

数据驱动的有机分子理化性质预测

孙一舟¹, 汤缪灵¹, 张硕卿^{1*}, 洪鑫^{1,2*}

1. 浙江大学化学系, 浙江大学化学前瞻技术研究中心, 杭州 310027

2. 河南师范大学化学化工学院, 新乡 453007

* 联系人, E-mail: angellasty@zju.edu.cn; hxchem@zju.edu.cn

2024-07-27 收稿, 2024-09-04 修回, 2024-09-18 接受, 2024-09-23 网络版发表

国家重点研发计划(2022YFA1504301)、国家自然科学基金(22122109, 22271253, 22103070)、浙江省自然科学基金(LDQ23B020002)、浙江大学上海高等研究院繁星科学基金(SN-ZJU-SIAS-006)、中国科学院青年交叉团队(JCTD-2021-11)、中央高校基本科研业务费专项资金(226-2022-00140, 226-2022-00224, 226-2023-00115, 226-2024-00003)、固体表面物理化学国家重点实验室项目(202210)、浙江省科技厅领军型创新创业团队项目(2022R01005)和河南师范大学化学化工学院开放研究基金(2024Z01)资助

摘要 分子的理化性质对于化学行为具有决定性影响, 其精确高效预测是化学和材料科学领域研究的长期热点之一. 随着理化性质数据的积累和分子人工智能(AI)建模方法的进步, 数据驱动分子理化性质预测迎来了跨越式发展, 在精度、广度、效率上取得了一系列突破. 在有机分子理化性质的AI预测方面, 本文介绍了相关数据库的发展, 围绕光谱性质、轨道能量、酸度系数(pK_a)、键解离能(BDE)、亲核性参数等代表性场景探讨了机器学习预测的相关进展, 并对该领域的现状进行了总结与展望.

关键词 分子性质预测, 分子数据库, 机器学习, AI化学

分子的理化性质, 如前线轨道能级、化学键能、光谱特征等, 构成了理解和预测分子化学行为的基础^[1,2]. 例如, 基于前线轨道的能级和轨道成分, 化学家能够有效地判断化学反应中的电子给体与受体, 并理性评估反应活性和选择性, 在包括周环反应在内的一系列有机分子转化中应用广泛^[3]. 化学键能提供了关于化学键切断与形成过程的核心热力学参数, 对于精确理解反应热力学及动力学行为至关重要^[4]. 而紫外、红外及核磁共振等谱学特征则为描述分子的立体结构与电子结构提供了有效的手段^[5]. 因此, 分子理化性质的精准认识, 是人类探索分子世界的重要途径之一, 不仅有助于深入理解化学反应的微观机制, 还能够有效指导新反应的设计与开发, 是化学科学研究的长期焦点之一^[2].

分子性质的经典获取途径依赖于实验测量和理论计算两大策略. 实验方法作为探索分子性质的第一手

途径, 为人类理解分子世界建立了可靠的数据基础^[6]. 然而, 实验测量往往受限于成本和效率的瓶颈, 对于样本的准备、精密测量技术的需求以及专业设备的依赖, 限制了分子性质实验测量的应用范围^[1,6]. 随着计算化学方法的不断发展以及计算机硬件性能的显著提升, 基于第一性原理的计算模拟已成为评估分子理化性质的有力补充^[7]. 这些方法通过模拟分子的电子结构和反应动力学, 拓展了分子性质研究的边界. 尽管如此, 今天的理论计算对于复杂分子体系同样也面临精度和效率难以兼得的问题, 在合成化学前沿的复杂体系中应用受限. 因此, 化学家不断追求更高效、经济的分子性质预测手段, 以适应社会发展对于功能分子日新月异的需求^[8,9].

随着化学数据的不断积累和人工智能技术的显著进步, 机器学习方法在化学中的应用取得了显著进展^[10~13], 在分子理化性质的预测上展现了重要潜

引用格式: 孙一舟, 汤缪灵, 张硕卿, 等. 数据驱动的有机分子理化性质预测. 科学通报, 2025, 70: 492–507

Sun Y-Z, Tang M, Zhang S, et al. Data-driven prediction of physicochemical properties of organic molecules (in Chinese). Chin Sci Bull, 2025, 70: 492–507, doi: [10.1360/TB-2024-0812](https://doi.org/10.1360/TB-2024-0812)

力^[14,15]。基于大规模的分子性质数据与创新的分子建模架构, 人工智能模型能够成功捕捉分子结构与理化性质之间的高维联系, 并据此准确预测分子性质^[16]。在QM9^[17]等代表性的分子性质数据集上, 前沿的机器学习模型不仅在预测精度上达到了媲美传统量化计算的水平, 同时在计算效率上实现了几何级数的增长^[18]。分子性质的智能预测为化学家认识和探索化学世界提供了全新的策略, 不仅将加速化学研究的步伐, 也为医药、材料等相关学科提供了新的助力^[8,19]。受限于篇幅, 本文将聚焦于有机分子的理化性质预测, 不涉及药化性质等生物活性分子的应用场景^[18,19], 从数据集、代表性场景等方面探讨该领域的前沿进展, 并就领域的研究现状和挑战进行总结和展望。

1 有机分子理化性质数据库与数据集

有机分子的结构复杂、性质多样, 共同组成了一个关于理化性质的庞大构效空间, 成为了机器学习建模的数据源泉。本节简要概述了代表性的有机分子理化性质数据库与数据集(表1)。在有机化学领域, 有许多数据库可以提供丰富的理化性质数据。例如, 由美国国家标准与技术研究院创建的NIST化学数据库(<http://webbook.nist.gov/chemistry>), 收集了实验测量数据、理论计算结果并整理归纳了已有文献, 包含超过7000种物质的热化学数据, 拥有超过8万不同的谱图数据等, 是化学领域著名的综合性数据库。由Nakata和Shimazaki^[20]所创建PubChemQC数据库(<http://nakatamaho.riken.jp/pubchemqc.riken.jp/>), 包含了基于B3LYP/6-31G*

水平密度泛函理论(DFT)计算的上百万个分子的基态电子结构与低位激发态数据, 构成了数据驱动化学研究的大规模第一原理电子结构数据库。由日本国立材料科学研究所创建的SDBS有机化合物光谱数据库(<http://sdb.sdb.aist.go.jp/>), 整理了3万多个有机分子的光谱数据, 是有机领域广泛应用的公开数据库。由清华大学程津培等人创建的iBonD数据库(<http://ibond.chem.tsinghua.edu.cn>), 系统整理了实验测定的化学键热力学参数, 是目前国际上化学键能领域综合性和收录数量均居首位的大型数据库, 包含2万多个有机化合物在各类溶剂中的3万多个酸度系数(pK_a)及7000多个R-H键的气相键解离能(BDE)参数。由Mobley和Guthrie^[21]创建的FreeSolv数据库(<http://www.escholarship.org/uc/item/6sd403pz>), 目前提供了数百种中性化合物的水合自由能的实验和计算数据。除了数据库以外, 一些代表性的理化性质数据集对机器学习在有机分子理化性质预测领域的应用亦发挥了关键性的促进作用。其中, 高精度的计算数据集作为实验数据的补充, 通常具有较大的数据规模。例如, QM9数据集包括了13万有机分子的计算几何、能量、电子和热力学性质^[17]; PubChemQC PM6数据集包含2.21亿个分子优化的几何形状和电子特性^[22]; Frag20数据集收录了超过150万个含有不超过20个重原子的分子的几何与电子特性数据, 进一步催生了高效的分子能量预测模型^[23]。

高质量有机分子理化性质的数据库和数据集为数据驱动的分子建模研究提供了宝贵的知识资源, 其中部分数据集已成为评估分子人工智能(AI)建模当前最

表1 代表性分子理化性质数据库与数据集

Table 1 Representative database of molecular physical and chemical properties

数据库名称	数据库网址	描述
NIST化学数据库	http://webbook.nist.gov/chemistry	包含7000多种物质的热化学数据、上万种化合物的近10万光谱数据等
PubChemQC ^[20]	http://nakatamaho.riken.jp/pubchemqc.riken.jp/	包含基于B3LYP/6-31G*级别的DFT的300万个分子的基态电子结构以及超过200万个分子的10个低位激发态
SDBS有机化合物光谱数据库	http://sdb.sdb.aist.go.jp/	包含3万多个有机分子的光谱数据, 红外光谱、核磁共振光谱、质谱和拉曼光谱等
iBonD	http://ibond.chem.tsinghua.edu.cn	包含约2万有机化合物的R-H键在各种溶剂中的约3万个平衡酸度和5000多个R-H键7600余个气相均裂能BDE
FreeSolv ^[21]	http://www.escholarship.org/uc/item/6sd403pz	包含600余种中性化合物实验和计算的水合自由能
数据集名称	数据集下载地址	描述
QM9 ^[17]	http://deepchemdata.s3-us-west-1.amazonaws.com/datasets/molnet_publish/qm9.zip	包含由C、H、O、N、F组成的13万多个稳定有机小分子的计算几何、能量、电子和热力学性质
PubChemQC PM6 ^[22]	http://nakatamaho.riken.jp/pubchemqc.riken.jp/pm6_datasets.html	包含2.21亿个分子优化的几何形状和电子特性
Frag20 ^[23]	http://yzhang.hpc.nyu.edu/IMA	包含多于150万个不超过20个重原子的分子的几何形状和电子特性

高水平(SOTA)的标准基准. 然而, 新的AI建模策略通常在不同数据集上进行基准测试, 这为评估新方法的质量带来了挑战. 2018年, MoleculeNet的发布标志着分子机器学习领域的一个基准的建立, 它统一管理了多个公共数据集并确立了评估标准, 详细阐述了包括QM7、QM7b、QM8、QM9、ESOL等在内的17个有机分子理化性质数据集的信息, 为数据集管理和基准衡量提供了一种可能的范例^[24]. 目前, 在广泛来源的数据支持下, 研究者能够评估和获取合适的数据集, 进而开展有机分子理化性质预测的研究. 尽管如此, 目前数据库和数据集的发展仍存在不足. 例如, 即使有计算数据集的数据补充, 公开可获得的实验数据的匮乏仍限制了部分研究的进展; 大多数商业大型数据库的底层数据仅通过一次一个查询的页面访问方式提供, 限制了机器学习方法的应用.

2 机器学习方法

自1956年达特茅斯会议首次提出人工智能概念, 至2010年深度学习技术的兴起, 机器学习方法持续演进与进步. 2022年, ChatGPT的推出进一步揭示了AI在众多领域的广泛应用潜力. 这些新兴的机器学习方法为有机分子理化性质的预测开辟了新的研究路径. 本节将概述有机分子理化性质预测中普遍采用的机器学习流程, 并介绍当前领域内一些代表性的分子编码、算法架构与机器学习策略.

2.1 机器学习流程简介

传统的判别式监督学习在有机分子理化性质预测中应用广泛, 通过在数据集上进行训练, 得到一个模型来联系数据集中的输入和输出, 进而建立分子结构与其理化性质之间的数学模型并用于新分子的性质预测. 图1描述了数据驱动理化性质预测的常见流程.

(1) 建立数据集: 首先, 根据理化性质目标, 需要收集、整理并清洗出相应的分子性质数据集作为建模的基础. 开源的分子数据库是常见的训练数据来源, 也可

以根据目标进一步通过实验或计算的手段补充数据. 同时可以考虑根据分子本身的化学特性, 进行数据增强^[25].

(2) 分子编码: 分子存在一维字符串、二维拓扑图、三维坐标等一系列不同但可以互通的表达形式^[26], 需要根据模型的需求转化为计算机能够识别的语言. 此外, 也可以根据用户指定的特征将分子单向编码为向量形式进行表达^[27]. 鉴于分子的原子键连关系与拓扑图天然的适配性, 分子图是当前分子性质建模预测中常用的编码架构之一^[28,29].

(3) 模型训练: 在数据集和分子编码的基础上, 即可进一步进行机器学习建模的训练. 线性模型, 支持向量机(SVM), 随机森林(RF)以及近年来兴起的大模型^[30]等架构在有机分子理化性质预测中都有涉及.

(4) 应用与分析: 基于训练好的模型, 需要在单独的测试集上评估模型的预测表现. 而该测试集的划分对于模型的科学评估至关重要, 往往需要从化学问题出发进行不同角度的集合划分^[31], 尤其是评估模型对于新分子的预测能力.

2.2 机器学习方法进展

机器学习的持续发展促进了有机分子理化性质预测的精确度和效率的显著提升. 目前, 多种机器学习模型已被应用于这一领域. 为了高效利用数据并提高模型的预测性能, 研究者探索了多种分子编码与机器学习算法^[2,27,32]. 本节简要概述领域前沿的分子编码、神经网络(NN)框架以及机器学习方法.

在分子编码方面, Weininger^[33]于1988年提出的简化分子线性输入规范(SMILES)至今仍为最广泛使用的分子编码之一. 近年来, 采用同一分子的多个SMILES表示来训练机器学习模型已成为一种增强数据、提升预测性能的常用策略^[25]; 分子指纹, 如扩展连接指纹(ECFP)和分子访问系统(MACCS)等基于序列的分子表示, 常作为传统机器学习模型的输入或与其他数据类型结合的辅助输入^[2]; 分子二维结构, 包括分子图和

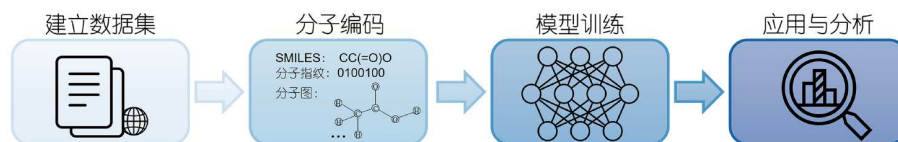


图1 (网络版彩色)分子性质机器学习建模的基本流程

Figure 1 (Color online) Workflow of machine learning modeling of molecular properties

分子图像数据等对机器学习极具价值,分子的拓扑结构可自然地以二维图形式表示,通常原子作为节点,键作为边,节点和边可携带表示原子类型、手性、键类型、方向和长度等特征的向量^[33];分子三维结构,如3D分子图和3D分子网格等提供了关于分子的详尽信息.目前,已开发出多种方法^[34,35]来构建3D几何的信息传递机制,以确保分子编码的图形表示遵循物理对称性,如平移和旋转的等变性.这一系列的分子编码提供了多种分子表示选择,支持了后续的建模以及预测任务的开展.

在算法架构方面,可以将常见的机器学习模型划分为传统机器学习与深度学习.传统机器学习,例如线性回归(LR)、核岭回归(KRR)、SVM、决策树(DT)等,虽然最早可以追溯到20世纪,但因其相对简单的流程和高效准确的预测性能,在有机分子理化性质预测领域仍具有广泛的应用.深度学习,尤其是神经网络模型,近年来受到了研究者的广泛关注^[36],包括卷积神经网络(CNN)、递归神经网络(RNN)、图神经网络(GNN)、Transformer等.深度学习架构在过去十年中迅速发展,为模型的创新、优势组合以及对模型细节的处理提供了新的思路和方法.例如,RNN有长短期记忆(LSTM)和门控循环单元(GRU)等变体;GNN衍生出的图卷积神经网络(GCN)等以及Transformer衍生出的BERT、Graphomer等在有机分子理化性质预测中应用广泛^[37~39].神经网络架构的多样性使其能够更好地适应工作任务的特定场景,与传统机器学习模型一起为有机分子理化性质预测的数据建模提供方法支持^[27].

在机器学习策略方面,为了更高效地利用数据并提升模型的预测性能,研究者发展了一系列策略方法,包括集成学习、迁移学习、对比学习、主动学习、多任务学习和置信度估计等.集成学习通过融合多种学习算法,实现了比单一算法更高的预测性能^[15];迁移学习策略涉及对预训练的模型进行调整,以便应用于新的相关任务,这种方法在实验数据不足的情况下尤为有用^[38];对比学习通过最大化相似输入之间的一致性来预训练模型,例如,通过比较同一分子的不同增强或描述符^[40];主动学习是一个迭代过程,通过不断从化学空间中识别有价值的数据来优化模型,从而在较少的标注数据下实现高效地探索和开发^[41];多任务学习策略同时训练多个任务,以提高模型的泛化能力^[42];置信度估计是对预测结果不确定性的量化,有助于评估模型的预测性能^[43].

3 具体机器学习模型在有机分子理化性质预测上的应用

围绕着分子理化性质的不同,其构效关系的化学内涵和复杂程度存在差异,且模型的输入输出结构也不尽相同.例如,在光谱性质预测中,通常采用神经网络模型,而在反应性相关的性质预测中,传统机器学习模型的应用则更广泛.近年来,围绕如光谱性质、轨道能量、 pK_a 、BDE、氧化还原电势、Mayr方程参数等代表性的有机分子理化性质,机器学习建模的精度和泛化能力取得了长足的进步.本节从应用场景、建模方法和模型精度等角度探讨其中的代表性工作.需要指出的是,所讨论的性质仅限有机化学中常用的理化性质,并非所有的分子本征性质.电子结构^[44]、原子化能^[45]等其他分子性质,同样有许多机器学习建模的精彩工作,限于本文的范畴不展开介绍.

3.1 光谱性质

分子光谱学,作为化学、物理和生物学研究中的核心领域之一,涵盖了红外光谱、拉曼光谱、紫外-可见吸收光谱、圆二色谱以及核磁共振谱等技术.这些光谱分析手段不仅能够揭示分子的能级结构,而且能够提供关于分子间相互作用和分子动态行为的深入见解^[46].这些信息对于深入理解分子的物理化学特性至关重要,对于化学合成、材料科学和生物医学研究等领域具有深远的影响^[47].随着机器学习技术的兴起,其在光谱性质预测中的应用,显著提升了数据处理效率和预测的准确性,为光谱学研究领域带来了新的活力.在红外光谱的预测方面,2020年,江俊课题组^[48]使用多层感知机(MLP),实现了蛋白质酰胺I带红外光谱的快速预测,平均绝对误差(MAE)仅为 1.761 cm^{-1} (图2(a));2024年,Hopkins课题组^[39]使用Graphormer,仅使用简化的SMILES字符串来预测红外光谱,平均测试光谱信息相似度(SIS μ)约为0.8449.除了直接预测红外光谱,一些研究则采取了不同的策略,通过预测分子的极化率来间接获取光谱性质,特别是拉曼光谱.2020年,Cari课题组^[49]用基于深度神经网络(DNN)的模型预测了液态水的极化率和拉曼光谱,并说明用相对少量的分子构型训练网络足以预测任意液体;2023年,Lewis等人^[50]基于三维电子响应的对称自适应学习(SALTER)预测分子和周期系统对均匀电场的真实空间密度响应,误差低于10%,并进一步导出了极化率张

量和拉曼光谱; 2024年, 何奕课题组^[51]使用基于神经进化势(NEP)的极化率模型, 在相似的小烷烃分子上训练极化率以预测较大烷烃的拉曼光谱, 证明了使用小分子有限训练数据准确有效地预测大分子拉曼光谱的潜力; 同年, 陈立朋课题组^[52]提出了一种高效的多任务机器学习替代模型——振动光谱神经网络(VSpecNN), 可根据通过机器学习增强的分子动力学模拟即时获得的偶极矩和极化率来准确计算红外光谱和拉曼光谱, 误差低于 5.87 cm^{-1} ; 同年, Komasa课题组^[53]使用预测张量属性的神经进化势(TNEP)模型预测了20种氨基酸的极化率, 通过与经典力场分子动力学相结合, 获得并研究了这些氨基酸的拉曼光谱(图2(b)).

除了伸缩振动光谱以外, 其他类型的光谱研究也有许多进展. 2020年, Mizoguchi课题组^[54]采用并扩展了之前报道的用于激发态机器学习的SchNarc方法, 并根据激发态能量和跃迁偶极矩计算紫外吸收光谱; 2021年, 江俊课题组^[55]使用嵌入原子神经网络(EANN), 比传统第一性原理快近4个数量级地模拟了蛋白质的圆二色谱, 与实验结果的Spearman等级相关性(ρ)不低于0.74; 2022年, Bande课题组^[56]对五种不同的GNN进行了基准测试和分析, 用于预测有机分子QM9数据集的激发光谱, 连续过滤卷积神经网络(SchNet)的表现最优, 相对标准误差(RSE)仅为0.023; 同年, Donadio课题组^[57]提出了一种通过量子化学和机器学习预测溶剂化

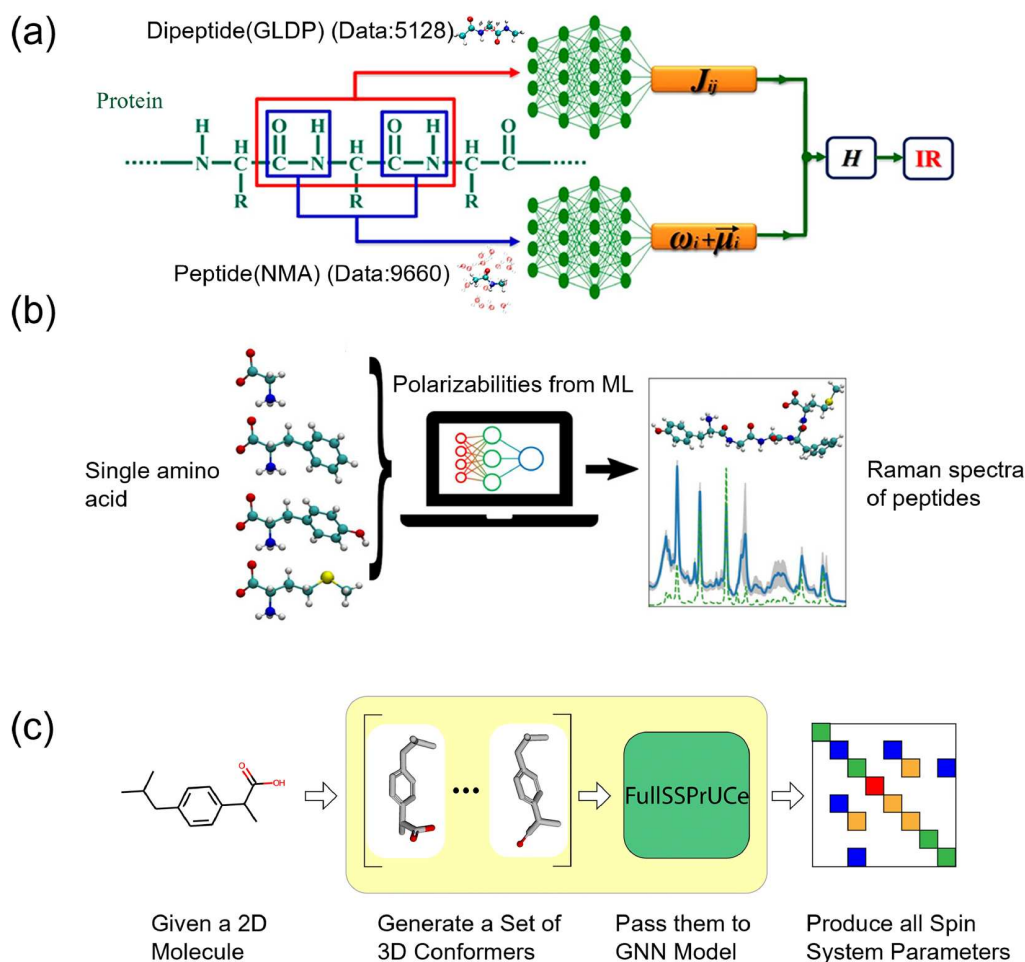


图2 (网络版彩色)光谱性质预测的部分最近进展. (a) NN模型快速预测酰胺I带红外谱图的工作流程^[48]. Copyright © 2020, American Chemical Society. (b) NN模型预测氨基酸的拉曼光谱^[53]. Copyright © 2024, American Chemical Society. (c) 全自旋系统精准预测的工作流程^[28]. Copyright © 2023, Royal Society of Chemistry

Figure 2 (Color online) Recent progress in the prediction of spectral properties. (a) Workflow of a NN model for rapid prediction of amide I IR spectra^[48]. Copyright © 2020, American Chemical Society. (b) NN model predicting the Raman spectra of amino acids^[53]. Copyright © 2024, American Chemical Society. (c) Workflow for accurate prediction in the full spin system^[28]. Copyright © 2023, Royal Society of Chemistry

芳香族分子的紫外可见吸收光谱的方法; 2023年, Williams和Jonas^[28]将不确定性感知深度学习与构象几何的快速估计相结合, 以实现具有不确定性的全自旋系统预测, 为溶液核磁共振谱的精确模拟提供支持, ^1H 和 ^{13}C 的化学位移预测精度分别达到0.209和1.213 ppm (图2(c)); 同年, Bultinck课题组^[58]采用全连接的前馈神经网络(FNN)实现了根据构象异构体几何形状预测构象异构体的振动圆二色谱, 较传统DFT计算模拟可以保持高于0.95阈值的相似性并提供6.6倍的加速; 同年, Bande课题组^[29]用特征归因来确定分子中各种原子对X射线吸收光谱(XAS)中观察到的峰的各自贡献, 将可解释性融入GNN模型, 用于XAS的预测; 同年, 胡伟及其合作者^[59]开发了DetaNet, 一种结合E(3)-等方差群和自注意力机制的深度学习模型, 可以高效率和高准确性地来预测分子光谱, 紫外吸收光谱MAE小于0.012 a.u., 核磁共振谱中 ^1H 和 ^{13}C 化学位移的MAE分别达0.054和0.5203 ppm.

3.2 分子轨道能量

分子的前线轨道, 特别是最高占据分子轨道(HOMO)和最低未占据分子轨道(LUMO), 对于分子的电子结构和化学活性至关重要. HOMO和LUMO轨道的能量水平是决定分子的亲电性和亲核性的关键, 这些轨道之间的能量差直接影响分子参与的化学反应类型和速率^[60]. 因此, 精确预测分子轨道能量对于理解分子的电子结构和化学性质, 以及设计新型分子实体具有重大意义^[61]. 近年来, 对于机器学习预测分子轨道能量的研究进展显著. 对于HOMO轨道能量的预测, 2019年, Rinke课题组^[62]用KRR在三个大型数据集专注预测HOMO轨道能量, 使用多体张量的描述符时, 预测HOMO能量的MAE低至0.09 eV; 2023年, Margraf课题组^[63]以有机分子中的HOMO能量为例探索了学习强化和局部属性的不同策略, 发现轨道加权平均(OWA)方法下, 所有机器学习权重与轨道位置的相关性都接近1, 该方法可被用于准确预测轨道能量和位置(图3(a)).

对于HOMO-LUMO能隙的预测, 2020年, Barnes及其合作者^[64]开发了一种CNN模型, 能够直接解析由空间点数据描述的分子的3D电子结构, 预测HOMO-LUMO能隙等理化性质, HOMO-LUMO能隙的均方根误差(RMSE)仅为0.351 eV; 2022年, Lilienfeld课题组^[31]通过划分训练集, 提升了HOMO-LUMO能隙的预测效率, MAE仅为0.1 eV; 同年, Choi等人^[37]对两个开

源大规模图数据集(PCQM4Mv2、AISD HOMO-LUMO)进行并行训练, 为HOMO-LUMO能隙这一重要量子属性构建GCN预测器, MAE仅为0.1 eV; 2024年, Patra课题组^[65]提出了分子快速响应(QR)码, 构建了基于分子QR码的卷积深度神经网络模型QRChEM, 预测了比热、焓、零点振动能和HOMO-LUMO能隙.

对于HOMO/LUMO轨道能量与能隙的综合预测, 2020年, Russo课题组^[66]证明机器学习方法可以利用DFT计算快速准确地估计重要的有机光伏(OPV)材料特性, 包括分子轨道能量与能隙等, HOMO、LUMO、HOMO-LUMO能隙的RMSE分别仅为0.21、0.23和0.30 eV(图3(b)); 2021年, 马晶及其合作者^[42]开发了一个名为DeepMoleNet的多级注意力神经网络(MLA-NN), 实现了化学精度范围内分子轨道能量与能隙等12个性质的预测, HOMO、LUMO、HOMO-LUMO能隙的MAE分别达到23.9、22.7和33.2 meV(图3(c)); 2022年, Tsai及其合作者^[67]使用LR、RF和SchNet方法来预测目标分子的分子轨道能量与能隙, SchNet表现最优, HOMO、LUMO、HOMO-LUMO能隙的MAE分别达到0.051、0.041和0.076 eV; 2023年, 余远斌课题组^[38]提出了一种深度迁移学习方法利用小数据来快速预测分子轨道能量与能隙, HOMO、LUMO、HOMO-LUMO能隙的MAE分别达到0.0774、0.0824和0.0549 eV.

3.3 pK_a

pK_a 是化学和生物化学领域中最重要基础数据之一, 是衡量化合物解离氢质子能力的热力学标度^[68]. 准确预测有机小分子的 pK_a 对于化学合成路径的规划、生物活性的预测以及药物设计的优化具有至关重要的作用^[14]. 同时, 对蛋白质 pK_a 值的预测有助于揭示其结构如何随pH变化而影响功能, 这是理解生物分子行为的关键^[69]. 随着技术的发展, 机器学习方法在准确预测有机分子 pK_a 值方面取得了显著进展, 极大地推动了相关领域的科学研究. 在小分子 pK_a 预测上, 2019年, Grzybowski课题组^[70]使用GCN预测了非水溶剂中有机分子内C-H的 pK_a . 在超过90%的情况下正确预测了最强的酸质子; 2020年, 程津培及其合作者^[68]基于iBond实验数据库, 采用基于结构和物理-有机参数的描述符建立了整体 pK_a 预测模型, NN和极端梯度提升(XGBoost)的机器学习模型表现最优, MAE值仅为0.87个 pK_a 单位(图4(a)); 2021年, Jang课题组^[71]将DFT与

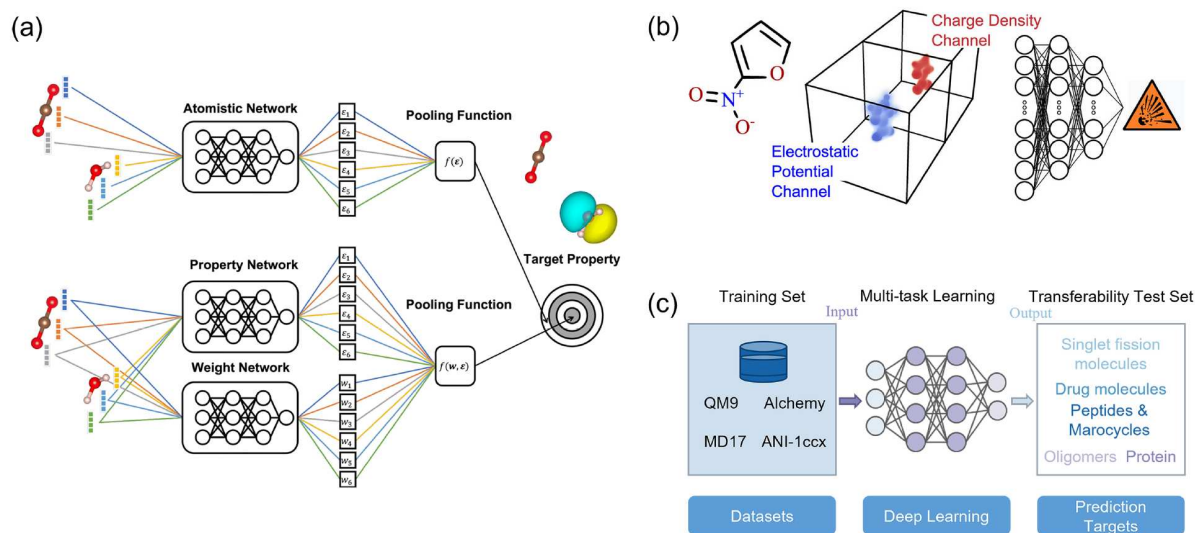


图3 (网络版彩色)分子轨道能量预测的部分最近进展。(a) 利用轨道加权平均方法预测轨道能量和位置的原子神经网络(AtNN)示意图^[63]。Copyright © 2023, Royal Society of Chemistry。(b) 机器学习方法快速准确地估计有机光伏材料特性^[66]。Copyright © 2020, Nastaran Meftahi et al。(c) DeepMoleNet实现分子性质的高效预测^[42]。Copyright © 2021, American Chemical Society

Figure 3 (Color online) Recent advances in molecular orbital energy prediction. (a) Schematic diagram of an atomic neural network using the orbital-weighted average method for predicting orbital energies and positions^[63]. Copyright © 2023, Royal Society of Chemistry. (b) Machine learning methods for rapid and accurate estimation of the characteristics of organic photovoltaic materials^[66]. Copyright © 2020, Nastaran Meftahi et al. (c) DeepMoleNet achieves efficient prediction of molecular properties^[42]. Copyright © 2021, American Chemical Society

KRR、高斯过程回归(GP)和人工神经网络(ANN)三种机器学习模型相结合,提出了一种新的预测多种酸中 pK_a 的方法,最低的MAE达到0.60个 pK_a 单位;同年,季长鸽课题组^[72]提出MolGpKa,一个使用GCN模型进行 pK_a 预测的网络服务器,MAE为0.43个 pK_a 单位;2022年,郑明月课题组^[73]编译了一个大规模 pK_a 数据集并在此基础上利用GNN建立了一种新颖的 pK_a 预测模型Graph- pK_a ,在宏观 pK_a 的测试数据集上MAE仅为0.55个 pK_a 单位;2023年,Johnston等人^[74]开发软件Epik7,使用GCN来预测复杂的药物分子的 pK_a 值和质子化状态分布,在七个测试集中预测 pK_a 值的MAE仅为0.42个 pK_a 单位;同年,Alcázar等人^[75]将低成本量子化学计算与定量构效关系(QSAR)方法和LR相结合,实现对各种含氮化合物 pK_a 的准确预测,标准误差为0.3066;2024年,邵学广课题组^[76]基于GNN和数据增强,提出了一种新的C-H酸 pK_a 值预测模型,MAE仅为2.03个 pK_a 单位;同年,邵学广课题组^[77]提出AttenGpKa,使用GNN进行溶剂化 pK_a 的预测,测试数据集上的RMSE仅为0.09;同年,Hutchison课题组^[78]开发QupKake,将GNN模型与半经验量子力学特征相结合以在微 pK_a 预测中实现卓越的准确性和泛化性,在五个外部测试集上的均方根误差在0.5~0.8个 pK_a 单位之间;同年,郑行课题组^[79]提出

Uni- pK_a 将热力学原理融入到机器学习建模中,实现了 pK_a 的高精度预测,在Novartis、SAMPL6和SAMPL7基准测试中,MAE分别为0.810、0.493和0.554个 pK_a ;同年,Moitessier课题组^[80]提出将化学知识传授给机器学习算法的策略并应用于 pK_a 的预测,MAE低于0.6。

近年来,基于深度学习的蛋白质 pK_a 预测进展显著。2021年,Kozłowski^[81]开发了一个使用深度学习和SVM模型的混合来预测等电点和 pK_a 值的网络服务器IPC2.0,对蛋白质和肽的均方根偏差(RMSD)分别为0.848与0.222个 pK_a 单位;同年,黄艳东课题组^[82]引入了一种基于深度学习的蛋白质 pK_a 预测器DeepKa,在2023年将该方法优化并应用于理论数据集PHMD549中,MAE仅为0.73个 pK_a 单位^[83],并在2024年开发了相应的网络服务器^[84];2022年,Isayev课题组^[85]将机器学习与原子环境矢量相结合,用神经网络势提供 pK_a 描述符,并使用RF模型进行训练以预测蛋白质 pK_a 值,对所有氨基酸类型MAE均低于0.5个 pK_a 单位(图4(b))。

3.4 BDE

BDE涉及化学键的均裂,是控制化学反应过程的核心本征要素。BDE的数值直接反映了化学键的强度,以及断裂该键所需的能量^[86]。因此,对BDE的准确预测

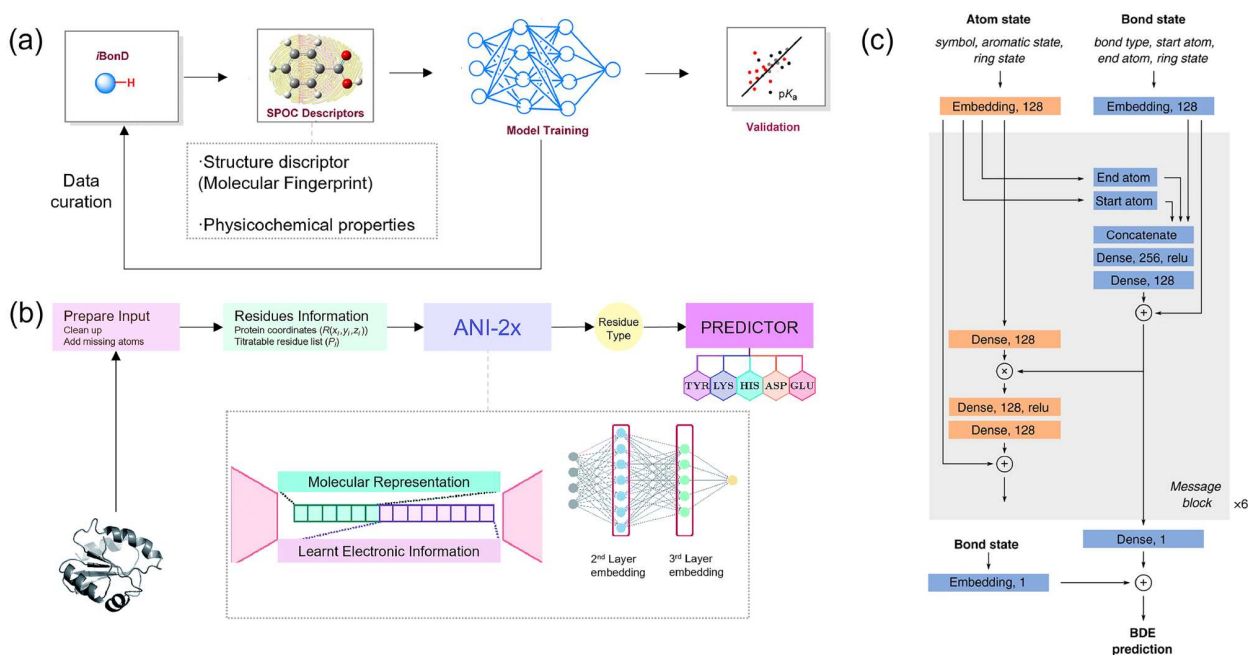


图 4 (网络版彩色)pK_a与BDE预测的部分最近进展. (a) 使用NN或XGBoost建立整体pK_a预测模型的工作流程^[68]. Copyright © 2020 Wiley-VCH GmbH. (b) 使用 ANI-2x 获得的神经网络特征进行蛋白质pK_a预测^[85]. Copyright © 2022, Royal Society of Chemistry. 使用NN或XGBoost建立整体pK_a预测模型的工作流程. Copyright © 2020 Wiley-VCH GmbH. (c) 预测BDE的GCN结构概述^[88]. Copyright © 2020, Peter C. St. John et al.

Figure 4 (Color online) Recent advances in pK_a and BDE prediction. (a) Workflow for establishing global pK_a prediction models using neural networks or XGBoost^[68]. Copyright © 2020 Wiley-VCH GmbH. (b) Prediction of protein pK_a using neural network features obtained from ANI-2x^[85]. Copyright © 2022, Royal Society of Chemistry. (c) Overview of the GCN structure for predicting BDE^[88]. Copyright © 2020, Peter C. St. John et al.

对于深入理解和评估分子在化学反应中的反应活性和稳定性具有极其重要的意义. 传统意义上, BDE的预测限于氢和周期表第二行的元素覆盖范围, 例如C-H键、C-O键、C-N键等. 在这一范畴内, 近年来有许多进展. 2019年, 江俊课题组^[87]通过使用原子种类和电荷信息作为描述符, 发展了一种利用NN快速预测有机分子键能的新技术, 其中表现最优的模型MAE仅为0.0425 eV; 2020年, Paton及其合作者^[88]开发了一种GCN模型, 能够在几分钟内准确预测有机分子的BDEs, MAE仅为0.0252 eV(图4(c)); 2021年, 张杰课题组^[89]介绍了一种名为片段基础图卷积神经网络(F-GCN)的通用定量结构-性质关系协议, 用于预测原子和原子间性质, 包括BDE等, C-H键的BDE估计的误差为0.117 eV; 同年, Persson课题组^[90]介绍了一种名为BonDNet的化学启发GNN机器学习模型, 用于快速准确地预测BDE, MAE达到0.022 eV, 显著优于化学精度(0.043 eV); 2023年, 王嵩及其合作者^[91]融合核磁共振和振动光谱的不同特征, 利用三层的ANN模型预测了有机分子中羟基的化学键性质, 包括键解离能、键长和 α -碳连接性, 对BDE预测的MAE为0.0539 eV; 2024年, Gelžinytė等

人^[92]探究了可转移机器学习原子间势在类药物分子键解离能预测中的应用, BDE预测的RMSE仅为0.0594 eV; 同年, 蒲雪梅课题组^[93]开发了一种高精度的含能材料(EM)键解离能机器学习预测模型, XGBoost模型实现了最佳预测精度, MAE仅为0.0912 eV; 同年, 乔娟课题组^[94]通过最小二乘法回归(OLS)、RF、XGBoost等机器学习模型, 实现了对具有饱和氮原子的杂环(HetSN)的环外C-N键键解离能的快速预测.

近年来, 研究者拓展BDE机器学习模型所预测化学键的类型范围, 也有一系列的成果. 2021年, Nemoto及其合作者^[95]将SMILES字符串转换为Avalon指纹并使用传统的弹性网络进行学习, 预测高价碘化合物的BDE, MAE达到0.0685 eV; 2023年, Paton及其合作者^[96]扩展了广泛的BDE数据库的大小、元素组成和键类型, 并训练了一个新的机器学习BDE模型, 其中包括S、Cl、F、P、Br和I等, 对C(sp²)-X碳卤键和C(sp³)-X碳卤键的预测MAE分别为0.0347和0.0520 eV; 2024年, 罗三中课题组^[97]基于*iBonD*实验BDE数据集和John等人计算的BDE数据集, 设计反映键均裂过程中分子结构和物理化学特征变化的微分结构和物理化学(D-

SPOC)描述符作为输入特征,建立了基于机器学习的综合BDE预测模型,MAE仅为0.0447 eV.

3.5 氧化还原电势

有机分子的氧化还原电势是决定其在多种应用领域性能的关键参数,这些应用领域包括传统的有机合成和催化过程,以及新兴的液流电池和太阳能电池技术^[98].氧化还原电势的精确预测不仅能够指导分子的设计,还能优化相关工艺的性能^[99].近年来,利用机器学习进行氧化还原电势预测的研究取得了显著成就.2019年,Aspuru-Guzik课题组^[100]结合量子化学计算和机器学习快速准确地预测羰基官能团还原为醇和胺的氧化还原电势,MAE可达0.02045 V;2020年,Jang课题组^[101]训练了ANN、梯度提升回归(GBR)和KRR这三种不同的机器学习模型,KRR的预测效果最佳,与DFT计算结果的平均相对误差达到3.94%;2021年,Zhang和Xu^[102]开发GP模型,根据电解质添加剂的分子结构特征来预测锂离子电池电解质添加剂的氧化还原电势,氧化电势和还原电势预测的MAE分别为0.1013和0.0522 V;2022年,Vanka课题组^[103]采用四种机器学习模型利用DFT来预测二甲醚中吩噻衍生物的氧化还原电位, R^2 大于0.74;同年,Arrigoni课题组^[104]使用基于机器学习的高效QSPR模型来预测蛋白质氧化还原电位,XGBoost模型的效果最优,误差低于1 kcal/mol;2023年,Mekonnen课题组^[105]开发了一种名为MolGAT的基于GNN的模型,用于使用分子结构、原子性质和键属性来预测有机分子的氧化还原电势,MAE达到0.033 V;同年,Tadger课题组^[106]尝试了KRR、DT、RF、极端随机树(ET)和GBR等多种机器学习模型预测醌的氧化还原电势, R^2 可达0.832;2024年,程俊课题组^[107]利用机器学习分子动力学加速了有机分子氧化还原电势的计算,MAE达0.33 V(图5(a));同年,Tognetti及其合作者^[108]建立了ORedOx159数据库,包含159个复杂结构的有机化合物,并应用基于图的机器学习方法进行氧化还原电势的预测,消息传递神经网络(MPNN)的预测效果最佳,对氧化电势与还原电势,MAE分别达到7.2和5.6 kcal/mol.

3.6 Mayr方程参数

在有机化学领域,化学反应速率与试剂的亲核性(N)和亲电性(E)之间的定量关系是一个重要的研究课题.为了描述这种关系,Mayr等人^[109]提出了一个适用

于亲核加成反应的量化公式,为亲核性和亲电性的定量评估和比较提供了一个标准化的分析框架.通过预测Mayr方程的参数,可以优化催化剂的选择和反应条件,进而提高反应的效率和产物的产率^[110].近年来,机器学习技术在预测Mayr方程参数方面取得了显著的进展.2020年,Gaüzère课题组^[111]评估了各种机器学习方法预测Mayr亲电性的能力,特别考虑了SVM与DT模型,在这项研究里,梯度提升决策树(GBDT)具有较好的准确性,MAE仅为0.72;2021年,Orlandi等人^[112]利用多元线性回归(MLR)分析方法来预测不同溶剂中化合物的亲核性参数,RMSE小于1.5;同年,Nguyen课题组^[113]报告了Mayr反应性参数数据库中四种最常见的亲核试剂中亲核性参数的机器学习预测,81.6%的预测在实验值的 ± 2.0 范围内;2022年,傅尧及其合作者^[114]提出一种基于SchNet的GNN模型有效整合分子结构信息和溶剂信息,预测Mayer亲核性参数,RMSE仅为1.63;2023年,Mora课题组^[115]使用Mayr的反应参数数据库创建了一个稳健的集成模型进行Mayr亲电性的预测,MAE仅为1.79;同年,程津培及其合作者^[115]通过机器学习方法开发了整体预测模型,使用ET算法训练的模型在预测Mayr的亲核性与亲电性参数方面表现出很高的准确性,N和E预测的MAE均为1.45(图5(b)).

3.7 有机分子理化性质预测应用汇总

过去五年中,机器学习技术广泛应用于预测有机分子的光谱性质、轨道能量、 pK_a 值、BDE值、氧化还原电位以及Mayr方程参数等理化性质,并取得了一系列重要研究成果,其中许多相关工作的代码已开源,为后续研究提供了宝贵的参考和借鉴(表S1).这些工作凸显了机器学习技术在有机分子理化性质预测领域的巨大潜力.与传统的理论化学计算相比,机器学习主导的性质预测通常具有较低的计算成本和更高的预测效率.例如,与传统的量子化学软件包相比,机器学习参与的光谱性质预测的计算效率可提升3~5个数量级^[59],机器学习预测模型在预测BDE时的效率远高于量子化学计算^[87,88,94,95].此外,端到端机器学习流程的自动化减少了人工干预,操作上更加便捷.预测精度方面,随着机器学习方法的不断进步,对有机分子理化性质预测的精度已接近量子化学计算,甚至在某些特定场景中超过了量子化学计算的精度^[116].结合量子化学计算与机器学习的混合方法,也展现出在降低计算成本的同时保持预测的准确性的应用价值.机器学习在平衡

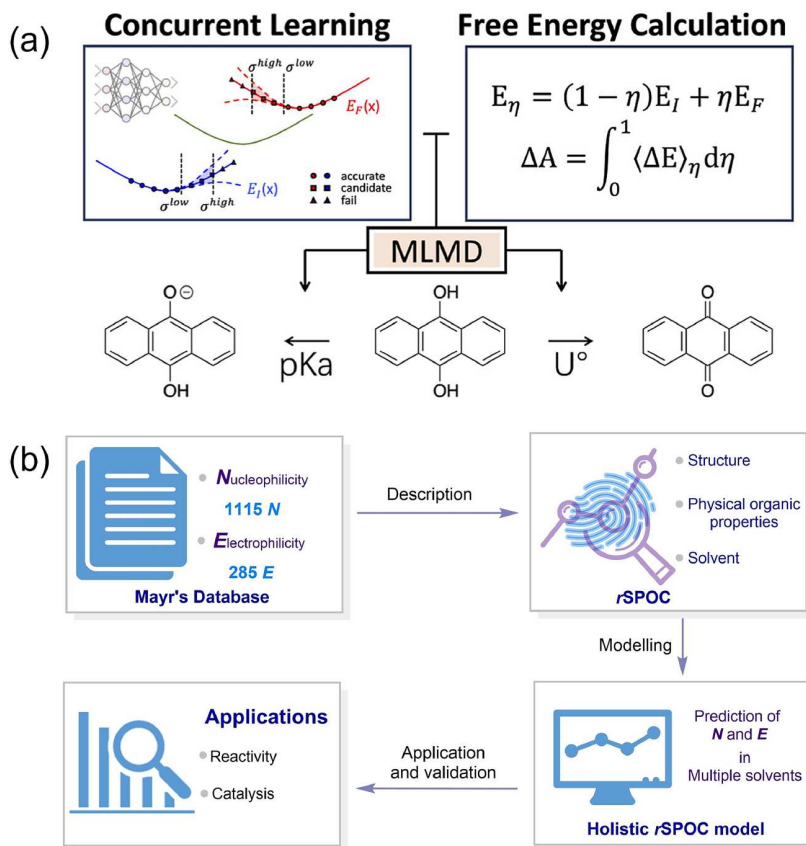


图5 (网络版彩色)氧化还原电势与Mayr方程参数预测的部分最近进展。(a) 机器学习加速氧化还原电势预测的流程^[107]。Copyright © 2024, American Chemical Society. (b) 通过机器学习研究亲核性和亲电性的一般工作流程^[115]。Copyright © 2023 Wiley - VCH GmbH

Figure 5 (Color online) Recent progress in predicting redox potentials and Mayer equation parameters. (a) The process of machine learning-accelerated prediction of redox potentials^[107]. Copyright © 2024, American Chemical Society. (b) A general workflow to study nucleophilicity and electrophilicity through machine learning^[115]. Copyright © 2023 Wiley - VCH GmbH

有机分子理化性质预测的精度与效率方面取得了显著成效,使得研究者能够从广泛的化学空间中快速、低成本地发现潜在的候选目标分子^[117]。

然而,目前机器学习在有机分子理化性质预测的应用中,仍然存在一些瓶颈。由于机器学习模型的预测性能受到训练数据的限制,模型的预测范围和精度通常受到制约。例如,在光谱性质预测中,通常只能对含有特定元素的分子进行建模,而对含有稀有官能团的分子进行预测时,精度会降低^[39];在分子轨道能量预测中,HOMO/LUMO对于大尺寸分子仍然是一个难题^[40];在 pK_a 与BDE的预测中,由于公开可获得的实验数据有限,因此难以获得足够的实验数据来支持模型的训练^[72,87]。模型的可解释性也是困扰研究者的一个问题。正如反应机理对于深入理解反应过程至关重要^[118-120],分子结构决定了分子的性质,机器学习模型通过分析

分子结构等化学信息来进行性质预测。然而,黑箱模型的结果可能超出了传统化学知识的解释范围,超参数的选择也往往难以用化学知识来判断。例如,在光谱性质预测中,FullSSPrUCe的研究工作中某些超参数的影响呈现出一些反常的结果^[28]。此外,模型架构的推广化、对分子环境的有效描述等可能成为该领域模型发展的潜在突破口。

4 总结与展望

近年来,随着化学数据的不断积累和人工智能技术的显著进步,数据驱动的有机分子理化性质预测经历了跨越式的发展。在光谱性质、轨道能量、 pK_a 、键解离能(BDE)、氧化还原电势以及Mayr参数等关键的分子性质方面,机器学习技术已经展示出了其卓越的预测能力。这些工作不仅能够部分例子上达到了与

传统量子化学计算相媲美的精度水平,而且在效率上实现了显著的提升,使得海量的分子性质预测成为了可能.这种新型的智能预测工具不仅将加速化学研究的步伐,而且为功能分子的设计和预测提供了强有力的支持,有望推动化学与材料科学的快速发展.

尽管分子性质的智能预测取得了显著进步,该领域仍面临一些关键的挑战和局限.首先,标准化且大规模的分子性质数据库的缺乏限制了机器学习模型的训练和验证.相对于理论上 10^{60} 数量级的分子化学空间,现有的数据量远远不足以覆盖广泛的化学多样性.其次,尽管机器学习模型在预测精度和效率上取得了一定的平衡,但这种平衡尚未达到完全可以取代传统的DFT计算的水平.特别是对于DFT难以处理的复杂体系,机器学习预测也难以胜任.最后,目前的预测建模工作与新颖功能分子设计的联系并不紧密,多数研究

停留在数据集本身的建模展示阶段,而未能实质性地应用于新功能分子的创造和设计中,因此在化学领域的影响力和共识还需进一步提升.

展望未来,为了充分利用机器学习在分子科学领域的潜力,可以从以下几个方向努力:一是构建更全面和标准化的分子性质数据库,以支持更复杂的模型训练和更广泛的化学空间探索.二是开发基于化学理论和原理的人工智能模型框架,而非仅仅对现有处理文字或图像的AI技术进行简单的修改.基于对化学过程本质的理解,将这些原理融合到算法中,以更精确地模拟和预测复杂的分子行为.三是加强预测模型与实验化学家的互动,将模型预测更直接地应用于新分子的设计和合成,以真正实现机器学习在化学创新中的应用.我们充分相信,在数据、建模和应用的协力发展下,分子科学将在AI时代迎来更大的突破.

参考文献

- 1 Fedik N, Zubatyuk R, Kulichenko M, et al. Extending machine learning beyond interatomic potentials for predicting molecular properties. *Nat Rev Chem*, 2022, 6: 653–672
- 2 Li Z, Jiang M, Wang S, et al. Deep learning methods for molecular representation and property prediction. *Drug Discov Today*, 2022, 27: 103373
- 3 Dannenberg J J. Using perturbation and frontier molecular orbital theory to predict diastereofacial selectivity. *Chem Rev*, 1999, 99: 1225–1242
- 4 Cheng G J, Zhang X, Chung L W, et al. Computational organic chemistry: Bridging theory and experiment in establishing the mechanisms of chemical reactions. *J Am Chem Soc*, 2015, 137: 1706–1725
- 5 Thomas N C. The early history of spectroscopy. *J Chem Educ*, 1991, 68: 631
- 6 Liu Q, Jiang Y, Zhang L, et al. A computational toolbox for molecular property prediction based on quantum mechanics and quantitative structure-property relationship. *Front Chem Sci Eng*, 2022, 16: 152–167
- 7 Louie S G, Chan Y H, da Jornada F H, et al. Discovering and understanding materials through computation. *Nat Mater*, 2021, 20: 728–735
- 8 Wei J, Chu X, Sun X, et al. Machine learning in materials science. *InfoMat*, 2019, 1: 338–358
- 9 López C. Artificial intelligence and advanced materials. *Adv Mater*, 2023, 35: 2208683
- 10 Li X, Zhang S, Xu L, et al. Predicting regioselectivity in radical C–H functionalization of heterocycles through machine learning. *Angew Chem Int Ed*, 2020, 59: 13253–13259
- 11 Li S W, Xu L C, Zhang C, et al. Reaction performance prediction with an extrapolative and interpretable graph model based on chemical knowledge. *Nat Commun*, 2023, 14: 3569
- 12 Xu L, Zhang S, Li X, et al. Towards data - driven design of asymmetric hydrogenation of olefins: Database and Hierarchical learning. *Angew Chem Int Ed*, 2021, 60: 22804–22811
- 13 Xu L C, Frey J, Hou X, et al. Enantioselectivity prediction of pallada-electrocatalysed C–H activation using transition state knowledge in machine learning. *Nat Synth*, 2023, 2: 321–330
- 14 Wu J, Kang Y, Pan P, et al. Machine learning methods for pK_a prediction of small molecules: Advances and challenges. *Drug Discov Today*, 2022, 27: 103372
- 15 Cuesta S A, Moreno M, López R A, et al. ElectroPredictor: An application to predict Mayr's electrophilicity E through implementation of an ensemble model based on machine learning algorithms. *J Chem Inf Model*, 2023, 63: 507–521
- 16 Hong X, Yang Q, Liao K, et al. AI for organic and polymer synthesis. *Sci China Chem*, 2024, 67: 2461–2496
- 17 Ramakrishnan R, Dral P O, Rupp M, et al. Quantum chemistry structures and properties of 134 kilo molecules. *Sci Data*, 2014, 1: 140022
- 18 Scalia G, Grambow C A, Pernici B, et al. Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction. *J Chem Inf Model*, 2020, 60: 2697–2717
- 19 Wang L, Zhou Z, Yang X, et al. The present state and challenges of active learning in drug discovery. *Drug Discov Today*, 2024, 29: 103985

- 20 Nakata M, Shimazaki T. PubChemQC Project: A large-scale first-principles electronic structure database for data-driven chemistry. *J Chem Inf Model*, 2017, 57: 1300–1308
- 21 Mobley D L, Guthrie J P. FreeSolv: A database of experimental and calculated hydration free energies, with input files. *J Comput Aided Mol Des*, 2014, 28: 711–720
- 22 Nakata M, Shimazaki T, Hashimoto M, et al. PubChemQC PM6: Data sets of 221 million molecules with optimized molecular geometries and electronic properties. *J Chem Inf Model*, 2020, 60: 5891–5899
- 23 Lu J, Xia S, Lu J, et al. Dataset construction to explore chemical space with 3D geometry and deep learning. *J Chem Inf Model*, 2021, 61: 1095–1104
- 24 Wu Z, Ramsundar B, Feinberg E N, et al. MoleculeNet: A benchmark for molecular machine learning. *Chem Sci*, 2018, 9: 513–530
- 25 Chen J H, Tseng Y J. Different molecular enumeration influences in deep learning: An example using aqueous solubility. *Brief BioInf*, 2021, 22: bbaa092
- 26 Zheng X, Tomiura Y. A BERT-based pretraining model for extracting molecular structural information from a SMILES sequence. *J Cheminform*, 2024, 16: 71
- 27 Dou B, Zhu Z, Merkurjev E, et al. Machine learning methods for small data challenges in molecular science. *Chem Rev*, 2023, 123: 8736–8780
- 28 Williams J, Jonas E. Rapid prediction of full spin systems using uncertainty-aware machine learning. *Chem Sci*, 2023, 14: 10902–10913
- 29 Kotobi A, Singh K, Höche D, et al. Integrating explainability into graph neural network models for the prediction of X-ray absorption spectra. *J Am Chem Soc*, 2023, 145: 22584–22598
- 30 Jablonka K M, Schwaller P, Ortega-Guerrero A, et al. Leveraging large language models for predictive chemistry. *Nat Mach Intell*, 2024, 6: 161–169
- 31 Mazouin B, Schöpfer A A, von Lilienfeld O A. Selected machine learning of HOMO–LUMO gaps with improved data-efficiency. *Mater Adv*, 2022, 3: 8306–8316
- 32 Townsend J, Micucci C P, Hymel J H, et al. Representation of molecular structures with persistent homology for machine learning applications in chemistry. *Nat Commun*, 2020, 11: 3230
- 33 Hu W, Liu B, Gomes J, et al. Strategies for pre-training graph neural networks. 2020, arxiv: 1905.12265v3
- 34 Satorras V G, Hoogeboom E, Welling M. E(n) equivariant graph neural networks. In: Proceedings of the 38th International Conference on Machine Learning, 2021. New York: Proceedings of Machine Learning Research, 2021. 9323–9332
- 35 Du W, Zhang H, Du Y, et al. SE(3) equivariant graph neural networks with complete local frames. In: Proceedings of the 39th International Conference on Machine Learning, 2022. New York: Proceedings of Machine Learning Research, 2022. 5583–5608
- 36 Zhang D, Xia S, Zhang Y. Accurate prediction of aqueous free solvation energies using 3D atomic feature-based graph neural network with transfer learning. *J Chem Inf Model*, 2022, 62: 1840–1848
- 37 Choi J Y, Zhang P, Mehta K, et al. Scalable training of graph convolutional neural networks for fast and accurate predictions of HOMO-LUMO gap in molecules. *J Cheminform*, 2022, 14: 70
- 38 Su A, Zhang X, Zhang C, et al. Deep transfer learning for predicting frontier orbital energies of organic materials using small data and its application to porphyrin photocatalysts. *Phys Chem Chem Phys*, 2023, 25: 10536–10549
- 39 Stienstra C M K, Hebert L, Thomas P, et al. Graphormer-IR: Graph transformers predict experimental IR spectra using highly specialized attention. *J Chem Inf Model*, 2024, 64: 4613–4629
- 40 Zhang R, Wu C, Yang Q, et al. MolFeSCue: Enhancing molecular property prediction in data-limited and imbalanced contexts using few-shot and contrastive learning. *Bioinformatics*, 2024, 40: btae118
- 41 Hao Z, Lu C, Huang Z, et al. ASGN: An active semi-supervised graph neural network for molecular property prediction. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020. New York: Association for Computing, 2020. 731–752
- 42 Liu Z, Lin L, Jia Q, et al. Transferable multilevel attention neural network for accurate prediction of quantum chemistry properties via multitask learning. *J Chem Inf Model*, 2021, 61: 1066–1082
- 43 Hirschfeld L, Swanson K, Yang K, et al. Uncertainty quantification using neural networks for molecular property prediction. *J Chem Inf Model*, 2020, 60: 3770–3780
- 44 Shao X, Paetow L, Tuckerman M E, et al. Machine learning electronic structure methods based on the one-electron reduced density matrix. *Nat Commun*, 2023, 14: 6281
- 45 Chu C, Xiao Q, He C, et al. A novel method for atomization energy prediction based on natural-parameter network. *Chin Chem Lett*, 2024, 35: 109186
- 46 Barone V, Alessandrini S, Biczysko M, et al. Computational molecular spectroscopy. *Nat Rev Methods Primers*, 2021, 1: 38
- 47 Teixeira dos Santos C A, Páscoa R N M J, Porto P A L S, et al. Raman spectroscopy for wine analyses: A comparison with near and mid infrared

- spectroscopy. *Talanta*, 2018, 186: 306–314
- 48 Ye S, Zhong K, Zhang J, et al. A machine learning protocol for predicting protein infrared spectra. *J Am Chem Soc*, 2020, 142: 19071–19077
- 49 Sommers G M, Calegari Andrade M F, Zhang L, et al. Raman spectrum and polarizability of liquid water from deep neural networks. *Phys Chem Chem Phys*, 2020, 22: 10592–10602
- 50 Lewis A M, Lazzaroni P, Rossi M. Predicting the electronic density response of condensed-phase systems to electric field perturbations. *J Chem Phys*, 2023, 159: 014103
- 51 Fang M, Tang S, Fan Z, et al. Transferability of machine learning models for predicting Raman spectra. *J Phys Chem A*, 2024, 128: 2286–2294
- 52 Chen Y, Pios S V, Gelin M F, et al. Accelerating molecular vibrational spectra simulations with a physically informed deep learning model. *J Chem Theor Comput*, 2024, 20: 4703–4710
- 53 Berger E, Niemelä J, Lampela O, et al. Raman spectra of amino acids and peptides from machine learning polarizabilities. *J Chem Inf Model*, 2024, 64: 4601–4612
- 54 Kiyohara S, Tsubaki M, Mizoguchi T. Learning excited states from ground states by using an artificial neural network. *npj Comput Mater*, 2020, 6: 68
- 55 Zhao L, Zhang J, Zhang Y, et al. Accurate machine learning prediction of protein circular dichroism spectra with embedded density descriptors. *JACS Au*, 2021, 1: 2377–2384
- 56 Singh K, Münchmeyer J, Weber L, et al. Graph neural networks for learning molecular excitation spectra. *J Chem Theor Comput*, 2022, 18: 4408–4417
- 57 Chen Z, Bononi F C, Sievers C A, et al. UV–visible absorption spectra of solvated molecules by quantum chemical machine learning. *J Chem Theor Comput*, 2022, 18: 4891–4902
- 58 Vermeyen T, Cunha A, Bultinck P, et al. Impact of conformation and intramolecular interactions on vibrational circular dichroism spectra identified with machine learning. *Commun Chem*, 2023, 6: 148
- 59 Zou Z, Zhang Y, Liang L, et al. A deep learning model for predicting selected organic molecular spectra. *Nat Comput Sci*, 2023, 3: 957–964
- 60 Woodward R B, Hoffmann R. The conservation of orbital symmetry. *Angew Chem Int Ed Engl*, 1969, 8: 781–853
- 61 Yu J, Su N Q, Yang W. Describing chemical reactivity with frontier molecular orbitals. *JACS Au*, 2022, 2: 1383–1394
- 62 Stuke A, Todorović M, Rupp M, et al. Chemical diversity in molecular orbital energy predictions with kernel ridge regression. *J Chem Phys*, 2019, 150: 204121
- 63 Chen K, Kunkel C, Cheng B, et al. Physics-inspired machine learning of localized intensive properties. *Chem Sci*, 2023, 14: 4913–4922
- 64 Casey A D, Son S F, Bilonis I, et al. Prediction of energetic material properties from electronic structure using 3D convolutional neural networks. *J Chem Inf Model*, 2020, 60: 4457–4473
- 65 Uthayakumar H, K R K, Jain R, et al. QRChEM: A deep learning framework for materials property prediction and design using QR codes. *ACS Eng Au*, 2024, 4: 91–98
- 66 Meftahi N, Klymenko M, Christofferson A J, et al. Machine learning property prediction for organic photovoltaic devices. *npj Comput Mater*, 2020, 6: 166
- 67 Ye Z R, Hung S H, Chen B, et al. Assessment of predicting frontier orbital energies for small organic molecules using knowledge-based and structural information. *ACS Eng Au*, 2022, 2: 360–368
- 68 Yang Q, Li Y, Yang J, et al. Holistic prediction of the pK_a in diverse solvents based on a machine-learning approach. *Angew Chem Int Ed*, 2020, 59: 19282–19291
- 69 Liu S, Yang Q, Zhang L, et al. Accurate protein pK_a prediction with physical organic chemistry guided 3D protein representation. *J Chem Inf Model*, 2024, 64: 4410–4418
- 70 Roszak R, Beker W, Molga K, et al. Rapid and accurate prediction of pK_a values of C–H acids using graph convolutional neural networks. *J Am Chem Soc*, 2019, 141: 17142–17149
- 71 Lawler R, Liu Y H, Majaya N, et al. DFT-Machine learning approach for accurate prediction of pK_a . *J Phys Chem A*, 2021, 125: 8712–8722
- 72 Pan X, Wang H, Li C, et al. MolGpka: A web server for small molecule pK_a prediction using a graph-convolutional neural network. *J Chem Inf Model*, 2021, 61: 3159–3165
- 73 Xiong J, Li Z, Wang G, et al. Multi-instance learning of graph neural networks for aqueous pK_a prediction. *Bioinformatics*, 2022, 38: 792–798
- 74 Johnston R C, Yao K, Kaplan Z, et al. Epik: pK_a and protonation state prediction through machine learning. *J Chem Theor Comput*, 2023, 19: 2380–2388
- 75 Alcázar J J, Misad Saide A C, Campodónico P R. Reliable and accurate prediction of basic pK_a values in nitrogen compounds: the pK_a shift in supramolecular systems as a case study. *J Cheminform*, 2023, 15: 90
- 76 An H, Liu X, Cai W, et al. Explainable graph neural networks with data augmentation for predicting pK_a of C–H acids. *J Chem Inf Model*, 2024, 64: 2383–2392

- 77 An H, Liu X, Cai W, et al. AttenGpKa: A universal predictor of solvation acidity using graph neural network and molecular topology. *J Chem Inf Model*, 2024, 64: 5480–5491
- 78 Abarbanel O D, Hutchison G R. QupKake: Integrating machine learning and quantum chemistry for micro- pK_a predictions. *J Chem Theor Comput*, 2024, 20: 6946–6956
- 79 Luo W, Zhou G, Zhu Z, et al. Bridging machine learning and thermodynamics for accurate pK_a prediction. *JACS Au*, 2024, 4: 3451–3465
- 80 Genzling J, Luo Z, Weiser B, et al. Leveraging our teacher's experience to improve machine learning: application to pK_a prediction. *ChemRxiv*, 2024, doi: 10.26434/chemrxiv-2024-bpd53
- 81 Kozlowski L P. IPC 2.0: Prediction of isoelectric point and pK_a dissociation constants. *Nucleic Acids Res*, 2021, 49: W285–W292
- 82 Cai Z, Luo F, Wang Y, et al. Protein pK_a prediction with machine learning. *ACS Omega*, 2021, 6: 34823–34831
- 83 Cai Z, Liu T, Lin Q, et al. Basis for accurate protein pK_a prediction with machine learning. *J Chem Inf Model*, 2023, 63: 2936–2947
- 84 Cai Z, Peng H, Sun S, et al. DeepKa Web Server: High-throughput protein pK_a prediction. *J Chem Inf Model*, 2024, 64: 2933–2940
- 85 Gokcan H, Isayev O. Prediction of protein pK_a with representation learning. *Chem Sci*, 2022, 13: 2462–2474
- 86 Blanksby S J, Ellison G B. Bond dissociation energies of organic molecules. *Acc Chem Res*, 2003, 36: 255–263
- 87 Feng C, Sharman E, Ye S, et al. A neural network protocol for predicting molecular bond energy. *Sci China Chem*, 2019, 62: 1698–1703
- 88 St. John P C, Guan Y, Kim Y, et al. Prediction of organic homolytic bond dissociation enthalpies at near chemical accuracy with sub-second computational cost. *Nat Commun*, 2020, 11: 2328
- 89 Gao P, Zhang J, Qiu H, et al. A general QSPR protocol for the prediction of atomic/inter-atomic properties: A fragment based graph convolutional neural network (F-GCN). *Phys Chem Chem Phys*, 2021, 23: 13242–13249
- 90 Wen M, Blau S M, Spotte-Smith E W C, et al. BondNet: A graph neural network for the prediction of bond dissociation energies for charged molecules. *Chem Sci*, 2021, 12: 1858–1868
- 91 Guo S, Jiang J, Ren H, et al. Fusion of multiple spectra for investigating chemical bonding properties via machine learning. *J Phys Chem Lett*, 2023, 14: 7461–7468
- 92 Gelžinytė E, Öeren M, Segall M D, et al. Transferable machine learning interatomic potential for bond dissociation energy prediction of drug-like molecules. *J Chem Theor Comput*, 2024, 20: 164–177
- 93 Gou Q, Liu J, Su H, et al. Exploring an accurate machine learning model to quickly estimate stability of diverse energetic materials. *iScience*, 2024, 27: 109452
- 94 Meng Q Y, Wang R, Shao H Y, et al. Precise regulation on the bond dissociation energy of exocyclic C–N bonds in various n-heterocycle electron donors via machine learning. *J Phys Chem Lett*, 2024, 15: 4422–4429
- 95 Nakajima M, Nemoto T. Machine learning enabling prediction of the bond dissociation enthalpy of hypervalent iodine from SMILES. *Sci Rep*, 2021, 11: 20207
- 96 S. V. S S, Kim Y, Kim S, et al. Expansion of bond dissociation prediction with machine learning to medicinally and environmentally relevant chemical space. *Digital Discov*, 2023, 2: 1900–1910
- 97 Liu Y, Li Y, Yang Q, et al. Prediction of bond dissociation energy for organic molecules based on a machine-learning approach. *Chin J Chem*, 2024, 42: 1967–1974
- 98 Francke R, Little R D. Redox catalysis in organic electrosynthesis: Basic principles and recent developments. *Chem Soc Rev*, 2014, 43: 2492–2521
- 99 Fedorov R, Gryn'ova G. Unlocking the potential: Predicting redox behavior of organic molecules, from linear fits to neural networks. *J Chem Theor Comput*, 2023, 19: 4796–4814
- 100 Jinich A, Sanchez-Lengeling B, Ren H, et al. A mixed quantum chemistry/machine learning approach for the fast and accurate prediction of biochemical redox potentials and its large-scale application to 315000 redox reactions. *ACS Cent Sci*, 2019, 5: 1199–1210
- 101 Allam O, Kuramshin R, Stoichev Z, et al. Molecular structure–redox potential relationship for organic electrode materials: Density functional theory–machine learning approach. *Mater Today Energy*, 2020, 17: 100482
- 102 Zhang Y, Xu X. Machine learning properties of electrolyte additives: A focus on redox potentials. *Ind Eng Chem Res*, 2021, 60: 343–354
- 103 Ghule S, Dash S R, Bagchi S, et al. Predicting the redox potentials of phenazine derivatives using DFT-assisted machine learning. *ACS Omega*, 2022, 7: 11742–11755
- 104 Galuzzi B G, Mirarchi A, Viganò E L, et al. Machine learning for efficient prediction of protein redox potential: The flavoproteins case. *J Chem Inf Model*, 2022, 62: 4748–4759
- 105 Chaka M D, Geffè C A, Rodriguez A, et al. High-throughput screening of promising redox-active molecules with MolGAT. *ACS Omega*, 2023, 8: 24268–24278
- 106 Kichev I, Borislavov L, Tadjer A, et al. Machine learning prediction of the redox activity of quinones. *Materials*, 2023, 16: 6687
- 107 Wang F, Ma Z, Cheng J. Accelerating computation of acidity constants and redox potentials for aqueous organic redox flow batteries by machine

- learning potential-based molecular dynamics. *J Am Chem Soc*, 2024, 146: 14566–14575
- 108 Jia L, Brémond É, Zaida L, et al. Predicting redox potentials by graph-based machine learning methods. *J Comput Chem*, 2024, 45: 2383–2396
- 109 Mayr H, Patz M. Scales of nucleophilicity and electrophilicity: A system for ordering polar organic and organometallic reactions. *Angew Chem Int Ed Engl*, 1994, 33: 938–957
- 110 Zhuo L, Liao W, Yu Z. A frontier molecular orbital theory approach to understanding the Mayr equation and to quantifying nucleophilicity and electrophilicity by using HOMO and LUMO energies. *Asian J Org Chem*, 2012, 1: 336–345
- 111 Hoffmann G, Balcilar M, Tognetti V, et al. Predicting experimental electrophilicities from quantum and topological descriptors: A machine learning approach. *J Comput Chem*, 2020, 41: 2124–2136
- 112 Orlandi M, Escudero-Casao M, Licini G. Nucleophilicity prediction via multivariate linear regression analysis. *J Org Chem*, 2021, 86: 3555–3564
- 113 Boobier S, Liu Y, Sharma K, et al. Predicting solvent-dependent nucleophilicity parameter with a causal structure property relationship. *J Chem Inf Model*, 2021, 61: 4890–4899
- 114 Nie W, Liu D, Li S, et al. Nucleophilicity prediction using graph neural networks. *J Chem Inf Model*, 2022, 62: 4319–4328
- 115 Liu Y, Yang Q, Cheng J, et al. Prediction of nucleophilicity and electrophilicity based on a machine-learning approach. *ChemPhysChem*, 2023, 24: e202300162
- 116 Zubatyuk R, Smith J S, Leszczynski J, et al. Accurate and transferable multitask prediction of chemical properties with an atoms-in-molecules neural network. *Sci Adv*, 2019, 5: eaav6490
- 117 Heid E, Greenman K P, Chung Y, et al. Chemprop: A machine learning package for chemical property prediction. *J Chem Inf Model*, 2024, 64: 9–17
- 118 Liu J R, Xu G X, Liu L G, et al. Recent advances in theoretical studies on Cu-mediated bond formation mechanisms involving radicals. *ACS Catal*, 2024, 14: 2429–2454
- 119 Wu H, Zhang S Q, Hong X. Mechanisms of nickel-catalyzed reductive cross-coupling reactions. *Chem Synth*, 2023, 3: 39
- 120 Qian G, Zhu L, Zeng Y, et al. Mechanism and origins of weak bonding-controlled selectivities in cinchoninium-catalyzed umpolung michael addition of imines. *CCS Chem*, 2024, doi: 10.31635/ccschem.024.202404401

补充材料

表 S1 代码开源的有机分子理化性质预测近5年代表性工作总结

本文以上补充材料见网络版csb.scichina.com. 补充材料为作者提供的原始数据, 作者对其学术质量和内容负责.

Summary for “数据驱动的有机分子理化性质预测”

Data-driven prediction of physicochemical properties of organic molecules

Yi-Zhou Sun¹, Miaojiang Tang¹, Shuoqing Zhang^{1*} & Xin Hong^{1,2*}

¹ Center of Chemistry for Frontier Technologies, Department of Chemistry, Zhejiang University, Hangzhou 310027, China

² School of Chemistry and Chemical Engineering, Henan Normal University, Xinxiang 453007, China

* Corresponding authors, E-mail: angellasty@zju.edu.cn; hxchem@zju.edu.cn

The physicochemical properties of molecules, such as frontier orbital energy levels, bond energies, and spectroscopic characteristics, form the basis for understanding and predicting molecular chemical behavior. For instance, based on the energy levels and arrangement of frontier orbitals, chemists can effectively determine electron donors and acceptors in chemical reactions, and rationally assess reaction activity and selectivity, widely applied in a series of organic molecular transformations including pericyclic reactions. Chemical bond energies provide core thermodynamic parameters for the process of bond breaking and formation, which are crucial for precisely understanding reaction thermodynamics and kinetics. UV, IR, and NMR spectroscopic features offer effective means to describe the geometric and electronic structures of molecules. Therefore, an accurate understanding of molecular physicochemical properties is one of the important strategies for humans to explore the molecular world, which not only helps in deeply understanding the microscopic mechanisms of chemical reactions but also effectively guides the design and development of new reactions, making it a long-term focus of chemical scientific research.

Traditionally, molecular properties were determined through experimental measurements and theoretical calculations. While experiments provide a reliable foundation, they are often constrained by cost, efficiency, and the need for specialized equipment. Computational chemistry, bolstered by advancements in computer hardware, has emerged as a complementary approach, extending the reach of molecular property research through simulations of electronic structure and reaction dynamics. However, computational methods also face challenges in achieving precision and efficiency in complex molecular systems.

With the continuous accumulation of chemical data and significant advancements in artificial intelligence (AI) technology, machine learning methods have made remarkable progress in chemistry, showing important potential in the prediction of molecular physicochemical properties. Based on large-scale molecular property data and innovative molecular modeling architectures, AI models can successfully capture the high-dimensional relationship between molecular structure and physicochemical properties, and accurately predict molecular properties accordingly. On representative molecular property datasets like QM9, cutting-edge machine learning models have achieved prediction accuracy comparable to traditional quantum calculations, while also achieving geometric progression in computational efficiency. Intelligent prediction of molecular properties provides a new strategy for chemists to understand and explore the chemical world, which will not only accelerate the pace of chemical research but also offer new assistance to related disciplines such as pharmaceuticals and materials.

This paper mainly introduces the recent progress in AI prediction of the physicochemical properties of organic molecules. The first part outlines the progress of related databases and datasets, including NIST, PubChemQC, SDBS, iBonD, FreeSolv, QM9, PubChemQC PM6, and Frag20, detailing their content and scope. The second part describes the basic workflow of machine learning modeling for molecular properties, outlining the specific steps of dataset establishment, molecular encoding, model training, application, and analysis, along with recent progress in machine learning methods. The third part discusses the relevant progress in machine learning prediction for representative scenarios such as spectroscopic properties, orbital energies, acid dissociation constant (pK_a) values, bond dissociation energies (BDEs), and nucleophilicity parameters. In these key aspects of molecular properties, machine learning technology has demonstrated exceptional ability in terms of precision and efficiency. Finally, the paper summarizes the current state of the field, discusses some of the key challenges and limitations faced by the field, and looks forward to the prospects for field development from three directions: constructing more comprehensive and standardized molecular property databases, promoting the interaction between prediction models and experimental chemists, and enhancing the interpretability and generalizability of AI models.

molecular property prediction, molecular databases, machine learning, AI for chemistry

doi: [10.1360/TB-2024-0812](https://doi.org/10.1360/TB-2024-0812)