文章编号:1671-4229(2024)06-0036-11

【功能材料设计与应用专题】　　专题主持人:乔智威

基于机器学习和大数据挖掘的金属 - 有机框架研究

明忠源[1,2]，乔智威[2*]，李　振[1]，李晓鹏[1]，赵　越[1**]，李和国[1***]
(1. 国民核生化灾害防护国家重点实验室，北京　100191；2. 广州大学 化学化工学院，广东 广州　510006)

摘　要: 金属 - 有机框架材料(Metal-Organic Frameworks,MOFs)因其多样化的化学结构而在气体储存与分离、催化、药物储存和递送等领域展现出广泛的应用潜力。随着 MOFs 种类和应用领域的快速扩展,传统的实验方法和分子模拟已经无法在短时间内充分评估新 MOFs 的所有性能。考虑到 MOFs 的数量庞大,且伴随着它们结构和性能方面的巨大数据量,将机器学习方法整合到 MOFs 的设计和开发中无疑将带来巨大的好处。通过构建机器学习模型,可有效揭示 MOFs 复杂的结构 - 性能关系,加速性能预测和材料设计过程。文章汇总和分析了利用机器学习方法进行 MOFs 吸附与分离领域的研究概述:①讨论了适用于机器学习工作流程的各种 MOFs 数据库、特征描述符、算法以及评价指标。②探讨了机器学习如何助力高通量计算筛选,加速对气体在 MOFs 中的吸附与分离研究。③综合讨论了机器学习在辅助基于大数据的 MOFs 气体吸附分离存储计算模拟中所面临的机遇与挑战。通过本工作的全面综述和分析,机器学习与大数据挖掘方法可以被更好地理解和应用,以加速 MOFs 的设计和开发,为相关领域提供新的研究方向和技术支持。

关键词: 金属 - 有机框架;气体吸附与分离;高通量计算;机器学习;结构 - 性能关系
中图分类号: O647.33;TB34　　**文献标志码:** A

Research advances in metal-organic frameworks with machine learning and big data mining

MING Zhong-yuan[1,2], *QIAO Zhi-wei*[2*], *LI Zhen*[1], *LI Xiao-peng*[1], *ZHAO Yue*[1**], *LI He-guo*[1***]
(1. State Key Laboratory of NBC Protection for Civilian, Beijing 100191, China;
2. School of Chemistry and Chemical Engineering, Guangzhou University, Guangzhou 510006, China)

Abstract: Metal-organic frameworks (MOFs), with their diverse chemical structures, exhibit broad application potential in fields such as gas storage and separation, catalysis, and drug storage and delivery. With the rapid expansion of MOFs varieties and application domains, traditional experimental methods and molecular simulations can no longer sufficiently evaluate the performance of new MOFs in a short time. Given the vast number of MOFs and the enormous amount of data related to their structures and properties, integrating machine learning methods into the design and development of MOFs

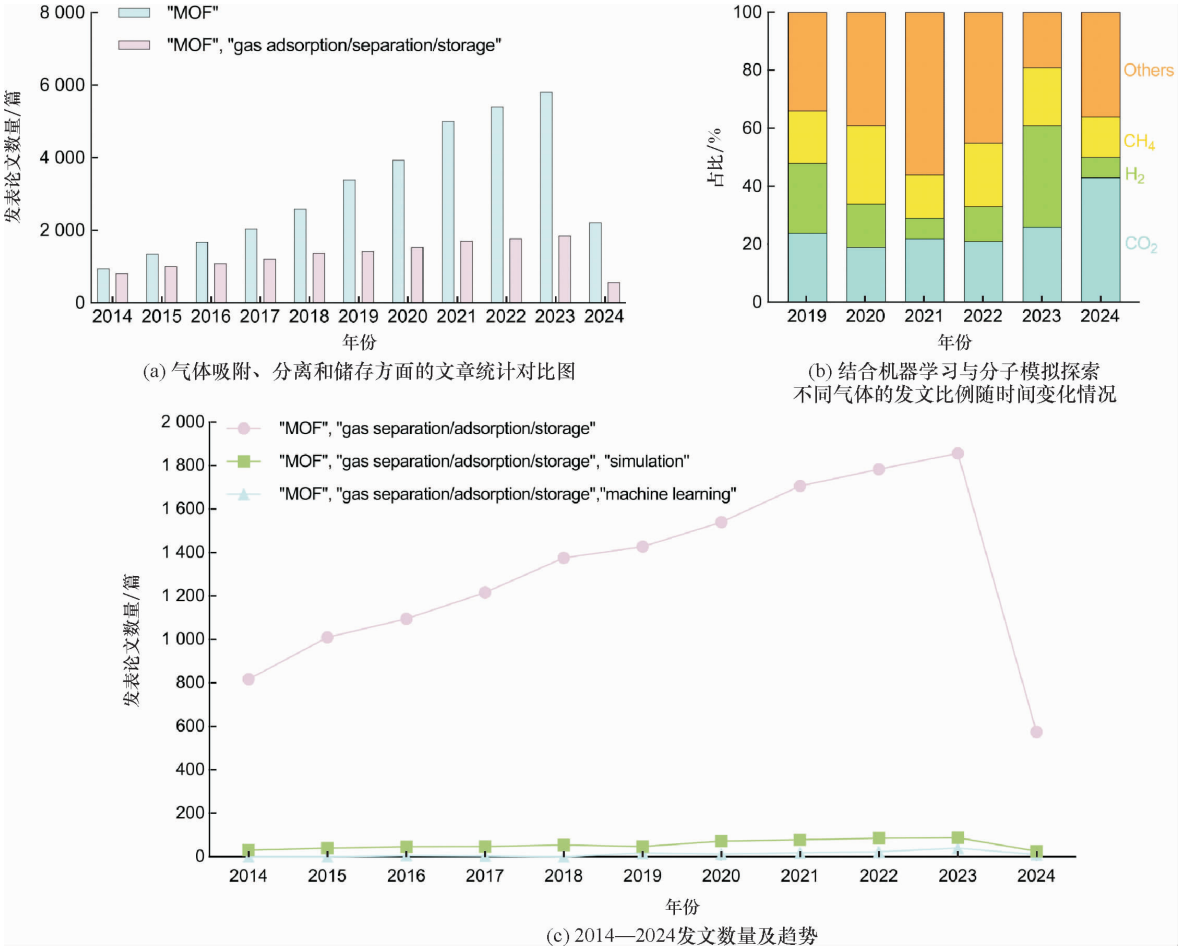
收稿日期: 2024 - 08 - 01;**修回日期:** 2024 - 09 - 13
基金项目: 国家自然科学基金资助项目(21978058);广东省自然科学基金项目(2023A1515240076,2022A1515011446)
作者简介: 明忠源(1998—),男,硕士研究生. E-mail:1312051877@qq.com
* 通信作者. E-mail:zqiao@gzhu.edu.cn
** 通信作者. E-mail:SA11226532@mail.ustc.edu.cn
*** 通信作者. E-mail:liheguo1972@126.com
引文格式: 明忠源, 乔智威, 李振, 等. 基于机器学习和大数据挖掘的金属 - 有机框架研究[J]. 广州大学学报(自然科学版),2024,23(6):36-46.

will undoubtedly bring significant benefits. By constructing machine learning models, the complex structure-property relationships of MOFs can be effectively elucidated, accelerating the performance prediction and material design processes. In this review, we comprehensively summarize and analyze research on MOFs adsorption and separation utilizing machine learning methods. First, various MOFs databases, feature descriptors, algorithms, and evaluation metrics suitable for machine learning workflows are discussed. Next, the role of machine learning in facilitating high-throughput computational screening and accelerating research on the adsorption and separation of gases such as CH_4, CO_2, and H_2 in MOFs are explored. Finally, this paper discusses the opportunities and challenges faced by machine learning in supporting big data-based computational simulations of MOFs gas adsorption, separation, and storage. Through this comprehensive review and analysis, researchers can better understand and apply machine learning and big data mining to accelerate the design and development of MOFs, providing new research directions and technical support for related fields.

Key words: metal-organic frameworks; gas adsorption and separation; high-throughput computational screening; machine learning; structure-performance analysis

金属有机框架(Metal-Organic Frameworks, MOFs)作为一类新型的有机 - 无机混合型聚合物材料[1],其结构由金属离子与有机配体组装形成,展示出极高的比表面积[2]、多样化的孔隙结构及可调控的化学性质。通过精确控制金属离子、有机配体及其功能基团的种类和组合,研究者能够有效地改变和优化 MOFs 的性质。具体方法包括:引入特定官能团[3]、更换框架中的金属中心或配体[4]、实现不同框架的相互穿插[5],以及调整配体的长度[6]等,这些措施极大地提升了 MOFs 的性能。因此,通过这些精细的调控手段,研究者成功开发出超过数十万种 MOFs 结构。MOFs 由于其独特的结构特性和化学多样性,在多个领域中展现出显著的应用前景。特别是在气体储存吸附与分离[7-9]、水净化[9]、生物医学[10]、化学传感[11]、催化[12]以及超级电容器[13]等方面,MOFs 的多功能性彰显了其在现代材料科学中的重要地位,为实现高效且可持续的技术发展提供了新的机遇。

在众多应用领域中,气体吸附、存储与分离占据了核心地位,如图 1(a)所示。MOFs 凭借其出色的孔隙率和比表面积,在吸附剂和膜材料的开发上展示了极大的潜力[14]。尽管 MOFs 的多样种类为开发新型吸附剂和膜材料提供了机遇,但同时也带来了挑战,即针对单一 MOF 的特定气体研究,也需要耗费数周时间进行实验与测试。因此,完全依赖传统实验方法评估所有 MOFs 的气体吸附和分离潜力显然不切实际。鉴于实验方法在评估 MOFs 的气体吸附分离性能方面的局限性,研究人员开始转向用分子模拟方法来模拟 MOFs 中的气体吸附分离等过程。在早期,分子模拟研究通常只针对少量 MOFs 进行。在此基础上,通过系统地研究不同 MOFs 对多种气体的吸附和分离效果,使分子模拟方法的应用范围得到了扩展[15-16]。这些工作不仅针对了特定气体如甲烷(CH_4)、二氧化碳(CO_2)、氢气(H_2)等,还探讨了混合气体体系的动态行为,为理解 MOFs 在实际应用中的表现提供了深入的观察[17]。此外,采用高通量计算筛选(High-Throughput Computational Screening, HTCS)技术,进一步优化了分子模拟的过程,提高了预测不同 MOFs 材料的气体吸附和分离性能的效率,从而加速新型 MOFs 的发现和评估[18]。HTCS 技术通过分子模拟预测大量 MOFs 的气体吸附分离性能,不仅评估了材料的应用潜力,而且确立了其结构 - 性能关系,从而为实验工作提供了有力的指导[18-20]。Lin 等[21]还强调了 HTCS 在筛选高效 MOF 气体分离材料方面的潜力。随着计算资源的增强,HTCS 能处理更大规模的数据,为探索新型高性能 MOFs 开辟了更广阔的道路[22]。



(a) 气体吸附、分离和储存方面的文章统计对比图

(b) 结合机器学习与分子模拟探索不同气体的发文比例随时间变化情况

(c) 2014—2024发文数量及趋势

图 1 MOF 在气体吸附和分离模拟方面的对比图和趋势图

Fig. 1 Comparison and trend graphs of MOF in gas adsorption and separation simulation

注:统计 Web of Science 文章数量截止于 2024 年 6 月。

随着 MOFs 数量的迅速增长,在探索成千上万的 MOFs 结构与性能关系时,传统的 HTCS 效率不高,同时也面临着大量时间与成本的投入,这些挑战促使研究者寻求更有效的解决方案。21 世纪,随着人工智能(Artificial Intelligence, AI)技术的迅猛发展,在材料科学领域,AI 的介入为研究带来了革命性的变革。面对 HTCS 效率不高和成本问题,机器学习(Machine Learning, ML)作为一种数据驱动方法被引入到 MOFs 研究中,这在材料科学领域实现了突破性进展,显著降低了对传统实验和模拟的依赖[23]。ML 技术不仅优化了研究流程,还被广泛地应用于材料发现、性质预测、结构分析及逆向设计等方面[24]。例如 Trickett 等[25]通过 ML 模型识别了 MOFs 的关键结构特征,进而提升了化学直觉的准确性,并验证了 MOFs 化学性质的可预测性。这种方法使得研究者能够更有效地整合和分析来自实验、理论和模拟的数据,促进了基于 ML 的新工作流程的发展,对材料研究产生了重要影响[26]。ML 模型在精确预测高性能材料方面表现卓越,特别是在处理复杂系统时,有效地揭示了 MOFs 的结构 - 性质 - 性能之间的复杂联系,从而大幅提高了筛选效率。当前,针对 MOFs 的 ML 算法主要应用于预测气体存储和分离性能[27-30]、氧化态[31]、热容[32],还包括 MOFs 原子部分电荷分配[33]以及预测其作为热泵的性能等方面[34-35],为 MOFs 研究和应用提供了重要工具和方法。图 1(b)展示了结合 ML 与分子模拟的不同气体研究在各年份的发文占比。图 1(c)展示了 2014 年起 ML 与模拟方法的年发文量,两者数量都在稳定上升,ML 方法亦展现出研究潜力。

本文系统探讨 ML 在 MOFs 气体吸附分离领域是如何辅助 HTCS,并展示了这些技术是如何推

动 MOFs 的快速发现与设计。

1　机器学习

传统的 HTCS 方法,如 GCMC[35]、MD[36]和密度泛函理论(Density Functional Theory, DFT)等[37],对于新型 MOF 材料的评估和设计至关重要。但随着 MOF 数量的不断增长,这些方法在评估材料性能时存在计算成本高、速度慢等缺点[38]。虽然 HTCS 取得显著进步,但生成的大量多维数据需借助大数据分析技术进行深入解析。近年来,ML 的兴起,为 HTCS 提供了强有力的支持[39]。

ML 模型构建的核心环节包括材料特征数据的选择、目标性能筛选、算法选取及性能评估,这 4 个步骤共同决定模型的准确性和可靠性。在选择材料特征数据时,首先选择合适的材料数据库,以确保排除结构错误的材料数据;其次,挑选合适的材料特征描述符进行数据预处理,如标准化处理,以消除不同 MOF 描述符之间的差异,保证数据的一致性和准确性;再次,模型的输出目标性质应与研究问题紧密相关(如气体吸附性能或整体性能),不同的数据类型和研究问题需要不同的算法,因此,采用多种成熟的算法进行训练,并在学习过程中优化参数是至关重要的;最后,对模型进行严格的性能评估,使用皮尔逊相关系数和均方误差等指标,以确保预测结果的准确性。通过这些步骤,不仅提高了 HTCS 的效率,还有助于揭示新材料的结构与性能之间的关系[40]。

1.1　材料特征数据的选择

在特定应用领域构建 ML 模型时,首要任务是选择合适的数据库。不同数据库构建的模型在特定气体吸附分离预测时存在显著差异。这些数据库主要分为两类:基于已实验合成材料构成的数据库(experimental MOFs, eMOFs),以及依据化学理论和计算设计构建的假设性数据库(hypothetical MOFs,hMOFs)[41-43]。最近,Bobbitt 等[43]推出的 MOFX-DB 数据库,囊括了超过 16 万个 MOFs 和 286 个沸石材料的详细结构和吸附数据,还包含 GCMC 数据、力场参数等,为模拟研究提供了复现基础。

此外,ML 模型的性能也取决于特征工程的优化,其中,关键在于挑选合适的特征描述符。在不同应用场景中,特征描述符的选择对模型性能至关重要。在 MOFs 计算模拟中,特征描述符通常被划分为 4 类,包括:结构描述符、化学描述符、拓扑描述符以及能量描述符,如图 2 所示。

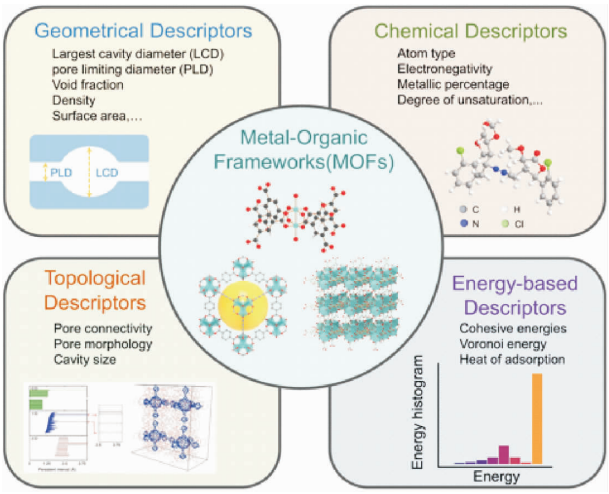


图 2　MOF 特征描述符的分类

Fig. 2　Classification of MOF Characteristic Descriptors

注:拓扑描述符来自 Smit 的研究[44]。

这些描述符从不同角度捕捉 MOFs 的关键属性,以确保模型能够全面理解和精确预测材料的性能。常用的几何描述符,如密度(Density, ρ)、孔容(Available Pore Volume, Va)、体积比表面积(Volumetric Surface Area, VSA)、重量比表面积(Gravimetric Surface Area, GSA)、孔隙率(Void Fraction, Ø)、最大孔直径(Largest Cavity Diameter, LCD)、受限孔直径(Pore Limited Diameter, PLD)[45]及其比例(LCD/PLD)等[46]。这些描述符可以使用分子探针获得,对于预测 MOFs 的吸附和分离性能具有重要作用。虽然能直观反映 MOFs 的基本性质,但它们本质上是一维的,限制了模型的泛化能力。因此,Fernandez 等[47]开始探索多维描述符,例如引入原子属性加权径向分布函数(Atomic Property Weighted Radial Distribution Functions, AP-RDF),更全面地捕捉 MOFs 的特性,从而提升模型性能。同时,构建模块的化学多样性也不容忽视。许多研究团队采用化学描述符,如原子电荷[48]、原子数量[49]、偶极矩[50]等,以丰富

模型的输入数据。Daglar 等[51]使用原子类型、金属百分比(Metallic Percentage, MP)、不饱和度(Degree of Unsaturation, DU)、氧与金属的比率(Oxygen-to-metal Ratio, O-to-M)等描述符,为 ML 提供了丰富的化学描述符。例如 Fanourgakis 等[52]基于结构中的原子类型描述符来预测 CO_2 和 CH_4 的吸附能力。与传统的构建块描述符相比,使用原子类型描述符的 ML 模型可以提供更准确的预测。相对于构建块方法,这种方法的优势在于更容易用于研究不同材料,最终结果清晰地展示模型的通用性和可迁移性。这些描述符可以揭示 MOFs 内部的化学相互作用,这类带有化学性质的数据通常具有广泛的分布范围,因此,需要进行特征转换,例如对亨利系数(Henry's coefficient, KH)值进行对数归一化处理。简单的化学描述符,如分子量或极性,通常不能准确描述材料在实际环境中的行为。因此,Bucior 等[53]开发了吸附剂-吸附质能量直方图来预测 MOFs 中气体吸附性能。这种能量描述符用于衡量 MOFs 和客体分子之间的相互作用强度,尽管计算成本较高,但与其他类型的描述符相比,这类描述符可以更好地捕捉主客体相互作用,从而提高对材料吸附分离等性能的预测准确度。此外,分子模拟输出的吸附热(Heat of Adsorption, Q_{st}^0)等能量描述符数据也常用于训练 ML 模型[54]。拓扑描述符通过量化 MOFs 的拓扑特性,捕捉了 MOFs 中金属节点与有机连接体的连接与排列方式,形成三维网络。利用这些描述符,研究人员能够精确地量化并比较不同 MOFs 结构间的相似性与差异性[44]。

Smit 团队使用 MOFs 晶体结构的邻接矩阵计算修正的自相关函数(Revised Autocorrelation, RACs)描述符,这种基于图的新型描述符成功预测了 MOFs 的 CO_2 和 CH_4 吸附性能[55]。鉴于 MOFs 的某些特征可能与关键目标属性(例如气体吸附量等)的相关性并不显著。因此,在 ML 模型构建过程中,根据具体研究场景精心选择恰当的特征描述符显得尤为重要。

在选择特征后,数据预处理也是 ML 模型开发中的关键步骤,旨在提高数据的一致性和可靠性。特征工程是在数据经过预处理后进一步优化模型性能的关键步骤。特征工程通过选择、转换或创建新特征,提取最具代表性的特征,增强模型的预测能力。然而,特征过多可能导致过拟合现象,即模型拟合了数据中的噪声,而非捕捉真实趋势。因此,在构建 MOFs 的机器学习模型时,特征描述符的选择应谨慎,避免引入不必要的复杂度和噪声。面对高维特征空间时,特征选择和降维是解决特征冗余和噪声的有效手段。高维特征不仅增加模型的复杂度和计算成本,还可能引入过多无关特征,导致模型泛化能力下降。因此,通过合理的特征选择技术剔除冗余或无关的特征,以确保模型的稳健性和泛化能力至关重要。常用的特征选择方法包括 Filter、Wrapper 和 Embedded 策略,分别从不同角度优化特征集。Filter 方法根据特征与目标变量的相关性筛选贡献较小的特征,而 Wrapper 方法则评估特征在多个模型中的表现,进而选择最佳特征组合。Embedded 方法则将特征选择与模型训练相结合,通过嵌入算法直接优化模型性能。Mukherjee 等[56]将特征描述符分为一阶和二阶,并使用上述策略进行特征选择,以提升模型的准确性和泛化性。

1.2 目标性质数据的筛选

ML 模型构建的关键之一在于精确地选择适应特定应用领域的目标性质和标签数据。鉴于气体储存和分离领域的数据资源丰富,大量 ML 研究集中于预测 MOFs 在暴露于单组分或多组分气体(如 CH_4、H_2 和 CO_2 等)环境下的吸附性能[57],在这些研究中,MOFs 通常作为吸附剂或膜的角色,ML 模型被用于预测多个关键性能指标,如气体的吸附量、选择性、吸附量与选择性的权衡值[58]、工作容量、渗透性及渗透选择性等[59]。这些性能指标的准确预测,不仅有助于理解 MOFs 的吸附机理,还为材料的设计和应用提供重要的数据支持。

1.3 机器学习算法的选择

ML 主要分为 4 大类:监督学习、无监督学习、半监督学习和强化学习。图 3 中列出了这些类别下的一些常用算法。

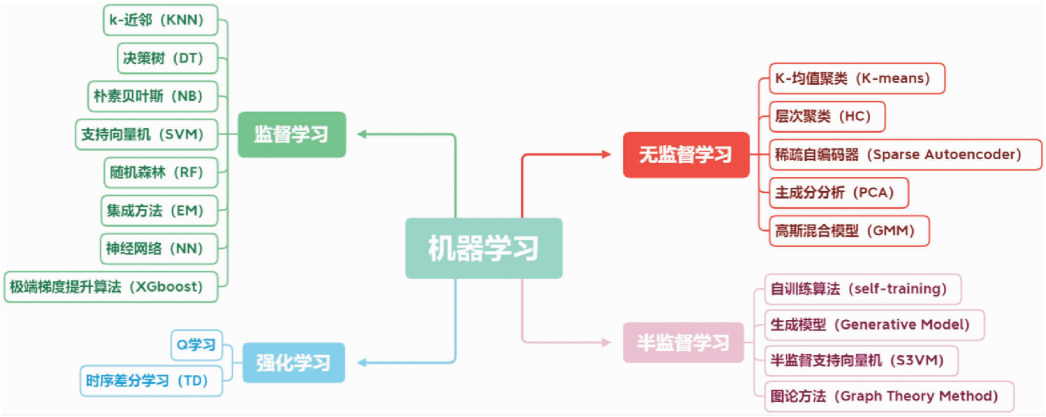


图 3　机器学习的主要分类及其常用算法概览

Fig. 3　Overview of Machine Learning Categories and Common Algorithms

1.3.1　监督学习

监督学习不仅是 ML 的核心组成部分，在 MOFs 吸附分离模拟领域也扮演着至关重要的角色。这种学习方式依赖于对精心标记的数据集进行模型训练，它的广泛应用体现在能够准确模拟和预测材料的吸附和分离性能，从而显著提高 HTCS 和实验设计的效率。在监督学习中，训练集包括输入变量 x 和对应的输出变量 y，即每个输入样本都有一个明确的标签或结果。通过这些数据，算法学习输入与输出之间的映射关系。训练完成后，这种映射关系可用于对新的、未知的数据进行分类[60]或回归分析。在监督学习中，模型旨在最小化预测值和实际值之间的误差，以优化模型的性能。常见的监督学习算法包括：神经网络（Neural Networks，NN）、决策树（Decision Trees，DT）、支持向量机（Support Vector Machines，SVM）、最近邻算法（K-Nearest Neighbors，KNN）、随机森林（Random Forest Regression，RF）和朴素贝叶斯分类器（Naive Bayes，NB）等，如图 4 所示。

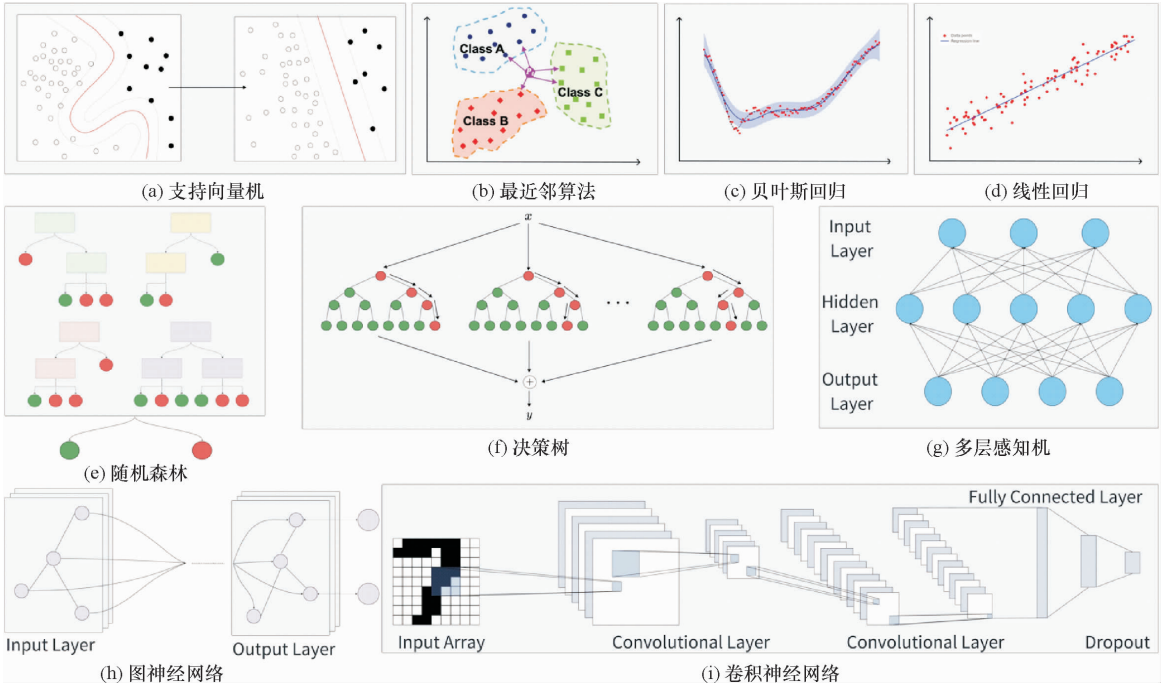


图 4　机器学习算法原理

Fig. 4　Schematic of machine learning algorithms

1.3.2 无监督学习

无监督学习是ML的另一种形式,它不依赖于预先标记的数据来训练模型。输入数据没有相关的输出标签,因此,其目标不是预测标签,而是探索数据的内在结构和模式。通过这种学习方法,算法尝试识别数据中的相似性或规律性,从而进行数据聚类、异常检测或者特征降维等任务。

常见的无监督学习技术包括聚类算法,如K均值(K-means)和层次聚类(Hierarchical clustering),这些算法能够将数据集合中的项分组到相似的集合中。

1.3.3 半监督学习

半监督学习结合了监督学习和无监督学习的特点,使用部分标记的样本和大量无标签样本进行训练。这种方法适用于标签获取成本高或困难的情况,通过利用有限的标签数据和探索未标记数据的结构,提高模型的预测能力。常用的半监督技术包括自训练和图基方法,标记自训练数据训练模型,通过再用预测结果逐步标记未标记数据;图基方法通过构建数据点间的关系图,利用图结构传播标签信息,增强学习效果。

1.3.4 强化学习

强化学习是ML的一个分支,它侧重于在不断变化的环境中优化决策策略。在强化学习框架中,一个智能体通过与环境的交互来学习最佳行为,目标是最大化获取的总奖励。智能体在每个时间步接收环境的状态信息,基于这些信息做出决策,执行动作,并接收新的状态信息。

这个学习过程主要依赖于试错机制和奖励信号,不需要预先定义的数据标签。智能体的最终目标是调整其决策策略,以识别并实施能够带来最大长期收益的行为。在强化学习中,主要的技术之一是Q学习。Q学习是一种无模型的强化学习算法,它不依赖于环境的具体模型,而是直接在与环境交互的过程中更新动作的价值(即Q值)。Q值表示在特定状态下采取某一动作的预期收益。通过反复更新这些Q值,智能体学习到每个状态下哪些动作能带来最高的长期奖励,从而逐步形成最优策略。另一种关键的强化学习技术是时序差分学习,即介于蒙特卡罗方法和动态规划之间的一种方法。它在从一个状态到下一个状态的转换中立即更新值函数,而不需要等到一个完整的序列结束。时序差分学习使用即时奖励加上对后续状态价值的估计来更新当前状态的价值,这种方法使得学习过程可以在不完全了解环境动态的情况下进行,从而增强了算法的适应性和效率。

1.4 机器学习算法评估指标

ML任务中,预测值与实际值之间的误差是衡量模型性能的关键。平均绝对误差(Mean Absolute Error, MAE)适用于回归问题,能够直观反映模型的平均误差,尤其在不希望放大异常值影响的场景中使用效果更好。均方误差(Mean Squared Error, MSE)和均方根误差(Root Mean Squared Error, RMSE)则在平方误差的基础上放大了异常值的影响,适用于对异常值更敏感的场景。平均绝对百分比误差(Mean Absolute Percentage Error, MAPE)用于不同数量级数据间的比较。皮尔逊相关系数则衡量预测值与实际值的线性相关性,适合用于评估模型与实际情况的一致性。

对于分类模型,混淆矩阵描述了模型的完整性能,适合分析模型在不同类别的表现。例如True Positives(TP)代表真阳性,False Negative(FN)代表假阳性。阳性率(True Positive Rate, TPR)和假阳性率(False Positive Rate, FPR)分别反映模型识别正例的能力和误识别负例的情况,适用于深入分析模型分类效果的场合。准确率(Accuracy, ACC)是分类任务中的常见评估指标,特别适用于类别平衡的二分类任务,而在类别不平衡的场景下,曲线下面积(Area Under the Curve of ROC, AUC)更为合适。召回率(Recall)和精准率(precision)则分别衡量模型识别出的正例比例和实际正例的比例,适用于需要权衡模型检出能力和准确度的场合。关于分类和回归等评估指标的详细信息见表1。

表 1　机器学习算法评估指标

Table 1　Evaluation metrics for machine learning algorithms

模型评估参数	指标定义
$\mathrm{MAE}(x,y) = \frac{1}{n}\sum_{i=1}^{n} \lvert x_i - y_i \rvert$	$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$
$\mathrm{MSE}(x,y) = \frac{1}{n}\sum_{i=1}^{n} (x_i - y_i)^2$	$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$
$\mathrm{RMSE}(x,y) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (x_i - y_i)^2}$	$\text{True Positive Rate} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$
$\mathrm{MAPE}(x,y) = \frac{100}{n}\sum_{i=1}^{n} \left\lvert \frac{x_i - y_i}{x_i} \right\rvert$	$\text{False Cositive Rate} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}$
$R(x,y)^2 = 1 - \sum_{i=1}^{n} \frac{(x_i - y_i)^2}{(y_i - \bar{y})^2}$	
$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$	
$\mathrm{AUC}(x,y) = \frac{1}{2}\sum_{i=1}^{n-1} (x_{i+1} - x_i)(y_{i+1} + y_i)$	

2　结论与展望

随着合成和计算生成的 MOFs 的数量以及潜在应用的迅速增长,传统的全面测试方法已经不再切实可行,因此,需要结合实验、计算和 ML 等一系列研究方法来确定特定应用的高性能材料。基于这些讨论,本文概括了结合实验、计算模拟与数据驱动方法时所面临的潜在机遇和挑战。此外,提出了针对现存问题的具体解决策略,并为未来的研究方向提供了预测和建议。

在数据驱动的研究方法中,确保所建立的模型具有透明度、可解释性和可验证性至关重要。在 ML 模型的开发过程中,经常会遇到数据集中存在缺失或不现实数值的问题。面对这些挑战,研究人员需要做出决策,如选择保留还是删除这些数值,以及如何处理未完整记录的数据。这些决策可能引入偏差,影响研究结果的可重复性。为了克服这些挑战,关键在于精心设计并透明化整个研究流程。这包括使用开源的计算工具,并以自动化方式产生可复现的结果,以确保研究者和其他相关方可以清楚地理解模型的构建过程及对数据的处理方式。此外,有一些高性能的 ML 模型因其“黑箱”特性而缺乏足够的解释性,致使模型的决策过程和机制难以被广泛理解和接受。因此,如何在模型的高性能与高解释性之间找到平衡,成为了一项关键挑战。

ML 辅助计算模拟与实验验证之间的研究仍处于起步阶段,部分原因在于 ML 模型的开发通常由专注于计算的团队进行,而改进 ML 模型往往需要实验带来的反馈,因此,这凸显了实验组和理论组之间合作的重要性。通过这种合作,可以进一步改进 ML 模型。这样的合作涉及跨组织的原始数据传输,同时需要解决可能深刻影响模型准确性和泛化能力的问题,例如模拟中的缺陷建模和非标准的合成、表征以及材料测试。因此,数据驱动方法不只依赖于数据采集和 ML 模型开发工具,还需要多学科科学家之间紧密的合作。这种合作既有助于理论家开发更优秀的 ML 模型,又能帮助实验者发现之前未察觉到的样品。ML 和分子模拟的初衷是辅助实验,但如今模拟研究涉及的材料种类大幅增加,而实验验证的速度却无法跟上。这也导致了高通量合成研究相对较少,因为其需要巨大的时间和金钱成本。因此,实验者、理论家和计算科学家之间的密切协作是推动这一领域进步的关键。

随着 OpenAI 开发的 ChatGPT 大语言模型引发广泛关注,各个研究领域正积极探索将大型语

言模型整合进自己的研究中。作为自然语言处理(Natural Language Processing, NLP)技术的一部分,大语言模型以其处理和生成大量文本数据的能力而著称。这些模型通常依赖于对大规模的数据集进行有效训练,能够从文本中学习语言的深层结构和语义。然而,在数据稀缺的领域,这种依赖可能构成挑战。因此,在MOFs吸附模拟领域,将大语言模型应用于数据分析和模式识别,成为了一个重要趋势。基于此,建立一个统一的行业标准并合力构建一个由多个团队维护的数据库,将极大地促进MOFs研究社区的协同工作。通过共享和分析大量的实验和模拟数据,不仅有助于加速材料创新和发现,还有助于开发并优化数据处理方法和算法,从而开启探索更广泛材料空间的新篇章。

参考文献:

[1] Yaghi O M, O'Keeffe M, Ockwig N W, et al. Reticular synthesis and the design of new materials[J]. Nature, 2003, 423(6941): 705-714.

[2] Furukawa H, Ko N, Go Y B, et al. Ultrahigh porosity in metal-organic frameworks[J]. Science, 2010, 329(5990): 424-428.

[3] Baumann A E, Han X, Butala M M, et al. Lithium thiophosphate functionalized zirconium MOFs for Li-S batteries with enhanced rate capabilities[J]. Journal of the American Chemical Society, 2019, 141(44): 17891-17899.

[4] Li L, Xue H, Wang Y, et al. Solvothermal metal metathesis on a metal-organic framework with constricted pores and the study of gas separation[J]. ACS Applied Materials & Interfaces, 2015, 7(45): 25402-25412.

[5] Ferguson A, Liu L, Tapperwijn S J, et al. Controlled partial interpenetration in metal-organic frameworks[J]. Nature Chemistry, 2016, 8(3): 250-257.

[6] Wu H, Gong Q, Olson D H, et al. Commensurate adsorption of hydrocarbons and alcohols in microporous metal organic frameworks[J]. Chemical Reviews, 2012, 112(2): 836-868.

[7] Fan W D, Zhang X K, Kang Z X, et al. Isoreticular chemistry within metal-organic frameworks for gas storage and separation[J]. Coordination Chemistry Reviews, 2021, 443: 213968.

[8] Wang C, Cheng P, Yao Y, et al. In-situ fabrication of nanoarchitectured MOF filter for water purification[J]. Journal of Hazardous Materials, 2020, 392: 122164.

[9] Mallakpour S, Nikkhoo E, Hussain C M. Application of MOF materials as drug delivery systems for cancer therapy and dermal treatment[J]. Coordination Chemistry Reviews, 2022, 451: 214262.

[10] Kreno L E, Leong K, Farha O K, et al. Metal-organic framework materials as chemical sensors[J]. Chemical Reviews, 2012, 112(2): 1105-1125.

[11] Yang D, Gates B C. Catalysis by metal organic frameworks: Perspective and suggestions for future research[J]. Acs Catalysis, 2019, 9(3): 1779-1798.

[12] Sundriyal S, Kaur H, Bhardwaj S K, et al. Metal-organic frameworks and their composites as efficient electrodes for supercapacitor applications[J]. Coordination Chemistry Reviews, 2018, 369: 15-38.

[13] Fan W, Zhang X, Kang Z, et al. Isoreticular chemistry within metal-organic frameworks for gas storage and separation[J]. Coordination Chemistry Reviews, 2021, 443: 213968.

[14] Qiao Z, Zhang K, Jiang J. In silico screening of 4764 computation-ready, experimental metal-organic frameworks for CO_2 separation[J]. Journal of Materials Chemistry A, 2016, 4(6): 2105-2114.

[15] 杨文远,梁红,乔智威. 高通量筛选金属-有机框架:分离天然气中的硫化氢和二氧化碳[J]. 化学学报,2018,76(10):785-792.

[16] 蔡铖智,李丽凤,邓小梅,等. 基于机器学习和高通量计算筛选金属有机框架的甲烷/乙烷/丙烷分离性能[J]. 化学学报,2020,78(5):427-436.

[17] Glover J, Besley E. A high-throughput screening of metal-organic framework based membranes for biogas upgrading[J]. Faraday Discussions, 2021, 231: 235-257.

[18] Daglar H, Keskin S. Recent advances, opportunities, and challenges in high-throughput computational screening of MOFs for gas separations[J]. Coordination Chemistry Reviews, 2020, 422: 213470.

[19] Colón Y J, Snurr R Q. High-throughput computational screening of metal-organic frameworks[J]. Chemical Society Reviews, 2014, 43(16): 5735-5749.

[20] Jablonka K M, Ongari D, Moosavi S M, et al. Big-data science in porous materials: Materials genomics and machine learning[J]. Chemical Reviews, 2020, 120(16): 8066-8129.

[21] Lin J, Liu Z, Guo Y, et al. Machine learning accelerates the investigation of targeted MOFs: Performance prediction, rational design and intelligent synthesis[J]. Nano Today, 2023, 49: 101802.

[22] 刘治鲁,李炜,刘昊,等. 金属有机骨架的高通量计算筛选研究进展[J]. 化学学报,2019,77(4):323-339.

[23] Borboudakis G, Stergiannakos T, Frysali M, et al. Chemically intuited, large-scale screening of MOFs by machine learning techniques[J]. NPJ Computational Materials, 2017. doi:10.1038/S41524-017-0045-8.

[24] Himanen L, Geurts A, Foster A S, et al. Data-driven materials science: Status, challenges, and perspectives[J]. Advanced Science, 2019, 6(21): 1900808.

[25] Trickett C A, Helal A, Al-Maythalony B A, et al. The chemistry of metal-organic frameworks for CO_2 capture, regeneration and conversion[J]. Nature Reviews Materials, 2017, 2(8): 1-16.

[26] Mallakpour S, Nikkhoo E, Hussain C M. Application of MOF materials as drug delivery systems for cancer therapy and dermal treatment[J]. Coordination Chemistry Reviews, 2022, 451: 214262.

[27] Yang D, Babucci M, Casey W H, et al. The surface chemistry of metal oxide clusters: From metal-organic frameworks to minerals[J]. ACS Central Science, 2020, 6(9): 1523-1533.

[28] Wang Q, Astruc D. State of the art and prospects in metal-organic framework (MOF)-based and MOF-derived nanocatalysis [J]. Chemical Reviews, 2019, 120(2): 1438-1511.

[29] Suresh K, Matzger A J. Enhanced drug delivery by dissolution of amorphous drug encapsulated in a water unstable metal-organic framework (MOF)[J]. Angewandte Chemie International Edition, 2019, 58(47): 16790-16794.

[30] Wang H S. Metal-organic frameworks for biosensing and bioimaging applications[J]. Coordination Chemistry Reviews, 2017, 349: 139-155.

[31] Kreno L E, Leong K, Farha O K, et al. Metal-organic framework materials as chemical sensors[J]. Chemical Reviews, 2012, 112(2): 1105-1125.

[32] Sheberla D, Bachman J C, Elias J S, et al. Conductive MOF electrodes for stable supercapacitors with high areal capacitance[J]. Nature Materials, 2017, 16(2): 220-224.

[33] Korolev V V, Mitrofanov A, Marchenko E I, et al. Transferable and extensible machine learning-derived atomic charges for modeling hybrid nanoporous materials[J]. Chemistry of Materials, 2020, 32(18): 7822-7831.

[34] Raza A, Sturluson A, Simon C M, et al. Message passing neural networks for partial charge assignment to metal-organic frameworks[J]. The Journal of Physical Chemistry C, 2020, 124(35): 19070-19082.

[35] Shi Z, Liang H, Yang W, et al. Machine learning and in silico discovery of metal-organic frameworks: Methanol as a working fluid in adsorption-driven heat pumps and chillers[J]. Chemical Engineering Science, 2020, 214: 115430.

[36] Adatoz E, Kesin S. Application of MD simulations to predict membrane properties of MOFs[J]. Journal of Nanomater, 2015,124:136867.

[37] Adatoz E, Avci A K, Keskin S. Opportunities and challenges of MOF-based membranes in gas separations[J]. Separation and Purification Technology, 2015, 152: 207-237.

[38] Altundal O F, Haslak Z P, Keskin S. Combined GCMC, MD, and DFT approach for unlocking the performances of COFs for methane purification[J]. Industrial & Engineering Chemistry Research, 2021, 60(35): 12999-13012.

[39] Ren E, Guilbaud P, Coudert F X. High-throughput computational screening of nanoporous materials in targeted applications [J]. Digital Discovery, 2022, 1(4): 355-374.

[40] 程洪,葛美伶,司天宇,等. 机器学习辅助金属材料力学性能预测[J]. 材料研究与应用,2023,17(6):1070-1077.

[41] Altintas C, Altundal O F, Keskin S, et al. Machine learning meets with metal organic frameworks for gas storage and separation[J]. Journal of Chemical Information and Modeling, 2021, 61(5): 2131-2146.

[42] 李炜, 梁添贵, 林元创, 等. 机器学习辅助高通量筛选金属有机骨架材料[J]. 化学进展, 2022, 34(12): 2619-2637.

[43] Bobbitt N S, Shi K, Bucior B J, et al. MOFX-DB: An online database of computational adsorption data for nanoporous materials[J]. Journal of Chemical & Engineering Data, 2023, 68(2): 483-498.

[44] Lee Y, Barthel S D, Dłotko P, et al. Quantifying similarity of pore-geometry in nanoporous materials[J]. Nature Communications, 2017, 8(1): 1-8.

[45] Haldoupis E, Nair S, Sholl D S. Efficient calculation of diffusion limitations in metal organic framework materials: A tool for identifying materials for kinetic separations[J]. Journal of the American Chemical Society, 2010, 132(21): 7528-7539.

[46] Liang H, Yang W, Peng F, et al. Combining large-scale screening and machine learning to predict the metal-organic frameworks for organosulfurs removal from high-sour natural gas[J]. APL Materials, 2019, 7(9):091101.

[47] Fernandez M, Trefiak N R, Woo T K. Atomic property weighted radial distribution functions descriptors of metal-organic frameworks for the prediction of gas uptake capacity[J]. The Journal of Physical Chemistry C, 2013, 117(27): 14095-14105.

[48] Cho E H, Deng X, Zou C, et al. Machine learning-aided computational study of metal-organic frameworks for sour gas sweetening[J]. The Journal of Physical Chemistry C, 2020, 124(50): 27580-27591.

[49] Wu X, Xiang S, Su J, et al. Understanding quantitative relationship between methane storage capacities and characteristic properties of metal-organic frameworks based on machine learning[J]. The Journal of Physical Chemistry C, 2019, 123(14): 8550-8559.

[50] Anderson R, Rodgers J, Argueta E, et al. Role of pore chemistry and topology in the CO_2 capture capabilities of MOFs: From molecular simulation to machine learning[J]. Chemistry of Materials, 2018, 30(18): 6325-6337.

[51] Daglar H, Keskin S. Combining machine learning and molecular simulations to unlock gas separation potentials of MOF membranes and MOF/polymer MMMs[J]. ACS Applied Materials & Interfaces, 2022, 14(28): 32134-32148.

[52] Fanourgakis G S, Gkagkas K, Tylianakis E, et al. A universal machine learning algorithm for large-scale screening of materials[J]. Journal of the American Chemical Society, 2020, 142(8): 3814-3822.

[53] Bucior B J, Bobbitt N S, Islamoglu T, et al. Energy-based descriptors to rapidly predict hydrogen storage in metal-organic frameworks[J]. Molecular Systems Design & Engineering, 2019, 4(1): 162-174.

[54] Li L, Duan Y, Liao S, et al. Adsorption and separation of propane/propylene on various ZIF-8 polymorphs: Insights from GCMC simulations and the ideal adsorbed solution theory (IAST)[J]. Chemical Engineering Journal, 2020, 386: 123945.

[55] Moosavi S M, Nandy A, Jablonka K M, et al. Understanding the diversity of the metal-organic framework ecosystem[J]. Nature Communications, 2020, 11(1): 1-10.

[56] Mukherjee K, Colón Y J. Machine learning and descriptor selection for the computational discovery of metal-organic frameworks[J]. Molecular Simulation, 2021, 47(10/11): 857-877.

[57] Yan Y, Zhang L, Li S, et al. Adsorption behavior of metal-organic frameworks: From single simulation, high-throughput computational screening to machine learning[J]. Computational Materials Science, 2021, 193: 110383.

[58] Shi Z, Yang W, Deng X, et al. Machine-learning-assisted high-throughput computational screening of high performance metal-organic frameworks[J]. Molecular Systems Design & Engineering, 2020, 5(4): 725-742.

[59] Daglar H, Keskin S. Recent advances, opportunities, and challenges in high-throughput computational screening of MOFs for gas separations[J]. Coordination Chemistry Reviews, 2020, 422: 213470.

[60] 王捍贫,张博闻. 分离逻辑的技术基础与研究现状[J]. 广州大学学报(自然科学版),2019,18(2):1-9.

【责任编辑: 陈 钢】