and reference site. *Aquat. Toxicol.* 95, 44–51.
- Khan, M.A., Chen, H.C., Wan, X.X., Tania, M., Xu, A.H., Chen, F.Z., Zhang, D.Z., 2013. Regulatory effects of resveratrol on antioxidant enzymes: A mechanism of growth inhibition and apoptosis induction in cancer cells. *Mol. Cells* 35, 219–225.
- Kumar, N., Willis, A., Satbhai, K., Ramalingam, L., Schmitt, C., Moustaid-Moussa, N., Crago, J., 2020. Developmental toxicity in embryo-larval zebrafish (*Danio rerio*) exposed to strobilurin fungicides (azoxystrobin and pyraclostrobin). *Chemosphere* 241, 124980.
- Lai, B., Zhang, L., Dong, L.Y., Zhu, Y.H., Sun, F.Y., Zheng, P., 2005. Inhibition of Qi site of mitochondrial Complex III with antimycin A decreases persistent and transient sodium currents via reactive oxygen species and protein kinase C in rat hippocampal CA1 cells. *Exp. Neurol.* 194, 484–494.
- Lapiente-Brun, E., Moreno-Loshuertos, R., Acín-Pérez, R., Latorre-Pellicer, A., Colás, C., Balsa, E., et al., 2013. Supercomplex assembly determines electron flux in the mitochondrial electron transport chain. *Science* 340, 1567–1570.
- Li, H., Zhao, F., Cao, F., Teng, M., Yang, Y., Qiu, L., 2019. Mitochondrial dysfunction-based cardiotoxicity and neurotoxicity induced by pyraclostrobin in zebrafish larvae. *Environ. Pollut.* 251, 203–211.

- Li, X.Y., Qin, Y.J., Wang, Y., Huang, T., Zhao, Y.H., Wang, X.H., Martyniuk, C.J., Yan, B., 2021. Relative comparison of strobilurin fungicides at environmental levels: Focus on mitochondrial function and larval activity in early staged zebrafish (*Danio rerio*). *Toxicology* 452, 152706.
- Li, Y., Dong, F., Liu, X., Xu, J., Li, J., Kong, Z., Chen, X., Liang, X., Zheng, Y., 2012. Simultaneous enantioselective determination of triazole fungicides in soil and water by chiral liquid chromatography/tandem mass spectrometry. *J. Chromatogr. A* 1224, 51–60.
- Liu, N., Dong, F., Xu, J., Liu, X., Zheng, Y., 2016. Chiral bioaccumulation behavior of tebuconazole in the zebrafish (*Danio rerio*). *Ecotox. Environ. Saf.* 126, 78–84.
- Liu, S., Sawada, T., Lee, S., Yu, W., Silverio, G., Alapatt, P., Millan, I., Shen, A., Saxton, W., Kanao, T., Takahashi, R., Hattori, N., Imai, Y., Lu, B., 2012. Parkinson's disease-associated kinase PINK1 regulates Miro protein level and axonal transport of mitochondria. *PLoS Genet.* 8, e1002537.
- Luo, L., Wang, F., Zhang, Y., Zeng, M., Zhong, C., Xiao, F., 2017. *In vitro* cytotoxicity assessment of roundup (glyphosate) in L-02 hepatocytes. *J. Environ. Sci. Health B* 52, 410–417.
- Mailloux, R.J., McBride, S.L., Harper, M.E., 2013. Unearthing the secrets of mitochondrial ROS and glutathione in bioenergetics. *Trends Biochem. Sci.* 38, 592–602.
- Mondal, S.N., Bhatia, A., Shilts, T., Timmer, L.W., 2005. Baseline sensitivities of fungal pathogens of fruit and foliage of citrus to azoxystrobin, pyraclostrobin, and fenbuconazole. *Plant Dis.* 89, 1186–1194.
- Mu, X., Pang, S., Sun, X., Gao, J., Chen, J., Chen, X., Li, X., Wang, C., 2013. Evaluation of acute and developmental effects of difenoconazole via multiple stage zebrafish assays. *Environ. Pollut.* 175, 147–157.
- Mu, X., Chai, T., Wang, K., Zhang, J., Zhu, L., Li, X., Wang, C., 2015. Occurrence and origin of sensitivity toward difenoconazole in zebrafish (*Danio rerio*) during different life stages. *Aquat. Toxicol.* 160, 57–68.
- Mu, X., Chai, T., Wang, K., Zhu, L., Huang, Y., Shen, G., Li, Y., Li, X., Wang, C., 2016. The developmental effect of difenoconazole on zebrafish embryos: A mechanism research. *Environ. Pollut.* 212, 18–26.
- Nickel, A., Kohlhaas, M., Maack, C., 2014. Mitochondrial reactive oxygen species production and elimination. *J. Mol. Cell Cardiol.* 73, 26–33.
- Oleg Trott, A.J.O., 2010. AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J. Comput. Chem.* 31 (2), 455–461.
- Pang, S., Guo, M., Zhang, X., Yu, L., Zhang, Z., Huang, L., Gao, J., Li, X., 2020. Myclobutanil developmental toxicity, bioconcentration and sex specific response in zebrafish (*Danio rerio*). *Chemosphere* 242, 125209.
- Park, H., Lee, J.Y., Lim, W., Song, G., 2021. Assessment of the in vivo genotoxicity of pendimethalin via mitochondrial bioenergetics and transcriptional profiles during embryogenesis in zebrafish: implication of electron transport chain activity and developmental defects. *J. Hazard. Mater.* 411, 125153.
- Pesticide Properties Data Base, 1992.** <http://sitem.herts.ac.uk/aeru/ppdb/en/Reports/293.htm>.
- Pfaffl, M.W., 2001. A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res.* 29, No. e40.
- Rigoulet, M., Yoboue, E.D., Devin, A., 2011. Mitochondrial ROS generation and its regulation: mechanisms involved in H₂O₂ signaling. *Antioxid. Redox. Signal* 14, 459–468.
- Rodrigues, E.T., Lopes, I., Pardal, M.Â., 2013. Occurrence, fate and effects of azoxystrobin in aquatic ecosystems: a review. *Environ. Int.* 53, 18–28.
- Sanchez, A.L.M., Vieira, B.H., Reghini, M.V., Moreira, R.A., Freitas, E.C., Espíndola, E.L., Daam, M.A., 2017. Single and mixture toxicity of abamectin and difenoconazole to adult zebrafish (*Danio rerio*). *Chemosphere* 188, 582–587.
- Sanchez, A.L.M., Daam, M.A., Freitas, E.C., Godoy, A.A., Meireles, G., Almeida, A.R., Domingues, I., Espíndola, E.L.G., 2018. Lethal and sublethal toxicity of abamectin and difenoconazole (individually and in mixture) to early life stages of zebrafish. *Chemosphere* 210, 531–538.
- Schwarz, K., Siddiqi, N., Singh, S., Neil, C.J., Dawson, D.K., Frenneaux, M.P., 2014. The breathing heart-mitochondrial respiratory chain dysfunction in cardiac disease. *Int. J. Cardiol.* 171, 134–143.
- Shen, J., Liu, P., Sun, Y., Xu, X., Guo, L., Rao, Q., Chen, M., Liu, X., 2021. Embryonic exposure to prothioconazole induces oxidative stress and apoptosis in zebrafish (*Danio rerio*) early life stage. *Sci. Total Environ.* 756, 143859.
- Souders II, C.L., Xavier, P., Perez-Rodriguez, V., Ector, N., Zhang, J.L., Martyniuk, C.J., 2019S. Sub-lethal effects of the triazole fungicide propiconazole on zebrafish (*Danio rerio*) development, oxidative respiration, and larval locomotor activity. *Neurotoxicol. Teratol.* 74, 106809.
- Souders II, C.L., Liang, X.F., Wang, X.H., Ector, N., Zhao, Y.H., Maryniuk, C.J., 2018. High-throughput assessment of oxidative respiration in fish embryos: Advancing adverse outcome pathways for mitochondrial dysfunction. *Aquat. Toxicol.* 199, 162–173.
- Sun, Y., Cao, Y., Tong, L., Tao, F., Wang, X., Wu, H., Wang, M., 2020. Exposure to prothioconazole induces developmental toxicity and cardiovascular effects on zebrafish embryo. *Chemosphere* 251, 126418.
- Teng, M., Zhu, W., Wang, D., Qi, S., Wang, Y., Yan, J., Dong, K., Zheng, M., Wang, C., 2018a. Metabolomics and transcriptomics reveal the toxicity of difenoconazole to the early life stages of zebrafish (*Danio rerio*). *Aquat. Toxicol.* 194, 112–120.
- Teng, M., Qi, S., Zhu, W., Wang, Y., Wang, D., Dong, K., Wang, C., 2018b. Effects of the bioconcentration and parental transfer of environmentally relevant concentrations of difenoconazole on endocrine disruption in zebrafish (*Danio rerio*). *Environ. Pollut.* 233, 208–217.
- Teng, M., Zhao, F., Zhou, Y., Yan, S., Tian, S., Yan, J., Meng, Z., Bo, S., Wang, C., 2019. Effect of propiconazole on the lipid metabolism of zebrafish embryos (*Danio rerio*). *J. Agric. Food Chem.* 67, 4623–4631.
- Tian, S., Teng, M., Meng, Z., Yan, S., Jia, M., Li, R., Liu, L., Yan, J., Zhou, Z., Zhu, W., 2019. Toxicity effects in zebrafish embryos (*Danio rerio*) induced by prothioconazole. *Environ. Pollut.* 255, 113269.
- Toni, C., Ferreira, D., Kreutz, L.C., Loro, V.L., Barcellos, L.J.G., 2011. Assessment of oxidative stress and metabolic changes in common carp (*Cyprinus carpio*) acutely exposed to different concentrations of the fungicide tebuconazole. *Chemosphere* 83, 579–584.
- Wang, X.H., Souders II, C.L., Zhao, Y.H., Martyniuk, C.J., 2018a. Paraquat affects mitochondrial bioenergetics, dopamine system expression, and locomotor activity in zebrafish (*Danio rerio*). *Chemosphere* 191, 106–117.
- Wang, X.H., Zheng, S.S., Huang, T., Su, L.M., Zhao, Y.H., Souders II, C.L., Martyniuk, C. J., 2018b. Fluazinam impairs oxidative phosphorylation and induces hyper/hypo-activity in a dose specific manner in zebrafish larvae. *Chemosphere* 210, 633–644.
- Wang, Y., Zhu, W., Wang, D., Teng, M., Yan, J., Miao, J., Zhou, Z., 2017. 1H NMR-based metabolomics analysis of adult zebrafish (*Danio rerio*) after exposure to diniconazole as well as its bioaccumulation behavior. *Chemosphere* 168, 1571–1577.
- Wang, Y., Xu, C., Wang, D., Weng, H., Yang, G., Guo, D., Yu, R., Wang, X., Wang, Q., 2020. Combined toxic effects of fludioxonil and triadimefon on embryonic development of zebrafish (*Danio rerio*). *Environ. Pollut.* 260, 114105.
- WBISS Consulting Co, Ltd, 2016. China triazole Fungicides Market Report edition.** <http://www.reportlinker.com/p03762688-summary/China-Triazole-Fungicides-Market-Report-Edition.html>.
- Weng, Y., Huang, Z., Wu, A., Yu, Q., Lu, H., Lou, Z., Lu, L., Bao, Z., Jin, Y., 2021. Embryonic toxicity of epoxiconazole exposure to the early life stage of zebrafish. *Sci. Total Environ.* 778, 146407.
- Wu, Y., Yang, Q., Chen, M., Zhang, Y., Zuo, Z., Wang, C., 2018. Fenbuconazole exposure impacts the development of zebrafish embryos. *Ecotoxicol. Environ. Saf.* 158, 293–299.
- Xu, X.M., Gao, L.Q., Yang, J.R., 2010. Are insensitivities of *Venturia inaequalis* to myclobutanil and fenbuconazole correlated? *Crop. Prot.* 29, 183–189.
- Yang, C., Lim, W., Song, G., 2020. Mediation of oxidative stress toxicity induced by pyrethroid pesticides in fish. *Comp. Biochem. Physiol. C Toxicol. Pharmacol.* 234, 108758.
- Yang, Y., Qi, S., Wang, D., Wang, K., Zhu, L., Chai, T., Wang, C., 2016. Toxic effects of thifluzamide on zebrafish (*Danio rerio*). *J. Hazard. Mater.* 307, 127–136.
- Zhang, Z.X., Zhang, J., Zhao, X.J., Gao, B.B., He, Z.Z., Li, L.S., Shi, H.Y., Wang, M.H., 2020. Stereoselective uptake and metabolism of prothioconazole caused oxidative stress in zebrafish (*Danio rerio*). *J. Hazard. Mater.* 396, 122756.
- Zhao, F., Cao, F., Li, H., Teng, M., Liang, Y., Qiu, L., 2020. The effects of a short-term exposure to propiconazole in zebrafish (*Danio rerio*) embryos. *Environ. Sci. Pollut. Res.* 27, 38212–38220.
- Zhao, R.Z., Jiang, S., Zhang, L., Yu, Z.B., 2019. Mitochondrial electron transport chain, ROS generation and uncoupling. *Int. J. Mol. Med.* 44, 3–15.
- Zhu, B., Liu, L., Gong, Y.X., Ling, F., Wang, G.X., 2014. Triazole-induced toxicity in developing rare minnow (*Gobiocypris rarus*) embryos. *Environ. Sci. Pollut. Res.* 21 (23), 13625–13635.
- Zoupa, M., Machera, K., 2017. Zebrafish as an alternative vertebrate model for investigating developmental toxicity-the triadimefon example. *Int. J. Mol. Sci.* 18 (4), 817.
- Zubrod, J.P., Bundschuh, M., Arts, G., Brühl, C.A., Imfeld, G., Knabel, A., Payraudeau, S., Rasmussen, J.J., Rohr, J., Scharmueller, A., Smalling, K., Stehle, S., Schulz, R., Schäfer, R.B., 2019. Fungicides: An overlooked pesticide class? *Environ. Sci. Technol.* 53, 3347–3365.



Research Paper

Unraveling the joint toxicity of transition-metal dichalcogenides and per- and polyfluoroalkyl substances in aqueous mediums by experimentation, machine learning and molecular dynamics

Guohong Liu^{a,b}, Xiliang Yan^{b,*}, Chengjun Li^{a,b,**}, Song Hu^{a,c}, Jiachen Yan^a, Bing Yan^{a,c,**}

^a Institute of Environmental Research at Greater Bay Area, Key Laboratory for Water Quality and Conservation of the Pearl River Delta, Ministry of Education, Guangzhou University, Guangzhou 510006, China

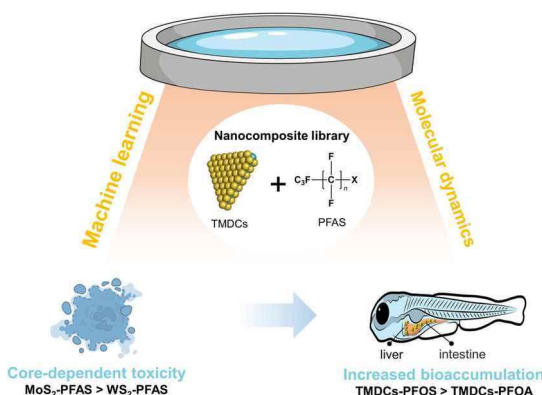
^b School of Agriculture and Biological Sciences, Qiannan Normal University for Nationalities, Duiyun 558000, China

^c School of Environmental Science and Engineering, Shandong University, Qingdao 266237, China

HIGHLIGHTS

- Nanocomposite library and high-throughput screening assays provided high-quality data for machine learning.
- Some critical physicochemical properties (e.g., size and surface chemistry) of TMDCs-PFAS were responsible for the cytotoxicity.
- The interaction of PFAS with cell membrane was an important driving force for the bioaccumulation of TMDCs-PFAS.
- The increased oxidative stress led to apparent histopathological alterations in zebrafish liver and intestine.

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: Youn-Joo An

Keywords:

Nanocomposite library
Cytotoxicity
Zebrafish
Machine learning
Molecular dynamics

ABSTRACT

The environmental fate of transition-metal dichalcogenides (TMDCs) may be further complicated by interacting with existing pollutants, especially per- and polyfluoroalkyl substances (PFAS). However, due to their sheer volume, it is impossible to explore all possible interactions by simply utilizing experimental methods. Herein, we used two model TMDC nanosheets, molybdenum disulfide (MoS₂) and tungsten disulfide (WS₂), and seven PFAS to explore their interactions and subsequent impacts on model cell lines and zebrafish. Utilizing experimental methods and machine learning approaches, we showed that TMDCs-PFAS interactions can pose unique challenges due to their interaction-specific toxicity niches towards cell lines. Further in vivo experiments, together with molecular dynamics simulation, suggested that TMDCs-PFAS interactions in aqueous environments significantly increased their bioaccumulation in zebrafish towards different target organs, mostly due to the differences in loading PFAS. Such enhanced bioaccumulation increased the oxidative stress in zebrafish liver and

* Corresponding author.

** Corresponding authors at: Institute of Environmental Research at Greater Bay Area, Key Laboratory for Water Quality and Conservation of the Pearl River Delta, Ministry of Education, Guangzhou University, Guangzhou 510006, China.

E-mail addresses: yanxiliang1991@gzhu.edu.cn (X. Yan), cli@gzhu.edu.cn (C. Li), drbingyan@gzhu.edu.cn (B. Yan).

<https://doi.org/10.1016/j.jhazmat.2022.130303>

Received 5 September 2022; Received in revised form 18 October 2022; Accepted 30 October 2022

Available online 1 November 2022

0304-3894/© 2022 Elsevier B.V. All rights reserved.

intestine, as demonstrated by the increased reactive oxygen species (ROS) level and other enzyme activities, which eventually led to obvious histopathological alterations in the liver and intestine. Our study highlights the importance of exploring interactions between emerging and existing contaminants with state-of-art techniques in aqueous environments and its significance in safeguarding aquatic environment health.

1. Introduction

Transition-metal dichalcogenides (TMDCs), e.g., molybdenum disulfide (MoS_2) and tungsten disulfide (WS_2), are a class of (2D) materials (Guiney et al., 2018) that have been extensively applied in various industrial sectors (Liu et al., 2021). Depending on their application scenarios and different physicochemical properties, TMDCs may be released into the atmosphere, soil and aquatic environments during their manufacture, transportation, and use (Dale et al., 2015). Approaching the end of their life cycle, most TMDCs such as MoS_2 and WS_2 might end up in landfills (Keller et al., 2013) and will eventually break down and leach out into surrounding soils and waters (Dale et al., 2015).

Understanding the environmental fate and relevant toxicity of TMDC nanomaterials, therefore, is crucial and a prerequisite for their safe applications. Previous studies have revealed that MoS_2 and WS_2 can induce toxicity to various organisms, including but not limited to microorganisms (Wu et al., 2019), phytoplankton (Zou et al., 2021), fish (Yu et al., 2018) and mammals (Scalisi et al., 2020). Meanwhile, their toxicity can also be greatly affected by environmental factors. For example, UV irradiation can significantly affect the toxicity of MoS_2 and WS_2 in aquatic environments towards *Escherichia coli* (Shang et al., 2017). Dissolved oxygen and visible light irradiation have also been reported to have similar impacts on the toxicity of MoS_2 in aquatic environments (Zou et al., 2019). Furthermore, due to their intrinsic properties (e.g., high surface-area-to-volume ratio), TMDCs can interact with other environmental pollutants, either natural or engineered (Xu et al., 2022). Zou et al. demonstrated that naturally secreted extracellular polymeric substances of algae can bind onto the surface and alter the toxicity profile of MoS_2 (Zou et al., 2021). Likewise, Yuan et al. (2020) reported a synergistic interaction between TMDCs and organic pollutants (i.e., triclosan and tris(1,3-dichloro-2-propyl)phosphate), where WS_2 nanosheets, even at noncytotoxic concentrations, could enhance the cytotoxicity of organic pollutants.

It is expected that released TMDCs can interact with existing pollutants in aquatic environments, especially persistent organic pollutants (POPs), such as per- and polyfluoroalkyl substances (PFAS) (Gagliano et al., 2020; Mahoney et al., 2022). PFAS had been used worldwide for their hydrophobic and oleophobic properties in industrial and commercial products and their pollution has been identified in almost every ecosystem on Earth (Ghisi et al., 2019; Sonne et al., 2022). Such widespread pollution of PFAS poses severe threats to both the environment and human health due to their persistent and bioaccumulative nature and toxicity (Sunderland et al., 2019). To further complicate this problem, existing PFAS may interact with emerging environmental pollutants such as TMDCs to cause joint toxic effects. However, due to their sheer volume, it is impossible to explore all combinations of environmental pollutants by experimental methods alone. Therefore, reliable quantitative structure-activity relationships (QSAR) or other predictive modeling approaches are urgently needed to address this issue. Recent advances in machine learning techniques hold great promise for handling high dimensional descriptors calculated from chemical structures, hence, resulting in enhanced predictive performance (Yan et al., 2020). Furthermore, molecular dynamics simulations are able to provide atomic-level structural insights into key molecular interactions, and thus facilitate our understanding of the mechanisms of joint toxicities (Gu et al., 2018).

In this study, we experimentally explored the interactions between two typical TMDCs (i.e., MoS_2 and WS_2 nanosheets) and seven PFAS and their environmental fate and toxicities in vitro and in vivo towards cell

models and zebrafish in aqueous mediums at neutral pH. MoS_2 and WS_2 are two of the most popular TMDCs with extensive industrial applications and thus were chosen as representatives of TMDCs in this study (Yuan et al., 2020; Zhang et al., 2020). Machine learning and molecular dynamics simulation were then used to explore the most important factors that affect their interactions, ingestion, bioaccumulation and subsequent toxicity. By combining experimental and computational results, we revealed the mode of interactions between PFAS and TMDCs (i.e., PFAS loading onto the surface of MoS_2 or WS_2) and their synergistic toxicities towards cells as well as zebrafish, mostly due to intrinsic properties of loading components. Further in vivo experiments, together with molecular dynamics simulation, demonstrated that their interactions altered their ingestion and accumulation behavior, and their toxicity towards aquatic organisms.

2. Methods

2.1. Synthesis and characterization

MoS_2 and WS_2 were custom synthesized by XFNANO (Nanjing, Jiangsu, China) according to Zeng et al. (2012). Since it is impossible to explore all PFAS due to the sheer number, seven representative PFAS were selected as the model PFAS and purchased from Sigma-Aldrich (St Louis, MO, USA), including trifluoroacetic acid (CF_3COOH), pentafluoropropionic acid ($\text{C}_2\text{F}_5\text{COOH}$), perfluoropentanoic acid ($\text{C}_4\text{F}_9\text{COOH}$), potassium nonafluoro-1-butanedisulfonate ($\text{C}_4\text{F}_9\text{SO}_3\text{K}$), tridecafluorohexane-1-sulfonic acid potassium salt ($\text{C}_6\text{F}_{13}\text{SO}_3\text{K}$), potassium perfluorooctanesulfonate ($\text{C}_8\text{F}_{17}\text{SO}_3\text{K}$, hereafter “PFOS”), and ammonium perfluorooctanoate ($\text{C}_7\text{F}_{15}\text{COONH}_4$, hereafter “PFOA”). These PFAS with different carbon chain lengths and thiol groups represent a relatively broad diversity of chemical structures, which is beneficial to explore the relationships between the toxicities of PFAS and their structures, and subsequent generalization and prediction of machine learning.

To prepare TMDCs-PFAS nanocomposites, MoS_2 and WS_2 nanosheets were mixed with the seven representative PFAS in aqueous solutions, respectively, at a fixed ratio set according to their environmental concentrations. Although the composition of PFAS in environmental water samples varies dramatically, their median concentrations have been reported to be ca. $30 \mu\text{g/L}$ or ca. 0.07 mmol/L equivalent, especially for PFOS and PFOA (Eriksen et al., 2011). Therefore, a universal molar concentration of 0.07 mmol/L was used for PFAS when preparing nanocomposites. As for nanosheets, there is no reference for their environmental water concentrations as of the writing of this paper. Therefore, we used an arbitrary concentration of 1 mg/L for both MoS_2 and WS_2 . Corresponding stock solutions were added to 200 mL flasks and shaken for 24 h at room temperature with a MaxQ 4450 orbital shaker (Thermo Fisher, MA, USA).

A total of 14 nanocomposites from all combinations between seven PFAS and MoS_2/WS_2 nanosheets were synthesized. Compared to traditional approaches of testing one nanomaterial at a time, such a nanocomposite library established a knowledge domain that covered diverse physicochemical properties of nanomaterials. The nanomaterial library strategy has been widely used in previous studies for nano-drug screening and nanosafety assessment (Yamankurt et al., 2019). This would bring us a closer, comprehensive understanding of nanocomposite-induced toxicity.

The morphology of MoS_2 and WS_2 , including size and thickness, was characterized with atomic force microscopy (AFM). Zeta potential and

hydrodynamic diameters of all nanomaterials were determined by dynamic light scattering in ultrapure water (18.2 MΩ) or the cell culture medium with 10% fetal bovine serum (FBS). Details can be found in **Method S1**.

2.2. In vitro toxicity on cell lines

Given the gastrointestinal tract is the first part of the human body that encounters environmental pollutants upon oral exposure, the gastric epithelium cell (GES-1) and colonic mucosa cell (FHC) lines were used to explore the in vitro toxicity of MoS₂ and WS₂ nanosheets, PFAS, and their nanocomposites. Specifically, GES-1 and FHC lines purchased from ATCC (Manassas, VA, USA), were grown in RPMI-1640 medium supplemented with 10% fetal bovine serum (Clarke bioscience, Webster, USA), 100 µg/mL penicillin, and 100 U/mL streptomycin (Wang et al., 2018). All cells were grown in a Thermo 3111 incubator (Thermo Fisher, MA, USA) at 37 °C (95% humidity and 5% CO₂). Then, GES-1 or FHC cell suspensions containing 5000 cells in 100 µL were seeded in 96-well plates. After incubation for 24 h, cells were incubated with a series of concentrations (10, 25, 50, 100, 200, 400, 600 and 800 µg/mL) of MoS₂/WS₂ and their corresponding nanocomposites for another 48 h. The CCK-8 kit (Dojindo, Japan) was used to detect the viability of GES-1 or FHC cells according to the manufacturer's instructions.

2.3. In vivo toxicity on zebrafish

Based on in vitro toxicity results, the selected in vivo toxicity of nanomaterials, i.e., MoS₂ and its PFOS and PFAS nanocomposites (MoS₂-PFOS and WS₂-PFOA; see justification in Results and discussion section about in vitro toxicity) was evaluated using zebrafish (*Danio rerio*) as the model species due to the advantages compared to other aquatic models (Pichler et al., 2003). In recent years, zebrafish has proven to be an excellent in vivo model because of its unique features, including high fecundity, fast and well-characterized development, and easy gene manipulation. More importantly, zebrafish shares a high degree of genetic similarity with humans, acting as an ideal model organism to study human disease. All animal procedures were carried out following the Guidelines for Care and Use of Laboratory Animals of Guangzhou University and approved by the Animal Ethics Committee of Guangzhou University.

2.3.1. Zebrafish maintenance and exposure

Two-month-old zebrafish (AB line) were maintained in freshwater tanks at 28 °C and pH 7.0, with a 12-h photoperiod. The fish were fed artificial diets with a 55:15:1.5:12:12:1.5 ratio of protein:fat:fiber:moisture:ash:phosphorus (Zeigler Bros., Inc., Gardners, USA) once a day. After two-week acclimation, zebrafish were transferred into fish tanks (30 cm × 20 cm × 18 cm) prefilled with 3 L exposure solutions. 11 exposure solutions containing either MoS₂ (0.1, 1, and 10 mg/L), PFOS/PFOA (0.03 mg/L), or their corresponding nanocomposites (see Table S1) were prepared 24 h before experiments to ensure adsorption equilibrium. Each exposure treatment had three replicates (i.e., tanks) while zebrafish kept in three fish tanks with freshwater instead of exposure solutions were set as the control group. Therefore, 12 exposure treatments (including the control) were involved in this study, and each had 45 zebrafish (15 zebrafish × 3 tanks), adding up to a total of 540 zebrafish. All tanks were kept in the above conditions for two weeks and the exposure solution in each tank was replaced daily to avoid potential aggregation of investigated nanomaterials.

2.3.2. Post-exposure quantification and localization

After two-week exposure, zebrafish were harvested and their gills, brains, muscles, intestines and livers were collected to quantify the bioaccumulation of nanocomposites in different organs and tissues by measuring Mo content with inductively coupled plasma mass spectrometry (ICP-MS). The localization of ingested nanomaterials was

observed by transmission electron microscopy (TEM) (see details in **Method S2**).

2.3.3. Histological and oxidative stress examination

Zebrafish intestines and livers collected after two-week exposure were also used for histological inspection. Sample tissues were firstly fixed in 4% (w/v) paraformaldehyde solution for 24 h, dehydrated in ethanol and then embedded in paraffin. Embedded samples were sectioned and stained with hematoxylin and eosin. Finally, obtained sections were observed to identify any potential histological changes using a Nikon ECLIPSE Ts2 light microscope (Nikon Corporation, Tokyo, Japan).

The level of reactive oxygen species (ROS), superoxide dismutase (SOD), catalase (CAT) and malondialdehyde (MDA) were tested on collected tissues using corresponding Beyotime Assay Kits (Beyotime, Shanghai, China) following the manufacturer's instructions. ROS, SOD, CAT, and MDA were tested using the ferric-reducing ability of plasma (FRAP) method (Benzie and Strain, 1996), the WST-8 method, the catalase test (Reiner, 2010), and the TBA method (Janero, 1990).

2.4. Machine learning and molecular dynamics simulation

To further explain the interactions between selected TMDCs and PFAS, and their potential health risks, machine learning approaches and molecular dynamics simulation were employed. Before the application of machine learning, a total of 211 descriptors including five experimental properties and 206 theoretical features were applied to describe the structural diversity of 16 synthesized nanomaterials. We then used a random forest (RF) regressor to construct the QSAR model. Scikit-learn (Version 0.19.2) was used to construct the RF regression model. The model's robustness was verified by the five-fold cross-validation. The determination coefficient (R^2) and root mean square error (RMSE) were used to evaluate the predictive ability of the generated model. The feature importance function of the RF algorithm was used to rank the effect of different input variables on the model (see details in **Method S3**). The main parameters and their ranges used for machine learning were listed in Table S2.

Molecular dynamics simulation was performed using CHARMM36 force field in Gromacs (Version 2020.4). The details of parameter settings can be found in **Method S4**. Water molecules simulated with the TIP3P model were used as solvents. The initial simulation box was set as 6 nm × 6 nm × 12 nm, which was large enough to avoid the interaction of compounds with their periodic images. Each simulation system included one MoS₂ nanosheet, one lipid bilayer, 40 PFOS/PFOA molecules, and 10,830 water molecules. Initially, 40 PFOS/PFOA molecules were randomly placed in the aqueous solution, while the MoS₂ nanosheet was positioned 0.3 nm above the pre-equilibrium lipid bilayer. To remove bad initial contacts, the entire system sequentially underwent energy minimization, NVT (moles, volume, and temperature are constant) and NPT (moles, pressure, and temperature are constant) ensemble equilibrium. Then, the unrestrained production of molecular dynamics simulation was performed for 500 ns with a timestep of 2 fs.

3. Results and discussion

The formation of nanocomposites between TMDCs and PFAS may significantly alter their intrinsic properties and thus present a novel toxicity profile compared to their original toxicity. Understanding the interactions between TMDCs and PFAS and the subsequent impact on living organisms is the key to evaluating their potential environmental and health risk. In this study, we explored the interactions between two typical TMDCs, i.e., MoS₂ and WS₂, and seven PFAS compounds with different carbon chain lengths and thiol groups (Fig. 1).

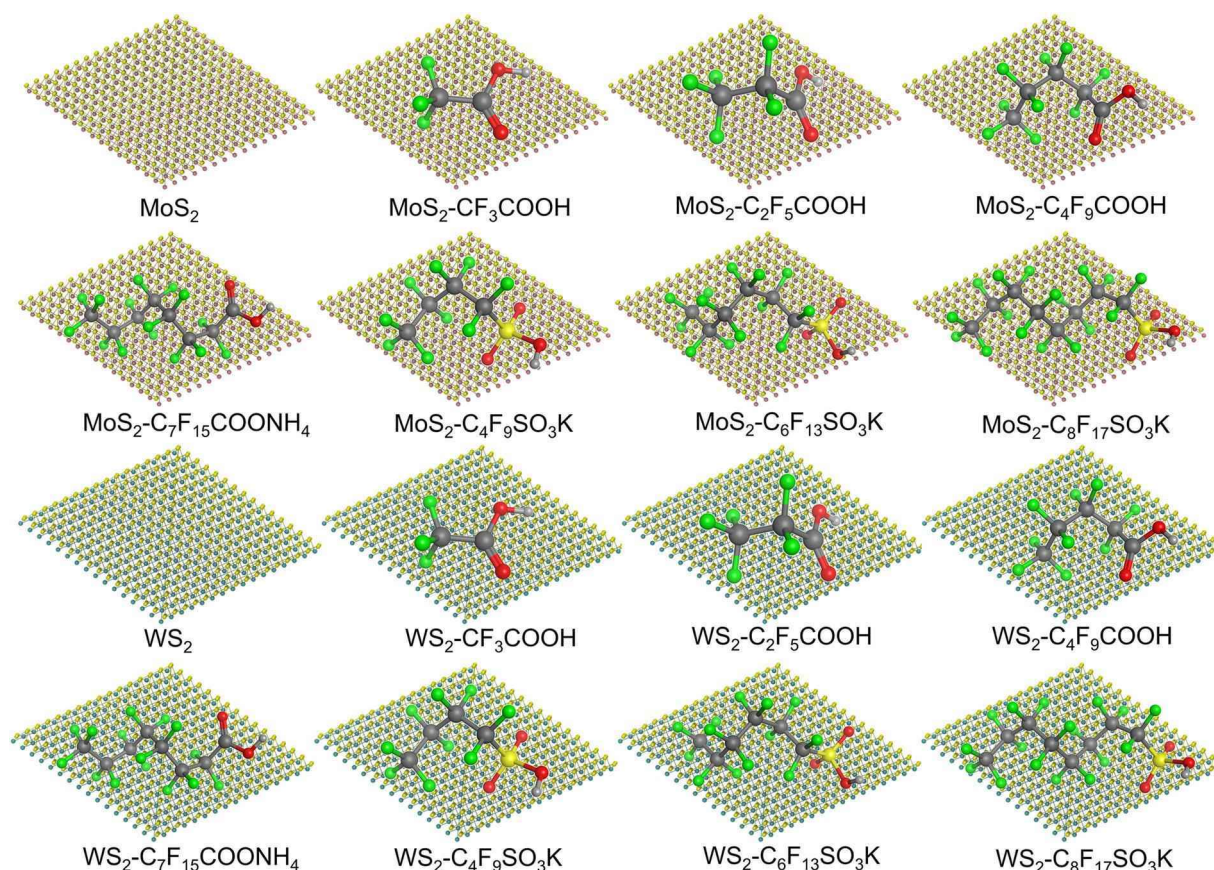


Fig. 1. A schematic of MoS₂ and WS₂ nanosheets and 14 nanocomposites with different loading PFAS. Seven representative PFAS compounds, as shown in the schematic, were loaded onto the surface of MoS₂ and WS₂, respectively. Color codes of atoms used in the schematic are listed as follows: Mo, purple; W, blue; C, black; F, green; O, red; S, orange.

3.1. Characterization of MoS₂, WS₂ and nanocomposites

Two typical TMDCs MoS₂ and WS₂ with high similarities in terms of their physical properties, including size, thickness and diameter were synthesized (Fig. S1). Specifically, AFM analysis of MoS₂ and WS₂ showed that both were nanosheets with a thickness of 1–2 nm (Fig. S1). Further size analysis showed that they also had similar dimensions, ranging between 100 and 200 nm (Fig. S1). Such results indicate that these two nanomaterials were comparable in their physicochemical properties except for their elemental composition.

MoS₂ and WS₂ nanosheets were then loaded with seven PFAS, forming 14 MoS₂-PFAS and WS₂-PFAS nanocomposites with similar zeta potential and hydrodynamic diameters (Fig. S2). The loading concentration of PFAS, as justified in the section on synthesis and characterization, was limited to 0.07 mmol per gram TMDCs to mimic their presence in environmental waters (Eriksen et al., 2011). Both WS₂-PFAS and MoS₂-PFAS were negatively charged in water, with the zeta potential ranging from −59.2 to −78.4 mV for MoS₂-PFAS and −47.3 to −69.4 mV for WS₂-PFAS (Fig. S2), suggesting that the formed nanocomposites were relatively stable since the absolute values of zeta potential were high. WS₂-PFAS and MoS₂-PFAS nanocomposites had a hydrodynamic diameter of 123.48–142.78 nm and 135–168.38 nm in water, respectively, indicating their uniformity in terms of their size.

By contrast, when suspended in a cell growth medium containing 10% FBS, obvious changes in the zeta potential and hydrodynamic diameter of nanocomposites were observed (Fig. S2). The zeta potential of WS₂-PFAS and MoS₂-PFAS in the growth medium ranged from −22.7 to −9.3 mV and −24.4 to −9.6 mV, respectively. Such changes in the zeta potential of nanocomposites were probably due to the absorption of protein molecules onto the surface of nanocomposites, thus leading to a

decrease in the absolute value of zeta potential (Walkey et al., 2014). Likewise, when measured in the FBS cell growth medium, their hydrodynamic diameters increased to 170.44–185.34 nm and 188.27–227.08 nm, respectively. Such results suggested that although protein molecules in the growth medium interacted with nanocomposites and altered their hydrodynamic diameters and zeta potential (Bai et al., 2020) there was no obvious aggregation of the synthesized nanomaterials.

3.2. In vitro toxicity on FHC and GES-1

GES-1 and FHC derived from the human stomach and intestine were used to explore the toxicity of all synthesized nanocomposites and their corresponding individual components, i.e., MoS₂, WS₂ and PFAS. Half maximal effective concentration (EC₅₀), a concentration leading to 50% of cell death, was used as an indicator of cytotoxicity (Patetsini et al., 2013) which was then further analyzed using machine learning approaches.

3.2.1. EC₅₀ induced by exposure to nanosheets and nanocomposites

As shown in Fig. 2A and B, MoS₂ had a lower EC₅₀ value than WS₂ in both cell lines (FHC: 225 ± 5 µg/mL of MoS₂ vs. 336 ± 8 µg/mL of WS₂; GES-1: 35 ± 3 µg/mL of MoS₂ vs. 111 ± 5 µg/mL of WS₂), indicating that MoS₂ nanosheets were more toxic to cells than WS₂. For PFAS, their toxicity increased as the carbon chain length increased, as indicated by their EC₅₀ values e.g., 11.36 ± 0.23 mmol/L for CF₃COOH and 6.46 ± 0.37 mmol/L for C₄F₉COOH in GES-1 (see more details Table S3), corresponding to previous reports suggesting that the toxicity of PFAS is carbon chain length dependent (Berntsen et al., 2017; Cai et al., 2019).

The toxicity of nanocomposites on GES-1 and FHC cells showed a

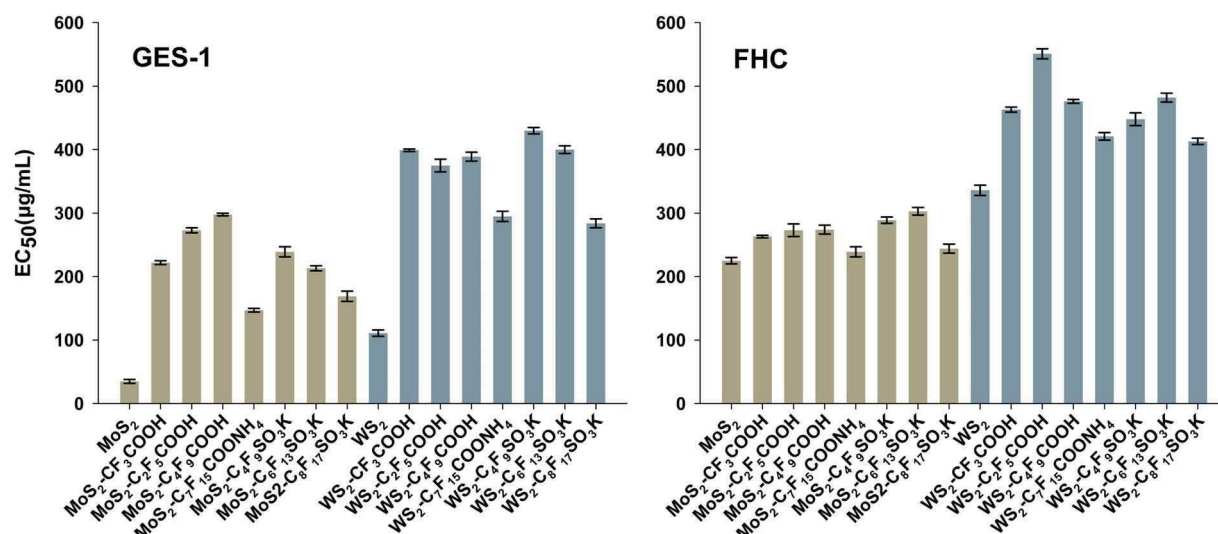


Fig. 2. EC_{50} values (\pm SE) of MoS_2 , WS_2 , PFAS, and their corresponding MoS_2 -PFAS and WS_2 -PFAS nanocomposites on GES-1 and FHC cells based on the content of core materials used in the experiment. Brown-colored bars indicate MoS_2 and its corresponding nanocomposites while cyan-colored bars suggest WS_2 -based nanomaterials.

more complicated pattern compared with bare MoS_2 and WS_2 , as demonstrated in Fig. 2A and B. After loading with PFAS, the toxicity of all formed nanocomposites on GES-1 cells tended to decrease compared with that of bare MoS_2 and WS_2 . Similar results have been reported by Chen et al. where PFAS-loaded nanoplastics showed a decrease in toxicity compared to bare core nanoplastics (Chen et al., 2020). Such results suggested that, at the concentration level used in our study, the introduction of PFAS and subsequent formation of nanocomposites may potentially decrease the in vitro toxicity of TMDCs. Such a difference in EC_{50} values might be because the loading of PFAS was achieved by absorption of reactive head groups in PFAS onto the surface of TMDCs whereas the perfluorinated carbon tails of PFAS remained unattached and served as relatively inert surfaces of nanocomposites compared with bare MoS_2 and WS_2 , creating an antagonistic effect that potentially reduced the overall toxicity towards cells (Chen et al., 2020).

The reduction in nanocomposite toxicity towards GES-1 (as indicated by the increased EC_{50} values) varied significantly between different PFAS loadings, especially when MoS_2 served as the core, suggesting that GES-1 might be more sensitive to the changes in pollutants absorbed onto the surface of nanosheets. By contrast, the above-mentioned trend was less obvious in FHC cells. WS_2 -based nanocomposites did show an increase in EC_{50} and thus a decreased toxicity on FHC cells; however, the toxicity of MoS_2 -based nanocomposites on FHC cells did not decrease dramatically compared to the bare MoS_2 nanosheets. There were also no obvious variations between nanocomposites with the same core material. Notably, there was a general trend that a relatively high toxicity level of nanocomposites was observed when the loading PFAS had a longer carbon chain length for both MoS_2 - and WS_2 -based nanocomposites, indicating that carbon chain length played an important role in determining the toxicity of PFAS and relevant nanocomposites. Indeed, a longer chain length in PFAS could result in a higher LogP value and a higher degree of mechanical damage to the cell membrane and other organelles (Berntsen et al., 2017; Bertanza et al., 2020; Cai et al., 2019). Such a length-dependent relationship, as demonstrated by our experimental results on PFAS (Table S3), was also obvious for the nanocomposites (Fig. 2).

3.2.2. In vitro toxicity elucidated by machine learning

To further explain the above-observed patterns in EC_{50} and gain insights into the quantitative relationships between the physicochemical properties of nanosheets and nanocomposites and their toxicities, we applied the random forest algorithm to construct QSAR models. Before

making predictions, the model parameters were tuned through the grid-search method to optimize the model performance. As shown in Fig. 3A and B, the constructed machine learning models exhibited acceptable performance in predicting the toxicity of nanocomposites towards two cell lines, i.e., $R^2 = 0.72$ for FHC and $R^2 = 0.63$ for GES-1. In addition, the deployment of five-fold cross-validation prevented overfitting of the machine learning models and thus increased their reliability. The results indicated that our machine learning model had the potential to be used for toxicity assessment of TMDCs-based nanocomposites. However, it should be noted that there were still several prediction outliers such as D_{out} data points (Fig. 3A), indicating the necessity to improve current nanodescriptors. Additionally, the generalization of machine learning models should be further verified by predicting more unseen data in current data sets.

We then further did feature importance analysis to reveal the most important factors contributing to the model. The top 10 important factors, including hydrodynamic diameter, core type, etc. are demonstrated in Fig. 3C and D. Based on the results of feature importance analysis, it can be concluded that the higher in vitro nanotoxicity of MoS_2 , as demonstrated in Fig. 2, was most likely due to the difference in its hydrodynamic diameter (relatively smaller; Fig. S2). Such size-dependent toxicity has been reported in many metal nanoparticles, e.g., silver and gold nanoparticles (Jia et al., 2017; Kim et al., 2012). By contrast, the in vitro toxicity of nanocomposites was determined by both the properties of the core and the loading components (i.e., PFAS). Interestingly, in vitro toxicity of synthesized nanosheets and nanocomposites towards the FHC cell line was mainly determined by the size and type of the core material (i.e., MoS_2 and WS_2) whereas core size and type, as well as other physicochemical properties featured by the loading PFAS, jointly determined the toxicity of nanocomposites towards the GES-1 cell line.

3.3. In vivo toxicity on zebrafish

Based on results of in vitro toxicity experiments using GES-1 and FHC cells, MoS_2 -based nanocomposites loading with PFOS and PFOA (hereafter MoS_2 -PFOS and MoS_2 -PFOA) showed relatively high toxicity and thus were selected for further in vivo experiments to explore the environmental fate and toxicity of TMDCs after their release into aquatic environments and interactions with existing pollutants. Nanocomposites with three different doses of MoS_2 , i.e., 0.1, 1, and 10 mg/L (hereafter “low”, “medium”, and “high”), loaded with an environmentally relevant concentration of PFOS/PFOA (0.03 mg/L) were used to assess the in

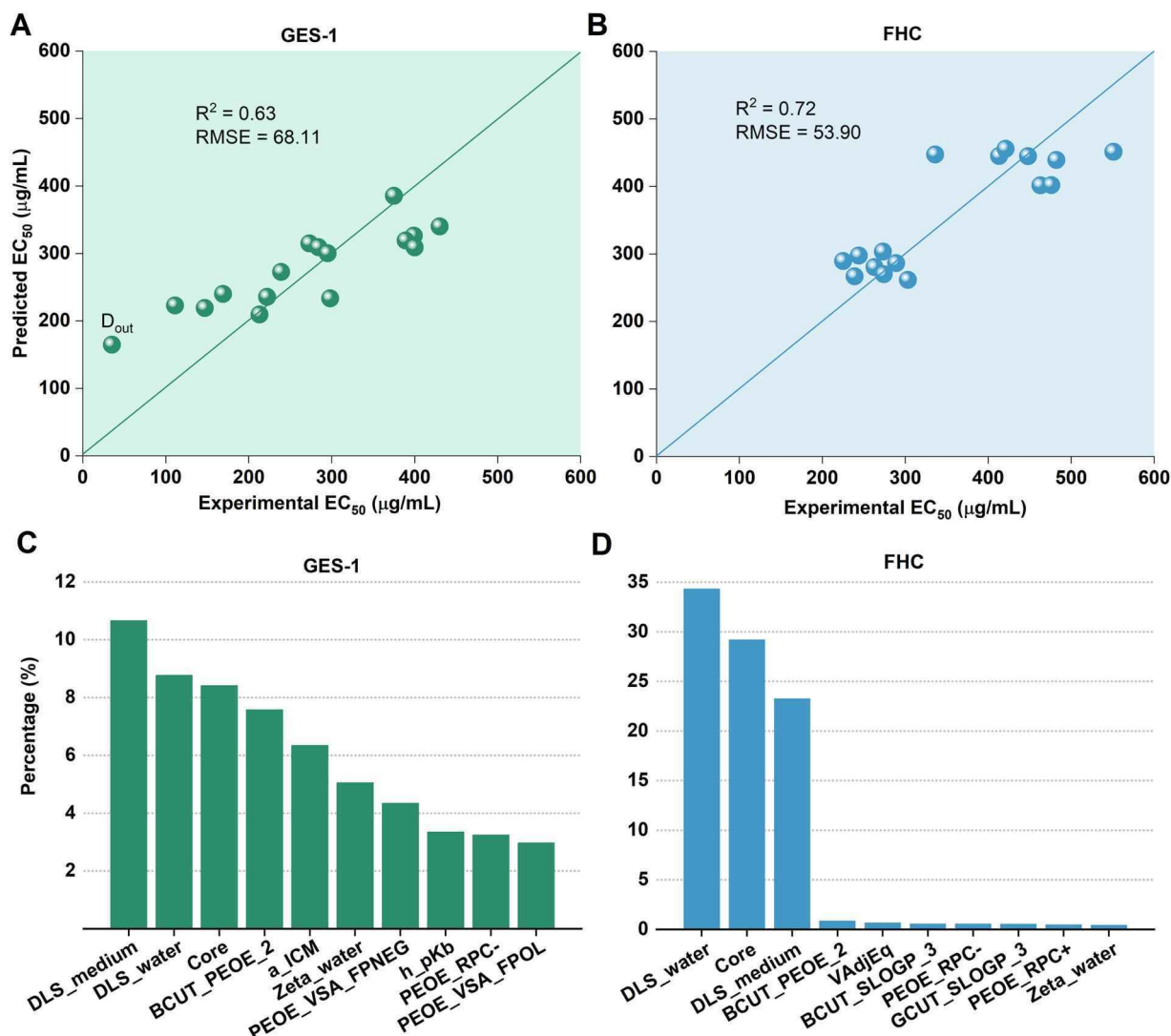


Fig. 3. Machine learning results using constructed QSAR models based on EC_{50} values demonstrated in Fig. 2. Experimental EC_{50} values are plotted against predicted ones for GES-1 (A) and FHC (B), respectively. The top 10 contributing factors to toxicity towards GES-1 and FHC revealed by feature importance analysis are demonstrated in C and D, respectively. The top three factors affecting the toxicity towards GES-1 are DLS medium (hydrodynamic diameter in the cell medium), DLS_water (hydrodynamic diameter in water), and Core (core of synthesized nanomaterials, i.e., WS_2 or MoS_2), which are also observed in FHC, but follow a slightly different order (DLS_water > Core > DLS_medium). Details of other factors can be found in Table S4.

vivo toxicity on zebrafish. The toxicity of bare PFOS, PFOA and MoS_2 with corresponding concentrations was also evaluated.

3.3.1. Bioaccumulation of nanocomposites in tissues

As shown in Fig. 4, with the help of TEM imaging and quantitative analysis of Mo contents, the presence of nanocomposites was detected in various organs of treated zebrafish, including gills, the liver, intestine, muscles, and brain. TEM images further confirmed the presence of nanocomposites in the intestine (Fig. 4A & E–F) and the liver (Fig. 4B & G–I) and demonstrated that most of the nanocomposites entered cellular organelles such as endosomes and lysosomes. Nanosheet-like structures with sizes similar to the model nanocomposites were also observed breaking into the cell membrane, suggesting that nanocomposites can enter the cell by cutting through the cell membrane. Mo element was also detected in the control group, indicating the natural presence of low concentrations of Mo in zebrafish.

Upon exposure, Mo content was relatively high in the gills compared to those in the brain and muscles (Fig. 4C). A relatively high concentration of Mo was observed in gills since they act as a multifunctional organ, participating in gas exchange, ion regulation, and

osmoregulation (Jönsson et al., 2009). The Mo content in the brain and muscles increased with an increasing exposure dosage. Such results indicate that TMDCs and related nanocomposites were able to enter these organs via blood vessels despite the presence of various barriers, e.g., the blood-brain barrier. Previous studies have demonstrated that various types of nanoparticles indeed can cross such barriers and lead to toxic effects on model species (Carmo et al., 2019). However, it is unclear how MoS_2 entered blood vessels and crossed the blood-brain barrier, eventually resulting in the bioaccumulation of Mo content in relevant organs, such as muscles and the brain.

After exposure to MoS_2 nanosheets and MoS_2 -PFOS nanocomposites, obvious bioaccumulation of nanocomposites in the intestine and liver was detected (Fig. 4D). Large amounts of Mo element in the intestine were detected, and the higher the exposure dose was, the greater the ingestion amount of Mo would be, suggesting positive dose-dependent ingestion of MoS_2 nanosheets and MoS_2 -PFOS nanocomposites. However, there was no significant difference in liver Mo content between the control group and the low-dose group. Similarly, after exposure to low and medium doses of MoS_2 or MoS_2 -PFOS, no significant difference was observed between different treatments in terms of Mo content in the

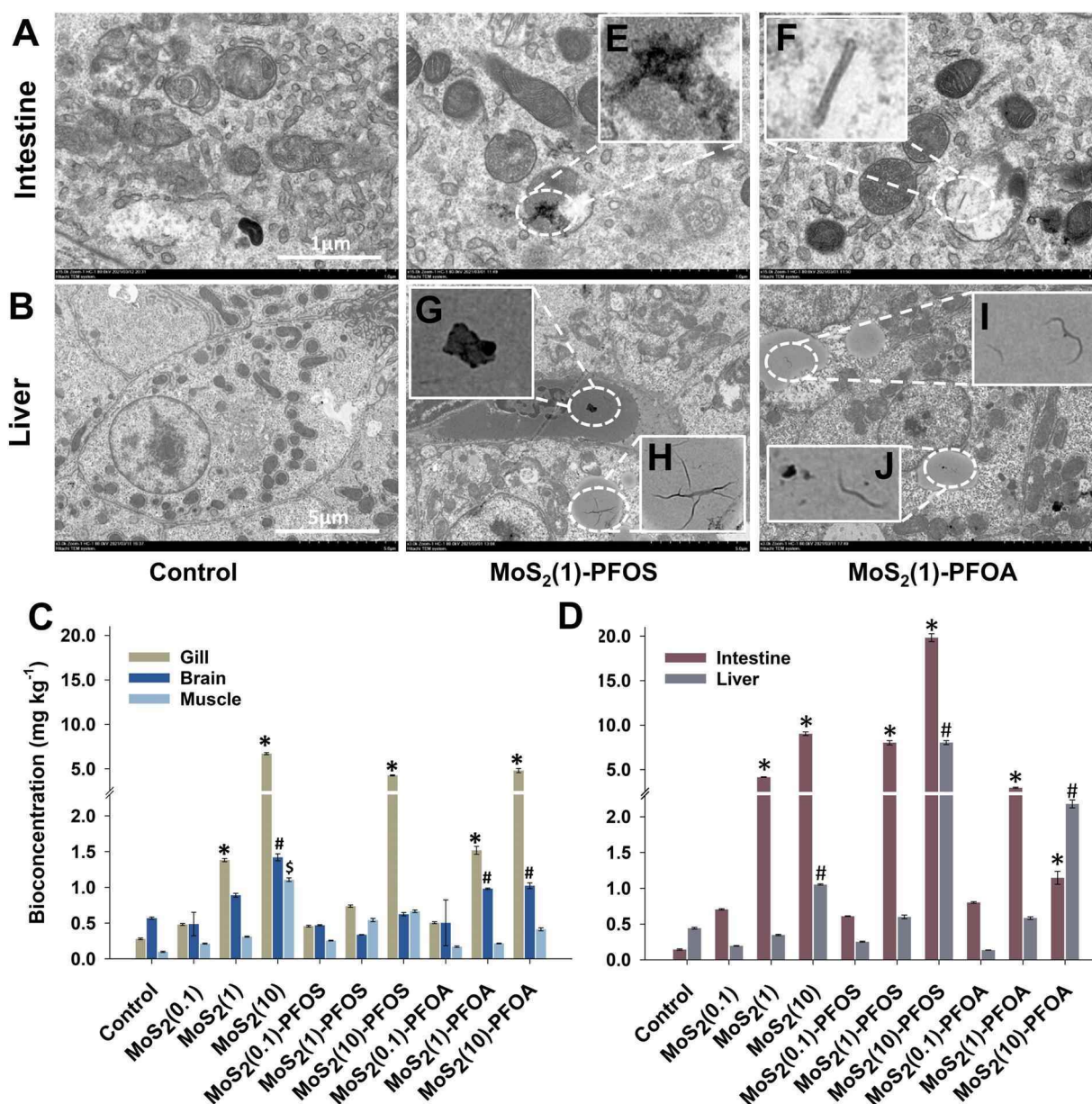


Fig. 4. Internalization and quantification of MoS₂-PFAS nanocomposites in the liver and intestine. TEM images of the liver (A) and intestine (B) demonstrated the obvious intake of nanocomposites in various organs. Quantification of MoS₂ and nanocomposites in different organs (C and D) were achieved by the analysis of Mo content using ICP-MS. Data are shown as means \pm SD ($n = 3$). Ellipses indicate areas where nanosheet-like structures with sizes similar to the model nanocomposites were identified, some of which were breaking into the cell membrane. Insets are enlarged views of such areas (E, F, G, H, I, and J). Significant differences ($P < 0.05$) compared to the control group are indicated by asterisks, dollar signs and hashtags, respectively.

liver. Nevertheless, a significantly higher Mo content was detected in the liver after exposure to the high dose compared with the control. As a matter of fact, after exposure to the high dose of MoS₂-PFOS, Mo content experienced a 138-fold and 17-fold increase in the intestine and the liver, respectively, compared with the control group. Such increases were still obvious compared with the high-MoS₂-dose treatment.

As for MoS₂-PFOA, an interesting variation pattern of Mo content in the zebrafish intestine and liver was observed after exposure to different doses of MoS₂-PFOA nanocomposites. Specifically, Mo content in the intestine experienced an initial increasing trend under low and medium doses of MoS₂-PFOA exposure and then decreased under the high-dose exposure. Moreover, such a peak value was 67 % lower than the peak value in the control group. By contrast, Mo content in the liver increased with the increase of exposure dose and then peaked at the high-dose exposure of MoS₂-PFOA, which was also observed in MoS₂-treated

groups although less obvious. The Mo content was 2.18 $\mu\text{g}\cdot\text{Kg}^{-1}$ at the high dose of MoS₂-PFOA, 2.1 times higher than it was when exposed to the high dose of MoS₂ alone. In summary, our results suggested that, compared with the exposure to MoS₂ alone, exposure to MoS₂-PFOA reduced the bioaccumulation of nanocomposites in the intestine but led to an increased level of nanocomposites (as indicated by Mo content) in the liver during the 14-day exposure period.

The above results suggested that both nanosheets and nanocomposites were able to cross the intestine barrier and thus ingested by intestine cells; however, it was obvious that interactions with existing pollutants (i.e., PFAS loading in our study) affected their abilities to cross biological barriers. For example, after loading with PFOS, the nanocomposite MoS₂-PFOS had an enhanced cross-barrier capability, thus leading to a hyperaccumulation of MoS₂-PFOS (as indicated by Mo content) in the intestine and liver. Similar results have also been

reported in previous studies where exposure to PFOS reduced the expression of the barrier function of intestinal cells, leading to increased intestinal permeability and subsequent enhanced PFOS translocation into the circulation (Diaz et al., 2021; Li et al., 2020). On the contrary, after loading with PFOA, the cross-intestinal-barrier capability of the formed MoS₂-PFOS seemed to be reduced; nevertheless, MoS₂-PFOS promoted the translocation of ingested MoS₂ from the intestine to the liver, thus leading to its bioaccumulation in the liver. Overall, different loadings of PFAS (i.e., PFOA and PFAS) on MoS₂ nanosheets play different and even opposite roles in MoS₂ crossing the intestinal barrier of zebrafish, thus leading to different levels of bioaccumulation of nanocomposites in different target organs.

To better explain our findings on the differences between MoS₂-PFOS

and MoS₂-PFOA in their ingestion and bioaccumulation in various organs, molecular dynamics was utilized to simulate the interactions between the cell membrane and nanomaterials. As shown in Fig. 5A and B, both MoS₂-PFOS and MoS₂-PFOA almost simultaneously initiated their cross-membrane activities at ~70 ns by the rotation of the nanosheet. At $t = \sim 300$ ns, the MoS₂-PFOS nanocomposite fully inserted into the cell membrane and remained attached throughout the entire simulation process (Fig. 5A and C). In the MoS₂-PFOA simulation system, the MoS₂ nanosheet wandered around the surface of the cell membrane, even though the PFOA molecules has fully entered the membrane (Fig. 5B and C). We further calculated the interaction energy between PFOA/PFOS molecules and the membrane, and identified stronger interactions between the membrane and PFOS molecules (Fig. 5D). Therefore, the PFOS

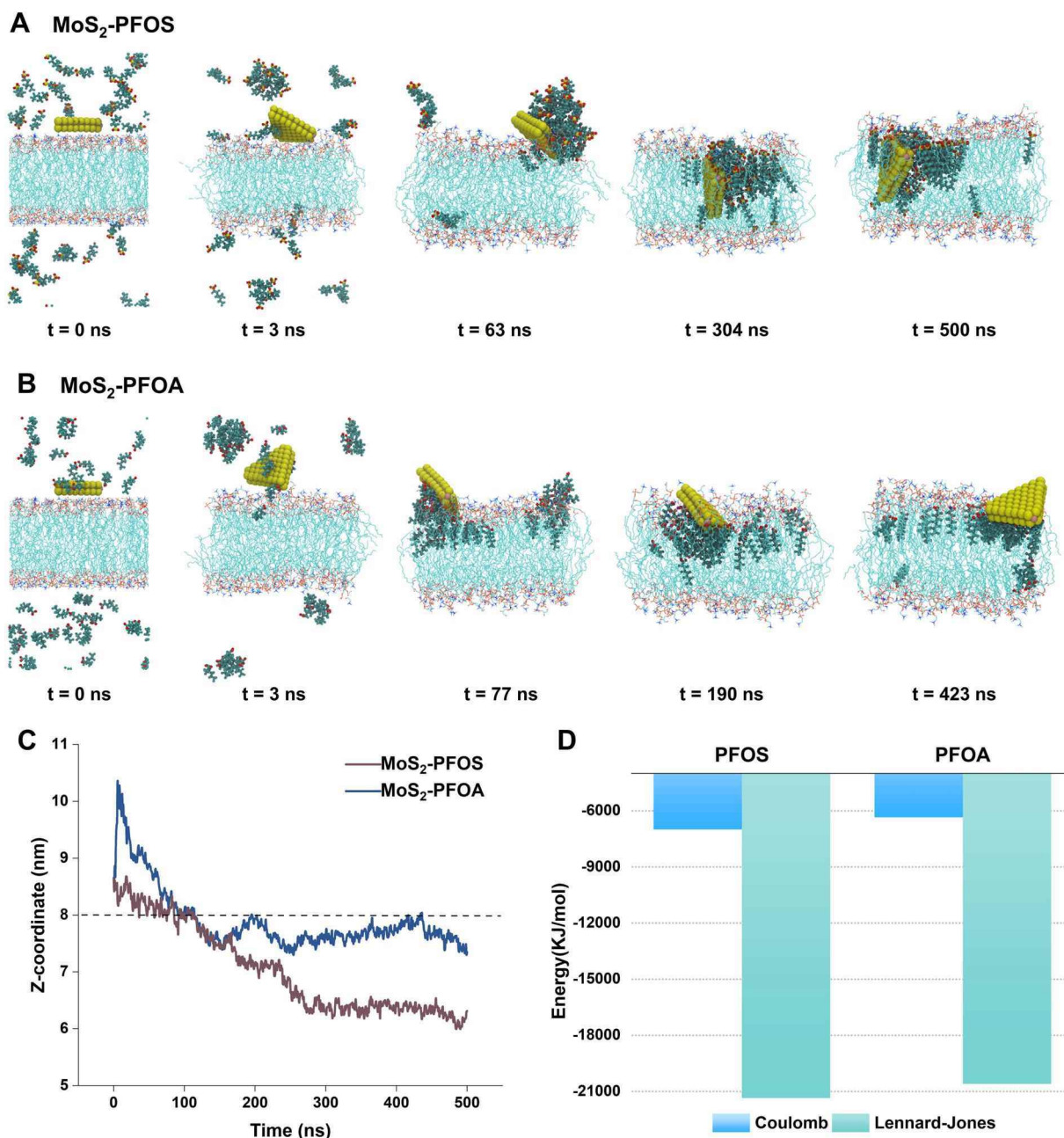


Fig. 5. Molecular dynamics simulation between nanocomposites and cell membranes by CHARMM36 force field in Gromacs. The simulation time was set as 500 ns for both MoS₂-PFOS and MoS₂-PFOA with six representative time points (A and B). Their movements throughout the entire process are demonstrated in C where the dashed line indicates the position of the cell membrane. The interaction between loading PFAS and the cell membrane was quantified by the Lennard-Jones (or van der Waals) and the Coulomb interaction energy (D). Further details about the setup of molecular dynamics simulation can be found in the method section as well as in Method S4.

molecule acted as a motor dragging the MoS₂ towards the membrane. By contrast, the interactions of PFOA with the membrane were not strong enough to allow the insertion of MoS₂ into the cell membrane during the 500 ns simulation. Furthermore, when PFOS interacts with the cell membrane, the Lennard-Jones (or van der Waals) interaction energy was much higher than the Coulomb interaction energy (Coul) or electrostatic interaction energy (Fig. 5D), thus serving as the driving force for MoS₂-PFOS insertion into the membrane.

In summary, the Lennard-Jones (or van der Waals) interaction energy played a predominant role in determining the cross-membrane activities of nanocomposites (Fig. 5D). Specifically, when the nanocomposites were far from the membrane, the electrostatic interactions between PFOS and phospholipid molecules can attract MoS₂-PFOS nanocomposites towards the membrane, after which the Lennard-Jones interactions between the PFOS and phospholipid molecules promote MoS₂-PFOS binding with the cell membrane. Such simulation results suggested that MoS₂-PFOS might have a greater cross-membrane capability than MoS₂-PFOA, thus leading to a higher concentration of MoS₂-

PFOS in the intestine and the liver, as observed in Fig. 4.

3.3.2. Alteration of histopathological and biological indices

The ingestion and subsequent bioaccumulation of TMDCs nanosheets and nanocomposites after exposure also caused histopathological and biological changes to the intestine, although the extent to which was different. For zebrafish exposed to PFAS (0.1 mg/L) or MoS₂ (0.1, 1 and 10 mg/L) alone, slight alterations of intestine structure and architecture were observed after 14 days of exposure, as can be seen from Fig. S3A–C. Similarly, most zebrafish treated with nanocomposites exhibited various degrees of structural alterations in the intestine, except those exposed to the low dose (Fig. 6A–G). In zebrafish treated with the medium dose of MoS₂-PFAS (1 mg/L), detachment and fusion of the epithelial cells, and beheaded villi were observed in the intestine (Fig. 6B & F). There was also an increase in the number of vacuole-like structures and goblet cells in the intestinal epithelium (Fig. 6B & F). In zebrafish treated with the high dose of MoS₂-PFOS/MoS₂-PFOS (10 mg/L), there was an obvious absence of regular structures in the serosa, muscular mucosae and

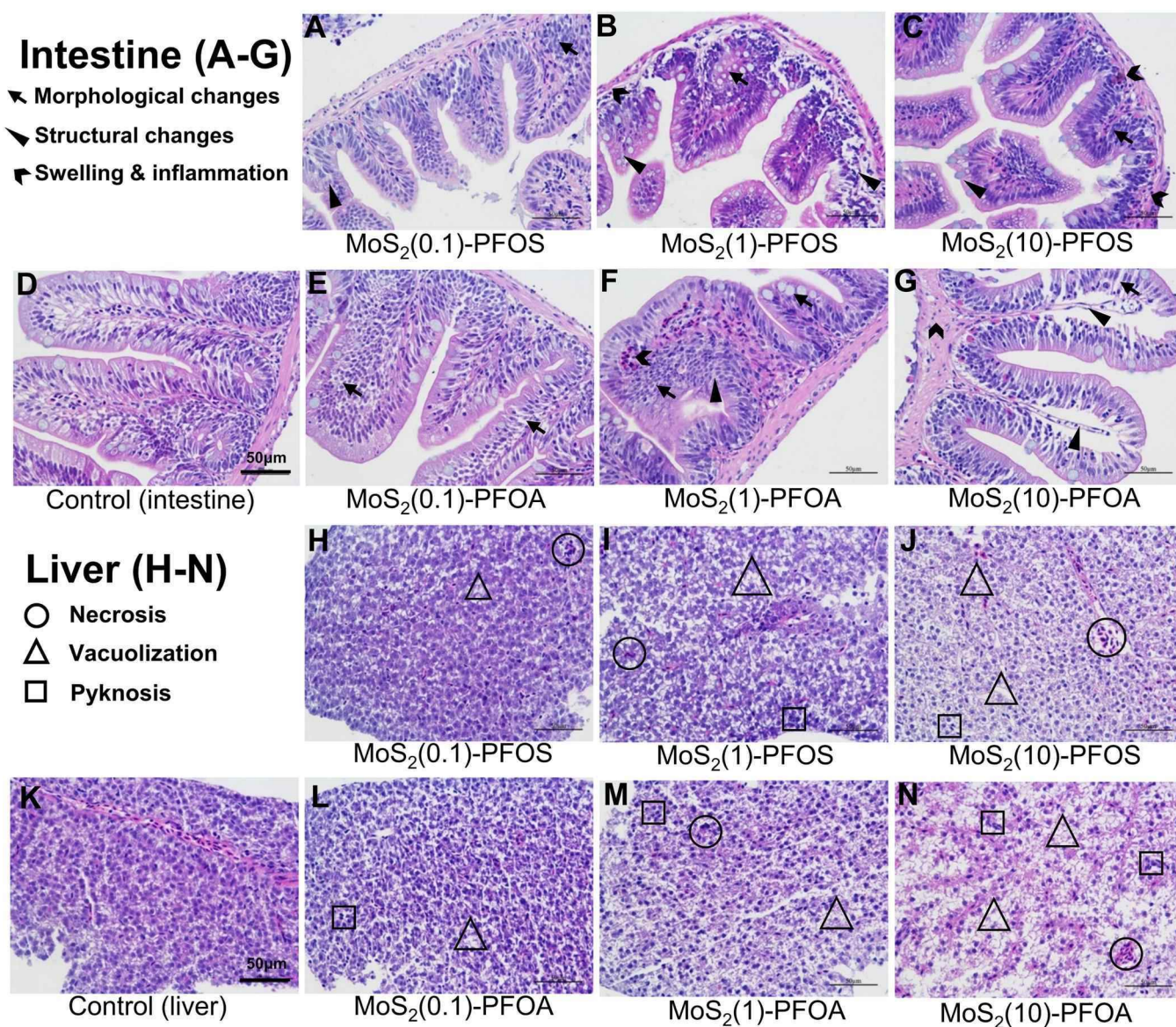


Fig. 6. Histopathological alterations observed in the intestine and liver after treatments with different doses of MoS₂-PFOS and MoS₂-PFOA nanocomposites, as indicated by the footnote. Histopathological alterations have been identified and marked with corresponding icons as shown in the figure. The numbers in the parentheses indicate the exposure dose based on the concentration of MoS₂ (0.1 mg/L, 1 mg/L and 10 mg/L). Corresponding examples of histopathological alterations after exposure to PFAS or nanosheets alone can be found in Fig. S3.

mucosa layer, as well as a lack of structural connections between them, due to the lesions observed in the above-mentioned tissues (Fig. 6C & G). An evident and complete detachment of the mucosal epithelium was also observed (Fig. 6C & G).

For the liver, a slightly different pattern of histopathological changes after exposure was observed (Fig. 6 & Fig. S3). In zebrafish treated with PFAS alone, hepatocytes were slightly enlarged and swollen, with congested hepatic sinusoids and blurred boundaries (Fig. S3G–I). Meanwhile, pyknosis was also observed in some cells (Fig. S3H–I). For exposure to MoS₂ alone, no obvious damage to liver tissue was observed at the low or medium dose after examining the microscopic results of zebrafish liver tissue sections despite identified necrosis and vacuolization (Fig. S3G & J–K). However, when exposed to the high dose of MoS₂, many localized lesions such as vacuole formation and necrosis were observed in the liver (Fig. S3L). For nanocomposite exposures, we observed more severe histopathological changes in zebrafish exposed to medium and high doses, especially in the high-dose group (Fig. 6H–N). Specifically, the nuclei of hepatocytes were severely deformed, atrophied, and deviated from the center of the cells, leading to the formation of vacuole-like structures. In some cases, the pyknosis or cytolysis of hepatocytes caused local necrosis of liver tissue (Fig. 6I–J & M–N).

Overall, both in the intestine and the liver, exposure to MoS₂-PFAS nanocomposites caused more severe structural damage to zebrafish tissues than exposure to MoS₂ or PFAS alone. In particular, exposure to medium and high doses of MoS₂-PFAS caused the most pronounced

structural changes and pathological damage. This is consistent with the results of the ingestion and bioaccumulation of TMDCs in the intestine and liver of zebrafish (as shown in Fig. 4). As for the zebrafish exposed to MoS₂ alone, the damage to the intestine was relatively mild, regardless of the exposure dose. By contrast, the high-MoS₂-dose exposure caused more severe liver damage in zebrafish liver, which might be caused by the higher accumulation of nanomaterials in the liver and subsequent difficulties in their metabolism and excretion (Yang et al., 2013).

3.4. Oxidative stress level after exposure

Reactive oxygen species (ROS) are a set of distinct molecular oxygen derivatives produced during normal aerobic metabolism; however, when exposed to pollutants, excessive ROS can be generated, posing oxidative stress and thus damage to living organisms, e.g., pyknosis and cytolysis as we observed and discussed in sections about histopathological changes (Hao and Chen, 2012). To ease the threats from such a non-equilibrium state of excessive ROS production, living organisms can produce various enzymes and metabolites involved in the body's antioxidant defense, especially superoxide dismutase (SOD), and catalase (CAT), (Lennicke and Cochemé, 2021) which was also observed in our study.

3.4.1. Increased intestinal oxidative stress

The exposure to MoS₂ (except the low dose), PFAS and their

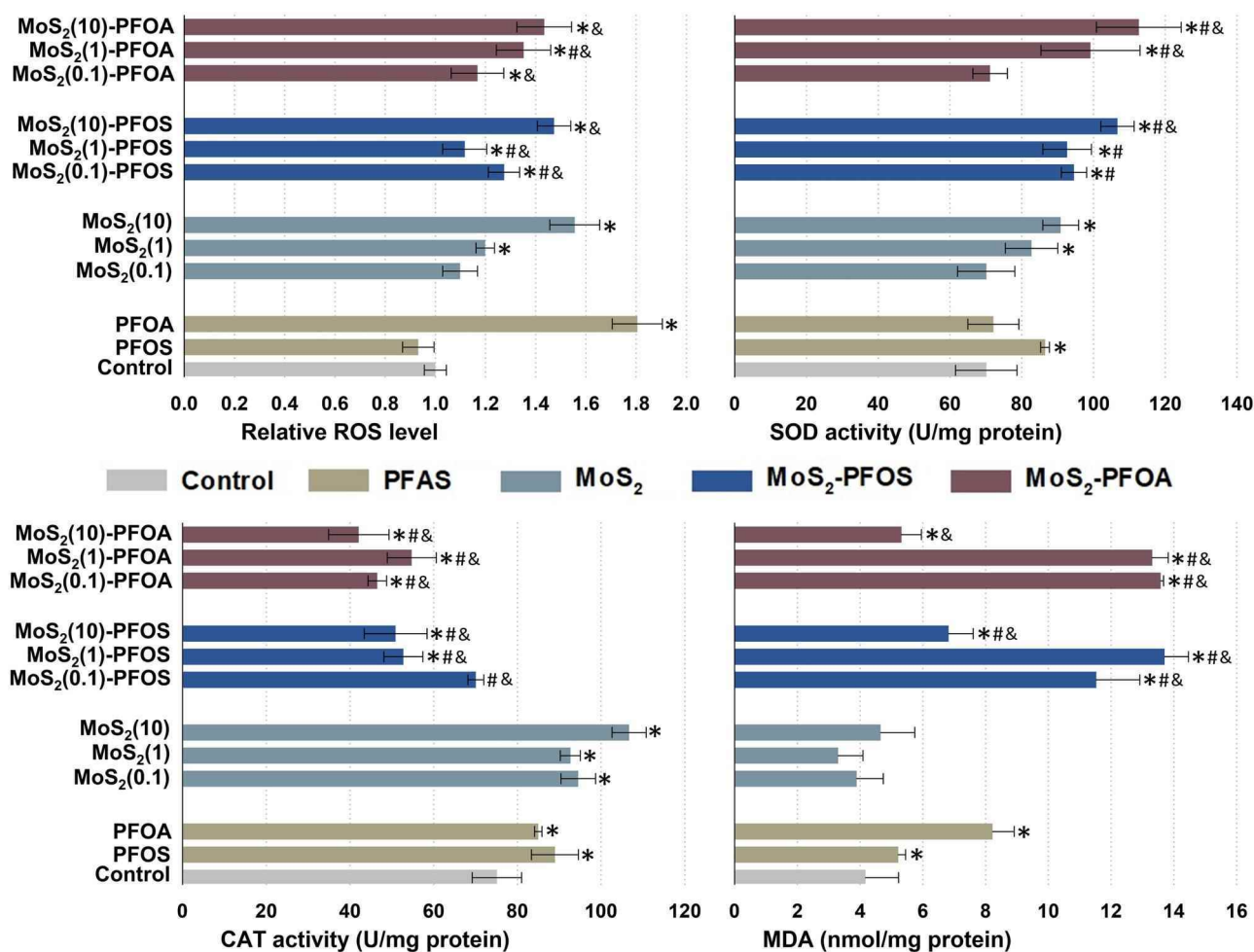


Fig. 7. Increased oxidative stress in the intestine as demonstrated by elevated ROS, MDA, SOD, and CAT levels after exposure to MoS₂, PFAS (PFOS and PFOA), and their corresponding nanocomposites. “*” indicates significant differences compared to the control group. “#” indicates significant differences compared to corresponding concentrations of MoS₂ alone. “&” indicates significant differences compared to corresponding PFAS alone (PFOA or PFOS). The numbers in the parentheses indicate the exposure dose based on the concentration of MoS₂ (0.1 mg/L, 1 mg/L and 10 mg/L). Significance level: $P < 0.05$.

nanocomposites significantly increased ROS levels in the intestine of zebrafish. Previous studies have shown that PFOA increases ROS production *in vitro* and *in vivo*, which in turn may lead to apoptosis or DNA damage (Xu et al., 2013). Similarly, when exposed to MoS₂-PFAS (i.e., MoS₂-PFOA and MoS₂-PFOS), ROS levels were significantly increased in all treatments and showed an increasing trend with the increase of exposure concentration, i.e., a dose-dependent effect (Fig. 7).

The excessive ROS production and subsequent oxidative stress in zebrafish after exposure to PFAS and their corresponding nanocomposites were further confirmed by the increased malondialdehyde (MDA) level. MDA is the end product of lipid peroxidation that represents oxidation and thus can be used as an indicator of changes in oxidative stress (Khoubnasabjafari et al., 2016). As shown in Fig. 7, exposure to PFAS and MoS₂-PFAS significantly increased MDA levels. However, the increase in MDA content was more significant in the low- and medium-dose groups and less obvious in the high-dose group, which may be related to the increased neutrophilic response and the inhibition of certain antioxidant functions in zebrafish exposed to the high dose (Ge et al., 2015). It should be noted that the highest ROS level induced by exposure to PFOA did not lead to the highest MDA level; in fact, the MDA level after exposure to PFOA was significantly lower than those of MoS₂-PFOA. This suggested that MoS₂-PFOA caused more peroxidative damage than PFOA alone. By contrast, there was no significant increase in MDA levels after MoS₂ exposure compared to the control group, indicating that MoS₂ exposure doses within the applied concentration

range did not cause obvious peroxidative damage to the zebrafish intestine, despite that the increased ROS levels in zebrafish after exposure to the medium or high dose of MoS₂ alone.

In our study, the detoxification procedure involving SOD and CAT was also pronounced in zebrafish cells (Fig. 7). SOD, generally recognized as the first and most powerful antioxidant in the cell, can catalyze the dismutation of superoxide anion (one of the most produced ROS) into molecular oxygen (O₂) and hydrogen peroxide (H₂O₂), which was further reduced to water (H₂O) and O₂ by CAT (Hao and Chen, 2012). As a result, SOD, together with CAT, can complete the detoxification process, reducing potential damage to cells caused by excessive ROS production (Hao and Chen, 2012; Lennicke and Cochemé, 2021). For SOD, zebrafish exposed to the low- and high-dose MoS₂-PFAS experienced enhanced intestinal SOD enzyme activity, especially in the high-dose group, with an activation rate of 52.2% [MoS₂(10)-PFOS] and 60.7% [MoS₂(10)-PFOA], respectively (Fig. 7). The increase of SOD enzyme activity, as a normal function of oxidative stress in organisms, helped maintain the balance of the oxidative-antioxidant system and ensured the normal operation of zebrafish after exposure (Xia et al., 2020). As for CAT, exposure to MoS₂ or PFAS alone significantly induced an increase in CAT activity. However, exposure to MoS₂-PFAS showed an overall inhibitory effect on zebrafish intestinal CAT activity, except for the low-dose treatment of MoS₂-PFOS (Fig. 7). The corresponding inhibition rate was 29.8% [MoS₂(1)-PFOS], 32.4% [MoS₂(10)-PFOS], 38.0% [MoS₂(0.1)-PFOA], 27.2% [MoS₂(1)-PFOA], and 43.9%

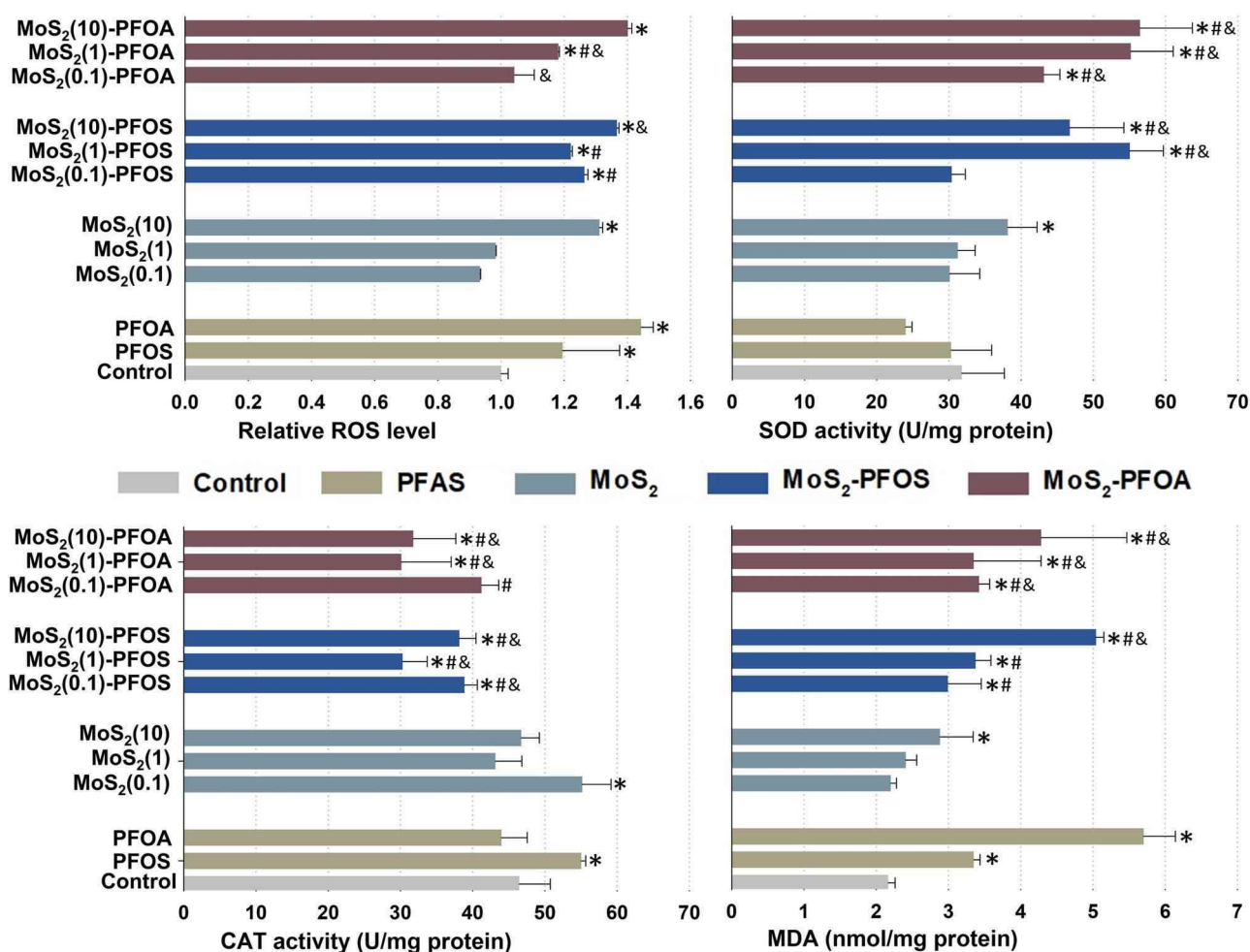


Fig. 8. Increased oxidative stress in the liver as demonstrated by elevated ROS, MDA, SOD, and CAT levels after exposure to MoS₂, PFAS (PFOS and PFOA), and their corresponding nanocomposites. “*” indicates significant differences compared to the control group. “#” indicates significant differences compared to corresponding concentrations of MoS₂ alone. “&” indicates significant differences compared to PFAS exposure alone (PFOA or PFOS). The numbers in the parentheses indicate the exposure dose based on the concentration of MoS₂ (0.1 mg/L, 1 mg/L and 10 mg/L). Significance level: $P < 0.05$.

[MoS₂(10)-PFOA], respectively, indicating that the exposure of MoS₂-PFAS nanocomposites caused toxic effects on zebrafish intestinal CAT enzyme activity. Such results suggested that when the toxicogenic effect of a pollutant is greater than the detoxification capability of the organism itself, the antioxidant enzyme activity will decrease, causing more severe damage to the organism (Hao and Chen, 2012).

3.4.2. Increased oxidative stress in the liver

The liver is the major metabolic organ in the body, and its physiological and biochemical indexes are sensitive to changes in the living environment when it is poisoned by external or internal pollutants (Jensen-Cody and Potthoff, 2021), which has also been demonstrated in this study. Fig. 8 shows the impact of exposure to MoS₂ nanosheets, PFAS contaminants and their nanocomposites on ROS levels and enzyme activities in the zebrafish liver.

The ROS activity was significantly increased in zebrafish liver after exposure to PFAS alone or the MoS₂-PFAS nanocomposites (except for the low dose of MoS₂-PFOA) (Fig. 8). It should be noted that, after exposure to MoS₂, changes in ROS levels were not obvious unless exposure to the high dose. By contrast, the MoS₂-PFAS exposure significantly enhanced the effect of MoS₂ on elevating ROS levels in zebrafish liver, resulting in significantly high ROS levels compared to the control group. Such results suggested that, compared to exposure to MoS₂, exposure to MoS₂-PFAS nanocomposites could enhance the oxidative stress damage of MoS₂ on zebrafish liver.

The above results were further confirmed by the MDA levels in the liver (Fig. 8). Specifically, exposure to MoS₂-PFAS under all doses significantly increased the MDA level compared to exposure to MoS₂, suggesting more severe damage to the cell membrane due to exposure to MoS₂-PFAS. Moreover, such damage showed a dose-dependent pattern since a higher exposure dose tended to increase the MDA level in the liver, as demonstrated in Fig. 8. Interestingly, unlike what we observed in the intestine, the MDA levels in the liver induced by exposure to PFAS were even higher than that induced by exposure to their corresponding nanocomposites, which was pronounced mostly for PFOA. This indicates that exposure to PFOA contaminants caused more severe peroxidative damage to the liver of zebrafish than MoS₂ as well as most of their nanocomposite counterparts, MoS₂-PFOA. Although PFOA has been suggested to induce severe peroxidative damage by decreasing the total antioxidant capacity (Wielsoe et al., 2015), it is still unclear why such a phenomenon was only observed in the liver in our study.

To reduce the potential damage posed by elevated ROS levels, the detoxification procedure involving SOD and CAT was initiated in the liver (Fig. 8). Specifically, the SOD level in zebrafish exposed to MoS₂-PFAS nanocomposites experienced an increasing trend. The increase in SOD activity was a normal physiological phenomenon in organisms exposed to oxidative stress, suggesting that a stronger antioxidant response was generated in the liver of zebrafish in the combined exposure group to maintain the oxidative-oxidative balance in vivo (Vale et al., 2016). However, it should be noted that the SOD activity was inhibited to some extent in zebrafish after exposure to PFAS alone (Wielsoe et al., 2015). The effect of MoS₂-PFAS exposure on CAT activity in zebrafish liver was generally inhibitory. Except for the low-dose group of MoS₂-PFOA, all other exposure treatments showed significant inhibitory effects, with an inhibition rate of 16% [MoS₂(0.1)-PFOS], 34.8% [MoS₂(1)-PFOS], 18.0% [MoS₂(10)-PFOS], 35.3% [MoS₂(1)-PFOA], and 31.6% [MoS₂(10)-PFOA], respectively. This indicated that MoS₂-PFAS exposure caused a moderate toxic response to CAT enzyme activity in zebrafish liver. When the toxicogenic effect of the contaminants was greater than the detoxification effect of the organism itself, the antioxidant enzyme activity decreased and caused more serious damage to the organism (Wielsoe et al., 2015). Such a pattern was to the trend of intestinal CAT enzyme activity in zebrafish described above.

4. Conclusions

Widespread applications of TMDCs will no doubt emit a significant amount of TMDCs (e.g., MoS₂ and WS₂) into aquatic environments, thus inevitably interacting with existing pollutants, especially persistent ones such as PFAS. In this study, we mimicked and explored the interactions between model TMDCs and PFAS and their environmental fates by loading PFAS onto the surface of TMDCs in aqueous mediums. With the help of both experimental and simulation methods, our study revealed that the interaction between TMDCs and PFAS enhanced their toxic effects both in vivo and in vitro, regardless of their compositions. However, the toxicity of the formed nanocomposites was cell dependent, demonstrating the necessity of toxicity evaluation on cell line panels when assessing the interactions between emerging and existing environmental pollutants and their potential health risks on living organisms in aquatic environments. More importantly, our findings through feature important analysis of machine learning models based on experimental results highlight the importance of and provide theoretical support for the prioritization of regulations on certain types of emerging or persisting pollutants that may play important roles in the interaction between environmental pollutants in aquatic environments.

CRediT authorship contribution statement

Xiliang Yan, Guohong Liu: Conceived the idea. **Xiliang Yan, Guohong Liu, Bing Yan:** Designed experiments. **Chengjun Li, Guohong Liu:** Performed experiments. **Song Hu:** Constructed the machine learning model. **Jiachen Yan:** Performed the molecular dynamics simulation. **Xiliang Yan, Guohong Liu, Chengjun Li, Bing Yan:** Interpreted the data. **Xiliang Yan, Guohong Liu, Chengjun Li, Bing Yan:** Co-wrote the manuscript and all authors discussed and approved the paper.

Environmental implications

Although there is an increasing interest in the potential hazards of transition-metal dichalcogenides (TMDCs) in conjunction with other environmental pollutants, it is still impossible to explore all interactions using straightforward experimental methods. Here, we solved the above problem by combining nanocomposite library, high-throughput screening assays, machine learning and molecular dynamics simulations. Our approach developed an efficient way to predict and provide an in-depth mechanism analysis of the joint toxicity of engineered nanomaterials with environmental pollutants.

Code availability

Codes in this study are available at <https://github.com/YanLabAI/TMDCs-PFAS>.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

Data will be made available on request.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (22106025, 22006025, and 22036002), the Introduced Innovative R&D Team Project under the “The Pearl River Talent Recruitment Program” of Guangdong Province (2019ZT08L387).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at [doi:10.1016/j.jhazmat.2022.130303](https://doi.org/10.1016/j.jhazmat.2022.130303).

References

- Bai, X., Wang, S., Yan, X., Zhou, H., Zhan, J., Liu, S., Sharma, V.K., Jiang, G., Zhu, H., Yan, B., 2020. Regulation of cell uptake and cytotoxicity by nanoparticle core under the controlled shape, size, and surface chemistries. *ACS Nano* 14 (1), 289–302. <https://doi.org/10.1021/acsnano.9b04407>.
- Benzie, I.F., Strain, J.J., 1996. The ferric reducing ability of plasma (FRAP) as a measure of “antioxidant power”: the FRAP assay. *Anal. Biochem.* 239 (1), 70–76.
- Berntsen, H.F., Bjørklund, C.G., Audinot, J.-N., Hofer, T., Verhaegen, S., Lentzen, E., Gutleb, A.C., Ropstad, E., 2017. Time-dependent effects of perfluorinated compounds on viability in cerebellar granule neurons: dependence on carbon chain length and functional group attached. *Neurotoxicology* 63, 70–83.
- Bertanza, G., Capoferri, G.U., Carmagnani, M., Icarelli, F., Sorlini, S., Pedrazzani, R., 2020. Long-term investigation on the removal of perfluoroalkyl substances in a full-scale drinking water treatment plant in the Veneto Region, Italy. *Sci. Total Environ.* 734, 139154.
- Cai, Y., Chen, H., Yuan, R., Wang, F., Chen, Z., Zhou, B., 2019. Toxicity of perfluorinated compounds to soil microbial activity: Effect of carbon chain length, functional group and soil properties. *Sci. Total Environ.* 690, 1162–1169.
- Carmo, T.L., Siqueira, P.R., Azevedo, V.C., Tavares, D., Pesenti, E.C., Cestari, M.M., Martinez, C.B., Fernandes, M.N., 2019. Overview of the toxic effects of titanium dioxide nanoparticles in blood, liver, muscles, and brain of a Neotropical detritivorous fish. *Environ. Toxicol.* 34 (4), 457–468.
- Chen, W., Yuan, D., Shan, M., Yang, Z., Liu, C., 2020. Single and combined effects of amino polystyrene and perfluorooctane sulfonate on hydrogen-producing thermophilic bacteria and the interaction mechanisms. *Sci. Total Environ.* 703, 135015.
- Dale, A.L., Casman, E.A., Lowry, G.V., Lead, J.R., Viparelli, E., Baalousha, M., 2015. Modeling nanomaterial environmental fate in aquatic systems. *Environ. Sci. Technol.* 49 (5), 2587–2593. <https://doi.org/10.1021/es505076w>.
- Diaz, O.E., Sorini, C., Morales, R.A., Luo, X., Frede, A., Kraus, A.M., Chávez, M.N., Wincent, E., Das, S., Villablanca, E.J., 2021. Perfluorooctanesulfonic acid modulates barrier function and systemic T-cell homeostasis during intestinal inflammation. *Dis. Model. Mech.* 14 (12), dmm049104.
- Eriksen, K.T., Sørensen, M., McLaughlin, J.K., Tjønneland, A., Overvad, K., Raaschou-Nielsen, O., 2011. Determinants of plasma PFOA and PFOS levels among 652 Danish men. *Environ. Sci. Technol.* 45 (19), 8137–8143. <https://doi.org/10.1021/es100626h>.
- Gagliano, E., Sgroi, M., Falciglia, P.P., Vagliasindi, F.G., Roccaro, P., 2020. Removal of poly-and perfluoroalkyl substances (PFAS) from water by adsorption: role of PFAS chain length, effect of organic matter and challenges in adsorbent regeneration. *Water Res.* 171, 115381.
- Ge, W., Yan, S., Wang, J., Zhu, L., Chen, A., Wang, J., 2015. Oxidative stress and DNA damage induced by imidacloprid in zebrafish (*Danio rerio*). *J. Agric. Food Chem.* 63 (6), 1856–1862.
- Ghisi, R., Vamerali, T., Manzetti, S., 2019. Accumulation of perfluorinated alkyl substances (PFAS) in agricultural plants: a review. *Environ. Res.* 169, 326–341.
- Gu, Z., Plant, L.D., Meng, X.-Y., Perez-Aguilar, J.M., Wang, Z., Dong, M., Logothetis, D.E., Zhou, R., 2018. Exploring the nanotoxicology of MoS₂: a study on the interaction of MoS₂ nanoflakes and K⁺ channels. *ACS Nano* 12 (1), 705–717. <https://doi.org/10.1021/acsnano.7b07871>.
- Guiney, L.M., Wang, X., Xia, T., Nel, A.E., Hersam, M.C., 2018. Assessing and mitigating the hazard potential of two-dimensional materials. *ACS Nano* 12 (7), 6360–6377. <https://doi.org/10.1021/acsnano.8b02491>.
- Hao, L., Chen, L., 2012. Oxidative stress responses in different organs of carp (*Cyprinus carpio*) with exposure to ZnO nanoparticles. *Ecotoxicol. Environ. Saf.* 80, 103–110.
- Janero, D.R., 1990. Malondialdehyde and thiobarbituric acid-reactivity as diagnostic indices of lipid peroxidation and peroxidative tissue injury. *Free Radic. Biol. Med.* 9 (6), 515–540.
- Jensen-Cody, S.O., Potthoff, M.J., 2021. Hepatokines and metabolism: deciphering communication from the liver. *Mol. Metab.* 44, 101138.
- Jia, Y.-P., Ma, B.-Y., Wei, X.-W., Qian, Z.-Y., 2017. The in vitro and in vivo toxicity of gold nanoparticles. *Chin. Chem. Lett.* 28 (4), 691–702.
- Jönsson, M.E., Brunström, B., Brandt, I., 2009. The zebrafish gill model: Induction of CYP1A, EROD and PAH adduct formation. *Aquat. Toxicol.* 91 (1), 62–70.
- Keller, A.A., McFerran, S., Lazareva, A., Suh, S., 2013. Global life cycle releases of engineered nanomaterials. *J. Nanopart. Res.* 15 (6), 1–17.
- Khoubnasabjafari, M., Ansarin, K., Jouyban, A., 2016. Critical review of malondialdehyde analysis in biological samples. *Curr. Pharm. Anal.* 12 (1), 4–17.
- Kim, T.H., Kim, M., Park, H.S., Shin, U.S., Gong, M.S., Kim, H.W., 2012. Size-dependent cellular toxicity of silver nanoparticles. *J. Biomed. Mater. Res. Part A* 100 (4), 1033–1043.
- Lennicke, C., Cochemé, H.M., 2021. Redox metabolism: ROS as specific molecular regulators of cell signaling and function. *Mol. Cell* 81 (18), 3691–3707. <https://doi.org/10.1016/j.molcel.2021.08.018>.
- Li, R., Tang, T., Qiao, W., Huang, J., 2020. Toxic effect of perfluorooctane sulfonate on plants in vertical-flow constructed wetlands. *J. Environ. Sci.* 92, 176–186.
- Liu, C., Zhang, B., Chen, W., Liu, W., Zhang, S., 2021. Current development of wearable sensors based on nanosheets and applications. *TRAC Trends Anal. Chem.* 143, 116334.
- Mahoney, H., Xie, Y., Brinkmann, M., Giesy, J.P., 2022. Next Generation Per-and Poly-Fluoroalkyl Substances: Status and Trends, Aquatic Toxicity, and Risk Assessment. *Eco-Environment & Health*.
- Patetsini, E., Dimitriadis, V., Kaloyianni, M., 2013. Biomarkers in marine mussels, *Mytilus galloprovincialis*, exposed to environmentally relevant levels of the pesticides, chlorpyrifos and penoxsulam. *Aquat. Toxicol.* 126, 338–345.
- Pichler, F.B., Laurensen, S., Williams, L.C., Dodd, A., Copp, B.R., Love, D.R., 2003. Chemical discovery and global gene expression analysis in zebrafish. *Nat. Biotechnol.* 21 (8), 879–883.
- Reiner, K., 2010. Catalase Test Protocol. American Society for Microbiology, pp. 1–6.
- Scalisi, E.M., Salvaggio, A., Antoci, F., Messina, A., Pecoraro, R., Cantarella, M., Gorrasi, G., Impellizzeri, G., Brundo, M.V., 2020. Toxicity assessment of two-dimensional nanomaterials molybdenum disulfide in *Gallus gallus domesticus*. *Ecotoxicol. Environ. Saf.* 200, 110772.
- Shang, E., Niu, J., Li, Y., Zhou, Y., Crittenden, J.C., 2017. Comparative toxicity of Cd, Mo, and W sulphide nanomaterials toward *E. coli* under UV irradiation. *Environ. Pollut.* 224, 606–614.
- Sonne, C., Xia, C., Lam, S.S., 2022. Ban fluorinated organic substances to spark green alternatives. *Eco-Environ. Health* 1 (2), 105–106.
- Sunderland, E.M., Hu, X.C., Dassuncao, C., Tokranov, A.K., Wagner, C.C., Allen, J.G., 2019. A review of the pathways of human exposure to poly-and perfluoroalkyl substances (PFASs) and present understanding of health effects. *J. Expo. Sci. Environ. Epidemiol.* 29 (2), 131–147.
- Vale, G., Mehennauti, K., Cambier, S., Libralato, G., Jomini, S., Domingos, R.F., 2016. Manufactured nanoparticles in the aquatic environment-biochemical responses on freshwater organisms: a critical overview. *Aquat. Toxicol.* 170, 162–174.
- Walkey, C.D., Olsen, J.B., Song, F., Liu, R., Guo, H., Olsen, D.W.H., Cohen, Y., Emili, A., Chan, W.C., 2014. Protein corona fingerprinting predicts the cellular interaction of gold and silver nanoparticles. *ACS Nano* 8 (3), 2439–2455.
- Wang, Y., Zhou, L., Wang, X., Liu, X., Jiang, L., Wang, J., Sun, H., Jiang, C., Xing, X., Zhang, Y., Pan, B., Yan, B., 2018. A human cell panel for evaluating safe application of nano-ZrO₂/polymer composite in water remediation. *Ecotoxicol. Environ. Saf.* 166, 474–481. <https://doi.org/10.1016/j.ecoenv.2018.09.098>.
- Wielsoe, M., Long, M., Ghisari, M., Bonefeld-Jørgensen, E.C., 2015. Perfluoroalkylated substances (PFAS) affect oxidative stress biomarkers in vitro. *Chemosphere* 129, 239–245.
- Wu, B., Chen, L., Wu, X., Hou, H., Wang, Z., Liu, S., 2019. Differential influence of molybdenum disulfide at the nanometer and micron scales in the intestinal metabolome and microbiome of mice. *Environ. Sci.: Nano* 6 (5), 1594–1606.
- Xia, X., Sun, M., Zhou, M., Chang, Z., Li, L., 2020. Polyvinyl chloride microplastics induce growth inhibition and oxidative stress in *Cyprinus carpio* var. larvae. *Sci. Total Environ.* 716, 136479.
- Xu, D., Li, C., Wen, Y., Liu, W., 2013. Antioxidant defense system responses and DNA damage of earthworms exposed to perfluorooctane sulfonate (PFOS). *Environ. Pollut.* 174, 121–127.
- Xu, Z., Liu, X., Peng, J., Qu, C., Chen, Y., Zhang, M., Liang, D., Lei, M., Tie, B., Du, H., 2022. Tungsten-humic substances complexation. *Carbon Res.* 1 (1), 1–12.
- Yamankurt, G., Berns, E.J., Xue, A., Lee, A., Bagheri, N., Mrksich, M., Mirkin, C.A., 2019. Exploration of the nanomedicine-design space with high-throughput screening and machine learning. *Nat. Biomed. Eng.* 3 (4), 318–327. <https://doi.org/10.1038/s41551-019-0351-1>.
- Yan, X., Sedyk, A., Wang, W., Yan, B., Zhu, H., 2020. Construction of a web-based nanomaterial database by big data curation and modeling friendly nanostructure annotations. *Nat. Commun.* 11 (1), 1–10.
- Yang, K., Gong, H., Shi, X., Wan, J., Zhang, Y., Liu, Z., 2013. In vivo biodistribution and toxicology of functionalized nano-graphene oxide in mice after oral and intraperitoneal administration. *Biomaterials* 34 (11), 2787–2795.
- Yu, Y., Yi, Y., Li, Y., Peng, T., Lao, S., Zhang, J., Liang, S., Xiong, Y., Shao, S., Wu, N., 2018. Dispersible MoS₂ micro-sheets induced a proinflammatory response and apoptosis in the gills and liver of adult zebrafish. *RSC Adv.* 8 (32), 17826–17836.
- Yuan, P., Zhou, Q., Hu, X., 2020. WS₂ nanosheets at noncytotoxic concentrations enhance the cytotoxicity of organic pollutants by disturbing the plasma membrane and efflux pumps. *Environ. Sci. Technol.* 54 (3), 1698–1709. <https://doi.org/10.1021/acs.est.9b05537>.
- Zeng, Z., Sun, T., Zhu, J., Huang, X., Yin, Z., Lu, G., Fan, Z., Yan, Q., Hng, H.H., Zhang, H., 2012. An effective method for the fabrication of few-layer-thick inorganic nanosheets. *Angew. Chem. Int. Ed.* 51 (36), 9052–9056.
- Zhang, X., Teng, S.Y., Loy, A.C.M., How, B.S., Leong, W.D., Tao, X., 2020. Transition metal dichalcogenides for the application of pollution reduction: a review. *Nanomaterials* 10 (6), 1012.
- Zou, W., Zhou, Q., Zhang, X., Hu, X., 2019. Dissolved oxygen and visible light irradiation drive the structural alterations and phototoxicity mitigation of single-layer molybdenum disulfide. *Environ. Sci. Technol.* 53 (13), 7759–7769.
- Zou, W., Wan, Z., Zhao, C., Zhang, G., Zhang, X., Zhou, Q., 2021. Impact of algal extracellular polymeric substances on the environmental fate and risk of molybdenum disulfide in aqueous media. *Water Res.* 205, 117708.

中国分析测试协会科学技术奖 CAIA 奖

证 书

为表彰2022年度中国分析测试协会科学技术奖 CAIA奖获奖者，特颁发此证书。

项目名称：纳米材料分离测定及其生物效应分析新方法

奖励等级：一等

获 奖 者：闫希亮

证书编号：2022-1-012-R02

国家科技奖励办公室登记证号：国科奖社证字第0032号



证书号第7834535号



专利公告信息

发明专利证书

发明名称：一种基于谱图分析的有机物生物毒性预测方法及系统

专利权人：广州大学

地址：510006 广东省广州市番禺区大学城外环西路230号

发明人：闫希亮;胡松;刘国红;颜嘉晨;周宏钰;周小霞;闫兵

专利号：ZL 2021 1 1270668.4

授权公告号：CN 114141316 B

专利申请日：2021年10月29日

授权公告日：2025年03月28日

申请日时申请人：广州大学

申请日时发明人：闫希亮;胡松;刘国红;颜嘉晨;周宏钰;周小霞;闫兵

国家知识产权局依照中华人民共和国专利法进行审查，决定授予专利权，并予以公告。

专利权自授权公告之日起生效。专利权有效性及专利权人变更等法律信息以专利登记簿记载为准。

局长
申长雨

申长雨



证书号第7060589号



专利公告信息

发明专利证书

发明名称：一种基于实验和计算的二维纳米复合物毒性评价方法

专利权人：广州大学

地址：510006 广东省广州市大学城外环西路230号

发明人：刘国红;闫希亮;李成俊;颜嘉晨;胡松;闫兵

专利号：ZL 2022 1 1358675.4

授权公告号：CN 116087190 B

专利申请日：2022年11月01日

授权公告日：2024年06月04日

申请日时申请人：广州大学

申请日时发明人：刘国红;闫希亮;李成俊;颜嘉晨;胡松;闫兵

国家知识产权局依照中华人民共和国专利法进行审查，决定授予专利权，并予以公告。
专利权自授权公告之日起生效。专利权有效性及专利权人变更等法律信息以专利登记簿记载为准。

局长
申长雨

申长雨



证书号第8055684号



专利公告信息

发明专利证书

发明名称：一种离子液体对乙酰胆碱酯酶的毒性预测方法及系统

专利权人：广州大学

地址：510006 广东省广州市大学城外环西路230号

发明人：颜嘉晨;闫希亮;刘国红;胡松;闫兵;何思源

专利号：ZL 2022 1 1092681.X

授权公告号：CN 115641489 B

专利申请日：2022年09月08日

授权公告日：2025年07月08日

申请日时申请人：广州大学

申请日时发明人：颜嘉晨;闫希亮;刘国红;胡松;闫兵;何思源

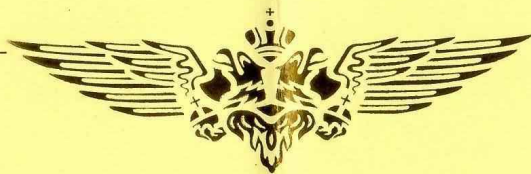
国家知识产权局依照中华人民共和国专利法进行审查，决定授予专利权，并予以公告。

专利权自授权公告之日起生效。专利权有效性及专利权人变更等法律信息以专利登记簿记载为准。

局长
申长雨

申长雨





荣誉证书

HONORARY CREDENTIAL

闫希亮博士：

因科研成果突出，被授予广州大学 2022 年“年度学术新锐”称号。

特发此证，以资鼓励。



中国毒理学会

Chinese Society of Toxicology

中国毒理学会

Chinese Society of Toxicology

证书

闫希亮

同志:

你在中国毒理学会第十次全国毒理学大会上报告的
论文《 基于机器学习似纳米材料与环境污染物复合毒性预测
》被评为优秀论文。特发此状，
以资鼓励。



二〇二三年四月十一日

中国毒理学会

Chinese Society of Toxicology

中国毒理学会

Chinese Society of Toxicology