

机器学习在新材料筛选方面的应用进展

戚兴怡[†] 胡耀峰[†] 王若愚 杨雅清 赵宇飞*

(北京化工大学化学学院 化工资源有效利用国家重点实验室 北京 100029)

摘要 新材料产业是许多相关领域技术变革的基础,也是新能源、航空航天、电子信息等高新技术产业发展的先导.传统研发手段由于成本高、效率低、商业化周期长等不利因素无法满足现代社会的发展需求.近年来大数据与人工智能不断深入结合,以数据驱动为核心的机器学习在新材料设计、筛选以及性能预测等方面取得巨大进展,极大促进了新材料的研发与应用.本综述总结了机器学习的基本过程及其在材料科学中常用的算法和相关材料数据库,重点介绍了机器学习在不同功能上的应用以及在催化剂材料、锂离子电池、半导体材料和合金材料等领域的性能预测和材料开发中的最新进展,并对其下一步在新材料应用方面提出展望.

关键词 机器学习; 材料科学; 材料基因组; 高通量计算

Recent Advance of Machine Learning in Selecting New Materials

Qi, Xingyi[†] Hu, Yaofeng[†] Wang, Ruoyu Yang, Yaqing Zhao, Yufei*

(State Key Laboratory of Chemical Resource Engineering, College of Chemistry, Beijing University of Chemical Technology, Beijing 100029)

Abstract The new material industry is the foundation of technological change in many related fields, and also the forerunner of the development of new energy, aerospace, electronic information and other high-tech industries. Traditional means cannot meet the development needs of modern society because of disadvantages such as high cost, low efficiency and long commercial cycle. In recent years, with the application of big data combined with artificial intelligence in a deeper degree, data-driven machine learning has made great progress in the design, screening and performance prediction of new materials, which has greatly promoted the development and application of new materials. In this review, the basic process of machine learning, the algorithms commonly used in materials science and the relevant materials database are summarized. This review focuses on the application of machine learning in different functions, as well as the performance prediction in the fields of catalyst materials, lithium-ion batteries, semiconductor materials and alloy materials, presenting the latest progress in materials development. Finally, machine learning in the application of new materials are analyzed and prospected.

Keywords machine learning; materials science; material genome; high throughput computing

1 引言

在知识经济新时代,材料、能源和信息并称为现代科学技术的三大支柱,其作用和意义尤为重要.传统材料主要依靠科学家的经验或者偶然,即试错法来发现,其通常存在效率低、成本高、研发周期长,无法满足生产与生活的实际需求^[1]等问题.随着第一性原理计算方法和工具的飞速发展以及超级计算机性能的指数级提升,材料科学家可以通过非实验的方法快速、真实地模拟计算材料的各种结构与性能,材料科学进入了计算科学的范畴(如图 1).然而随着理论与实验数据的不断积累,如何高效地利用目前已知信息缩短材料从筛选到应用的开发周期成为了材料科学领域和工程应用领域发展的瓶颈.

数据驱动科学是科学研究的第四范式,作为其核心技术的机器学习(Machine Learning, ML)在物理科学、游戏开发、基因组计划、医疗健康、气象监测以及金融市场分析^[2]等众多科学和工程领域具有关键作用.机器学习是发现高维数据背后隐藏的统计规律的强有力武器^[3],其为新材料领域发展提供了新的契机^[4].2011年,“材料基因组计划(Materials Genome Initiative, MGI)”的提出,标志着材料信息学的正式形成.材料基因组计划在大量数据的基础上建立了诸多数据库平台^[5-7],基于这些材料数据库,机器学习针对特定材料属性建立相对应的算法模型,可快速实现对材料性能的预测,从而有望加速新材料的筛选和设计,缩短材料研发周期^[8].机器学习正在利用化学和材料数据改变材料研发领域^[9],

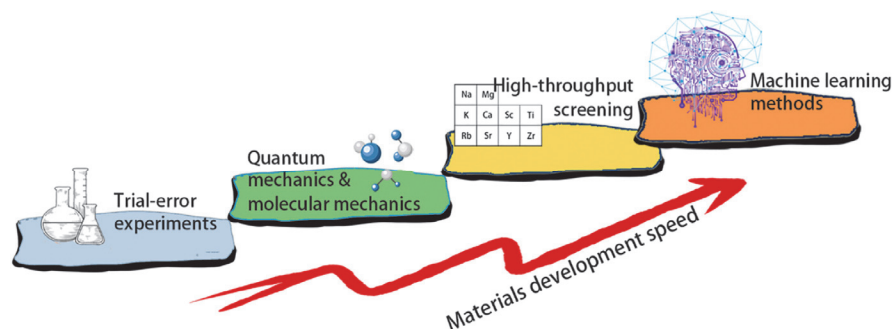
* E-mail: zhaoyufei@mail.buct.edu.cn

[†] 同等贡献(These authors contributed equally to this work)

Received November 1, 2022; published December 23, 2022.

Project supported by the National Natural Science Foundation of China (Nos. 21922801, 22090032, 22090030) and the Natural Science Foundation of Beijing (No. 2202036).

项目受国家自然科学基金(Nos. 21922801, 22090032, 22090030)和北京自然科学基金(No. 2202036)资助.

图1 加速新材料探索方法的过程^[3]Figure 1 Accelerate the process of exploring new materials^[3]

在指导化学合成、辅助多维材料表征、获取新材料设计方法等方面展现出了巨大潜力,为新材料领域带来了强劲的发展动力,其必将对材料科学产生变革性的影响^[10-11]。另一方面,作为一种新兴的研究手段,机器学习技术虽然为材料科学领域带来了重大机遇,但其在发展过程中也同样面临着许多亟待解决的难题。

2 机器学习简介

2.1 简介

机器学习是一门研究机器如何获取新知识和新技能,并分析现有数据信息的学科。机器学习的相关概念自20世纪50年代中叶提出,经过几十年的蓬勃发展,当前已经发展成一门综合应用数学、计算机科学、生物学与工程学相关知识的多领域交叉学科,并成为人工智能(Artificial Intelligence, AI)领域中的关键技术。机器学习主要研究机器(计算机)模拟或实现人类学习行为的过程,即计算机通过模拟人类理解、思考和创造的方式,重新组织其已有的知识结构并积累经验,并不断进行自我修正和改善自身性能的过程。近年来,机器学习受到了人们的广泛关注,尤其是谷歌公司开发的人工智能Alpha Go在与围棋世界冠军的人机大战中频频取胜,将机器学习的巨大潜力展现在人们的视野当中。机器学习已被广泛应用于计算机视觉、自然语言处理、大数据挖掘、机器人、战略类游戏、医学诊断、金融证券市场分析等领域,并取得了巨大的成就。

机器学习技术可根据人们所需要的性能快速地预测出相应候选材料,目前已被世界各国的材料研究者广泛应用于新材料的筛选中。Xu等^[12]结合实验所得规律计算了一个超过300个尖晶石氧化物的数据集,并将其用于训练机器学习模型,最终预测得 $[\text{Mn}]_{\text{T}}[\text{Al}_{0.5}\text{Mn}_{1.5}\text{O}_4]$ 为一种高活性的析氧反应(Oxygen Evolution Reaction, OER)催化剂;Deringer等^[13]通过机器学习+GAPMD (Gaussian approximation potential molecular-dynamics)实现了由100000个原子(10 nm长度)组成的Si体系的液相-非晶相、非晶相-非晶相之间的组织转变过程。

此外,国内开展的关于机器学习的研究也在稳步进行^[14]。这几年来,中国的一些科研院所和单位也对其进行了开创性的研究,如杨炯团队^[15]在机器学习筛选热电材料中用主动学习预测了类金刚石热电材料的功率因子。该工作基于前期高通量计算的158个类金刚石热电材料的功率因子,用主动学习的框架结合机器学习和第一性原理计算,建立高精度的外推模型;李晓刚团队^[16]对合金设计理念进行创新,以腐蚀大数据、机器学习和相图计算作为指导合金设计的切入点,高通量筛选出具有合成可能性的高强度、抗腐蚀、抗氢脆断合金材料。但目前来看,我国与美国等先进国家相比依然存在较大差距,比如缺乏自主研发的高通量材料计算程序、高通量材料检测设备和较为实用完备的材料基因数据库^[17]等。

2.2 算法介绍

选择合适的机器学习算法是构建机器学习系统的关键步骤,它对预测精度和泛化能力有很大的影响。每一种算法都有自己的适用范围,因此,没有一种算法可以适用于所有的需求^[4]。机器学习算法种类很多,本综述将简要介绍几种影响最大的算法。

2.2.1 支持向量机

支持向量机(support vector machine, SVM)是Vapnik等提出的一种基于结构风险最小化原理的统计学习方法。近年来,支持向量机被认为是解决各种学习、分类、回归和预测问题的最先进的技术^[18]。SVM尝试寻找一个最优的决策边界(如图2),从二维空间扩展到多维空间中时,将A和B完全正确地划分开的 $kx+b=0$ 直线就成了一个超平面。以最大间隔把两类样本分开的超平面,也称之为最大间隔超平面。两类样本分割在该超平面的两侧,并且两侧距离超平面最近的样本点到超平面的距离被最大化了。求解线性可分的SVM的步骤中还涉及到构造拉格朗日函数、强对偶性转化、序列最小优化算法等数学问题。此外,核函数的引入可以处理非线性分类和回归任务,并且减少了计算量和存储数据的内存使用量,但当训练时间较长,支持向量的数量较大时,预测计算的复杂度也随之升高。

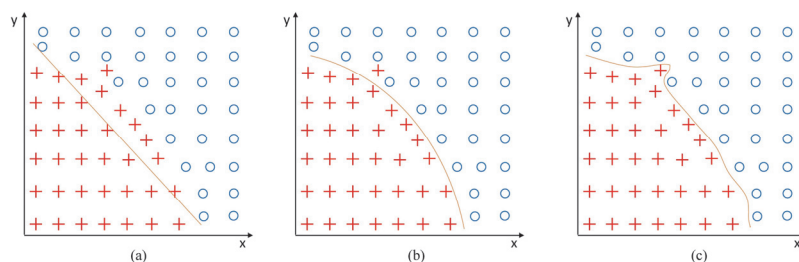


图2 ML 研究的结果主要包括(a)欠拟合, (b)正确拟合和(c)过拟合^[19]

Figure 2 The results of ML researches which mainly include (a) underfitting, (b) correct fitting and (c) overfitting^[19]

2.2.2 决策树

决策树(Decision tree, DT)是一种基本的分类与回归方法,以输入的数据为树的根,选取数据的特征属性为节点即决策点,将数据集划分为不同子集^[20](如图3),直到子集数据的类型相同,此时到达叶节点,该过程结束。其核心思想是对递归选择最优特征进行分类^[21]。不同的决策树代表了复杂程度不同的分类模型。完整的决策树构建通常有三个步骤:特征选择、决策树的生成、决策树的修剪。在ML的应用中,定量评价方法的特征选择各不相同,导致运用不同算法的决策树的产生,如ID3(信息增益特征选择)、C4.5(信息增益比特征选择)和CART(基尼指数特征选择)^[19]。这种方法非常强大,通常不需要反复调整或缩放输入数据就能得到很好的结果,且计算复杂度不高,输出结果易于理解。决策树精度不高,很少单独使用,目前常用的是基于决策树的随机森林。

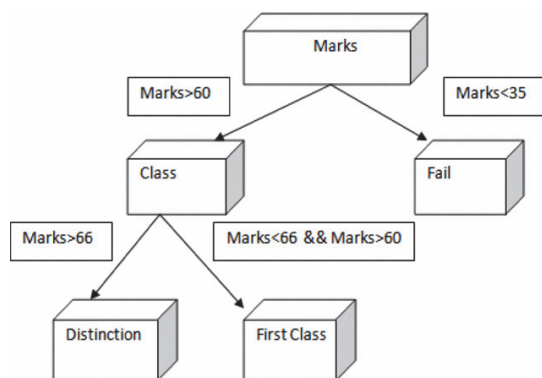


图3 决策树的示例^[22]

Figure 3 Example of a decision tree^[22]

2.2.3 人工神经网络

人工神经网络(Artificial Neural Network, ANN)由大量非线性的、相互连接紧密的元素组成,这些元素即为神经元,每一层的神经元都以逻辑并行的方式运行。因其模仿了生物系统中神经元的工作方式,所以得名人工神经网络。神经网络的基本结构通常由三个不同的层组成(如图4):输入层(数据引入神经网络)、隐藏层(对数据进行非线性映射)、输出层(产生神经网络结果)。神经网络的设计是通过在神经元之间放置权值,使用一个控制神经元输出的激活函数,并通过预测值和真实值之差

做最小化来定义一个神经元输入权值的相对重要性。在训练中,ANN通过一个向后传播过程,对权重进行调整^[23]。ANN的优势在于它能使多重非线性元素通过许多隐藏的神经元层进行转换,在这个过程中,越来越复杂和抽象的特征可以通过添加更多隐藏层或者是增加每层神经元的数量来表示^[24],因此神经网络方法更适用于物理模型难以描述的非线性复杂系统的建模。与传统的线性回归方法相比,人工神经网络数值模型具有更好的预测效果。传统的ANN采用经验风险最小化原则来最小化训练数据的误差,SVM则遵循结构风险最小化原则来建立泛化误差的上界^[25]。

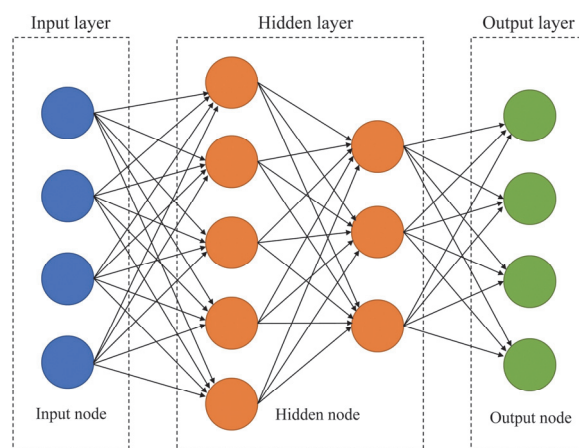


图4 神经网络算法的工作原理^[19]

Figure 4 The working principle of the algorithm of ANN^[19]

2.3 主要步骤

2.3.1 构建样本

构成样本的数据的数量和质量直接决定了预测模型的优劣。数据主要源于实验、文献、计算模拟以及材料数据库。在实际情况下,收集数据的过程会遇到很多问题,比如数据缺乏、数据重复、数据超出范围、数据异常等。如何从繁杂的数据中筛选有用的数据并确定合适的特征属性,是构建样本的关键问题。目前,材料科学的数据大致可以分成四种类型:从实验和模拟得到的物理、化学、热力学、动力学等性质;速率和化学反应温度等数据;扫描电镜的图像及材料表面照片;其他文献资料。这些数据有离散(如文本)、连续(如向量和张量)

或加权图形等多种形式^[26]. 把图像转换成数字数据是一项相当复杂的任务, 同一组图像可能产生不同的解读. 用不同的表征方法获得的同一表面的图像往往揭示了不同的信息. 因此, 需要将这些不同的信息片段正确组合在一起^[27].

2.3.2 建立模型

建立模型是指使用一组特定的非线性或线性函数来链接输入数据和输出数据, 建立条件因子与目标属性之间的关系^[28]. 此阶段通常包括选择合适的算法、调整模型结构参数、训练及测试等过程^[21]. 在典型的材料科学研究中, 条件因子和目标属性之间往往存在着复杂的关系, 传统的方法很难处理. 机器学习可以提供一种方法, 使用目标函数的示例来查找某一映射函数逼近的结果, 尽可能接近目标函数^[28]. 机器学习可以分为监督学习、无监督学习、半监督学习和强化学习, 各自又包含不同算法. 不同的算法适用对象及范围不同, 优势也不同. 正确选择算法才能建立合适模型, 对数据进行训练并做出准确的预测, 以建立条件因子和目标属性的关系.

2.3.3 模型评估

一般来说, 我们可以对模型的泛化误差进行评估, 测试并选用结果最好的一个. 只有选用与问题相匹配的评估方法, 才能快速发现模型选择或训练中存在的问题, 迭代地对模型进行优化. 常用的评估方法有

Holdout 验证、 k -fold 交叉验证、Bootstrapping 验证等^[28]. Holdout 验证是留出一定比例的随机数据作为测试集, 剩余的数据作为训练集, 然后根据测试集对模型进行评估; k -fold 交叉验证是将原始数据随机分成 k 个部分, 在每一轮中使用 $k-1$ 个部分进行模型训练, 同时保留一个部分作为模型验证, 因此所有部分都将用于模型验证, k -fold 交叉验证相比 Holdout 验证消除了随机性, 但计算成本相对较高, 对于大型数据集耗时较长; Bootstrapping 验证每次从包含 m 个样本的数据集中随机抽取一个样本, 然后放回数据集中, 取 m 次样本后形成一个新的数据集作为训练集, 这种方法可以有效测试数据^[19]. Bootstrapping 优点是适用于数据量很少的情况, 缺点是改变了数据的原始分布.

2.4 材料数据库

21 世纪初, 麻省理工学院(MIT)的 Ceder 受到人类基因组计划的启发, 考虑借鉴基因研究者的经验, 建立材料基因组数据平台, 利用大量数据驱动材料研究. 建立公共数据库以探究材料性质的新方法有效地缩短了新材料的开发周期, 显著降低了新材料的开发成本, 并改变了传统材料的发现方式, 大幅提高了发现新材料的可能性. 目前较著名的材料基因组大规模高通量计算平台和材料数据库包括 Materials Project、Aflowlib 等(如表 1).

表 1 部分材料数据库平台简介

Table 1 Part of the material database platform introduction

数据库名称	数据库简介	数据库网址	所属机构
Materials Project (MP)	MP 提供了超过 13 万个无机化合物的结构信息以及性质, 存储了几乎所有的晶体材料和高通量材料的性质计算信息(包括能带态密度信息、电池材料的充放电曲线、相图等). 但 MP 中目前仅有 50% 的结构拥有电子结构信息.	https://materialsproject.org/	美国麻省理工学院等
The Open Quantum Materials Database (OQMD)	OQMD 可以清楚地显示出材料的相图, 目前该数据库钙钛矿数据居多. OQMD 可快速预测任意化合物的热力学性质并推测其稳定性. OQMD 包含大量易于计算的合金材料, 但目前对离子化合物的相图刻画非常不全面, 且数据质量参差不齐.	http://oqmd.org/	美国西北大学
Aflowlib	目前 AFLOWLIB 拥有超过 100 万种的材料. 其包含成千上万的假想材料, 其中许多在实际中仅能存在极短时间. 该系统可以以已知的晶体结构为基础, 自动预测新的晶体结构.	http://aflowlib.org/	美国杜克大学
Atomly	Atomly 拥有 30 余万种无机晶体材料的详细电子结构信息(能带、态密度等)以及近 4 万组热力学相关的相图. 其可迅速预测新材料热力学数据等, 进而指导新材料设计、预测合成难度等. 目前平台仍有大量材料的介电函数、声子谱等数据尚未进行分析计算.	https://atomly.net	中国科学院物理研究所
机数大材库 (dcaiku)	机数大材库包含 9448 万分子基础化合物及相应性能数据, 同时也包括经理论计算获得的部分理论性质. 该数据库也整合了分子结构检索功能, 方便用户直观快捷的结构搜索, 与此同时还提供基于 WebGL 技术的分子 3D 模型展示, 用于实时交互式观察分子结构.	http://www.dcaiku.com/#/	中国合肥机数量子科技有限公司
非线性光学材料数据库	非线性光学材料数据库包含材料的化学式、空间群、ICSD 编号, 用 DFT 方法计算的带隙、折射率、双折射、倍频系数等性质. 此外, 该数据库提供综合检索功能, 可通过元素组成、材料体系、材料性能等检索出满足要求的材料.	http://nlo.hbu.cn	中国河北大学

此外,由剑桥晶体数据中心(CCDC)创办的剑桥结构数据库是一个世界范围的小分子有机和金属有机晶体结构数据库;MatWeb 拥有超过 130000 种金属、塑料、陶瓷和复合材料的数据;Total Materia 以 74 项国际标准为基础,提供超过 1500 万种材料数据;ASM International 可允许所有人进行访问,并提供可信的材料信息.还有一部分较为小众的数据库也各有特点,如 The Superconducting Material Database (SuperCon)包含从文献中记录的超导材料(金属、合金、氧化物、非晶态材料等)的超导特性及其相关特性(导热率、霍尔系数、热电势等);Interatomic Potential (force field) Databases 提供了原子间势(力场)相关文件和评估工具的来源,以帮助研究人员获得原子间模型并判断其质量和适用性;Computational Materials Property Databases 拥有用于高级材料设计和加工的预测性合成系统以推进围绕材料合成方法的计算学习,进行材料合成;The Materials Data Facility 运营两项云托管服务(数据发布和数据发现),促进了开放数据共享,自助数据发布和管理以及鼓励数据重新利用等^[29].目前所有的数据库仍在持续发展中,更多的化合物数据被不断收录,大量的计算结果也不断地被修正或替换,数据误差逐渐减小.世界上的许多科研团队已经使用平台中的材料数据进行了科学研究,并取得了重大进展和突破^[30].

我国长期依赖于欧美的材料数据库,无法及时获得关键材料数据,因此学科发展受到制约,慢人一步.发展具有我国自主知识产权的材料库,可补全我国的材料研发链条,健全材料研发完整的生态体系,深度提升我国的材料研发生产力.中国科学院物理研究所和松山湖材料实验室领衔研发了 Atomly 材料数据库.当前数据库中的无机晶体材料已扩充到了 30 余万个,并且以每年上万种材料的速率增加,Atomly 的数据库内容条目已经超越了世界上的其他类似数据库.数据规模和质量已经跻身世界顶级水准,并且具有自主知识产权.机数量子依托中科大微尺度物质科学国家研究中心,率先建立了材料基因创新研究平台,开发了“国内创新性材料知识图谱”,目前其已建成综合性材料数据库平台——机数大材料库.该库目前已包含 9448 万分子基础化合物、40 万材料性能数据、90 万条含磷化合物,同时又扩展了 1120 万的化合物分子数据,包含 mol 或者 sdf 格式的原子坐标参数,基本化学性质内容,以及经理论计算获得的部分理论性质.国产数据库的发展为我国物质科学发展提供了优质的基础数据及平台,提升了整个行业的生产力、效率及竞争力.

3 机器学习在新材料方面的应用进展

3.1 不同功能上的应用

3.1.1 指导发现新材料

ML 方法已被证明在加速材料开发方面具有重大潜

力,对于材料科学中的模拟,重要的是在不同时间(从 fs 到 ns)和尺寸范围(从 pm 到 nm)内实现高精度计算,同时限制计算成本.利用机器学习开发新材料通常包含以下步骤:先利用高效率的机器学习做海选,再利用高精度的理论计算对海选出来的材料做优选或辅以证明,最后对优选出来的材料进行实验制备和验证.

利用高效率的机器学习做海选是发现新材料的第一步.李波研究小组^[31]基于高通量密度泛函理论(DFT)计算的机器学习,研究多硫化物的吸附模式并筛选负载型的单原子催化剂(SAC).研究人员首先利用机器学习对上千种氮掺杂碳材料负载的过渡金属单原子催化剂进行了筛选.通过构建 7 个氮掺杂石墨烯的不同结构并进行 DFT 计算以获得 LiPS 的几何结构和能量,随后将其分为训练集、验证集和测试集用于机器学习.利用晶体图卷积神经网络(CGCNN)框架进行了机器学习分类和回归,在对材料性能进行评估方面取得巨大成功,便于后续使用高精度的理论计算进行优选.

第二步是利用高精度的理论计算对海选出来的材料做优选或辅以证明.孙宽课题组^[32]使用机器学习模型预测 1296 个可能的供体-受体组合相应的有机太阳能电池的能量转换效率(PCE)值,筛选得到 22 个 PCE 预测值高于 17% 的全新的 Y6 衍生物受体分子.进一步借助量子化学计算分析了 5 种不同端基受体单元对分子性质的影响,发现主要是由于前线分子轨道能级和分子表面静电势的分布的差异,继而导致其具有不同光电特性.

最后一步是对优选出来的材料进行实验制备和验证.Li 等^[33]通过机器学习研究了含有 5 个阳离子元素的 227273 个高熵钙钛矿材料的氧化还原热力学,产生了约 20000 种有希望的氧载体候选物,继而进行 DFT 计算以及用大量的实验数据对机器学习的适用性和有效性进行了验证.DFT 预测结果显示 113 种材料适用于化学循环空气分离(CLAS),而 85 种材料适用于基于 CO₂ 裂解的化学循环(CL).在实验验证环节,预测的化合物中有 43 种在以往报道中得到验证,显示出优异的性能.此外,研究人员还制备并评估了 25 种模型预测材料,其中 23 个氧载体表现出令人满意的性能,15 个氧载体表现出优于大多数之前报道的氧载体的性能(如图 5 所示).

通过上述例子我们可以大致了解到机器学习应用于发现新材料的研究范式,借助于高精度理论计算的机器学习,相比传统实验方式,可以减少冗长的实验过程,不再仅凭化学直觉或经验开发新材料,在未来将有更大的潜力促进新材料的发展.

3.1.2 预测及优化材料性能

机器学习在新材料方面的重要应用之一是建立结构与性能之间的关系.它试图在材料指纹(包括组成元素的特征、原子结构信息以及这些特征的任意组合)和我们感兴趣的目标属性之间建立预测关系^[34].由于不可能通过人工实验手段逐一探索数量庞大的材料的性

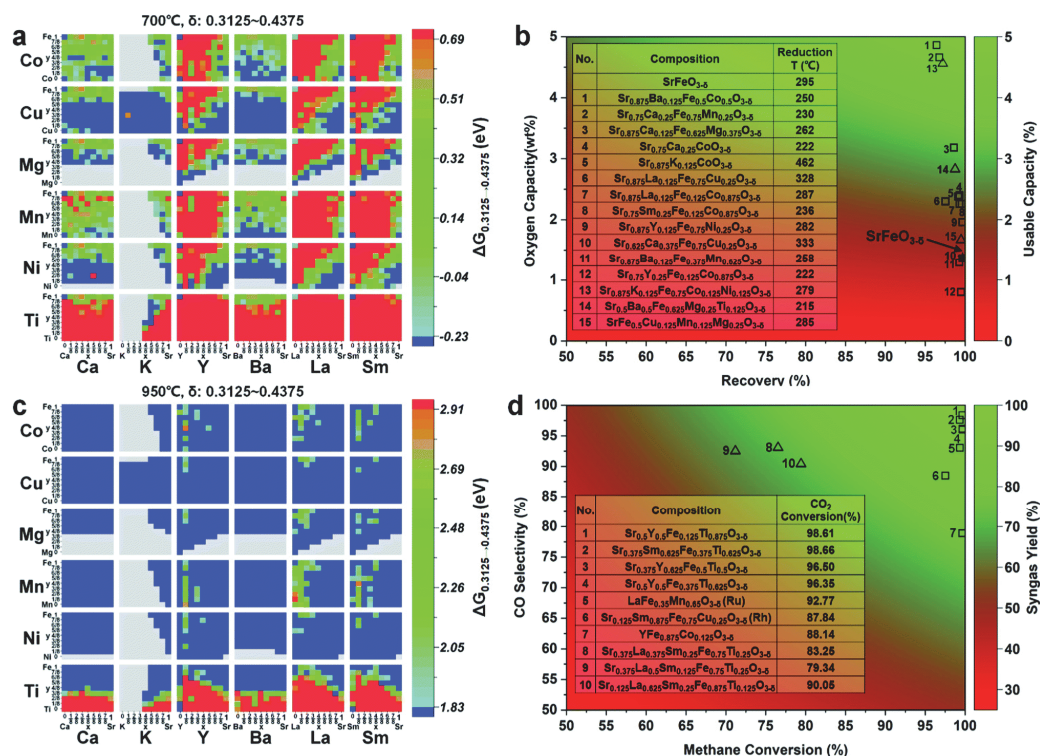


图 5 高通量筛选结果和实验验证^[33]. (a) 筛选出的适用于 CLAS 的热图; (b) CLAS 测试样品的氧容量、回收率和可用容量的实验结果; (c) 在 950 °C 下筛选出的适用于基于 CL 的 CO₂/H₂O 裂解材料的热图; (d) CO₂ 裂解测试样品的合成气产量和 CO₂ 转化率实验结果

Figure 5 High throughput screening results and experimental validations^[33]. (a) A heatmap of the screened candidates for CLAS; (b) experimental oxygen capacity, recovery and usable capacity of the samples tested for CLAS; (c) a heatmap of the screened promising candidates for CL CO₂/H₂O splitting at 950 °C; (d) experimental syngas yield and CO₂ conversion of the samples tested for CL CO₂ splitting

能, 因此材料数据库和机器学习的方法在预测现有和理论上存在的材料的性能方面具有巨大优势.

热力学稳定性通常是筛选新材料的第一标准, 需要用形成能、结合能、声子谱等进行定量描述, 其中形成能较为常用. Faber 等^[35]提出用 Ewald 算法矩阵、扩展类库仑矩阵和在库仑矩阵上发展的正弦矩阵来表示晶体结构. 利用上述矩阵作为特征, 从 MP 数据库中提取 3938 个晶体结构数据集, 训练核岭回归(KRR)模型进而预测晶体的形成能及稳定性. Nolan 等^[36]应用机器学习方法探究锂离子电池中固态化学物质和正极材料之间界面的稳定性, 界面稳定性高的材料可作为正极包覆涂层或固体电解质的理想选择. 但最近 Bartel 等^[37]提出了新的看法, 即准确的形成能预测并不意味着准确的稳定性预测. 该研究团队仔细研究了基于形成能的 ML 模型是否能够区分相似材料的相对稳定性, 并提出了消除量化 ML 模型中的错误的方法. 尽管如此, 预测形成能短期内仍将是判断热力学稳定性的重要标准之一.

带隙在光催化和光电器件等方面的应用中具有重要的意义, 可用于预测电导率、太阳能电池效率等特性. 王金兰团队^[38]结合 ML 和 DFT, 使用梯度提升回归树(GBRT)模型进行测试表现出色(如图 6a, 6b), 预测的带隙分布十分接近原始输入数据集, 成功预测了 5000 余种潜在有机无机杂化钙钛矿材料(HOIPs)的带隙. 潘锋团队^[34]在高通量计算得到的数据库上利用机器学习来

预测带隙, 以了解原子结构与电子结构之间的关系. 通过研究发现 AgO₂F 具有不同寻常的组成, 即含有正三价氧化态的银离子和过氧根离子. ML 方法预测带隙还可能使用多种 ML 算法相结合的方式, 进而建立并选择合适的模型. Rajan 等^[39]通过机器学习精确估计带隙(如图 6c)以研究 MXene 材料的电学性能, 将沸点、熔点、原子体积、相、键长等作为输入特征, 使用 bagging 算法进而通过机器学习筛选出与带隙大小相关的主要特征, 其中原子体积、沸点影响最大(如图 6d). 并且采用了核岭回归(KRR)、支持向量回归(SVR)、高斯过程(GPR)等算法高精度地预测带隙, 同时也开发了精确度达 94%的金属-半导体分类模型, 可在几分钟内估计整个 MXene 数据库的带隙.

超导性自一个多世纪前发现以来, 一直是大量研究工作的焦点, 也是最令人费解的科学课题之一. 然而, 理解超导性与材料化学结构的关系, 具有重要的理论和实际意义^[40]. 机器学习在超导性上的预测主要是对超导转变温度的预测. Owolabi 等^[41]应用支持向量机(SVM)算法, 使用晶格参数等作为特征预测 Fe 基和掺杂 MgB₂ 的超导体的超导转变温度. 前者在测试集上的均方根误差为 1.104 K, 后者的百分比偏差为 5.99%~10.46%. Stanev 等^[40]开发了几种机器学习方案, 用于模拟超过 12000 种已知超导体和候选材料的临界温度(T_c), 发现了 30 多个非铜基、铁基氧化物为潜在的

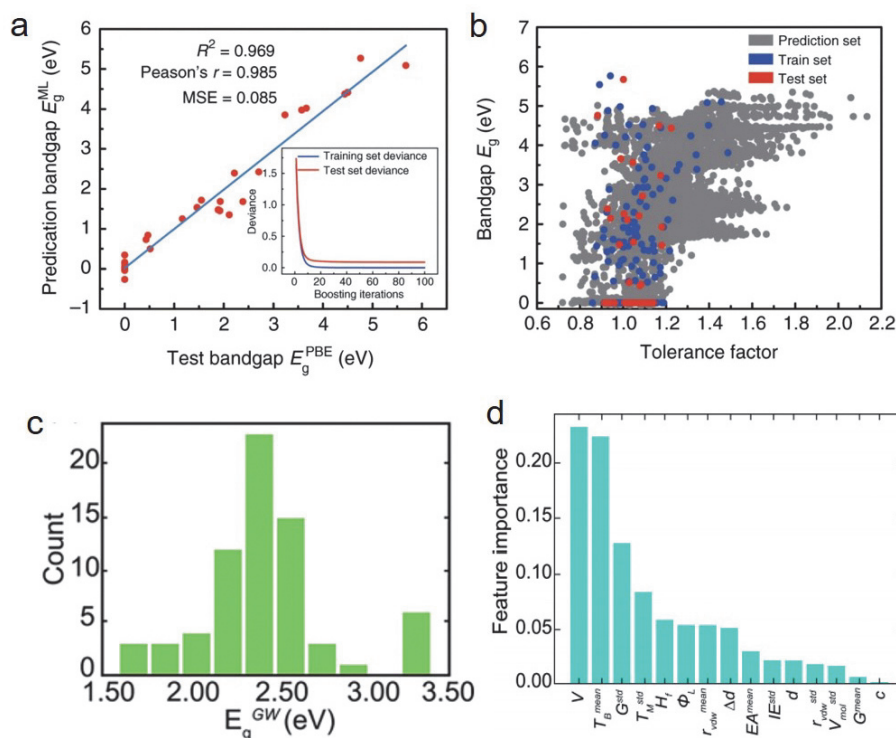


图6 (a)测试带隙 E_g^{PBE} 与预测带隙 E_g^{ML} 的拟合情况^[38]; (b) ML 模型的带隙容忍因子散点图^[38]; (c)机器学习数据集对应的 G_0W_0 带隙分布^[39]; (d) bagging 算法计算主要特征的重要性^[39]

Figure 6 (a) The fitting results of test bandgaps E_g^{PBE} and predicted bandgaps E_g^{ML} ^[38]; (b) scatter plots of tolerance factors against the bandgaps for the prediction dataset from trained ML model^[38]; (c) the distribution of G_0W_0 band gaps corresponding to the data set for machine learning^[39]; (d) primary feature importance calculated through bagging method^[39]

超导材料. Xie 等^[42]提出可通过机器学习算法改进原有的计算临界温度的方法^[43], 强调需要一种新的描述符来描述超导性. 利用这种隐含的表征费米面特性的描述符, 是预测超导性的一种有潜力的方法.

将 ML 方法和 DFT 方法相结合来预测及优化材料性能, 可以在保持 DFT 计算精度的同时大大节约计算成本, 已经成为一种越来越常见的方法. 从主要特征中筛选出重要特征的方法有很多, LASSO 是最简单的一种, 常用的还包括欧阳润海提出的 SISSO^[44]、以及前向筛选、置换重要性等方法, 用以映射结构-性质关系. 然而在实际的应用中, 新材料的性能要求会更加多样化, 使任务难度大大增加. 尽管随着技术的发展, 机器学习正面临越来越多的挑战, 但通过不断的升级、改良和丰富, 其仍具有极大的潜力.

3.1.3 辅助表征

物质的微观结构直接影响其宏观性质, 因此对物质的微观表征是科学界永恒的话题. 传统分子和材料的结构往往受限于仪器灵敏度等诸多限制, 误差较大且并不完美. 而机器学习可有效提高表征的精准性, 为材料微观表征提供了全新思路^[10]. 例如, 在团簇物理学中, 由于势能表面上的局部最小值数量随团簇大小呈指数增长, 确定中型和大型团簇的基态 3D 结构仍是一个挑战.

潘锋团队^[45]通过构建基于拓扑的 ML 模型以揭示锂簇中隐藏的立体结构-能量关系, 在工作中预测 Li_{20} 的最低能量结构由三个中心三角棱柱和五个额外的封顶原子组成, 其相对于其他理论预测结构在能量方面具有更大优势. 而 Li_{40} 的最低能量结构是具有五个缺失顶点原子的 45 原子多二十面体, 其预测的全局稳定结构与文献中的一致. 通过基于拓扑的 ML 模型有效地降低了 DFT 计算成本, 并极大地提高了对于 Li 团簇能量的预测精度. 此项工作开创了多元素簇预测和复杂晶体表面吸附结构预测的先河.

探究多相催化剂的 3D 结构表征目前仍是研究人员的巨大挑战, 而机器学习可助力 3D 物质微观结构的表征. Frenkel 团队^[46]使用 X 射线吸收近边缘结构(XANES)光谱和有监督的机器学习(SML)来完善多相金属催化剂的 3D 构型(铂纳米粒子的平均尺寸、形状和形态). 低温电子显微镜和超分辨率光学显微镜等方法的广泛应用极大地扩展了我们对纳米级软物质的结构和动力学的认识. 但是, 这些技术无法在具有纳米分辨率的天然液体环境中提供结构和功能动力学的实时成像. 陈倩团队^[47]利用基于机器学习的液相透射电子显微镜(TEM)视频分析和研究纳米胶体的自组织途径, 并首次得到了一系列纳米粒子系统自组织途径的 3D 成像(如图 7). 将

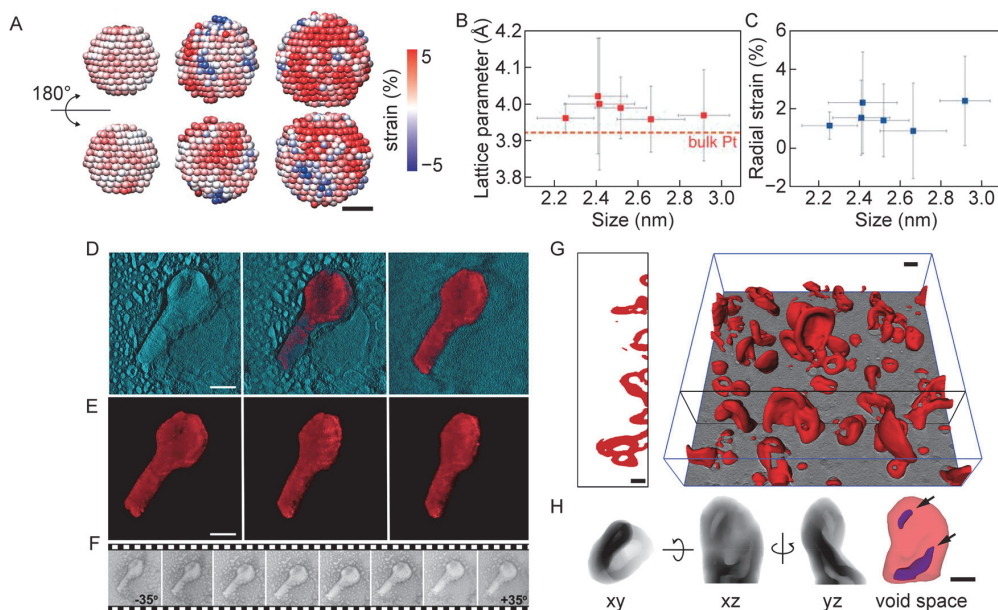


图7 纳米级物质的电子断层扫描. (A~C)单晶 Pt 颗粒的径向应变图(A)、拟合晶格参数(B)和平均径向应变值(C); (D~F)噬菌体的分割. (D)图像系列和(E)密度图显示噬菌体(洋红色)从液体背景(青色)的分割. (F)噬菌体颗粒的倾斜图像. (G)重建后聚酰胺膜的 2D 切片和 3D 视图. (H)解决各种投影和内部空隙(紫色)的 3D 数据集中分割出的单个皱褶^[46]

Figure 7 Electron tomography for nanoscale matter. (A~C) Radial strain maps (A), fitted lattice parameters (B), and averaged radial strain values (C) of single-crystalline Pt particles. (D~F) Segmentation of a bacteriophage. (D) Image series and (E) density maps showing the segmentation of a bacteriophage (magenta) from the liquid background (cyan). (F) Tilt images of a phage particle. (G) 2D slice and 3D view of the polyamide membrane after reconstruction. (H) Individual crumples segmented from the 3D data set resolving various projections and interior void (purple)^[46]

ML 与 TEM 相结合对软物质的研究范式转变具有启发作用, 推动了表征技术在纳米技术研究领域的应用发展.

随着机器学习的飞速发展和广泛应用, 图像识别技术也得到史无前例的发展契机^[48]. 近年来扫描透射电子显微镜(STEM)和扫描探针显微镜(SPM)精度进一步提高, 其数据集的大小和质量也随之呈指数性增长, 但关于各种原子构型和晶格缺陷的结构信息仍相对有限和缺乏. Ziatdinov 团队^[49]为扫描透射电子显微镜原子分辨图像的深度学习(DL)提出了一种“弱监督”方法, 以此确定了与进入石墨烯空位中的硅原子相关的多个缺陷结构, 并观察了表面缺陷的各种化学和结构转变(如图 8a~8d), 最后确定了缺陷的位置和化学结构(如图

8e~8h).

反射高能电子衍射(RHEED)可现场观察分子束外延生长过程中的样品表面行为, 具有广泛的应用前景, 但 RHEED 模式分析对操作人员具有较高要求, 需要其拥有大量的专业知识积累和操作经验. 因此, 将其应用于实时反馈控制是十分困难的. Kwoen 团队^[50]使用卷积神经网络(CNN)的机器学习方法识别输入数据库中的特征点, 其适用于具有可变性的图像分类. Kwoen 团队提出了一种在连续旋转过程中识别 GaAs 衬底的 RHEED 模式测量方法, 并建立了生长条件的数据集. 通过使用 CNN 训练深度学习模型来建立分类模型, 准确率超过 99%, 可预测 GaAs 在 III~V 族化合物中有巨大潜力.

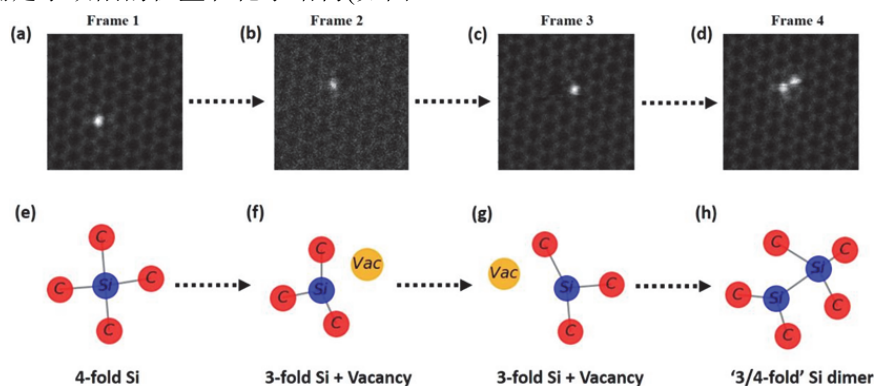


图8 跟踪复杂缺陷在表面上的转换. (a~d)基于 STEM 的石墨烯表面缺陷转换跟踪; (e~h)通过将全卷积网络、高斯拉普拉斯算子和图形表示应用于(a~d)中数据进而提取出的相应单缺陷结构^[49]

Figure 8 Tracking complex defect transformations on the surface. (a~d) STEM based tracking of defect transformations on graphene surface; (e~h) corresponding single defect structures extracted by applying FCN, LoG and graph representations to data in (a~d)^[49]

除此之外,机器学习尤其是深度学习在材料表征方面还有诸多应用如降维处理^[51]、数据增强^[52]等极大地促进了人们对材料的理解与应用开发.随着人们对于材料性能及材料基础理论研究的需求增加,在大量实验数据库不断的建立和完善下,基于机器学习技术的材料微观结构表征势必会给构效关系的阐述带来全新的解释^[21].

3.1.4 加速量子化学发展

量子化学(QC)应用量子力学(QM)来深入了解原子系统的物理化学性质.然而量子化学的计算成本过高,目前只有在非常小的原子系统中才可以通过求解薛定谔方程来找到正确的波函数.在大多数实际情况下,研究人员不得不使用各种近似的QC方法以牺牲准确性为代价来减少计算时间.近年来,由于ML算法开发的进步和硬件资源的改进,ML在许多领域取得了显著进步.用ML来预测所需的QC属性可以大大提高计算速度并且不会失去准确性.将ML校正添加到计算成本低廉的近似传统量子方法中,经训练后可以非常准确地预测能量、自由能、熵和电子相关能(如图9)^[53-55].

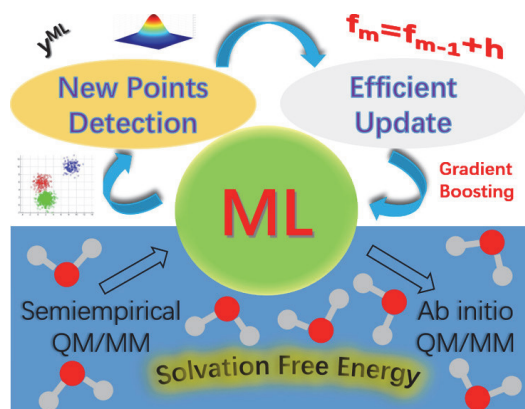


图9 量子力学/分子力学和机器学习模型的溶剂化自由能计算^[55]
Figure 9 Solvation free energy calculations with quantum mechanics/molecular mechanics and machine learning models^[55]

目前在很多研究领域需要大量的量子化学计算,现在已经有多种机器学习算法被应用在量子化学领域,例如神经网络(NN)、高斯过程(GP)、核岭回归(KRR)、支持向量机(SVM)等^[54].

机器学习改进的半经验哈密顿(ML-SQC)比传统的重新参数化更有优势. Dral 等^[56]用ML模型作为参数来研究提高半经验量子化学(SQC)方法的准确性,ML-SQC方法允许自动调整单个分子的SQC参数,从而提高准确性,并且不会降低分子描述符与训练集中的分子的转移性.在研究过程中,他们用一组6095个结构异构体C₇H₁₀O₂对OM2等半经验模型方法论证了该方法的性能,与标准OM2结果相比,ML-OM2结果显示更高的平均准确度,并且误差范围大大减小,原子化焓的平均绝对误差从26.4 kJ/mol降至7.1 kJ/mol.所以说,ML-SQC方法有望对材料和分子进行快速、合理的高通量筛选.

机器学习可以用来加速DFT计算,并发展新的DFT方法.求解DFT的Kohn-Sham(KS)方程已成为现代材料、化学科学研究与开发组合的重要组成部分,但由于KS方程造成的计算瓶颈,常规DFT计算通常被限制在几百个原子内^[57-58]. Chandrasekaran 等^[58]利用ML开发了一种新的计算方案,它可以学习DFT的KS方程的行为,有效吸收KS方程的功能,并绕过求解KS方程,快速、准确地预测材料或分子的电子结构.在研究过程中,他们利用了一种新的旋转变换表示法,围绕网格点绘制原子环境图,以显示该网格点的电子密度和状态的局部密度.这种方案可以节省大量的计算时间.

机器学习可以与非绝热分子动力学(NAMD)结合应用到实验探究分析中去.非辐射电子-空穴复合是铜锌锡硫化(CZTS)太阳能电池中电荷和能量损失的主要途径. Zhang 等^[59]利用非绝热分子动力学(NAMD)和深度神经网络(DNN)相结合的方法来研究纯CZTS、掺杂Ag和掺杂Ag+Cd的CZTS的非辐射电子空穴复合.在纯CZTS、掺杂Ag和掺杂Ag+Cd的系统中,与预先的NA耦合数据相比,用相邻的带隙作为描述符,DNN分析可以预测额外的NA耦合,加速量子动力学模拟.模拟结果表明,与纯净系统相比,掺有Ag和Ag+Cd的CZTS对电子空穴复合的抑制作用为1.8和3.2倍.研究表明,掺杂Ag或Ag+Cd几乎不会改变CZTS的基本带隙,并保留了优异的光电性能,同时延缓了非辐射电子-空穴复合,增强了CZTS太阳能电池的性能. NAMD与机器学习的结合研究表明,掺杂诱导的电荷局域化在抑制CZTS系统中的非辐射电子-空穴复合中起着主导作用,并为设计高性能CZTS太阳能电池提供了可靠的策略.

机器学习支持针对复杂材料进行大规模原子模拟.材料科学中的许多现象都涉及分子/材料在纳米结构尺度上的化学键破坏和缺陷形成,因理论方法的不完善和高效计算软件的缺乏,即便使用超算设施,直接在原子水平上模拟材料纳米结构的转变过程仍然具有高度挑战性.对于具有数千个原子的纳米材料系统,传统的原子模拟技术要么计算效率低下,要么无法准确描述势能面(PES).刘智攀课题组开发的LASP软件整合了基于SSW方法的全局PES探索^[60-61]和使用全局神经网络(G-NN)的快速PES探索^[62-63]两种模拟工具,可以做到从零开始对复杂材料的多维PES进行无偏全局采样,且更具稳定性和通用性,从而解决了从材料结构确定到反应路径预测等具有挑战性的问题^[64].

机器学习还可以助力分子动力学模拟.长期以来,分子模拟界一直面临着原子间作用力建模的精度与效率之间的矛盾.其中,从头算分子动力学(AIMD)具有密度泛函理论(DFT)的准确性^[65],但DFT在评估PES和原子间作用力时的计算成本过高,将其应用限制在数百到数千个原子的系统尺寸和100 ps的时间尺度上.另一方面,基于empirical force fields (EFFs)的分子动力学方

法^[66]可以进行更大系统、更长时间的模拟. 然而, EFFs 的准确性经常受到质疑. 为了解决这种矛盾, 鄂维南团队^[67]在 2018 年提出一种基于机器学习的分子动力学方法 DeePMD-kit. 使用由 AIMD 方法得到的数据训练 DNN 网络, 将训练好的 DNN 网络作为原子的势函数, 提供给 LAMMPS 等分子动力学模拟软件进行计算. 在加快了模拟速度的同时, 保证了模拟结果可以达到 AIMD 方法的精度.

可以说量子化学是机器学习的另一关键应用. 机器学习强大的数据处理能力使它能够解决量子化学中的许多问题, 是预测分子复杂特性和结构、研究量子多体系统和发现新材料的有力工具^[26].

3.2 不同材料上的应用

3.2.1 半导体材料

半导体材料处于产业链的上游, 是半导体产业发展的基础, 是支撑 5G、工业互联网、人工智能、特高压、新能源汽车等领域的关键基础材料^[68]. 世界各国的研究团队对机器学习在半导体材料上的应用也做了相关研究.

其中探究最多的是金属卤化物钙钛矿材料, 稳定性问题是目前钙钛矿研究和应用的核心问题和难点, 也是商业化面临的主要障碍. 然而在过去的一个世纪里, 测

定钙钛矿稳定性一直依赖于从物理直觉中推导出的半经验模型, 如常用 Goldschmidt 容忍因子(t 容忍因子)^[69]. 尹万健课题组^[70]提出了一种结合 ML 和 DFT 计算的策略来设计稳定的卤化物钙钛矿(如图 10). 通过选取 354 种钙钛矿的 DFT 结果作为训练集, 建立了钙钛矿稳定性与组成离子半径之间的映射, 表现出比 t 容忍因子和最近提出的 $(\mu+t)\eta$ 描述符更好的性能. 符号回归(SR)是一种监督学习方法, 搜索最优数学公式和该函数中的参数集, 因此可以用来得到简单的描述子, 指导实验合成新材料. 在另一项工作中, 尹万健课题组^[71]利用符号回归方法, 得到氧化物钙钛矿催化析氢反应(OER)的简单描述符 μ/t , 提高了该反应的活性, 并成功地合成了 5 种新的氧化物钙钛矿材料. 冯宏剑课题组^[72]也结合了 ML 和 DFT 计算, 通过机器学习预测了 285 个潜在光伏材料, 用 DFT 计算进行进一步筛选, 预测四种理想无机光伏材料 $\text{Ba}_4\text{Te}_{12}\text{Ge}_4$, $\text{Ba}_8\text{P}_8\text{Ge}_4$, $\text{Sr}_8\text{P}_8\text{Sn}_4$, $\text{Y}_4\text{Te}_4\text{Se}_2$ 的理论光电转换效率大于 26%, 并且选取 GBRT 模型进行训练避免出现过拟合现象. 在合成方面, Sargent 团队^[73]构建高通量合成钙钛矿单晶的框架, 并借助基于卷积神经网络(CNN)的图像识别晶体是否生长, 然后用机器学习模型预测了合成新的钙钛矿单晶的最佳条件, 合成了一种新型钙钛矿材料 $(3\text{-PLA})_2\text{PbCl}_4$ 晶体, 为加速发现新型钙钛矿提供了一条新途径.

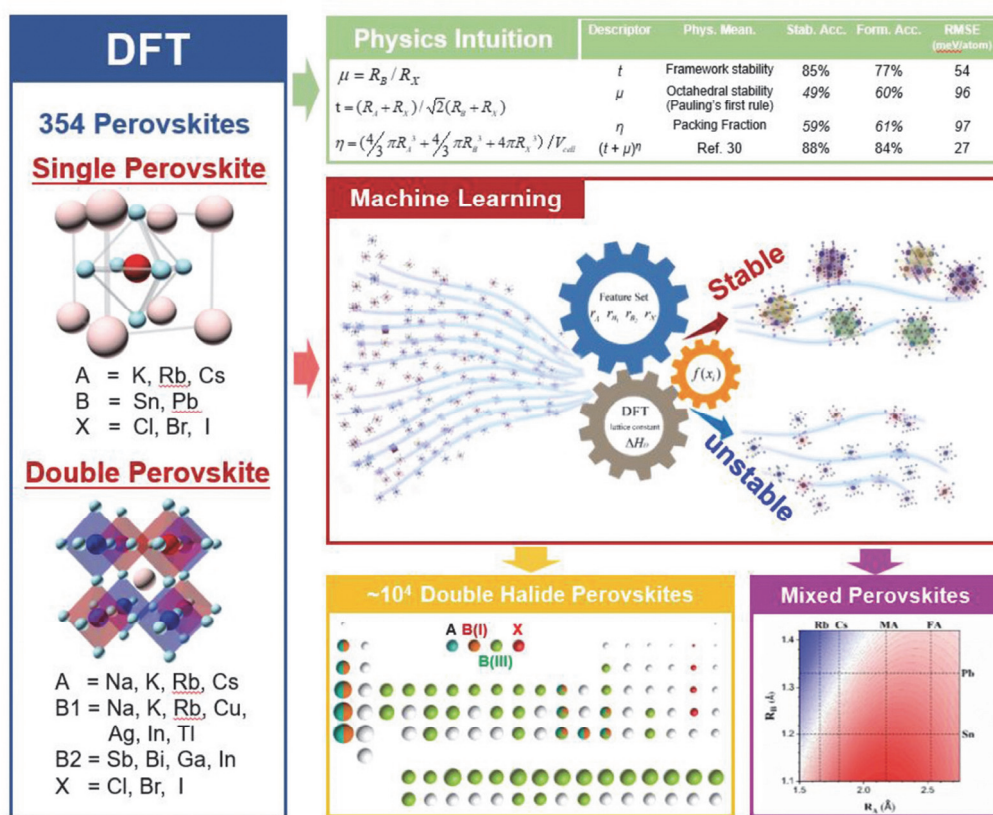


图 10 基于 DFT 的 ML 在卤化钙钛矿稳定性工程中的原理及策略^[70]

Figure 10 Schematic strategy of ML based on DFT for high-throughput stability engineering of halide perovskites^[70]

除了无机半导体材料,有机半导体材料同样引发了关注.金属有机框架(MOFs)材料由于其丰富的化学组成、可控的光电性质和低成本制备,使人们对它的研究兴趣也在半导体材料领域发展起来^[74].克莱姆森大学 Saha 研究团队^[75]通过引入一种新的蝴蝶形富电子的四硫富烯配体,构建了一种可以导电的新型双螺旋 MOFs 体系结构,可以用比无机半导体更节能的方式制造出来.美国桑迪亚国家实验室^[74]通过结合机器学习和第一性原理计算方法,筛选了 2932 种 MOFs 材料,确定 6 种有电子传导性,其中有 5 种结构已经合成并有文献报道,但其电学特性尚未见报道.这证明了 ML 在材料科学中的潜在力量.

预测半导体中杂质可否掺入及其电子能级的变化能力,对于控制其导电性至关重要.用实验和计算测定杂质存在许多困难,这使得数据驱动的机器学习成为一种合适的方法.美国阿贡国家实验室^[76]通过将半导体+杂质系统转换为数值描述符(元素属性、晶胞缺陷特性、元素属性+晶胞缺陷特性),进而建立了杂质生成焓和电荷跃迁能级的回归模型,这些回归模型可以较准确地预测混合阴离子 CdX 化合物(其中 X 是 Te、Se 和 S 的组合)的杂质性质,揭示了机器学习模型在快速筛选影响半导体光电行为的杂质方面的能力.

3.2.2 聚合物材料

聚合物材料具有优异的化学、物理、生物和机械性能,因此其在能源储存、建筑、医药、航空航天领域应用十分广泛.研究人员利用传统试错法合成新型特定功能材料效率低下,无法满足生产生活的巨大需求^[77].

尽管在已有的实验和计算数据的支持下,机器学习已经成为分子特性预测和设计的有力工具.但将机器学习

应用于聚合物材料目前依旧面临着高分子系统的化学、拓扑及形态的复杂性等障碍,这使得相关数据的获得及整合相比较小分子体系尤为困难^[78].这些原因使得机器学习在聚合物材料中的发展受到了极大限制. Ferguson 小组^[79]结合粗粒度分子模拟、变分自编码器、高斯过程回归和贝叶斯优化,定义了一个迭代主动学习协议,以有效识别能够组装具有良好堆叠的电子活性 π 核的伪一维纳米聚集体的分子.这项工作使用贝叶斯优化实验设计对 DXXX-OPV3-XXXD 家族的 π 共轭肽的组装情况进行了全面研究(如图 11),以指导微秒时间尺度上昂贵的粗粒度分子模拟.其工作平台亦可扩展至其他肽和肽样系统的设计.李国强等^[80]通过在预训练过程中学习包含药物分子的大型数据集,能够有效实现将固化热固性形状记忆聚合物(TSMP)映射到更接近高斯分布的隐藏空间.使用这种方法,作者创造了一个大的组成空间,并能够发现 5 种具有所需特性的新型紫外线固化 TSMP,其中一种已通过实验验证.该方法可以利用数据稀缺(和有偏差)的 TSMP 目标和数据丰富的药物源之间的相关性来有效地学习 TSMP 特征,并且结果比采用直接标签编码和 Morgan 编码的支持向量机方法的基准集更准确且稳健.

大多数关于聚合物的机器学习计算数据来自于单体或小分子量聚合物的密度函数理论计算.然而聚合物的结构和构象的复杂性主要在于其大分子量.从模拟的角度来看,聚合物材料通常是大型无序系统,其特性取决于弱相互作用和构象异质性,第一原理表征对聚合物特征描述相对困难.因此,高分子特性的计算往往是粗粒度经典建模的领域,使用机器学习对粗颗粒度(CG)模拟本身进行替代以牺牲化学分辨率来换取计算的可

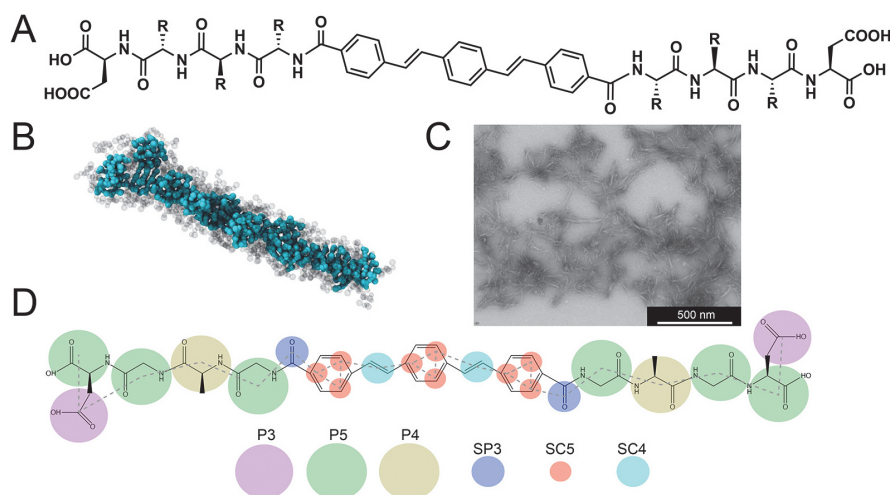


图 11 DXXX-OPV3-XXXD 系统. (a) DXXX-OPV3-XXXD 肽单体的化学结构; (b) 由 DVAA-OPV3-VAAD 肽单体自发结合成线性堆叠形成的自组装伪一维纳米聚集体的分子模拟快照; (c) DFFG-OPV3-GFFD 肽在酸性环境中形成的自组装原纤维的实验透射电子显微镜(TEM)图像; (d) 从 DGAG-OPV3-GAGD 全原子结构到进行本工作模拟的粗粒度表示的映射说明^[79]

Figure 11 DXXX-OPV3-XXXD system. (a) Chemical structure of the prototypical DXXX-OPV3-XXXD peptide monomer; (b) molecular simulation snapshot of a self-assembled pseudo-1D nanoaggregate formed by the spontaneous association of DVAA-OPV3-VAAD peptide monomers into a linear stack; (c) experimental transmission electron microscopy (TEM) image of self-assembled fibrils formed by DFFG-OPV3-GFFD peptides in an acidic environment; (d) illustration of the mapping from the DGAG-OPV3-GAGD all-atom structure to the coarse-grained representation at which the simulations in this work are conducted^[79]

操作性^[81]. Ketten 等^[82]建立了一种粗粒度方法,将能量重整化方法与分子动力学模拟的高斯过程代理模型相结合. 环氧树脂的 CG 模型揭示了在同一树脂中结合使用一种刚性硬化剂和一种更柔韧的硬化剂会产生优异的机械和弹道响应效果. Ketten 团队采用功能校准来匹配密度、德拜-沃勒因子 $\langle u^2 \rangle$ 、杨氏模量和网络在固定温度 $T=300\text{ K}$ 下任何程度的交联时的屈服应力.

3.2.3 锂电池材料

锂离子电池被视为新型储能技术的重大突破,随着应用领域的不断拓宽,近年来锂离子电池行业呈现稳步快速增长的态势. 仅在 2016 年到 2018 年三年期间,关于锂离子电池的研究论文就发表了超出 53000 篇,未曾深入了解此领域的科学家需要耗费大量的时间用于文献消化. 2019 年, Springer Nature 和法兰克福大学基于相似性的聚类分析共同开发了一种先进算法 Beta Writer,并出版了第一本由机器学习生成的书籍——《锂离子电池: 机器生成的当前研究摘要》^[83],其包含 2016~2018 年发表过的 150 多篇权威研究论文,节省了大量人工筛选的时间. 通过机器学习及大数据不仅可以了解当前研究前沿方向,同时也对当前的前沿研究起到推进作用. 早在 2010 年 Ceder^[84]便提出将高通量计算和大数据库运用于锂离子电池材料的开发中,此后机器学习在锂离子电池材料应用方面大放光彩,例如 Zhu 和夏勇课题组^[85]使用机器学习算法建立锂离子电池单元的高精度计算模型,为电动汽车电池的设计和安全性提供了重要的技术指导; Chueh 和 Braatz 课题组^[86]使用机器学习方法建立了三种新的电池寿命早期预测模型,可以及时、准确地对电池寿命进行可靠预测; Homma 团队^[87]

使用机器学习来优化导电材料的锂离子材料配比,有效增强混合体系的锂离子电导率. 机器学习正在高速发展的锂离子电池领域担任着愈发重要的角色.

尽管近年来,国内外对锂电池材料的研究开发如火如荼,但许多基本原理和机制还没有得到足够的理解. 高容量、可逆的锂金属阳极对于下一代储能系统而言是必需的^[88]. 目前绝大多数锂离子电池均使用石墨阳极,其工作原理非常复杂. 但此过程决定了电池充电速度的上限,并且对电池的使用寿命起着至关重要的作用. Panosetti 团队^[89]提出此过程的仿真模型,将粒子群优化(PSO)与 ML 程序结合高斯过程回归(GPR),在最广泛的电荷状态范围内(包括极稀的电荷状态)探究 LiC_x 化合物的嵌入机制相关的计算,对色散作用进行了重要描述. Wang 课题组^[90]提出了一种新的固态聚合物电解质(SPE)设计方法,该方法将粗粒度分子动力学(CGMD)与机器学习集成在一起,对锂电导率与分子水平的固有材料特性之间的关系(如分子大小和分子间相互作用)进行了全面的描述. 在工作中通过控制单一变量进行理论计算发现诸多影响电导率因素的规律变化. 例如,粒子的大小对电导率的影响最大,而与阳离子相关的相互作用强度紧随其后. 相比之下,链间相互作用较弱. 赵克杰等^[91]通过基于相位衬度机制的高分辨率硬 X 射线纳米断层成像技术并结合机器学习算法实现了对锂复合电极微观结构的具有统计相关性的分析(如图 12a, 12b). 该工作揭示了颗粒脱离的程度与充电速率呈正相关,并且较小的颗粒在从碳/粘合剂基质脱离时表现出更高程度的不确定性.

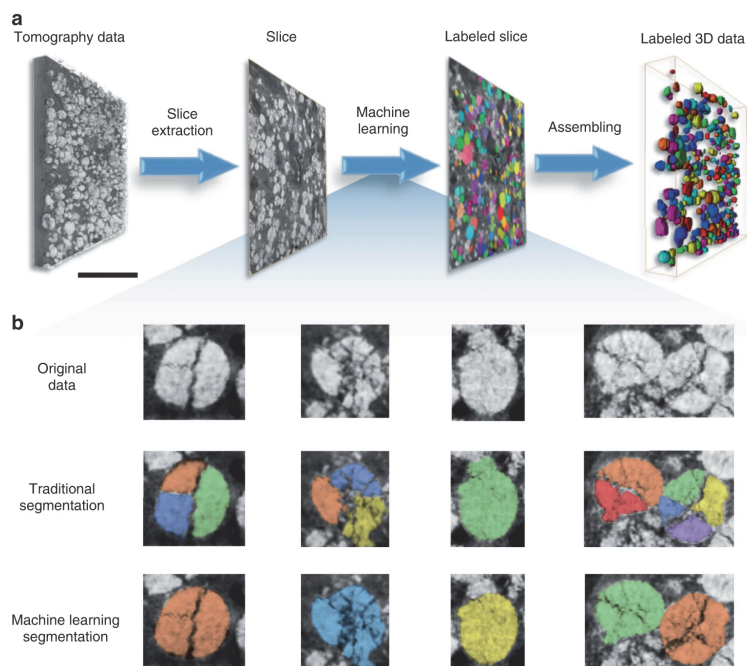


图 12 (a) 基于机器学习的分割的工作流程; (b) 一些代表性粒子的常规分割结果与机器学习辅助分割结果的比较^[91]

Figure 12 (a) Workflow of the machine learning-based segmentation; (b) comparison of conventional segmentation results and the machine-learning-assisted segmentation results for a few representative particles^[91]

由于对锂离子传导的结构与性质的关系了解甚少,因此在过去的几十年中,数以万计的锂离子相关物质单纯依靠人力进行可靠电极材料的筛选,不仅浪费大量的人力、物力和财力,而且出错率较高. Reed 团队^[92]通过基于机器学习的方法来筛选 12000 多种固体电解质,发现了室温下单晶锂-硼-硫(Li-B-S)固体锂离子电解质具有高离子电导率等优点,同时发现了许多在室温下新型锂离子晶体固体材料^[93]. Mueller 课题组^[94]使用将分子动力学和机器学习模型结合筛选出了两种未报道过的锂离子电池涂层材料,以及其他几种可以用作增强离子传导的掺杂材料. 工作中将分子动力学和动态机器学习相结合使得计算效率提高了 7 个数量级,显著降低了计算迁移能的不确定性,并提高了与实验测得活化能的一致性.

当前,开发高离子电导率的固态电解质是当今锂离子电池研究的一大方向. 通过机器学习数据驱动等方式可低成本、快速筛选具有高离子电导率的新型电池材料,同时利用数据循环等方法预测电池的安全性以及其使用寿命. 不仅如此,机器学习还可以利用先进的数据提取和收集技术促进锂离子电池的开发和利用,相信在不久的将来机器学习将在锂电池材料研究中为我们带来更多的惊喜.

3.2.4 催化剂材料

催化是一个复杂、多维、多尺度的研究领域. 近年来,数据科学界的革命对工业和学术界的传统催化研究产生了重大影响,加速了催化剂的发展^[95]. 机器学习有助于建立更好的模型,了解催化研究并扩展催化知识^[96].

电催化是催化的一个重要分支,机器学习可以辅助电催化剂的合理设计^[97],在其中的应用主要集中于 CO_2 还原(CO_2RR)、氧还原反应(ORR)和析氧反应(OER)电催化剂的研究. 二氧化碳电还原(CO_2ER)是实现二氧化碳循环和能量再生的最有潜力的方法之一,推广这项技术的关键是开发高性能的电催化剂^[98]. Tran 等^[99]通过筛选 31 种不同元素的合金,进行了 42785 次吸附能计算,已在 54 种合金中识别出 131 个用于 CO_2 还原的电催化剂材料,但不足之处在于其严重依赖于描述符. 两年后在另一工作中该课题组^[100]进行了更深入的探索,列举了 228969 个含铜金属间化合物晶体的吸附位点,然后在这些位置的子集上进行了 DFT 模拟,以计算它们的 CO_2 吸附能(如图 13a, 13b). 结果表明 Cu-Al 合金提供了多个位点和表面取向,并有效地将 CO_2 还原为乙烯,实现了迄今为止报道的最高法拉第效率,感应电流的效率超过 80%(纯铜约 66%). 此外, ORR 和 OER 对于能量转换和存储设备(如燃料电池、金属空气电池和电热槽)的电化学性能至关重要. Batchelor 等^[101]提出了一种寻找活性合金电催化剂的理论方法,并应用于氧还原反应(ORR). 利用机器学习方法预测了反应中间体的吸附能,并研究

了合金表面的吸附位点,发现计算值与预测值符合较好,对 *OH 和 *O 吸附能的均方根偏差(RMSD)分别为 0.063 和 0.076 eV. 在描述符方面, Liu 等^[102]总结了 ORR 和 OER 过程中的基准活性描述符(如图 13c),对于杂化材料,描述符在界面重构、约束效应、多位点吸附等利于揭示催化机理的方面将显示出更强的预测能力. 因此,机器学习和高通量模拟对于发现催化领域新的多重描述符和反应机制是至关重要的.

随着能源危机的到来,人们对于开发可再生能源的需求日益迫切,而光催化反应为人类利用太阳能提供一种选择. 光催化分解水反应是一个典型的光催化反应,自 1972 年由 Fujishima 和 Honda^[103]首次在紫外灯下实现以来便受到了广泛关注. Zhang 等^[104]建立了基于晶格参数和表面积的高斯过程回归(GPR)模型,来预测掺杂了 TiO_2 的光催化剂的带隙 E_g , 模型表明 E_g 的预测值与实验值的相关系数高达 99.99%, 预测精度极高. 该模型可以广泛应用于各种方法制备的 TiO_2 , 便于设计和理解多掺杂的 TiO_2 光催化剂. 另外,近年来,钙钛矿在光催化分解水领域也有着越来越多的应用. Can 等^[105]以 540 个案例为基础,构建并分析了钙钛矿光催化水分解的数据库. 对数据进行挖掘找出了导致高产氢率的因素,并验证了文献综述得出的结论,后采用决策树和随机森林算法进行建模分析(如图 13d). 在 20 世纪 90 年代以来,某些有机材料也被证明是析氢反应(HER)的光催化剂. 利用机器学习和高通量计算相结合的方法,将共聚光电材料的特性与催化活性相结合,结果证明光催化活性与个别特性的相关性要弱得多,这支持了光催化活性是一种真正的复合特性,依赖于大量独立因素的观点^[106].

机器学习在其他催化材料上也有着重要的应用. 例如,沸石是一种多孔性铝硅酸盐晶体,是传统的催化材料之一,合成历史悠久,广泛应用于催化等工业领域. 寻找和表征新型沸石并将其应用于催化领域是一个巨大挑战,而机器学习方法可以帮助发现新型沸石. 早在 2006 年, Corma 等^[107]提出了一种新的高温合成沸石的方法,总结了 X 射线衍射特征数据,利用人工神经网络和决策树算法成功地建模. Chaikittisilp 等^[108]应用聚类算法识别主要结构描述符,构建相似性网络,进一步明确合成与结构间的关系,指导沸石材料合成. MOFs 材料其具有在分子水平上可控的特性,这种独特的特性使其适合作为催化剂^[95]. 在温室气体 CO_2 的捕获存储分离和高附加值催化转化方面, Smit 等^[109]对包含超过 30 万个 MOFs 的计算数据库进行挖掘筛选,可以识别出不同类别的 CO_2 强结合位点. 在多相催化领域,刘智攀课题组^[110]综述了多相催化中的 ML 方法及其应用现状,重点介绍了基于 ML 的原子模拟,阐述了 SSW-NN 方法在 LASP 软件中的应用,突出了算法的关键特征和主要功能. 从确定反应活性位点到理解反应机理,神经网络在多相催化领域大放异彩. 用神经网络势函数结合势能面

搜索方法来对势能面结构进行搜索是解决复杂催化体系的有力工具,它可以显著降低催化体系的计算代价,而维持与密度泛函理论同一级别的计算精度^[11].在双原子催化剂方面,申威峰课题组^[12]用机器学习技术来识别双原子催化剂上 ORR 的活性位点,其中 CoNi/NC 表现出最佳活性,极限电位高达 0.88 V,为开发新型高效催化剂提供了有效方针.

表示催化活性的催化剂描述符在机器学习方面一直具有挑战性。Taniike 等^[113]利用机器学习设计了甲烷氧化偶联(OCM)反应的催化剂描述符, 揭示了在该反应中代表乙烯/乙烷选择性(C₂S)的五个关键物理量, 其中机器学习预测了三种催化剂具有较高的 C₂S 值。实验结果也证明了这一点, 催化反应过程中仍能保持其活性和选择性。其中 La-Pr-Hf-BaO 取得重要进展, C₂ 选择性为 84%, C₂ 产率为 15.6%。高旺等^[114]对不同合金体系的反应描述符进行了总结和归纳, 强调了设计普适性描述符的重要性。尽管短期内无法发现所有材料和反应的通用描述符, 但现有的描述符及其改进对于最终实现全局通用描述符至关重要。

此外, 有关 ML 在吸附能的预测、表面相图及 d 带中心等催化剂特性表征方面, 研究者们也开展了许多工作. Foster 团队^[115]分析了 3 种纳米系统催化析氢反应的

结构描述符的性能, 分析得到其中一种描述符原子位置平滑重叠(SOAP)具有显著优越的预测性能, 可以作为吸附能预测的工具. Nørskov 小组^[116]利用高斯过程回归(GPR)模型, 对吸附位点所有可能的吸附质覆盖的自由能进行了预测, 可以用较少的计算资源生成精确的自由能图. Takigawa 小组^[117]基于 ML 方法预测金属和双金属化合物的 d 带中心, 利用 GBRT 方法, 仅用 6 个描述符就可以很好地预测 d 带中心.

近期, 罗毅、江俊课题组^[118]介绍了一种具有科学思维的人工智能化学家, 可在前人研究的基础上提出科学假说并制定实验方案, 通过2台移动机器人和15个智能化学工作站完成化学实验全流程, 使智能模型融入复杂的化学实验. 在发光材料和OER等催化剂的筛选中, 该平台可以在任务要求下得到最佳材料, 向人们展示了人工智能化学家的巨大潜力, 为新材料和发现和制造提供了一个新的思路.

ML 在图像分析领域近年来取得了惊人的进展, 利用各种形式的催化研究中的显微镜图像可能是未来的一种趋势^[96]. 我们相信, ML 方法与现有化学和物理模型的整合将大大加速催化剂和催化过程的自动化设计和优化^[95].

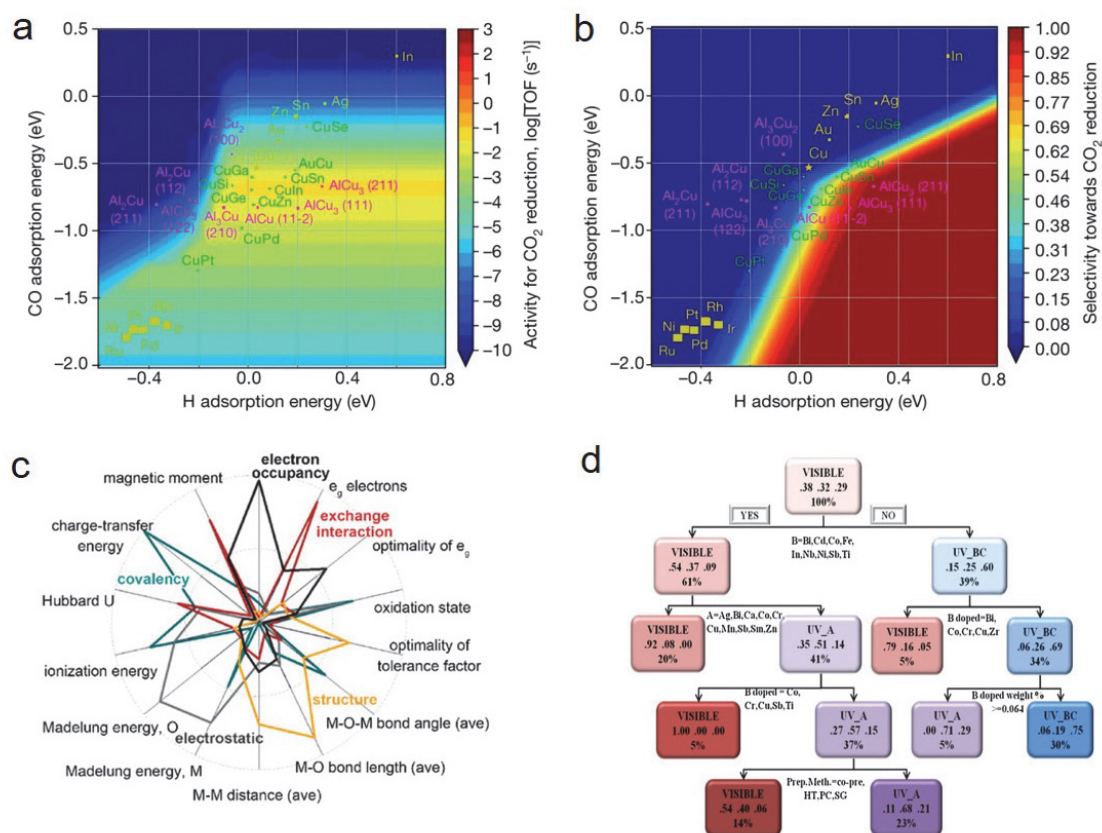


图 13 (a) CO₂ 还原的活度图^[100]; (b) CO₂ 还原的选择性图^[100]; (c) 通过因子分析得到 14 个描述符的载荷大小^[102]; (d) 带隙分析结果: 为带隙构建决策树^[105]

Figure 13 (a) A two-dimensional activity volcano plot for CO₂ reduction^[100]; (b) a two-dimensional selectivity volcano plot for CO₂ reduction^[100]; (c) loading magnitudes for the 14 descriptors obtained by factor analysis^[102]; (d) results of the analysis of band gaps: decision tree constructed for band gaps^[105]

4 未来展望

传统研究手段由于实验周期长、成本高,难以满足材料科学的要求。机器学习方法可以节约大量的时间和资源^[119],并且在许多实验中的平均表现优于有丰富经验的工作者^[120-121],在新材料中的应用^[122]中彰显出巨大潜力。然而,当前机器学习技术并不是解决一切问题的万能钥匙。

数据驱动是机器学习的重要特点,同样也是它的局限性所在。当前,可用于材料领域的机器学习的数据规模仍然很小,限制着材料科学的发展,从每种算法模型的应用经验来看,数据集的规模仍然是模型预测准确性的主要因素^[29]。伴随数据集规模而来的还有数据收集的高成本问题,大部分研究人员实验得来的数据难以得到有效利用,而且高质量数据的筛选也是一大难题。利用技术和策略来充分利用有限的数据集将是未来研究的热点问题^[123]。其次,机器学习模型通常是连接输入和输出的“黑匣子”,因此很难从机器学习模型中提取知识,然后将其概括为一般案例的科学定律。

尽管存在许多问题,但不可否认的是,作为一门正在发展初期的学科,机器学习在新材料应用方面有许多有前景的发展方向^[124]。目前来看,机器学习未来的发展道路可能集中于以下几个方向:

(1)建立更为全面的数据库平台。机器学习是一种基于统计学的研究方法^[125],其学习能力本质上与数据集的质量有关。虽然材料基因组计划的发展促使了多个数据库的发展,但是更多的公开数据没有在数据库体现。科技的飞速发展导致数据爆炸式增长,将来研究人员可以结合理论与实验进一步发展数据库平台^[3]。

(2)完善更为快捷的学习模型与算法。针对材料中的数据量相对较少的问题,在保证精度以及普适性的前提下,建立和发展精确的小数据分析模型^[126-127]是目前研究者亟待解决的问题,例如迁移学习^[128]、元学习^[129]、神经网络图灵机^[130]、贝叶斯框架^[131]等多种学习模型^[132-135]。

(3)促进分子和材料的性能化导向设计^[10]。通过深度理解材料组成-结构-性质(CSP)三者关系,由特定的物理化学性能反推到材料中原子的空间构型排布^[136],设计合成一些特殊材料如高稳定性、高容量的储能材料、高效率的催化剂材料、高电导率的电池材料等。

(4)结合高通量合成和表征技术,助力大尺度物质研究。随着每一个新的成分或材料参数的增加,机器学习预测难度将呈指数增长。机器学习用于大尺度材料将会是一个兼具挑战与机遇的研究领域^[137-138]。

机器学习相比于传统研究方式,为人们研究开发新材料提供了一种新视角和思路。在“第四次工业革命”的大时代背景下,高新技术不断发展,新型材料的需求量持续加大。机器学习不仅可在理论上促进科学家对材料性质的理解,同时也可指导新材料的合成。我们期待

未来机器学习在新材料的应用中绽放的无限可能。

作者简介



戚兴怡,北京化工大学化学学院化学专业在读研究生。2022年6月于北京化工大学化学学院应用化学专业获得学士学位。



胡耀峰,北京化工大学化学学院化学专业在读本科生。



赵宇飞,北京化工大学化学学院教授,博士生导师。入选中国科协“青年人才托举工程”计划,2019年度国家自然科学基金优秀青年科学基金获得者。工作围绕水滑石基二维插层材料的可控合成及精细结构表征,面向高值精细化学品的光/电催化合成。近年来以第一/通讯作者在 Chem. Soc. Rev.、J. Am. Chem. Soc.、Angew. Chem.、Adv. Mater.、Chem.、Joule、Ind. Eng. Chem. Res.等期刊上发表SCI收录论文30余篇;累计SCI他引4500余次;5篇入选ESI高被引论文,已授权国家发明专利19项。

References

- [1] Correa-Baena, J.-P.; Hippalgaonkar, K.; van Duren, J.; Jaffer, S.; Chandrasekhar, V. R.; Stevanovic, V.; Wadia, C.; Guha, S.; Buonassisi, T. *Joule* **2018**, 2, 1410.
- [2] Feng, N.; Wang, H. J.; Li, M. *Inf. Sci.* **2014**, 256, 57.
- [3] Chen, A.; Zhang, X.; Zhou, Z. *InfoMat* **2020**, 2, 553.
- [4] Ramprasad, R.; Batra, R.; Pilania, G.; Mannodi-Kanakkithodi, A.; Kim, C. *Npj Comput. Mater.* **2017**, 3, 1.
- [5] Hachmann, J.; Olivares-Amaya, R.; Atahan-Evrenk, S.; Amador-Bedolla, C.; Sánchez-Carrera, R. S.; Gold-Parker, A.; Vogt, L.; Brockway, A. M.; Aspuru-Guzik, A. *J. Phys. Chem. Lett.* **2011**, 2, 2241.
- [6] Kirklin, S.; Saal, J. E.; Meredig, B.; Thompson, A.; Doak, J. W.; Aykol, M.; Rühl, S.; Wolverton, C. *Npj Comput. Mater.* **2015**, 1, 1.

- [7] Curtarolo, S.; Setyawan, W.; Hart, G. L.; Jahnatek, M.; Chepulskii, R. V.; Taylor, R. H.; Wang, S.; Xue, J.; Yang, K.; Levy, O. *Comput. Mater. Sci.* **2012**, *58*, 218.
- [8] Sun, Z. T.; Li, Z. Z.; Cheng, G. J.; Xu, Q. C.; Hou, Z. F.; Yin, W. J. *Chin. Sci. Bull.* **2019**, *64*, 3270 (in Chinese). (孙中体, 李珍珠, 程观剑, 徐其琛, 侯柱峰, 尹万健, 科学通报, **2019**, *64*, 3270.)
- [9] Suh, C.; Fare, C.; Warren, J. A.; Pyzer-Knapp, E. O. *Annu. Rev. Mater. Res.* **2020**, *50*, 1.
- [10] Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. *Nature* **2018**, *559*, 547.
- [11] Zhu, B. Y.; Wu, R. L.; Yu, X. *Acta Chim. Sinica* **2020**, *78*, 1366 (in Chinese). (朱博阳, 吴睿龙, 于曦, 化学学报, **2020**, *78*, 1366.)
- [12] Sun, Y. M.; Liao, H. B.; Wang, J. R.; Chen, B.; Sun, S. N.; Ong, S. J. H.; Xi, S. B.; Diao, C. Z.; Du, Y. H.; Wang, J. O.; Breese, M. B. H.; Li, S. Z.; Zhang, H.; Xu, Z. C. *J. Nat. Catal.* **2020**, *3*, 554.
- [13] Deringer, V. L.; Bernstein, N.; Csanyi, G.; Ben Mahmoud, C.; Ceriotti, M.; Wilson, M.; Drabold, D. A.; Elliott, S. R. *Nature* **2021**, *589*, 59.
- [14] Cai, C. Z.; Li, L. F.; Deng, X. M.; Li, S. H.; Liang, H.; Qiao, Z. W. *Acta Chim. Sinica* **2020**, *78*, 427 (in Chinese). (蔡铨智, 李丽凤, 邓小梅, 李树华, 梁红, 乔智威, 化学学报, **2020**, *78*, 427.)
- [15] Sheng, Y.; Wu, Y. S.; Yang, J.; Lu, W. C.; Villars, P.; Zhang, W. Q. *Npj Comput. Mater.* **2020**, *6*, 1.
- [16] Luo, H.; Sohn, S. S.; Lu, W. J.; Li, L. L.; Li, X. G.; Soundararajan, C. K.; Krieger, W.; Li, Z. M.; Raabe, D. *Nat. Commun.* **2020**, *11*, 1.
- [17] Lin, H.; Zheng, J. X.; Lin, Y.; Pan, F. *Energy Stor. Sci. Technol.* **2017**, *6*, 990 (in Chinese). (林海, 郑家新, 林原, 潘锋, 储能科学与技术, **2017**, *6*, 990.)
- [18] Pei, J. F.; Cai, C. Z.; Zhu, Y. M.; Yan, B. *Macromol. Theory Simul.* **2013**, *22*, 52.
- [19] Juan, Y.; Dai, Y.; Yang, Y.; Zhang, J. *J. Mater. Sci. Technol.* **2021**, *79*, 178.
- [20] Wang, H.; Ji, Y.; Li, Y. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2019**, *10*, e1421.
- [21] Mi, X. X.; Tang, A. T.; Zhu, Y. C.; Kang, L.; Pan, F. S. *Mater. Rep.* **2021**, *15*, 1 (in Chinese). (米晓希, 汤爱涛, 朱雨晨, 康靓, 潘复生, 材料导报, **2021**, *15*, 1.)
- [22] Somvanshi, M.; Chavan, P.; Tambade, S.; Shinde, S. V. *ICCUBEA*, **2016**, pp. 1-7.
- [23] Diamantopoulou, M. J.; Papamichail, D. M.; Antonopoulos, V. Z. *Oper. Res.* **2005**, *5*, 115.
- [24] Goh, G. B.; Hodas, N. O.; Vishnu, A. *J. Comput. Chem.* **2017**, *38*, 1291.
- [25] Fang, S.; Wang, M.; Qi, W.; Zheng, F. *Comput. Mater. Sci.* **2008**, *44*, 647.
- [26] Wei, J.; Chu, X.; Sun, X. Y.; Xu, K.; Deng, H. X.; Chen, J.; Wei, Z.; Lei, M. *InfoMat* **2019**, *1*, 338.
- [27] Kalidindi, S. R.; De Graef, M. *Annu. Rev. Mater. Res.* **2015**, *45*, 171.
- [28] Liu, Y.; Zhao, T.; Ju, W.; Shi, S. J. *Materiomics* **2017**, *3*, 159.
- [29] Liu, Y.; Niu, C.; Wang, Z.; Gan, Y.; Zhu, Y.; Sun, S.; Shen, T. *J. Mater. Sci. Technol.* **2020**, *57*, 113.
- [30] Liu, Z. L.; Li, W.; Liu, H.; Zhuang, X. D.; Li, S. *Acta Chim. Sinica* **2019**, *77*, 323 (in Chinese). (刘治鲁, 李炜, 刘昊, 庄旭东, 李松, 化学学报, **2019**, *77*, 323.)
- [31] Lian, Z.; Yang, M.; Jan, F.; Li, B. *J. Phys. Chem. Lett.* **2021**, *12*, 7053.
- [32] Zhang, Q.; Zheng, Y. J.; Sun, W.; Ou, Z.; Odumbaku, O.; Li, M.; Chen, S.; Zhou, Y.; Li, J.; Qin, B.; Sun, K. *Adv. Sci.* **2022**, *9*, 2104742.
- [33] Wang, X.; Gao, Y.; Krzystowczyk, E.; Iftikhar, S.; Dou, J.; Cai, R.; Wang, H.; Ruan, C.; Ye, S.; Li, F. *Energy Environ. Sci.* **2022**, *15*, 1512.
- [34] Jie, J.; Hu, Z.; Qian, G.; Weng, M.; Li, S.; Li, S.; Hu, M.; Chen, D.; Xiao, W.; Zheng, J.; Wang, L.; Pan, F. *Sci. Bull.* **2019**, *64*, 612.
- [35] Faber, F.; Lindmaa, A.; von Lilienfeld, O. A.; Armiento, R. *Int. J. Quantum. Chem.* **2015**, *115*, 1094.
- [36] Nolan, A. M.; Liu, Y.; Mo, Y. *ACS Energy Lett.* **2019**, *4*, 2444.
- [37] Bartel, C. J.; Trewartha, A.; Wang, Q.; Dunn, A.; Jain, A.; Ceder, G. *Npj Comput. Mater.* **2020**, *6*, 1.
- [38] Lu, S.; Zhou, Q.; Ouyang, Y.; Guo, Y.; Li, Q.; Wang, J. *Nat Commun.* **2018**, *9*, 1.
- [39] Rajan, A. C.; Mishra, A.; Satsangi, S.; Vaish, R.; Mizuseki, H.; Lee, K.-R.; Singh, A. K. *Chem. Mater.* **2018**, *30*, 4031.
- [40] Stanev, V.; Oses, C.; Kusne, A. G.; Rodriguez, E.; Paglione, J.; Curtarolo, S.; Takeuchi, I. *Npj Comput. Mater.* **2018**, *4*, 1.
- [41] Owolabi, T. O.; Akande, K. O.; Olatunji, S. O. *Adv. Phys. Theor. Appl.* **2014**, *35*, 12.
- [42] Xie, S.; Stewart, G.; Hamlin, J.; Hirschfeld, P.; Hennig, R. *Phys. Rev. B* **2019**, *100*, 174513.
- [43] Allen, P. B.; Dynes, R. *Phys. Rev. B* **1975**, *12*, 905.
- [44] Bartel, C. J.; Sutton, C.; Goldsmith, B. R.; Ouyang, R.; Musgrave, C. B.; Ghiringhelli, L. M.; Scheffler, M. *Sci. Adv.* **2019**, *5*, eaav0693.
- [45] Chen, X.; Chen, D.; Weng, M.; Jiang, Y.; Wei, G. W.; Pan, F. *J. Phys. Chem. Lett.* **2020**, *11*, 4392.
- [46] Timoshenko, J.; Lu, D.; Lin, Y.; Frenkel, A. I. *J. Phys. Chem. Lett.* **2017**, *8*, 5091.
- [47] Ou, Z.; Liu, C.; Yao, L.; Chen, Q. *Acc. Mater. Res.* **2020**, *1*, 41.
- [48] LeCun, Y.; Bengio, Y.; Hinton, G. *Nature* **2015**, *521*, 436.
- [49] Ziatdinov, M.; Dyck, O.; Maksov, A.; Li, X.; Sang, X.; Xiao, K.; Unocic, R. R.; Vasudevan, R.; Jesse, S.; Kalinin, S. V. *ACS Nano* **2017**, *11*, 12742.
- [50] Kwoen, J.; Arakawa, Y. *Cryst. Growth Des.* **2020**, *20*, 5289.
- [51] Nada, H. *ACS Omega* **2018**, *3*, 5789.
- [52] Fortunato, M. E.; Coley, C. W.; Barnes, B. C.; Jensen, K. F. *J. Chem. Inf. Model.* **2020**, *60*, 3398.
- [53] Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. *J. Chem. Theory Comput.* **2015**, *11*, 2087.
- [54] Dral, P. O. *J. Phys. Chem. Lett.* **2020**, *11*, 2336.
- [55] Zhang, P.; Shen, L.; Yang, W. *J. Phys. Chem. B* **2018**, *123*, 901.
- [56] Dral, P. O.; von Lilienfeld, O. A.; Thiel, W. *J. Chem. Theory Comput.* **2015**, *11*, 2120.
- [57] Brockherde, F.; Vogt, L.; Li, L.; Tuckerman, M. E.; Burke, K.; Müller, K.-R. *Nat. Commun.* **2017**, *8*, 1.
- [58] Chandrasekaran, A.; Kamal, D.; Batra, R.; Kim, C.; Chen, L.; Ramprasad, R. *Npj Comput. Mater.* **2019**, *5*, 1.
- [59] Zhang, Z.; Zhang, Y.; Wang, J.; Xu, J.; Long, R. *J. Phys. Chem. Lett.* **2021**, *12*, 835.
- [60] Shang, C.; Liu, Z.-P. *J. Chem. Theory Comput.* **2013**, *9*, 1838.
- [61] Shang, C.; Zhang, X.-J.; Liu, Z.-P. *Phys. Chem. Chem. Phys.* **2014**, *16*, 17845.
- [62] Huang, S.-D.; Shang, C.; Zhang, X.-J.; Liu, Z.-P. *Chem. Sci.* **2017**, *8*, 6327.
- [63] Huang, S.-D.; Shang, C.; Kang, P.-L.; Liu, Z.-P. *Chem. Sci.* **2018**, *9*, 8644.
- [64] Huang, S. D.; Shang, C.; Kang, P. L.; Zhang, X. J.; Liu, Z. P. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2019**, *9*, e1415.
- [65] Car, R.; Parrinello, M. *Phys. Rev. Lett.* **1985**, *55*, 2471.
- [66] Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I. *J. Comput. Chem.* **2010**, *31*, 671.
- [67] Wang, H.; Zhang, L.; Han, J.; Weinan, E. *Comput. Phys. Commun.* **2018**, *228*, 178.
- [68] Yang, S. X. *World. Nonferrous. Met.* **2020**, *14*, 154 (in Chinese). (杨素心, 世界有色金属, **2020**, *14*, 154.)
- [69] Goldschmidt, V. M. *Naturwiss.* **1926**, *14*, 477.
- [70] Li, Z.; Xu, Q.; Sun, Q.; Hou, Z.; Yin, W. *J. Adv. Funct. Mater.* **2019**, *29*, 1807280.
- [71] Weng, B.; Song, Z.; Zhu, R.; Yan, Q.; Sun, Q.; Grice, C. G.; Yan, Y.; Yin, W. *J. Nat Commun.* **2020**, *11*, 3513.
- [72] Feng, H.-J.; Wu, K.; Deng, Z.-Y. *Cell Rep. Phys. Sci.* **2020**, *1*, 100179.
- [73] Kirman, J.; Johnston, A.; Kuntz, D. A.; Askerka, M.; Gao, Y.; Todorović, P.; Ma, D.; Privé, G. G.; Sargent, E. H. *Matter* **2020**, *2*, 938.
- [74] He, Y.; Cubuk, E. D.; Allendorf, M. D.; Reed, E. J. *J. Phys. Chem. Lett.* **2018**, *9*, 4562.
- [75] Gordillo, M. A.; Benavides, P. A.; Panda, D. K.; Saha, S. *ACS Appl. Mater. Interfaces* **2020**, *12*, 12955.
- [76] Mannodi-Kanakithodi, A.; Toriyama, M. Y.; Sen, F. G.; Davis, M. J.; Klie, R. F.; Chan, M. K. *Npj Comput. Mater.* **2020**, *6*, 1.
- [77] Chen, G.; Shen, Z.; Iyer, A.; Ghumman, U. F.; Tang, S.; Bi, J.; Chen, W.; Li, Y. *Polymers* **2020**, *12*, 163.
- [78] Gormley, A. J.; Webb, M. A. *Nat. Rev. Mater.* **2021**, *6*, 642.
- [79] Shmilovich, K.; Mansbach, R. A.; Sidky, H.; Dunne, O. E.; Panda, S. S.; Tovar, J. D.; Ferguson, A. L. *J. Phys. Chem. B* **2020**, *124*, 3873.
- [80] Yan, C.; Feng, X.; Li, G. *ACS Appl. Mater. Interfaces* **2021**, *13*, 60508.
- [81] Webb, M. A.; Jackson, N. E.; Gil, P. S.; de Pablo, J. J. *Sci. Adv.* **2020**, *6*, eabc6216.
- [82] Giuntoli, A.; Hansoge, N. K.; van Beek, A.; Meng, Z.; Chen, W.; Keten, S. *npj Comput. Mater.* **2021**, *7*, 1.
- [83] Writer, B.; Batteries, L.-I. *A Machine-generated Summary of Current Research*, Springer, Cham, Switzerland, **2019**.

- [84] Ceder, G. *MRS bull.* **2010**, 35, 693.
- [85] Li, W.; Zhu, J.; Xia, Y.; Gorji, M. B.; Wierzbicki, T. *Joule* **2019**, 3, 2703.
- [86] Severson, K. A.; Attia, P. M.; Jin, N.; Perkins, N.; Jiang, B.; Yang, Z.; Chen, M. H.; Aykol, M.; Herring, P. K.; Fraggadakis, D.; Bazant, M. Z.; Harris, S. J.; Chueh, W. C.; Braatz, R. D. *Nat. Energy* **2019**, 4, 383.
- [87] Homma, K.; Liu, Y.; Sumita, M.; Tamura, R.; Fushimi, N.; Iwata, J.; Tsuda, K.; Kaneta, C. *J. Phys. Chem. C* **2020**, 124, 12865.
- [88] Dixit, M. B.; Verma, A.; Zaman, W.; Zhong, X.; Kenesei, P.; Park, J. S.; Almer, J.; Mukherjee, P. P.; Hatzell, K. B. *ACS Appl. Energy Mater.* **2020**, 3, 9534.
- [89] Panosetti, C.; Annies, S. B.; Grosu, C.; Seidlmayer, S.; Scheurer, C. *J. Phys. Chem. A* **2021**, 125, 691.
- [90] Wang, Y.; Xie, T.; France-Lanord, A.; Berkley, A.; Johnson, J. A.; Shao-Horn, Y.; Grossman, J. C. *Chem. Mater.* **2020**, 32, 4144.
- [91] Jiang, Z.; Li, J.; Yang, Y.; Mu, L.; Wei, C.; Yu, X.; Pianetta, P.; Zhao, K.; Cloetens, P.; Lin, F. *Nat. Commun.* **2020**, 11, 1.
- [92] Sendek, A. D.; Antoniuk, E. R.; Cubuk, E. D.; Ransom, B.; Francisco, B. E.; Buettner-Garrett, J.; Cui, Y.; Reed, E. J. *ACS Appl. Mater. Interfaces* **2020**, 12, 37957.
- [93] Sendek, A. D.; Cubuk, E. D.; Antoniuk, E. R.; Cheon, G.; Cui, Y.; Reed, E. J. *Chem. Mater.* **2018**, 31, 342.
- [94] Wang, C.; Aoyagi, K.; Wisesa, P.; Mueller, T. *Chem. Mater.* **2020**, 32, 3741.
- [95] Toyao, T.; Maeno, Z.; Takakusagi, S.; Kamachi, T.; Takigawa, I.; Shimizu, K.-i. *ACS Catal.* **2019**, 10, 2260.
- [96] Kitchin, J. R. *Nat. Catal.* **2018**, 1, 230.
- [97] Sun, Z.; Yin, H.; Liu, K.; Cheng, S.; Li, G. K.; Kawi, S.; Zhao, H.; Jia, G.; Yin, Z. *SmartMat.* **2022**, 3, 68.
- [98] Chen, A.; Zhang, X.; Chen, L.; Yao, S.; Zhou, Z. *J. Phys. Chem. C* **2020**, 124, 22471.
- [99] Tran, K.; Ulissi, Z. W. *Nat. Catal.* **2018**, 1, 696.
- [100] Zhong, M.; Tran, K.; Min, Y.; Wang, C.; Wang, Z.; Dinh, C.-T.; De Luna, P.; Yu, Z.; Rasouli, A. S.; Brodersen, P. *Nature* **2020**, 581, 178.
- [101] Batchelor, T. A.; Pedersen, J. K.; Winther, S. H.; Castelli, I. E.; Jacobsen, K. W.; Rossmeisl, J. *Joule* **2019**, 3, 834.
- [102] Liu, J.; Liu, H.; Chen, H.; Du, X.; Zhang, B.; Hong, Z.; Sun, S.; Wang, W. *Adv. Sci.* **2020**, 7, 1901614.
- [103] Fujishima, A.; Honda, K. *Nature* **1972**, 238, 37.
- [104] Zhang, Y.; Xu, X. *ACS Omega* **2020**, 5, 15344.
- [105] Can, E.; Yildirim, R. *Appl. Catal., B* **2019**, 242, 267.
- [106] Bai, Y.; Wilbraham, L.; Slater, B. J.; Zwijnenburg, M. A.; Sprick, R. S.; Cooper, A. I. *J. Am. Chem. Soc.* **2019**, 141, 9063.
- [107] Corma, A.; Moliner, M.; Serra, J. M.; Serna, P.; Díaz-Cabañas, M. J.; Baumes, L. A. *Chem. Mater.* **2006**, 18, 3287.
- [108] Muraoka, K.; Sada, Y.; Miyazaki, D.; Chaikitilap, W.; Okubo, T. *Nat. Commun.* **2019**, 10, 1.
- [109] Boyd, P. G.; Chidambaram, A.; García-Díez, E.; Ireland, C. P.; Daff, T. D.; Bounds, R.; Gładysiak, A.; Schouwink, P.; Moosavi, S. M.; Maroto-Valer, M. M.; Reimer, J. A.; Navarro, J. A. R.; Woo, T. K.; García, S.; Stylianou, K. C.; Smit, B. *Nature* **2019**, 576, 253.
- [110] Ma, S.; Liu, Z.-P. *ACS Catal.* **2020**, 10, 13213.
- [111] Ma, S. C.; Liu, Z.-P. *Chem. Ind. Eng. Prog.* **2020**, 39, 3433 (in Chinese). (马思聪, 刘智攀, 化工进展, **2020**, 39, 3433.)
- [112] Deng, C.; Su, Y.; Li, F.; Shen, W.; Chen, Z.; Tang, Q. *J. Mater. Chem. A* **2020**, 8, 24563.
- [113] Ishioka, S.; Fujiwara, A.; Nakanowatari, S.; Takahashi, L.; Taniike, T.; Takahashi, K. *ACS Catal.* **2022**, 12, 11541.
- [114] Yang, Z.; Gao, W. *Adv. Sci.* **2022**, 9, 2106043.
- [115] Jäger, M. O.; Morooka, E. V.; Canova, F. F.; Himanen, L.; Foster, A. S. *Npj Comput. Mater.* **2018**, 4, 1.
- [116] Ulissi, Z. W.; Singh, A. R.; Tsai, C.; Nørskov, J. K. *J. Phys. Chem. Lett.* **2016**, 7, 3931.
- [117] Takigawa, I.; Shimizu, K.-i.; Tsuda, K.; Takakusagi, S. *RSC Adv.* **2016**, 6, 52587.
- [118] Zhu, Q.; Zhang, F.; Huang, Y.; Xiao, H.; Zhao, L.; Zhang, X.; Song, T.; Tang, X.; Li, X.; He, G.; Chong, B.; Zhou, J.; Zhang, Y.; Zhang, B.; Cao, J.; Luo, M.; Wang, S.; Ye, G.; Zhang, W.; Chen, X.; Cong, S.; Zhou, D.; Li, H.; Li, J.; Zou, G.; Shang, W.; Jiang, J.; Luo, Y. *Natl. Sci. Rev.* **2022**, nwac190.
- [119] Balachandran, P. V.; Young, J.; Lookman, T.; Rondinelli, J. M. *Nat. Commun.* **2017**, 8, 1.
- [120] Shields, B. J.; Stevens, J.; Li, J.; Parasram, M.; Damani, F.; Alvarado, J. I. M.; Janey, J. M.; Adams, R. P.; Doyle, A. G. *Nature* **2021**, 590, 89.
- [121] Raccuglia, P.; Elbert, K. C.; Adler, P. D.; Falk, C.; Wenny, M. B.; Mollo, A.; Zeller, M.; Friedler, S. A.; Schrier, J.; Norquist, A. J. *Nature* **2016**, 533, 73.
- [122] Wang, A. Y.-T.; Murdock, R. J.; Kauwe, S. K.; Oliynyk, A. O.; Gurlu, A.; Brgoch, J.; Persson, K. A.; Sparks, T. D. *Chem. Mater.* **2020**, 32, 4954.
- [123] Krishnamurthy, D.; Weiland, H.; Barati Farimani, A.; Antono, E.; Green, J.; Viswanathan, V. *ACS Energy Lett.* **2018**, 4, 187.
- [124] De Luna, P.; Wei, J.; Bengio, Y.; Aspuru-Guzik, A.; Sargent, E. *Use machine learning to find energy materials*, Nature Publishing Group, **2017**.
- [125] Schmidt, J.; Marques, M. R.; Botti, S.; Marques, M. A. *npj Comput. Mater.* **2019**, 5, 1.
- [126] Zhang, Y.; Ling, C. *Npj Comput. Mater.* **2018**, 4, 1.
- [127] Trivedi, H.; Shvartsman, V. V.; Medeiros, M. S.; Pullar, R. C.; Lupascu, D. C. *Npj Comput. Mater.* **2018**, 4, 1.
- [128] Luo, W.; Shen, F.; Bommier, C.; Zhu, H.; Ji, X.; Hu, L. *Acc. Chem. Res.* **2016**, 49, 231.
- [129] Jankowski, N.; Duch, W.; Grabczewski, K. *Meta-learning in computational intelligence*, Springer, **2011**.
- [130] Graves, A.; Wayne, G.; Danihelka, I. *arXiv preprint arXiv:1410.5401*. **2014**.
- [131] Lake, B. M.; Salakhutdinov, R.; Tenenbaum, J. B. *Science* **2015**, 350, 1332.
- [132] Kauwe, S. K.; Welker, T.; Sparks, T. D. *Integr. Mater. Manuf. I.* **2020**, 9, 213.
- [133] Ouyang, R.; Ahmetcik, E.; Carbogno, C.; Scheffler, M.; Ghiringhelli, L. M. *J. Phys. Mater.* **2019**, 2, 024002.
- [134] Jackson, N. E.; Bowen, A. S.; Antony, L. W.; Webb, M. A.; Vishwanath, V.; de Pablo, J. J. *Sci. Adv.* **2019**, 5, eaav1190.
- [135] Sutton, C.; Boley, M.; Ghiringhelli, L. M.; Rupp, M.; Vreeken, J.; Scheffler, M. *Nat. Commun.* **2020**, 11, 1.
- [136] Wu, W.; Sun, Q. *Sci. China: Phys., Mech. Astron.* **2018**, 48, 58 (in Chinese). (吴炜, 孙强, 中国科学: 物理学 力学 天文学, **2018**, 48, 58.)
- [137] Wu, S.; Kondo, Y.; Kakimoto, M.-a.; Yang, B.; Yamada, H.; Kuwajima, I.; Lambard, G.; Hongo, K.; Xu, Y.; Shiomi, J. *Npj Comput. Mater.* **2019**, 5, 1.
- [138] Jia, W.; Wang, H.; Chen, M.; Lu, D.; Liu, J.; Lin, L.; Car, R.; Zhang, L. *arXiv preprint arXiv:2005.00223*. **2020**, 1.

(Cheng, B.)