工程科学学报
Chinese Journal of Engineering

机器学习助力发现高效缓蚀剂分子

龚海燕 马菱薇 张达威

Machine learning aids in the discovery of efficient corrosion inhibitor molecules

GONG Haiyan, MA Lingwei, ZHANG Dawei

引用本文:
龚海燕, 马菱薇, 张达威. 机器学习助力发现高效缓蚀剂分子[J]. 北科大：工程科学学报, 2025, 47(6): 1228–1239. doi: 10.13374/j.issn2095–9389.2024.10.31.002
GONG Haiyan, MA Lingwei, ZHANG Dawei. Machine learning aids in the discovery of efficient corrosion inhibitor molecules[J]. *Chinese Journal of Engineering*, 2025, 47(6): 1228–1239. doi: 10.13374/j.issn2095–9389.2024.10.31.002

在线阅读 View online: https://doi.org/10.13374/j.issn2095–9389.2024.10.31.002

工程科学学报, 第 47 卷, 第 6 期: 1228−1239, 2025 年 6 月
Chinese Journal of Engineering, Vol. 47, No. 6: 1228−1239, June 2025
https://doi.org/10.13374/j.issn2095-9389.2024.10.31.002; http://cje.ustb.edu.cn

机器学习助力发现高效缓蚀剂分子

龚海燕[1,2,3,4)]，马菱薇[1,2)]，张达威[1,2,4)]✉

1) 北京科技大学北京材料基因工程高精尖创新中心，北京 100083　2) 北京科技大学国家材料腐蚀与防护科学数据中心，北京 100083
3) 北京科技大学顺德创新学院，佛山 528399　4) 北京科技大学新材料技术研究院，北京 100083
✉通信作者，E-mail: dzhang@ustb.edu.cn

摘　要　缓蚀剂是用于防止金属材料腐蚀的化学物质，其有效性对于延长设备寿命、降低维护成本至关重要. 然而，传统的缓蚀剂分子筛选方法，如失重测量和电化学测试，通常需要大量实验和大量时间，成本高昂. 基于机器学习技术可以分析已知缓蚀剂分子数据，从而学习和预测新分子的缓蚀性能. 该方法可以提高筛选效率，揭示传统方法可能忽略的分子结构和性质，但其局限性也不容忽视. 首先，缓蚀剂分子筛选模型的化合物搜索空间有限. 其次，模型在实际应用中面临着与计算资源和时间成本相关的挑战. 在讨论了机器学习技术的应用和局限性之后，本文介绍了分子生成技术在发现新的高效缓蚀剂分子方面的应用以及挑战. 例如，生成模型需要大量高质量数据进行训练，生成的结果需要实验验证. 此外，生成模型在生成新分子时必须考虑分子稳定性、可合成性、环境影响等多种因素，使得模型的设计和优化更加复杂. 总体而言，机器学习技术在缓蚀剂分子研究中具有广阔的应用前景，但也面临着重大挑战. 通过不断优化机器学习算法并结合实验验证，有望在未来实现缓蚀剂分子的高效高精度发现，从而为材料科学和工业应用带来突破.

关键词　机器学习；深度学习；缓蚀剂分子；分子筛选；分子生成
分类号　TG174.42

Machine learning aids in the discovery of efficient corrosion inhibitor molecules

GONG Haiyan[1,2,3,4)], *MA Lingwei*[1,2)], *ZHANG Dawei*[1,2,4)]✉

1) Beijing Advanced Innovation Center for Materials Genome Engineering, University of Science and Technology Beijing, Beijing 100083, China
2) National Materials Corrosion and Protection Data Center, University of Science and Technology Beijing, Beijing 100083, China
3) Shunde Innovation School, University of Science and Technology Beijing, Foshan 528399, China
4) Institute for Advanced Materials and Technology, University of Science and Technology Beijing, Beijing 100083, China
✉Corresponding author, E-mail: dzhang@ustb.edu.cn

ABSTRACT　In recent years, machine learning (ML) has demonstrated significant potential in corrosion inhibitor molecule research and has emerged as a powerful tool for scientists to explore new and efficient corrosion inhibitors. Corrosion inhibitors are chemical substances used to prevent the corrosion of metallic materials, and their effectiveness is crucial for extending equipment lifespan and reducing maintenance costs. However, traditional methods for screening corrosion inhibitor molecules, such as weight loss measurements and electrochemical testing, typically require extensive experiments and considerable time, making them costly. Consequently, the application of ML technology in this field has garnered widespread attention. This review provides an overview of the application of ML technology in screening corrosion inhibitor molecules. Artificial intelligence technologies, particularly deep learning and machine learning, can analyze vast amounts of data on known corrosion inhibitor molecules, to learn and predict the corrosion inhibition performance of new molecules. These technologies not only enhance screening efficiency but also uncover molecular

收稿日期: 2024−10−31
基金项目: 国家重点研发资助项目（2024ZD0607500）；广东省基础与应用基础研究基金资助项目（2021B1515130009）；国家资助博士后研究人员计划资助项目（GZC20230239）；中国博士后科学基金资助项目（2023M740219）

structures and properties that traditional methods may overlook. Specifically, ML models can extract key information and construct predictive models through feature extraction and pattern recognition using existing data. These models can rapidly identify potential high-efficiency corrosion inhibitor molecules, thereby significantly accelerating research. However, despite the numerous advantages of ML technology in screening corrosion inhibitor molecules, its limitations cannot be ignored. First, the current compound search space for corrosion inhibitor molecule screening models remains limited. Second, these models face challenges related to computational resources and time costs in practical applications. After discussing the applications and limitations of ML technology, this study further explores the concept of molecular generation technology and its application in generating corrosion inhibitor molecules. Molecular generation technology employs deep learning techniques for automatically generating new molecular structures, often based on generative models such as generative adversarial networks (GANs) and variational autoencoders (VAEs). These technologies can learn the rules of molecular generation from existing corrosion inhibitor molecule data and generate new molecules with specific properties. Molecular generation technology can help researchers discover new and efficient corrosion-inhibitor molecules and accelerate the development of new materials. Finally, this paper highlights the challenges faced by generative machine learning models in the discovery of efficient corrosion inhibitor molecules. Although generative models have shown great potential for molecule generation and screening, their application in the discovery of corrosion inhibitors still faces many challenges. For example, generative models require large amounts of high-quality data for training, and the generated results require experimental validation. Moreover, when generating new molecules, generative models must consider various factors, such as molecular stability, synthesizability, and environmental impact, making the design and optimization of these models more complex. Overall, ML technology holds broad application prospects in the research on corrosion inhibitor molecules; however, it also faces significant challenges. Continuously optimizing ML algorithms and combining them with experimental validation should contribute to the efficient and high-precision discovery of corrosion inhibitor molecules in the future, leading to breakthroughs in materials science and industrial applications.

KEY WORDS machine learning; deep learning; corrosion inhibitor molecules; molecular screening; molecular generation

腐蚀是导致金属材料性能降低和结构破坏的主要原因之一, 在世界范围内造成巨大的经济损失. 例如, 李晓刚团队在 2017 年提到全球每年因腐蚀而损失的 GDP 高达 2.5 万亿美元, 约占全球 GDP 的 3.4%[1]. 减缓金属材料腐蚀的方法包括使用防护涂层、阳极保护、阴极保护、耐蚀合金设计、添加缓蚀剂等. 其中缓蚀剂是抑制金属腐蚀的一种有效方法, 通过添加化学物质到腐蚀介质中, 形成保护膜或改变腐蚀过程, 从而减缓材料的腐蚀速率, 具有成本低、操作简单、效率高等优点[2].

缓蚀剂的有效性通常通过缓蚀效率(Inhibition efficiency, IE)来评估, IE 值越高表明缓蚀剂的腐蚀抑制效果越好, 且与分子结构、浓度以及金属底物和腐蚀环境的变化密切相关[3]. 传统的实验评估 IE 的方法, 如失重测量[4]、电化学测试[5], 虽然广泛应用, 但通常需要在多个实验条件下逐一测定特定浓度下缓蚀剂的 IE. 这些方法不仅耗时且成本高昂, 特别是在需要从广泛的化学空间中筛选出高效缓蚀剂及其合适浓度时, 实验的高消耗和低效率成为了一大瓶颈. 除了实验方法, 理论工具, 如密度泛函理论(Density functional theory, DFT)计算和分子动力学(Molecular dynamics, MD)模拟, 已广泛应用于缓蚀剂的研究[6−8]. DFT 计算方法提供了缓蚀剂分子与金属表面之间电荷共享(Donor-acceptor)相互作用的重要信息[9], 从而描述了缓蚀剂结构性质对腐蚀过程的影响[10−11]. 但是, DFT 计算得到的量子化学参数众多, 仅靠理论筛选缓蚀剂耗时且无法保证准确性. MD 计算机模拟用于模拟缓蚀剂/表面系统, 可视化吸附过程, 并确定其相互作用的能量, 以阐明介观水平(1 到 100 纳米)、原子和分子水平上的缓蚀机制[12−16]. 但其计算成本高, 且难以处理大规模的数据集, 限制了其在广泛筛选中的应用.

随着人工智能技术的发展, 机器学习(Machine learning, ML)尤其在缓蚀剂分子筛选、腐蚀机理研究[17−18] 等领域显示出了巨大的潜力. ML 算法通过分析大量已有数据, 能够识别缓蚀剂分子特征与其缓蚀性能之间的复杂关系, 从而挖掘出传统方法难以发现的规律. 与传统方法相比, ML 不仅能够处理高维数据, 还能综合分析多种因素(例如环境条件、缓蚀剂浓度、分子结构特征等)对缓蚀性能的影响. 因此, ML 为缓蚀剂性能预测提供了一种高效、经济且创新的方法, 推动了缓蚀剂研究和应用的发展. 例如多元线性回归分析[19−21]、神经网络[22−25]、支持向量机[26−27]、随机森林[28]、K-近邻算法[29] 等机器学习技术[30−32] 被应用于构建定量结构–

活性/性能关系(Quantitative structure-activity/property relationships, QSAR/QSPR), 从而建立缓蚀剂的 IE 与结构参数(电负性、极化率、范德华体积等)之间的相关性, 以预测同系列分子的缓蚀性能. 然而, 最近的文献表明, DFT 衍生参数与 IE 之间的相关性具有误导性, 或者对于大型缓蚀剂数据集来说, 由于参数选择错误, 或者计算方法选择错误, 导致相关性太弱而无法定量分析[33−34]. 因此, 已有作者借助消息传递神经网络(Message passing neural network, MPNN)等深度学习技术[35−36]建立分子结构与缓蚀剂分子的 IE 之间关系, 可有效提高 IE 预测精度和模型泛化性能, 从而快速筛选缓蚀剂.

但是上述方法均是从已有化学空间中筛选缓蚀剂, 筛选效率过低, 且无法生成新的缓蚀剂分子. 生成建模是机器学习的一个子领域, 专注于开发能够生成新数据样本的算法, 这些样本类似于给定训练数据集的数据分布, 有期克服从超大型化学库中进行缓蚀剂筛选的局限性. 特别是条件式分子生成模型, 可生成具有特定所需特性的分子, 有助于生成特定腐蚀环境、特定缓蚀效率的缓蚀剂分子. 本文将首先对应用于缓蚀剂分子筛选的机器学习和深度学习技术进行全面分析, 并讨论已有技术的局限性; 然后简要介绍分子生成建模技术概念, 给出其在高效缓蚀剂生成的可能应用以及挑战.

1　缓蚀剂分子筛选方法

如图 1 所示, 通过对已有缓蚀剂分子库进行分子特征计算, 可分别作为 IE 预测模型和分子生成模型的输入, 从而得到筛选后缓蚀剂分子, 最后将经过实验验证的缓蚀剂分子放入缓蚀剂分子库中, 此过程形成一个研究闭环. 其中, 基于 ML 的缓蚀剂分子筛选主要以分子描述符、物理参数、量子化学参数或分子结构特征、以及环境因素(例如温度、环境 pH 值、缓蚀剂浓度等)作为预测模型的输入特征来预测缓蚀效率(IE). 常见的方法包括以分子描述符、物理参数、量子化学参数作为特征输入到 ML 模型进行 IE 预测的 QSAR/QSPR 分析, 以及以分子结构特征作为输入进行机器学习或神经网络建模的 IE 预测方法.

上述方法均通过建立分子特征(包括分子结构特征)与 IE 之间的关系, 来实现缓蚀剂分子筛选, 如图 2 所示, 其建模步骤如下文列出.

(1)准备输入数据集: 收集并整理缓蚀剂分子特定环境(针对特定基体材料、腐蚀介质)下的 IE 数据、缓蚀剂浓度、介质浓度和分子特征数据.

(2)获取分子特征: IE 预测输入的分子描述符主要包括利用 RDkit 软件包[37]计算得到的化学属性(例如分子量、辛醇/水分配系数(Log*P*)、拓扑分子极性表面积(TPSA)等)和使用化学信息学软件或工具(例如 Dragon[38]、PaDEL-Descriptor[39]、COD-ESSA[40]、Gaussian、VASP、Materials Studio 等)生成的计算参数, 这些参数定量表示分子的结构特征、物理化学性质或基于分子结构得到的数据. 由于腐蚀抑制在很大程度上取决于分子在表面上的吸附能力[23], 因此在缓蚀剂分子缓蚀效率预测算

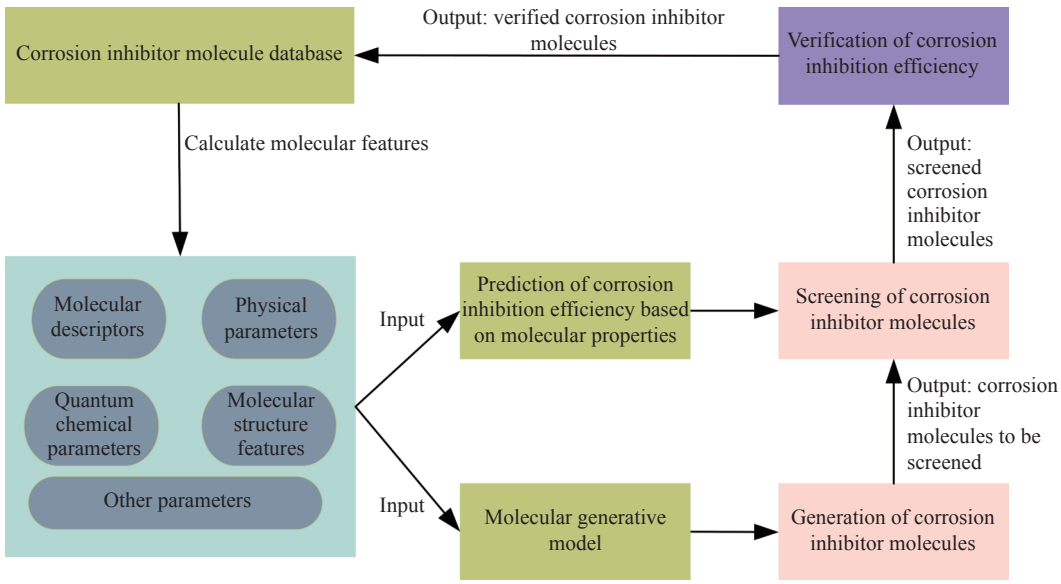

图 1　ML 助力缓蚀剂分子研究的思路

Fig.1　Machine learning ideas for assisting research on corrosion inhibitor molecules

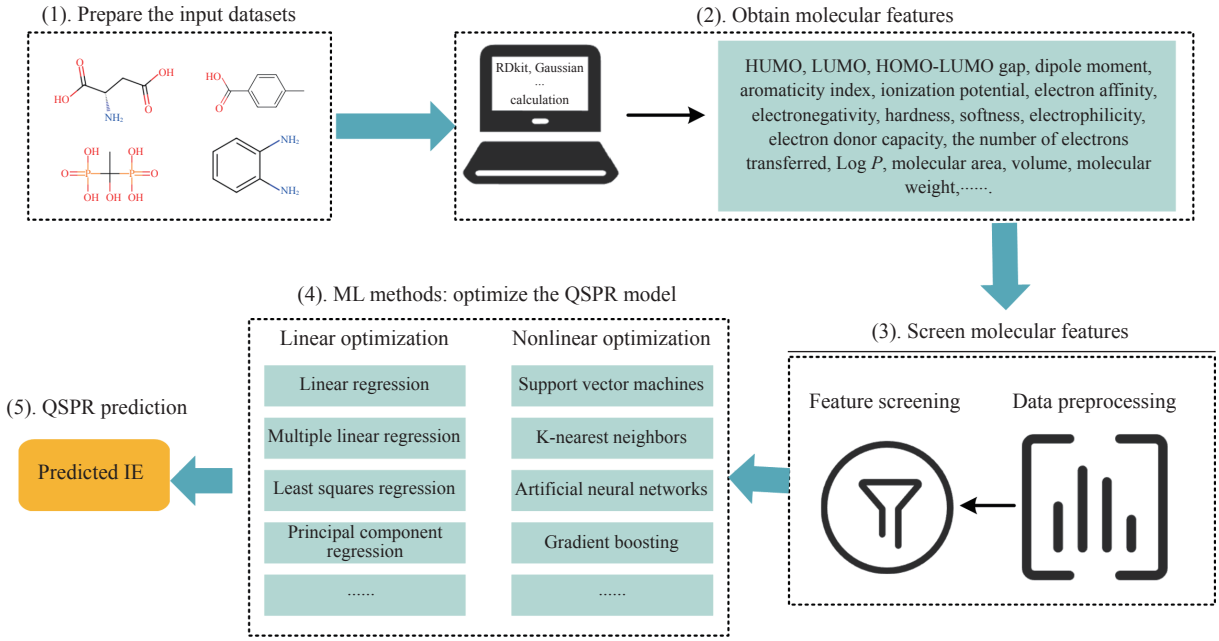

图 2 面向 IE 预测的 QSAR/QSPR 建模流程

Fig.2 Quantitative structure activity/property relationship (QSAR/QSPR) modeling process for inhibition efficiency (IE) prediction

法的特征选择中，通常选择与控制吸附能力相关的参数，包括偶极矩、极化率、最高占位分子轨道(Highest occupied molecular orbital, HOMO)、最低未占分子轨道 (Least unoccupied molecular orbital, LUMO)、HOMO–LUMO 间隙、电离势、电子亲和性、电负性、硬度、软度、范德华表面积、范德华体积等属性. 其中，分子的电子亲和性、电负性、硬度、软度等参数与分子的化学活性和反应性密切相关；HOMO–LUMO 间隙反映了分子的稳定性和反应性，从而影响分子与金属表面的相互作用；偶极矩与分子的极性相关，影响其在金属表面的吸附能力[41].

(3)分子特征筛选：Rajan[42] 在书籍《Informatics for Materials Science and Engineering: Data–Driven Discovery for Accelerated Experimentation and Application》中提到，分子特征的选择对于 QSAR/QSPR 建模至关重要，不恰当的分子特征的选择可能导致错误解释或者模型过拟合. 因此，分子特征的选择对于 QSAR/QSPR 建模精度起到至关重要的作用，可以通过专家经验、统计学方法或机器学习算法，筛选出与目标性质相关性较高的分子特征，去除冗余和无关的特征，提高模型的准确性和效率.

(4)QSPR 模型优化：选择适当的线性优化(如多元线性回归)或非线性优化的机器学习算法(如人工神经网络)，并通过交叉验证和参数调优等方法优化模型性能，确保模型的泛化能力和预测精度. 一般常利用均方误差(Mean squared error, MSE)，均方根误差(Root mean square error, RMSE)和拟合优度指标 R^2 几个指标评估 QSPR 模型性能并进行验证.

(5)QSPR 模型预测：使用优化后的模型对新分子进行性质预测.

由于特征工程和模型优化两个部分对于 IE 预测精度影响较大，因此，第 2 节和第 3 节将针对两部分进行综述.

2 特征工程–分子描述符筛选

在进行 IE 预测模型构建过程中，特征工程严重影响了模型的性能. 特征工程主要包括数据预处理、特征选择、特征降维和特征构建几个方面.

2.1 数据预处理

数据预处理是确保机器学习模型能够有效学习和预测 IE 的重要步骤. 数据预处理的主要步骤和方法包括：

(1)数据清理. 主要包括缺失值处理、异常值处理. 其中缺失值处理可通过删除法、插值法(例如线性插值)、填充法(例如均值、中位数、众数进行填充)、模型预测法预测缺失值. 异常值处理可通过使用箱线图、z-score 等统计方法识别和处理异常值，或使用孤立森林和局部离群因子(Local outlier factor, LOF)[43] 等算法检测异常值，或结合领域知识手动检查和处理异常值.

(2)数据转换. 在进行模型训练和测试前，输

入变量之间可能存在量纲上的差别, 为了尽可能减少量纲差别对各个输入特征的影响, 可采用 z-score 标准化、min–max 归一化、标签编码、独热编码(例如不同缓蚀剂分类、不同腐蚀介质分类)等方式对数据进行转换, 以保证模型能正确理解这些特征.

(3)数据增强. 由于 ML 建模需要较大的数据量, 而缓蚀剂数据较少, 因此可通过数据增强的方法对数据集进行扩充. 例如 Sutojo 等[29] 为了克服缓蚀剂化合物数据集的小样本问题, 使用虚拟样本生成 (Virtual sample generation, VSG) 方法生成虚拟样本并将其添加到训练集. 利用六个小数据集验证了在训练数据中加入虚拟样本, 有助于 KNN 算法识别特征–目标关系模式, 从而增加与缓蚀效率相关的量子化学描述符的数量.

(4)数据拆分. 在 ML 模型建立中, 需要按比例随机划分训练集和测试集数据集, 如 80% 用于训练, 20% 用于测试. 也可采用如 k 折交叉验证, 通过多次训练和测试确保 IE 预测模型的稳定性和泛化能力.

2.2　分子特征筛选

在 IE 预测建模过程中, 常见的用于对分子描述符进行特征选择、特征构建、特征降维或者回归建模的统计分析和 ML 方法主要包括主成分分析法(Principle component analysis, PCA)[41]、遗传算法(Genetic algorithm, GA)[21,23,44] 等方法.

(1)主成分分析(PCA)是一种常用的无监督降维技术, 旨在通过线性变换将高维数据投影到低维空间, 同时保留尽可能多的数据方差. PCA 通过找到数据中方差最大的方向作为主成分来实现降维. 将高维的分子特征数据降维到较低维度, 可更好地进行可视化、建模或分析. 通过选择最重要的主成分来构建新的特征, 用于后续的回归任务. 例如 Hadisaputra 等 [41] 对 HOMO、LUMO、HOMO–LUMO 间隙、偶极矩、电离电位、电子亲和力、硬度、软度、电负性、电子转移分数、亲电性指数、Log*P*、临界体积和相对分子质量几个分子特征进行 PCA 分析, 研究呋喃衍生物缓蚀剂对低碳钢的腐蚀抑制性能, 得出如下结论: LUMO、HOMO–LUMO 间隙、偶极矩、软度、电子转移分数、亲电性指数和反馈能对主成分 1 有显著贡献; 相对分子质量和临界体积对主成分 2 有显著贡献; HOMO、电离电位和 Log*P* 对主成分 3 有显著贡献.

(2)遗传算法(GA)是一种模拟生物进化过程的优化算法, 通过模拟自然选择、交叉和变异等过程来搜索问题的解空间. 在特征选择或特征优化中, 遗传算法可以用来探索最优特征子集的组合. 例如 Ser 等[23] 利用 GA 研究了酸性介质中, 41 种吡啶和喹啉 N-杂环化合物对铁合金的缓蚀性能影响因素. 通过结合人工神经网络和 GA 建模得出最优九个变量(HOMO、LUMO、HOMO–LUMO 间隙、电负性、柔软度、亲电性、电子供体容量、N 原子电荷和吸附能指数), 揭示了吸附能、物理尺寸等参数对金属腐蚀抑制的重要性, 提出了关键的腐蚀抑制设计原则.

3　面向 IE 预测的模型优化与预测

根据模型输入是分子特征还是分子结构进行分类, 可将建模划分为 QSAR/QSPR 模型构建和基于分子结构的 IE 预测模型构建.

3.1　QSAR/QSPR 模型构建

基于分子特征进行 QSAR/QSPR 模型构建方法可分为三大类, 即线性优化方法(多元线性回归和偏最小二乘回归)、非线性优化方法(人工神经网络、支持向量机、K 最近邻算法、决策树、随机森林等)以及混合优化方法. 通过调研已有 IE 预测的 QSAR/QSPR 模型, 发现主要采用多元线性回归[20–21,44–47]、人工神经网络[24,46–48]、梯度增强[27,31,41,49–50] 几种方法, 输入的分子特征主要为物理参数和量子化学参数.

支持向量机(Support vector machine, SVM)是一种基于结构风险最小化原理的统计学习方法. 近年来, 它已成功应用于解决众多领域的分类和回归问题, 并在实际应用中提供最先进的性能. 支持向量机的预测结果具有快速学习、全局优化和出色的泛化能力等诸多优点. 另一方面, SVM 在小样本数据集上表现出色. Zhao 等[27] 针对在 1 mol·L^{-1} 盐酸中, 浓度为 0.01 mol·L^{-1} 的氨基酸对铁腐蚀的动电位极化曲线获得的 IE 数据, 基于 SVM 模型, 以 HOMO、LUMO、偶极矩(μ)、电离势(I)、电子亲和力(A)、电负性(χ)、硬度(η)、软度(σ)、分子体积(V)、转移电子分数(ΔN)、自然电荷(Qtotal)、Mulliken 电荷(Ztotal)等物理化学参数作为模型输入, 构建了非线性 QSAR 模型, 模型预测和实验 IE 之间的差异(RMSE)为 1.48. 为了对比不同阶段的相关程度, 给出在气相和水相条件下的相关系数绝对值($|R|$), 发现$|R|$均小于 0.3, 表明氨基酸的缓蚀性能与气相和水相中单个量子化学参数的相关性较低.

多元线性回归(Multiple linear regression, MLR)

是一种用于建立因变量与多自变量之间线性关系的模型, 一般采用最小二乘法, 即最小化残差平方和来估计模型参数. 由于 MLR 模型相对简单, 回归系数直接反映了每个自变量对因变量的影响, 易于理解, 计算复杂度较低, 适合处理大规模数据集, 且可以通过加入交互项和非线性项扩展到更复杂的模型. 因此, MLR 广泛被应用于 QSAR/QSPR 建模, 用于预测和解释多个分子描述符因素对 IE 变量的共同影响. 例如, Quadri 等[24]用 20 种吡哒嗪衍生物在酸溶液中对低碳钢的缓蚀效果进行了评价. 根据化合物的化学结构和电子结构的不同, IE 从 65% 到 97% 不等. 包括 HOMO–LUMO 能隙、HOMO、LUMO、μ、I、A、χ、η、σ、ΔN、总能量(TE)在内的分子的结构性质, 由 GaussView 5.0 查看结构后使用 Gaussian 进行几何优化进行确定. 利用缓蚀剂浓度(Conc)、温度(Temp)、实验参数和理论参数建立了一个 MLR 模型来预测观察到的 IE, MSE、RMSE 和平均绝对百分比误差的预测结果分别为 111.5910、10.5637 和 10.2362. IE 与常规键级之和(Sum of conventional bond orders, SCBD)、软度(σ)、LUMO、电离势(Mi)和分子量(MW)密切相关. 但 MLR 也存在一定局限性, 例如(1)MLR 假设因变量与自变量之间的关系是线性的, 因此无法捕捉非线性关系; (2)虽然模型简单, 但在高维空间中解释性可能下降, 尤其当自变量数量较多时, 模型容易过拟合. 因此, 需要通过适当的数据预处理和模型改进, 克服一些局限性, 提高模型的适用性和预测能力.

人工神经网络(Artificial neural network, ANN)是一种模拟生物神经网络行为的计算模型, 由大量的人工神经元(节点)组成, 这些节点通过加权连接相互作用. ANN 通过学习和调整权重来从输入数据中提取复杂的模式和非线性关系. 在建立分子描述符与 IE 关系的 QSAR/QSPR 模型中, ANN 能够处理高度非线性和复杂的数据关系, 理论上可以逼近任意复杂的函数. 具体过程包括: 将分子描述符作为输入层节点, 通过多个隐藏层进行非线性变换, 最后在输出层得到预测的 IE 值. 通过反向传播算法(Backpropagation), 调整网络权重以最小化预测误差, 从而实现精确的 QSAR/QSPR 模型. 例如 Ser 等[23]以 9 个输入分子描述符(计算吸附能指数、HOMO、LUMO、HOMO–LUMO 能隙、电负性、柔软度、亲电性、电子供体容量和 N 原子电荷)作为输入, 建立基于遗传算法–人工神经网络(GA–ANN)方法的 QSPR 模型. 通过与基于遗传算法–偏最小二乘(GA–PLS)的 QSPR 模型进行性能对比, 结果显示 GA–ANN 方法在训练集、测试集和验证集上的平均 RMSE 优于线性的 GA–ANN 方法, 得到的平均 RMSE(训练/测试/验证) 为 8.8%. 但 ANN 也存在一定局限性, 例如: (1)数据量不足时容易导致模型过拟合或欠拟合; (2)可能会遇到梯度消失或梯度爆炸问题, 尤其在深层网络中, 这会导致训练不稳定或难以收敛; (3)神经网络的“黑箱”性质使得很难解释模型的内部机制和预测结果.

梯度增强(Gradient boosting, GB)是一种迭代的集成学习方法, 通过逐步构建多个弱学习器(通常是决策树)来优化模型性能. GB 通过决策树的分裂过程, 可自动选择重要特征, 减少手动特征选择的工作量, 且相比 ANN 更易于解释. 另一方面, 由于逐步减少残差的过程, 梯度增强方法对噪声数据具有一定的鲁棒性. 例如, Akrom 等[31]设计了包含 50 种天然有机化合物的数据集, 以 11 种量子化学性质(HOMO、LUMO、HOMO–LUMO 能隙、电离势、电子亲和力、全局硬度、全局柔软度、电负性、偶极矩、亲电性、转移电子分数)和化合物浓度作为输入特征, 以 IE 值作为目标变量. 为了提高机器学习模型的预测精度, 在训练过程中使用核密度估计函数生成虚拟样本, 并测试了 k 最近邻(K-Nearest neighbors, KNN)、随机森林(Random forest, RF)和 GB 三种不同的模型的 IE 预测性能. 结果表明, 由于引入了虚拟样本, 有效地提高了输入特征与目标值之间的相关性, 模型的预测性能得到了显著提高. GB、RF 和 KNN 模型的 R2 值分别从 0.557 增加到 0.996、0.522 增加到 0.999、0.415 增加到 0.994. 此外, 每个模型都显示 RMSE 值的显著降低, 分别从 1.41 过渡到 0.19、1.27 过渡到 0.10 和 1.22 过渡到 0.16. 可见, 基于 GB 方法构建的模型更有利于该数据集进行 IE 准确预测. 但 GB 方法也同样有一定局限性, 例如: (1)GB 方法涉及多个超参数(如学习率、弱学习器数量、树的深度等), 需要仔细调优才能达到最佳性能; (2)在弱学习器数量过多或模型过复杂时, GB 方法容易发生过拟合, 需要使用正则化方法(如缩减树的深度、增加学习率衰减等)来防止过拟合.

上述方法各有特点, 适用于不同的数据集和预测需求, 但都展示了在缓蚀剂分子筛选中的潜力和应用前景. 通过结合适当的数据预处理和正则化方法, 选择合适的模型和分子描述符, 可以有效地建立分子描述符和缓蚀性能之间的关系, 为

实验和理论研究提供了重要的支持和指导. 为了对比上述四个模型对于缓蚀剂分子的缓蚀效率(IE)预测准确度, 本文采用随机森林(RF)、梯度增强(GB)、人工神经网络(NN)和多元线性回归(MLR)四个算法对 Gong 等[51] 提供的特定环境条件下(环境温度: 25 ℃ ~ 30 ℃, HCl 浓度: 1 mol·L^{-1})的缓蚀剂分子数据集(包括缓蚀剂浓度、HOMO、LUMO、偶极矩、电离电位、电子亲和力、硬度、软度、电负性、范德华体积、范德华表面积等特征参数)进行 IE 预测模型构建. 为确保实验的公平性, 四个模型均采用 8∶2 的比例将数据集划分为训练集和验证集, 并通过十折交叉验证进行模型训练. 如图 3 所示, NN 模型在验证集上的 RMSE 为 0.018, 表现出最高的准确度; RF 和 GB 模型的表现相近, 而 MLR 模型的准确度最低, RMSE 为 0.15. 因此, 针对缓蚀效率的预测, NN、GB 和 RF 模型展现出更优的预测准确性.

3.2　基于分子结构的 IE 预测模型构建

在基于分子结构的 IE 预测建模中, 需要对缓蚀剂分子进行分子表示后进行建模. 分子表示方法包括以字符串进行编码的 SMILES 形式、分子指纹、图的形式. SMILES 格式[52] 基于一定规则的语法词典进行编码, 适用于循环神经网络、transformer 等自然语言模型. 分子指纹一般采用摩根编码为 1024 或 2048 位的 0、1 向量. 令分子图表示为 G=(V, E), 则 G 的表示按照分子粒度区分, 细粒度表示可以以原子为节点(V), 化学键为边(E), 可精确描述每个原子和键的变化; 粗粒度表示将分子分割为多个片段或官能团, 以片段或官能团为节点, 连接关系为边来构造图数据, 可以简化输入数据, 提供 ML 模型训练速度和效果. 例如图 4 中 SMILES 为 BrCCOc1cccc2cccnc12 的缓蚀剂 8-(2-bromoethoxy)quinoline(QN-C2Br), 按照 Br, Ethoxy 和 Benzopyridine 环三个片段作为节点可得到粗粒度

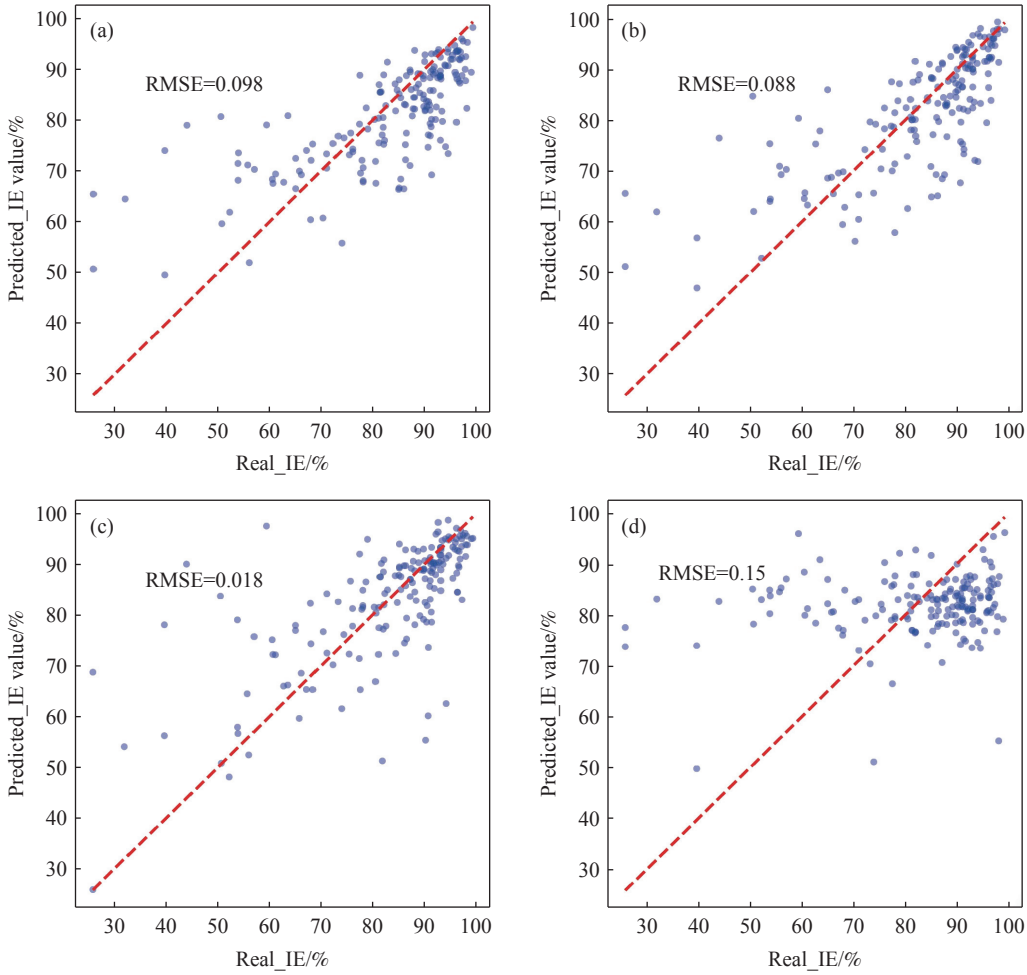

图 3　不同模型预测缓蚀剂在 1 mol·L^{-1} 盐酸溶液中的缓蚀效率准确度对比图. (a) 随机森林模型; (b) 梯度增强模型; (c) 人工神经网络模型; (d) 多元线性回归模型

Fig.3　Prediction accuracies compared for corrosion inhibition efficiency of inhibitors in 1 mol·L^{-1} hydrochloric acid solution, using different models: (a) random forest; (b) gradient boosting; (c) artificial neural network; (d) multiple linear regression

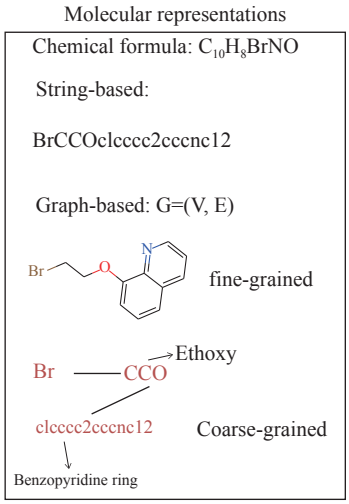

图 4 分子表示示意图

Fig.4 Schematic of molecular representations

的图结构.

传统的深度卷积神经网络(Convolutional neural network, CNN)和递归神经网络(Recurrent neural networks, RNN)只能处理文本、音频、图像、视频等欧几里得数据. 与图像和文本不同, 图数据包含基本的结构信息. 图神经网络(Graph neural network, GNN)是直接从由节点和边组成的图数据中学习的框架, 其中节点和边可以很好地用于表示分子结构中的原子和键[53]. 因此, GNN 可用于处理非欧几里得数据, 如化学分子结构和蛋白质, 并从其结构预测分子的性质[54], 包括量子力学特征, 如能量、电子和热力学性质[55−57]; 理化性质, 如疏水性、水中的水合自由能、辛醇/水分配系数(Log*P*)[58] 和毒性[59].

缓蚀剂的缓蚀效率不仅与其内部结构参数(如杂化程度、每个原子的键数、每个原子的价电子数和键类型)密切相关, 而且与整体分子特征(如分子量、芳环数、受体数和给体数)密切相关. 消息传递神经网络(Message passing neural network, MPNN)是监督图学习的通用框架, 它简单地抽象了几种最有前途的 GNN 模型之间的共性, 能够直接从分子图中学习原子级和化学键级特征, 并预测分子性质. 因此, Dai 等[35] 通过检索 116 篇以“缓蚀剂”为关键词研究盐酸溶液中不同缓蚀剂分子对碳钢影响的论文, 得到特定环境条件下(环境温度: 25 ℃ ~ 30 ℃, 缓蚀剂浓度: 1 mmol·L^{-1}, HCl 浓度: 1 mmol·L^{-1})的缓蚀剂名称、类别、分子结构和实验 IE 值, 共 270 条数据. 基于该数据, Dai 等[35] 提出了一个基于 DMPNN 框架的三级直接消息传递神经网络(3L-DMPNN)模型, 通过结合原子水平特征、化学键水平特征和分子水平特征来筛选缓蚀剂. 具体而言, 使用简化分子输入行输入系统(Simplified molecular-input line-entry system, SMILES)[52] 作为唯一输入, 将分子结构视为图, 并使用开源的 RDKit 包从 SMILES 中提取原子和化学键特征[39]. 随后, 将消息传递模块后的新分子图向量与全局分子特征相结合, 通过前馈神经网络预测分子的 IE. 而 Ma 等[36] 在 3L-DMPNN 模型的基础上提出 2D3DMol-CIC 模型, 对数据集进行扩充, 得到包含缓蚀剂名称、分子的 SMILES、分子中的原子坐标、缓蚀剂浓度和 IE 值的数据 1241 条, 然后结合 2D–3D 分子图和缓蚀剂浓度进行建模, 验证了三维特征对 IE 预测的影响以及模型的泛化能力. 与已有的预测模型相比, 所构建的 2D3DMol-CIC 模型不仅可以更准确地预测特定浓度下跨类缓蚀剂的 IE, 而且可以预测不同浓度下缓蚀剂的 IE, 从而确定产生高 IE(>90%) 的最小浓度.

4 局限性分析与展望

虽然通过提取定量分子描述符等结构化数据, 可以建立 IE 预测模型, 在几分钟甚至更短的时间内预测 IE, 从而基于 IE 进行缓蚀剂分子筛选. 然而仍存在以下问题: (1)基于 ML 建立的 QSPR 模型在很大程度上依赖于分子特征的选择, 仅限于预测特定浓度下某类缓蚀剂的 IE; (2)目前已有文献[60] 提到电离势、HOMO 或 LUMO 能量或任何其他量子化学衍生的描述符与腐蚀效率之间基本上没有相关性, 因此, 利用量子化学参数作为 IE 预测模型输入参数是否准确有效, 仍有待研究; (3)利用深度学习技术, 建立基于分子结构的 IE 预测模型可一定程度提高 IE 预测精度和模型泛化性能, 能较准确地预测缓蚀剂及其浓度, 为缓蚀剂及其浓度的筛选提供了一种低成本、快速的方法, 但是仍然只能在有限的化合物空间进行搜索.

在未来研究中, 一方面, 可尝试利用高通量计算技术, 生成大批量缓蚀剂分子, 并利用神经网络模型建立基于分子特征的缓蚀效率预测模型; 另一方面, 基于分子生成模型生成新的缓蚀剂分子有可能克服从大型化合物空间进行缓蚀剂分子搜索的局限性. 例如利用条件生成模型可按照特定属性进行分子设计, 从而缩小化合物搜索空间. 目前, 在药物发现领域, 已有研究利用生成式建模得到现有化学库以外的、经过实验验证的化合物[61−62]. 针对 SMILES、graph 两种常用分子表示方法的分

子输入数据, 分子生成模型可主要分为自回归模型(Autoregressive models)、生成对抗网络(Generative adversarial networks, GANs)、变分自编码器(Variational autoencoders, VAEs)和规范流模型(Normalizing flows, NFs)四类. VAEs 包括一个经过训练的编码器, 用于参数化潜在变量 z 的分布, 以及一个经过训练的解码器, 用于使用编码器定义的分布中的样本重建输入, 因此适用于缓蚀剂分子生成建模. 在缓蚀剂分子生成方面, Gong 等[51] 利用 Ma 等[36] 文献提供的基于碳钢材料的 1368 条缓蚀剂数据, 基于 RDKit 软件计算了包括药物相似性定量估计 (Quantitative estimate of drug-likeness, QED), 分配系数的对数(ALOGP, Log$P$), 分子量 (Molecular weight, MolWt), 氢键受体 (Hydrogen bond acceptors, HBA), 氢键供体 (Hydrogen bond donors, HBD), 拓扑极性表面积 (Topological polar surface area, TPSA), 可旋转键数量 (Number of rotatable bonds, NumRotBonds) 和 N、O、S、P 原子数量 11 个分子特征, 以这 11 个分子特征和缓蚀剂浓度作为缓蚀剂分子生成模型的输入条件, 建立条件式 VAEs 模型. 如图 1 所示, 基于生成模型生成的新分子可作为缓蚀剂分子筛选的候选集, 然后利用 IE 预测模型得到筛选后的缓蚀剂分子, 从而减少缓蚀剂分子筛选的化合物空间.

尽管已有众多可供选择的分子生成方法来针对缓蚀剂分子进行分子生成模型训练, 但在缓蚀剂分子生成领域仍存在较多的问题亟待解决.

(1)不同缓蚀剂针对不同基体材料、不同腐蚀环境所体现的缓蚀效率不同, 因此, 在进行缓蚀剂分子生成建模时, 需要针对特定腐蚀环境、特定基体材料, 收集相关缓蚀剂分子数据, 但是目前已有数据集稀少, 可能导致生成有效分子的比例过低. 因此, 需要利用文献挖掘、高通量实验等方式增加数据集. 例如, Wang 等[63] 利用自然语言处理流程对高温合金的化学成分、属性数据进行了自动提取. Ren 等[64] 针对高通量实验技术进行腐蚀研究进行了综述, 提到将高通量腐蚀实验与电化学高通量表征相结合, 可以进一步提高测量的效率.

(2)利用条件分子生成模型构建与给定属性集相对应的缓蚀剂分子结构, 可通过事先了解分子的 SMIELS、片段(或官能团)和目标特性之间的关系, 将采样限制在非常特定的分子上. 但这样可能会阻止模型发挥其潜力.

(3)要想获得具有高效率的新缓蚀剂分子, 需要从新生成的分子中进行缓蚀效率预测, 或者直接以与缓蚀效率相关的分子性质作为属性条件, 建立条件生成模型, 但目前常用于缓蚀效率预测的量子化学参数仍不能在本研究领域达成一致, 将导致以已有的量子化学参数作为条件生成模型的条件输入, 可能生成无法满足特定缓蚀效率的分子.

(4)生成的缓蚀剂分子需要经过实验验证其实际缓蚀性能, 这样的实验验证步骤往往费时费力. 因此, 如何高效地筛选和验证生成分子的实际缓蚀效果是一个挑战. 除了缓蚀效率, 生成的分子还需要满足其他性质(如环保性、稳定性、成本等). 多目标优化在生成模型中的应用及其实现难度较大.

(5)生成的分子既要多样化又要具备创新性, 避免生成与已有分子相似或重复的结构. 这需要模型在生成时能够探索化学空间的广度, 同时保持一定的创新性. 另外, 目前许多生成模型都属于“黑箱”模型, 缺乏解释性. 理解模型是如何生成缓蚀剂分子的, 以及生成的分子为何具有高效的缓蚀性能, 对于提升模型和设计新分子具有重要意义.

虽然仍存在上述问题需要解决, 但是分子生成模型在药物等研究领域的验证成功证明了其可行性, 借助文本挖掘等技术整理特定腐蚀环境下的缓蚀剂分子数据集, 可有效解决数据集缺失问题. 将分子生成模型和已有缓蚀剂分子筛选方法进行闭环结合, 可极大缩小高效缓蚀剂分子研究的化合物空间. 另一方面, 利用高通量实验和高通量计算技术, 对筛选到的缓蚀剂分子进行实验验证, 可进一步加速高效缓蚀剂研究.

5　总结

本文首先总结了当前基于 ML 的缓蚀剂分子筛选方法, 指出仅依赖缓蚀效率预测模型进行分子筛选的局限性. 利用分子生成模型生成具有某种分子特征的缓蚀剂分子, 并结合分子筛选方法对生成的分子进行缓释效率预测, 将大大缩小缓蚀剂化合物筛选的范围. 因此, 本文进一步简要介绍了现有的缓蚀剂分子生成模型. 但由于缓蚀剂分子研究的特殊性, 仍有许多问题亟待解决. 本文通过总结缓蚀剂分子发现研究中存在的问题, 并展望未来的发展方向, 以期为材料腐蚀研究人员提供参考.

参 考 文 献

[1] Hou B R, Li X G, Ma X M, et al. The cost of corrosion in China. *NPJ Mater Degrad*, 2017, 1: 4

[2] Finšgar M, Jackson J. Application of corrosion inhibitors for steels in acidic media for the oil and gas industry: A review. *Corros Sci*, 2014, 86: 17

[3] Fazal B R, Becker T, Kinsella B, et al. A review of plant extracts as green corrosion inhibitors for CO_2 corrosion of carbon steel. *NPJ Mater Degrad*, 2022, 6: 5

[4] Elqars E, Oubella A, Hachim M E, et al. New 3-(2-methoxyphenyl) -isoxazole-carvone: Synthesis, spectroscopic characterization, and prevention of carbon steel corrosion in hydrochloric acid. *J Mol Liq*, 2022, 347: 118311

[5] Zou Y, Wang J, Zheng Y Y. Electrochemical techniques for determining corrosion rate of rusted steel in seawater. *Corros Sci*, 2011, 53(1): 208

[6] Bahlakeh G, Ramezanzadeh B, Ramezanzadeh M. Cerium oxide nanoparticles influences on the binding and corrosion protection characteristics of a melamine-cured polyester resin on mild steel: An experimental, density functional theory and molecular dynamics simulation study. *Corros Sci*, 2017, 118: 69

[7] Boucherit L, Al-Noaimi M, Daoud D, et al. Synthesis, characterization and the inhibition activity of 3-(4-cyanophenylazo) -2, 4-pentanedione (L) on the corrosion of carbon steel, synergistic effect with other halide ions in 0.5 M H_2SO_4. *J Mol Struct*, 2019, 1177: 371

[8] Gece G. The use of quantum chemical methods in corrosion inhibitor studies. *Corros Sci*, 2008, 50(11): 2981

[9] Verma D K, Aslam R, Aslam J, et al. Computational modeling: Theoretical predictive tools for designing of potential organic corrosion inhibitors. *J Mol Struct*, 2021, 1236: 130294

[10] Obot I B, MacDonald D D, Gasem Z M. Density functional theory (DFT) as a powerful tool for designing new organic corrosion inhibitors. Part 1: An overview. *Corros Sci*, 2015, 99: 1

[11] Wen C, Tian Y W, Yang D Y, et al. Controlled release mechanism and inhibition performance of smart inhibitor LDH-NO_2 in the reinforced concrete structures. *Chin J Eng*, 2022, 44(8): 1368
(文成, 田玉琬, 杨德越, 等. 智能阻锈剂 LDH-NO_2 在钢筋混凝土中的控释机制及缓蚀性能. 工程科学学报, 2022, 44(8): 1368)

[12] Obot I B, Gasem Z M. Theoretical evaluation of corrosion inhibition performance of some pyrazine derivatives. *Corros Sci*, 2014, 83: 359

[13] Tang Y M, Yang X Y, Yang W Z, et al. A preliminary investigation of corrosion inhibition of mild steel in 0.5 M H_2SO_4 by 2-amino-5-(n-pyridyl) -1, 3, 4-thiadiazole: Polarization, EIS and molecular dynamics simulations. *Corros Sci*, 2010, 52(5): 1801

[14] Haris N I N, Sobri S, Yusof Y A, et al. An overview of molecular dynamic simulation for corrosion inhibition of ferrous metals. *Metals*, 2021, 11(1): 46

[15] Chen X S, Chen Y, Cui J J, et al. Molecular dynamics simulation and DFT calculation of "green" scale and corrosion inhibitor. *Comput Mater Sci*, 2021, 188: 110229

[16] Verma C, Lgaz H, Verma D K, et al. Molecular dynamics and Monte Carlo simulations as powerful tools for study of interfacial adsorption behavior of corrosion inhibitors in aqueous phase: A review. *J Mol Liq*, 2018, 260: 99

[17] Zhang M, Fu D M, Zhang D W, et al. Extraction of important variables and mining of dependencies of atmospheric corrosion of carbon steel based on a comprehensive intelligent model. *Chin J Eng*, 2023, 45(3): 407
(张明, 付冬梅, 张达威, 等. 基于综合智能模型的碳钢大气腐蚀重要变量提取和依赖关系挖掘. 工程科学学报, 2023, 45(3): 407)

[18] Yin Z B, Wang S S, Zhu Z H, et al. Key parameters of soil corrosivity and a model for predicting the corrosion rate of Q235steel in Beijing. *Chin J Eng*, 2023, 45(11): 1939
(尹志彪, 王莎莎, 祝振洪, 等. 北京地区土壤腐蚀性关键参量与 Q235 钢腐蚀速率预测模型研究. 工程科学学报, 2023, 45(11): 1939)

[19] Fernandez M, Breedon M, Cole I S, et al. Modeling corrosion inhibition efficacy of small organic molecules as non-toxic chromate alternatives using comparative molecular surface analysis (CoMSA). *Chemosphere*, 2016, 160: 80

[20] Gutiérrez E, Rodríguez J A, Cruz-Borbolla J, et al. Development of a predictive model for corrosion inhibition of carbon steel by imidazole and benzimidazole derivatives. *Corros Sci*, 2016, 108: 23

[21] Camacho-Mendoza R L, Feria L, Zárate-Hernández L Á, et al. New QSPR model for prediction of corrosion inhibition using conceptual density functional theory. *J Mol Model*, 2022, 28(8): 238

[22] Winkler D. Predicting the performance of organic corrosion inhibitors. *Metals*, 2017, 7(12): 553

[23] Ser C T, Žuvela P, Wong M W. Prediction of corrosion inhibition efficiency of pyridines and quinolines on an iron surface using machine learning-powered quantitative structure-property relationships. *Appl Surf Sci*, 2020, 512: 145612

[24] Quadri T W, Olasunkanmi L O, Akpan E D, et al. Development of QSAR-based (MLR/ANN) predictive models for effective design of pyridazine corrosion inhibitors. *Mater Today Commun*, 2022, 30: 103163

[25] Quadri T W, Olasunkanmi L O, Fayemi O E, et al. Multilayer perceptron neural network-based QSAR models for the assessment and prediction of corrosion inhibition performances of ionic liquids. *Comput Mater Sci*, 2022, 214: 111753

[26] Du L, Zhao H X, Hu H X, et al. Quantum chemical and molecular dynamics studies of imidazoline derivatives as corrosion inhibitor and quantitative structure–activity relationship (QSAR) analysis using the support vector machine (SVM) method. *J Theor Comput*

Chem, 2014, 13(2): 1450012

[27] Zhao H X, Zhang X H, Ji L, et al. Quantitative structure–activity relationship model for amino acids as corrosion inhibitors based on the support vector machine and molecular design. *Corros Sci*, 2014, 83: 261

[28] Alamri A H, Alhazmi N. Development of data driven machine learning models for the prediction and design of pyrimidine corrosion inhibitors. *J Saudi Chem Soc*, 2022, 26(6): 101536

[29] Sutojo T, Rustad S, Akrom M, et al. A machine learning approach for corrosion small datasets. *NPJ Mater Degrad*, 2023, 7: 18

[30] Galvão T L P, Novell-Leruth G, Kuznetsova A, et al. Elucidating structure–property relationships in aluminum alloy corrosion inhibitors by machine learning. *J Phys Chem C*, 2020, 124(10): 5624

[31] Akrom M, Rustad S, Dipojono H K. A machine learning approach to predict the efficiency of corrosion inhibition by natural product-based organic inhibitors. *Phys Scr*, 2024, 99(3): 036006

[32] Pham T H, Le P K, Son D N. A data-driven QSPR model for screening organic corrosion inhibitors for carbon steel using machine learning techniques. *RSC Adv*, 2024, 14(16): 11157

[33] Kokalj A, Lozinšek M, Kapun B, et al. Simplistic correlations between molecular electronic properties and inhibition efficiencies: Do they really exist? *Corros Sci*, 2021, 179: 108856

[34] Kokalj A. Molecular modeling of organic corrosion inhibitors: Calculations, pitfalls, and conceptualization of molecule–surface bonding. *Corros Sci*, 2021, 193: 109650

[35] Dai J X, Fu D M, Song G X, et al. Cross-category prediction of corrosion inhibitor performance based on molecular graph structures *via* a three-level message passing neural network model. *Corros Sci*, 2022, 209: 110780

[36] Ma J B, Dai J X, Guo X, et al. Data-driven corrosion inhibition efficiency prediction model incorporating 2D–3D molecular graphs and inhibitor concentration. *Corros Sci*, 2023, 222: 111420

[37] Landrum G. Rdkit documentation. *Release*, 2013, 1: 4

[38] Mauri A, Consonni V, Pavan M, et al. Dragon software: An easy approach to molecular descriptor calculations. *Match*, 2006, 56(2): 237

[39] Yap C W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *J Comput Chem*, 2011, 32(7): 1466

[40] Katritzky A R, Perumal S, Petrukhin R, et al. Codessa-based theoretical QSPR model for hydantoin HPLC-RT lipophilicities. *J Chem Inf Comput Sci*, 2001, 41(3): 569

[41] Hadisaputra S, Irham A D, Purwoko A A, et al. Development of QSPR models for furan derivatives as corrosion inhibitors for mild steel. *Int J Electrochem Sci*, 2023, 18(8): 100207

[42] Rajan K. *Informatics for Materials Science and Engineering: Data-Driven Discovery for Accelerated Experimentation and Application*. Oxford: Butterworth-Heinemann, 2013.

[43] Cheng Z Y, Zou C M, Dong J W, et al. Outlier detection using isolation forest and local outlier factor // *Proceedings of the Conference on Research in Adaptive and Convergent Systems*. Chongqing, 2019: 161

[44] Khaled K F. Modeling corrosion inhibition of iron in acid medium by genetic function approximation method: A QSAR model. *Corros Sci*, 2011, 53(11): 3457

[45] Chafai N, Salhi H, Hadjira A, et al. Development of new models to predict the corrosion inhibition efficiency as functions of some molecular descriptors using statistical analysis. *J Indian Chem Soc*, 2023, 100(9): 101073

[46] Quadri T W, Olasunkanmi L O, Fayemi O E, et al. Computational insights into quinoxaline-based corrosion inhibitors of steel in HCl: Quantum chemical analysis and QSPR-ANN studies. *Arab J Chem*, 2022, 15(7): 103870

[47] Quadri T W, Olasunkanmi L O, Fayemi O E, et al. Predicting protection capacities of pyrimidine-based corrosion inhibitors for mild steel/HCl interface using linear and nonlinear QSPR models. *J Mol Model*, 2022, 28(9): 254

[48] Iyer R S, Iyer N S, Rugmini Ammal P, et al. Harnessing machine learning and virtual sample generation for corrosion studies of 2-alkyl benzimidazole scaffold small dataset with an experimental validation. *J Mol Struct*, 2024, 1306: 137767

[49] Akrom M, Rustad S, Saputro A G, et al. Data-driven investigation to model the corrosion inhibition efficiency of Pyrimidine-Pyrazole hybrid corrosion inhibitors. *Comput Theor Chem*, 2023, 1229: 114307

[50] Akrom M, Rustad S, Saputro A G, et al. A combination of machine learning model and density functional theory method to predict corrosion inhibition performance of new diazine derivative compounds. *Mater Today Commun*, 2023, 35: 106402

[51] Gong H Y, Fu Z H, Ma L W, et al. Inhibitor_Mol_VAE: A variational autoencoder approach for generating corrosion inhibitor molecules. *NPJ Mater Degrad*, 2024, 8: 102

[52] Weininger D. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *J Chem Inf Comput Sci*, 1988, 28(1): 31

[53] Hao Z K, Lu C Q, Huang Z Y, et al. ASGN: An active semi-supervised graph neural network for molecular property prediction // *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Virtual Event, 2020: 731

[54] Asif N A, Sarker Y, Chakrabortty R K, et al. Graph neural network: A comprehensive review on non-euclidean space. *IEEE Access*, 2021, 9: 60588

[55] Wu Z Q, Ramsundar B, Feinberg E N, et al. MoleculeNet: A benchmark for molecular machine learning. *Chem Sci*, 2017, 9(2): 513

[56] Yang K, Swanson K, Jin W G, et al. Analyzing learned molecular representations for property prediction. *J Chem Inf Model*, 2019, 59(8): 3370

[57] Gilmer J, Schoenholz S S, Riley P F, et al. Neural message passing for quantum chemistry // *Proceedings of the 34th International*

Conference on Machine Learning, PMLR 70. Sydney, 2017: 1263

[58] Wang X F, Li Z, Jiang M J, et al. Molecule property prediction based on spatial graph embedding. *J Chem Inf Model*, 2019, 59(9): 3817

[59] Withnall M, Lindelöf E, Engkvist O, et al. Building attention and edge message passing neural networks for bioactivity and physical–chemical property prediction. *J Cheminf*, 2020, 12(1): 1

[60] Winkler D A, Breedon M, Hughes A E, et al. Towards chromate-free corrosion inhibitors: Structure–property models for organic alternatives. *Green Chem*, 2014, 16(6): 3349

[61] Zhavoronkov A, Ivanenkov Y A, Aliper A, et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat Biotechnol*, 2019, 37(9): 1038

[62] Ren F, Ding X, Zheng M, et al. AlphaFold accelerates artificial intelligence powered drug discovery: Efficient discovery of a novel CDK20 small molecule inhibitor. *Chem Sci*, 2023, 14(6): 1443

[63] Wang W R, Jiang X, Tian S H, et al. Automated pipeline for superalloy data by text mining. *NPJ Comput Mater*, 2022, 8: 9

[64] Ren C H, Ma L W, Zhang D W, et al. High-throughput experimental techniques for corrosion research: A review. *Mater Genome Eng Adv*, 2023, 1(2): e20