基于机器学习的材料设计

陆文聪[1],吴炎森[1],刘太昂[2],卢天[2],纪晓波[1],邢雷凯[1]

(1.上海大学 理学院化学系,上海 200444;2.上海数之微信息科技有限公司,上海 200444)

摘 要:材料创新一直是推动人类文明进步的重要引擎,随着现代科技对高性能材料日益增长的需求,材料科学的重要性也逐渐凸显.传统的"试错法"和第一性原理应用于复杂的材料设计还有较大的局限性.机器学习已发展成为材料科学研究的新范式,通过分析大量数据,可实现高通量筛选、性能预测、晶体结构预测和材料配方优化等方面的功能.机器学习结合第一性原理方法的材料设计,为材料研究带来了崭新的思路.回顾了机器学习在材料设计领域的应用,探讨了其加速材料创新、降低试错成本、定制化材料设计等方面的应用潜力,展望了其对材料科学领域带来的机遇和挑战.

关键词:机器学习;材料设计;配方优化

中图分类号:O69 **文献标志码**:A **文章编号**:1000-2367(2024)04-0120-12

材料科学是一门与工程技术密不可分的应用科学,主要研究材料的合成、性能、结构和应用.从古至今,材料的发现和创新一直是人类文明进步的重要驱动力.当代社会,现代科技和工业对于新型、高性能材料的需求与日俱增,材料科学的重要性日益凸显.材料科学的发展为科技生活注入了新的活力,如高性能轻质金属材料、新型钙钛矿材料、高温超导材料、半导体材料等,在能源、通信、交通、医疗等各领域都发挥着重要作用[1].在过去几十年里,材料科学家通常依赖实验室试错和基于第一性原理的理论计算来探索新材料[2].然而,随着科技的不断进步和人类对于材料性能要求的提高,仅用传统"试错法"和第一性原理的材料研究方法的局限性也日益显现出来.实验试错法耗时费力,尤其在大量候选材料中筛选最佳时十分耗费时间和资源.基于第一性原理的理论计算法在某些情况下难以解决复杂的材料性能预测问题,因为某些材料的性能受多个因素共同影响,难以完全充分考虑[3].此外,虽然密度泛函理论(DFT)和分子动力学模拟在研究中发挥作用,但它们也存在局限性[4],如体系大小限制、自相互作用误差、时间和空间尺度限制以及势函数的试错成本较高等[5].

随着机器学习技术的迅速发展,其被视为科学研究的第四范式,为材料科学带来了新的机遇[6].在材料科学领域,机器学习被广泛应用于高通量筛选、性能预测、晶体结构预测和材料优化等方面.通过学习海量实验和模拟数据,机器学习算法能够发现隐藏在数据中的规律,为材料研究提供新的洞察力和设计思路[7].机器学习与第一性原理方法结合,已成为材料科学的研究热点(图 1),有助于预测材料性能和稳定性,减少试错成本和时间[8].这一革命性进展为材料科学带来了新的突破.

本文简要介绍了材料科学领域中常用的机器学习方法、机器学习方法在材料科学领域应用的一般流程以及机器学习方法在金属材料、钙钛矿材料等科学领域中的具体应用案例,最后对机器学习方法在材料化学领域可能遇到的机遇和挑战进行了展望.

收稿日期:2023-11-16;**修回日期**:2023-12-19.

基金项目:国家自然科学基金(52102140);云南贵金属实验室重大科技专项(YPML-2023050205;YPML-2023050208).

作者简介(通信作者):陆文聪(1964—),男,浙江慈溪人,上海大学教授,博士生导师,主要从事基于数据挖掘/机器学习的材料设计和工业优化等研究工作,E-mail:wclu@shu.edu.cn.

引用本文:陆文聪,吴炎森,刘太昂,等.基于机器学习的材料设计[J].河南师范大学学报(自然科学版),2024,52(4):120-131.(Lu Wencong,Wu Yanmiao,Liu Taiang,et al.Machine learning-based materials design[J].Journal of Henan Normal University(Natural Science Edition),2024,52(4):120-131.DOI:10.16366/j.cnki.1000-2367.2023.11.16.0003.)

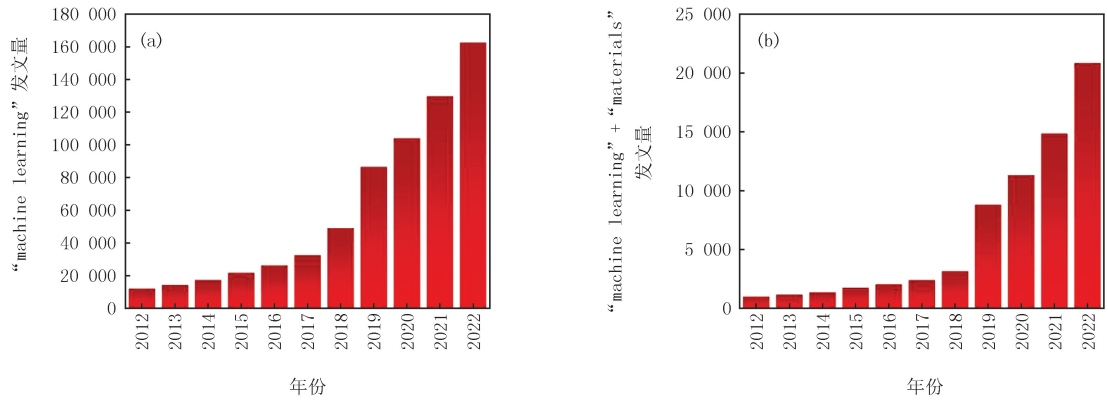


图1　近10年Web of science上关键词“机器学习”(a)和“机器学习”+“材料”(b)发文量总结

Fig. 1　Summary of publication quantity on Web of science in the past 10 years for "machine learning"(a) and "machine learning" + "materials"(b)

1　材料领域中常用的机器学习方法简介

在材料领域,机器学习方法的应用主要体现在定性问题和定量问题的探索上.定性问题涉及分类,如材料是否形成的类别判断;定量问题则涉及回归(连续数值的定量拟合),如材料性能的定量预测.机器学习建模用于建立函数(或映射)关系,实现数据可视化、正向预测材料性能以及逆向设计材料配方和加工条件.机器学习算法旨在寻找全局最优解,并保证高效求解过程[9].材料科学中常用的机器学习算法的特点和适用场合见表 1.

表 1　材料科学领域常见的机器学习算法

Tab. 1　Common machine learning algorithms in the field of materials science

机器学习算法	类型	简要说明
K-最近邻(KNN)	回归/分类	KNN 是一种基于欧氏距离的成熟模式识别方法,通过考虑未知样本的 K 个最近邻的类别来进行“少数服从多数”的类别预测.
偏最小二乘法(PLS)	回归/分类	偏最小二乘法是一种多变量统计方法,通过降维和提取主成分的方式,有效处理输入变量间的多重共线性,常用于建立回归或分类模型.
岭回归(RR)	回归	岭回归是一种用于处理回归问题的线性模型,通过在最小二乘法中引入正则化项,有效解决多重共线性,提高模型的稳定性和泛化能力.
决策树(DT)	回归/分类	决策树是一种可解释性强的监督学习算法,通过特征分裂的信息增益、信息增益率或 Gini 系数生成一颗二叉树,表示为 if-else 规则的集合.
集成学习	回归/分类	集成学习是一种新的机器学习范式,通过多个学习器协同解决同一个问题,其中 AdaBoost 注重错分点,Bagging 注重结果的平均化,随机森林和梯度提升树是常见的集成学习方法,以有效提高学习系统的泛化能力.
随机森林(RF)	回归/分类	并行式集成学习策略,算法包含多个决策树,小噪声情况下可以防止过拟合.
线性回归分析	回归	线性回归分析主要采用多元线性回归(MLR)方法,通过最小二乘原理求解待定系数,广泛应用于科学工程领域.
支持向量机(SVM)	回归/分类	SVM 是建立在 Vapnik 的统计学习新理论基础上的机器学习方法,包括支持向量分类和支持向量回归算法,特别适用于小样本集的建模预测.
人工神经网络(ANN)	回归/分类	人工神经网络是模拟生物神经网络的信息处理系统,具有强大的非线性拟合能力,仅需 3 层即可拟合任意非线性函数关系.
深度学习	回归/分类	深度学习与人工神经网络的区别在于深度学习包含多个深层隐藏层,是人工神经网络的一种扩展,主要应用于图像、声音、语言数据处理,常见的深度学习模型包括卷积神经网络,循环神经网络、图神经网络等.

续 表

机器学习算法	类型	简要说明
卷积神经网络(CNN)	回归/分类	卷积神经网络是一种专门设计用于处理网格结构数据(如图像)的深度学习模型,通过卷积层、池化层等操作有效提取和学习数据的空间层级特征.
循环神经网络(RNN)	回归/分类	循环神经网络是一类具有循环连接的深度学习模型,专门用于处理序列数据,能够捕捉数据中的时间依赖关系,广泛应用于自然语言处理和时序数据分析.
图神经网络(GNN)	回归/分类	图神经网络是一类深度学习模型,专门用于处理和学习图状数据结构,通过节点和边上的信息传递和聚合,从而实现对图结构中节点关系的复杂特征学习.
K-均值算法(K-means)	聚类	K-均值算法是一种无监督的聚类算法,根据对象的特征变量将其分为 K 个类别,同一聚类中对象相似度高,采用欧氏距离进行相似度评判.
层次聚类	聚类	层次聚类是一种无监督学习方法,通过逐步合并或分裂样本以构建层级结构,从而将数据集中的样本组织成多个层次化的簇.
谱聚类	聚类	谱聚类是一种基于数据的图论方法,通过分析样本数据的谱结构,将数据集划分为不同的簇,适用于发现非凸形状簇和处理高维数据.

表 2 列出了几种常见的机器学习工具,包括 MATLAB 和基于 Python 编程的 scikit-learn.此外,本文作者所在的实验室还自主研发了材料数据挖掘在线计算平台 OCPMDM(online computation platform for materials data mining),旨在满足材料化学领域同行的计算需求.这些机器学习工具在材料科学领域的应用发展迅速,为材料研究人员提供了强大的工具和资源,加速了新材料的发现和优化过程.

表 2 若干机器学习工具的比较

Tab. 2 The comparison of several machine learning tools

工具对比	简介	网址
MATLAB	MATLAB 是一种高性能的编程语言和环境,主要用于数值计算、数据分析和算法开发.	https://www.mathworks.com/[10]
Weka	强大的开源机器学习软件工具,提供丰富的数据挖掘算法和用户友好的界面,用于探索、分析和建模数据.	https://cs.waikato.ac.nz/ML/weka[11]
OCPMDM	拥有多种机器学习算法,可用于各种领域的数据分析、数据挖掘.	http://materials-datamining.com/ocpmdm[12]
TensorFlow	TensorFlow 是由 Google 开发的开源深度学习框架,用于构建和训练各种机器学习模型,支持灵活的模型定义和高效的部署.	https://tensorflow.google.cn/[13]
scikit-learn	scikit-learn 是一个开源的 Python 机器学习库,提供了丰富而强大的工具和算法,用于数据挖掘、数据分析和机器学习模型的构建与评估.	http://scikitlearn.org[14]

2 机器学习应用在材料领域的一般流程

机器学习在材料领域的应用旨在快速高效地发现已有材料中目标性能与其影响因素之间的关系,即机器学习模型,用以指导新材料的发现和材料性能的优化.材料目标性能变量在机器学习建模中也可称为目标变量或因变量.材料目标性能的影响因素也可称为自变量、特征变量或描述符,包括元素组成、结构信息和实验条件等.材料领域中机器学习方法应用的一般流程包括数据准备、特征工程、模型筛选(选择和评估),以及模型应用等步骤,如图 2 所示.

2.1 数据准备

数据集的准备是机器学习中不可或缺的关键步骤.数据收集方式主要包括实验室积累的实验数据、期刊文献中的可利用数据以及可用的数据库信息.实验室数据具有较高的可信度,但直接通过实验获得数据可能成本较高.期刊文献整理的数据需要经过筛选和清洗,收集过程可能耗时且数据一致性不够理想.现有各种数据库提供大量数据,但可靠的数据需从权威数据库获取.材料科学领域常用的数据库整理在表 3 中.模型构建效果与数据集质量密切相关.精准可靠的数据集有助于揭示目标变量与自变量之间的潜在关系,而粗糙误差大的数据集则难以实现良好预测效果.为避免“垃圾进,垃圾出”,必须对数据进行清洗处理,包括处理包

含缺失值、异常值或误差较大数据.保证数据集质量对于机器学习至关重要,细致准备和清洗为后续模型构建和预测奠定坚实基础.

2.2　特征工程

特征变量是与材料性质相关的自变量.特征工程是从原始数据中创建特征变量的过程,包括特征变量构建和特征变量选择两个部分.特征变量通常来自实验参数和通过计算得到的固定的数字指纹信息,包括原子参数和分子参数等.常用的 Python 工具包如 DScribe[27]、ASE[28]和 RDKit[29]用于获取原子指纹信息.特征变量过多可能导致模型过拟合,影响预测的准确性.因此,特征变量筛选是特征工程的关键步骤,旨在保留与目标变量最相关的特征变量,其数量应小于数据样本数量的三分之一.机器学习模型所用的特征变量应准确描述材料特性,对目标变量敏感且易获得,以避免不必要的计算成本和时间消耗.精心构建和筛选特征变量有助于提高材料机器学习模型(目标性能与其特征变量之间关系)的理解,提高模型预测性能和应用效果.在材料科学领域,特征工程的重要性不可忽视,为机器学习的成功应用奠定坚实基础,并在新材料的设计和发现中发挥关键作用[30].

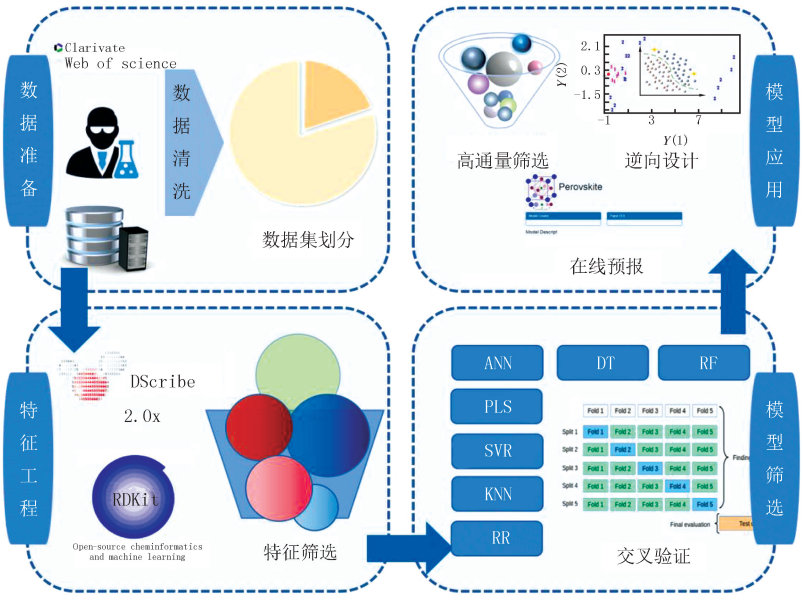


图2　机器学习方法在材料科学领域应用的一般流程示意图

Fig. 2　The general workflow diagram of the application of machine learning methods in the field of materials science

2.3　模型筛选

材料机器学习算法种类繁多,根据数据集是否具有标签,即目标变量,机器学习模型分为监督学习、半监督学习和无监督学习.在材料科学中,常用的是监督学习模型,用于回归问题和分类问题,分别对应定量问题和定性问题.在模型选择时,需综合考虑问题特点和数据性质,并根据实际情况进行选择.常用的模型评价指标包括均方根误差(RMSE)、平均绝对误差(MAE)、决定系数(R^2)、皮尔逊相关系数(R)等用于回归问题,以及准确率(accuracy)、精确率(precision)、召回率(recall)、F1 分数、ROC 曲线和 AUC 曲线等用于分类问题.选择合适的模型和评价指标对机器学习的成功应用至关重要,需综合考虑问题特点、数据性质和模型性能指标.

在材料科学领域,模型的性能评估是衡量其泛化能力的关键指标之一.一般而言,出色的模型应当表现出较小的训练误差和预测误差.为了评估模型的性能,通常会采用 3 种主要方法:独立测试集法、交叉验证法以及自助法.独立测试集法通过将原始数据集划分为训练集和测试集,其中训练集用于模型的训练,而测试集用于验证模型的泛化能力.通过比较模型在测试集上的误差,可以评估模型的泛化能力,误差越小则表明模型的泛化能力越强.在材料科学领域,由于数据获取困难,常见的训练集和测试集划分比例为 8∶2 或 7∶3.不过,在充分训练模型之后,可以根据实际情况调整划分比例,以充分利用有限的数据集.

交叉验证法将数据集划分为 k 个互斥子集,每个子集依次作为验证集,其余子集作为训练集,经过 k 次迭代后,将每次训练的结果取平均作为交叉验证的最终结果.常用的 k 值为 5、10 或 20,不同的 k 值可能会导致略微不同的结果,但差异通常不会太大.当 k 的值等于数据集大小时,称为留一法交叉验证(LOOCV),尽管其结果准确率最高,但计算量较大,尤其是在数据量较大时会消耗大量时间[31].自助法适用于数据较少的情况,它通过有放回地均匀抽样,从已有的训练集中随机抽取样本并放回到训练集中,重复此过程直到抽取的样本数等于原始训练集的样本数.通过得到多个模型的训练结果,并对所有模型求取平均值来评估性能.

然而,自助法引入了重复样本,改变了数据的原有分布,可能会引入估计偏差,因此在使用自助法时需要谨慎考虑.

综合考虑这些评估方法的优缺点和实际情况,进而选择适合的模型评估方法来准确评估机器学习模型的性能,并优化其泛化能力,从而为材料科学领域的研究和应用提供可靠的指导[32].

表 3 常用的材料数据库

Tab. 3 The common material databases

数据库	简介	网址
Materials Project(MP)	已知和假设材料的计算数据	https://materialsproject.org[15]
The Inorganic Crystal Structure Database (ICSD)	无机晶体结构的实验表征数据	https://icsd.fiz-karlsruhe.de/index.xhtml[16]
Cambridge Structural Database(CSD)	由剑桥晶体学数据中心收集的基于 X 射线和中子衍射实验的小分子和金属有机分子晶体的结构数据库	https://www.ccdc.cam.ac.uk/[17]
Aflow-Automatic-FLOW for Materials Discovery(AFLOW)	高通量从头计算的无机材料的结构和性质的数据存储库	http://www.aflowlib.org[18]
Crystallography Open Database(COD)	有机、无机和金属有机化合物和矿物的结构数据	http://cod.ensicaen.fr[19]
Open Quantum Materials Database (OQMD)	主要是假设材料的理论模拟计算数据	http://www.oqmd.org/[20]
Springer Materials	世界上最大的材料数据资源	https://materials.springer.com[21]
GDB Databases	有机分子数据库	http://gdb.unibe.ch/downloads[22]
ZINC	二维和三维形式的有机分子	https://zinc15.docking.org/[23]
Materiae	拓扑材料数据库	http://materiae.iphy.ac.cn/
Materials Cloud	二维材料的结构计算数据	https://www.materialscloud.org/discover/2dstructures/dashboard/ptable[24]
The Perovskite Database Project(PDP)	钙钛矿材料数据库	https://www.perovskitedatabase.com[25]
Materials Project(MP)	材料属性数据库	https://materialsproject.org/[15]
Materials Platform for Data Science(MPDS)	无机材料数据库	https://mpds.io/#modal/menu[26]

2.4 模型应用

建立模型的主要目的在于探索目标变量与特征变量之间的内秉关系,从而指导新材料设计乃至生产应用.机器学习模型在材料设计中的应用包括正向设计和逆向设计.其中,高通量筛选是一种常见的基于机器学习的材料正向设计应用方法.该方法首先根据特征变量的分布区间和变化步长,构建大量的虚拟样本,然后利用所建的机器学习模型预测所有虚拟样本的属性或性能,从中筛选出所需属性或性能最佳的虚拟样本,供实验合成参考[33].尽管这种方法在特征较多的情形下可能涉及较大的计算量,但它为新材料探索提供了行之有效的筛选方法[34].除了高通量筛选方法,笔者所在实验室还开发了独特的基于机器学习模型的材料逆向设计方法,包括定性的模式识别逆投影方法[35-36]和定量的主动渐进式搜索方法[37].模式识别逆投影方法是一种基于模式识别技术的逆向搜索方法,该方法首先通过本课题组研发的模式识别最佳投影技术获得材料不同属性(如性能优类或劣类)样本在计算机屏幕上可视的最佳分类图,然后在优类样本分布区或其变化趋势上选取“虚拟样本”,进而根据机器学习模型和优类样本的边界条件求解出“虚拟样本”对应的特征变量,为发现性能更好的新材料提供指导.主动渐进式搜索方法是一种利用机器学习模型结合优化模型的逆向搜索方法,该方法首先利用已知的实验样本构建机器学习模型,然后通过虚拟样本构建的优化模型主动选择下一步最有价值的实验样本,优化模型经过不断迭代逐步推荐性能逼近需求的虚拟样本,直至机器学习模型预测的虚拟样本的材料性能定量满足预先设定的需求[38].

3　机器学习在材料科学方面的应用

自从机器学习技术应用于材料领域以来,利用机器学习方法辅助新材料设计和性能优化的研究逐年增加,该方法已广泛应用于金属材料、钙钛矿材料、陶瓷材料、高分子材料和纳米材料等领域.在金属材料方面,机器学习被用于预测金属合金的性质、相图和晶体结构,有助于设计具有特定性能的金属材料.此外,机器学习还用于优化金属的机械性能、耐腐蚀性和热稳定性等方面.针对具有优异光电性能的钙钛矿材料,机器学习在钙钛矿太阳能电池的设计和优化中发挥关键作用,帮助发现新型的钙钛矿材料并提高光电转换效率.在陶瓷材料方面,机器学习用于预测陶瓷材料的机械性能、热性能和导热性能等关键性能参数,从而优化陶瓷材料的性能和稳定性.机器学习在高分子材料的设计和合成中具有重要应用,可以预测高分子材料的力学性能、热性能和电学性能等,为高分子材料的功能化设计提供指导.针对具有独特尺寸效应和表面效应的纳米材料,机器学习用以预测纳米材料的性能和稳定性,帮助优化纳米材料的应用性能[39-40].综上所述,机器学习技术在材料领域的应用范围广泛且不断扩大,为新材料设计和性能优化提供了有力的工具和方法,有望推动材料科学领域的进一步发展与创新.

3.1　机器学习在金属材料方面的应用

近年来,机器学习在金属材料领域的应用广泛而深入,涵盖了高熵合金、铝合金、铜合金、镁合金等多个合金类型[41-42].这些不同类型的合金具备各自独特的特性,适用于不同的应用场景.

高熵合金由 5 种或更多种合金元素组成,其成分比例通常在 5%到 35%之间,因系统熵值较高而被冠以“高熵”之名.近年来,高熵合金因其卓越的耐腐蚀性、高硬度以及热稳定性,引起了广泛关注[43].在这一领域,RAO 等学者[44]采用主动学习机器学习策略成功实现了高熵因瓦合金的设计(图 3).研究过程首先涉及使用生成模型对 699 个原始数据集进行训练,然后结合无监督学习和随机采样方法,生成了 1 000 个初步筛选出的候选样本.随后,他们应用多层感知机和梯度提升决策树集成模型,在仅包含原子特征的数据集基础上,对候选样本进行进一步筛选,最终选择排名前 10～30 的候选样本.随后,通过整合包含 DFT 计算和分子动力学模拟特征的模型,再次对新生成的候选样本进行筛选,最终选择出排名前 3 的样本进行实验合成.经过 6 轮迭代,研究团队成功确定了两种具有极低热膨胀系数的高熵因瓦合金.

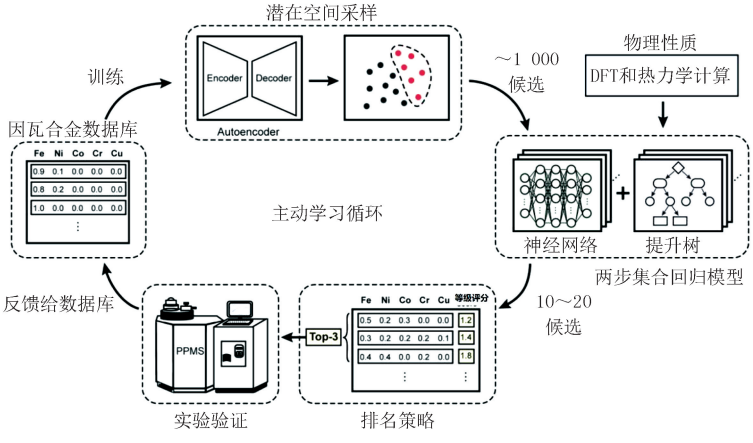


图3　主动学习机器学习策略发现高熵因瓦合金流程示意图[44]

Fig.3　Active learning machine learning strategy for discovering high-entropy ceramic alloys process flow diagram[44]

另一方面,本课题组[35]提出了一个综合的机器学习合金设计系统(图 4),该系统涵盖数据库建立、模型构建、成分优化和实验验证,旨在指导高硬度高熵合金的理性设计.该研究过程基于一个包含 370 条铸态高熵合金成分和维氏硬度数据的数据库,通过特征工程提取了 142 个特征.随后,通过 4 步法特征筛选,确定了影响高熵合金硬度的 5 个关键特征.利用支持向量机和这 5 个关键特征构建了高熵合金硬度预测模型,该模型在测试集和 LOOCV 中的相关系数均达到了 0.94.接下来,通过模式识别逆投影方法和高通量筛选方法,成功设计出了 3 个新的候选样本.其中一个候选样本的硬度值比原始数据集中的最高硬度合金提升了 24.8%.这项研究在一定程度上验证了机器学习在辅助高熵合金成分设计方面的可行性,为设计高硬度高熵合金提供了有力的指导.

铝合金是一类以铝元素为基础,并添加了一定量其他元素的合金.这种合金具有出色的导电性、高导热

性和强大的焊接能力,使其在航空航天、交通运输、机电工程和日用品等多个领域得到广泛应用.尽管铝合金在各行各业中发挥着重要作用,但在选择具有期望的疲劳性能和其他机械性能的铝合金材料方面仍然存在一些挑战.因此,合成具有特定疲劳性能的铝合金类型对于各种工程应用都至关重要.

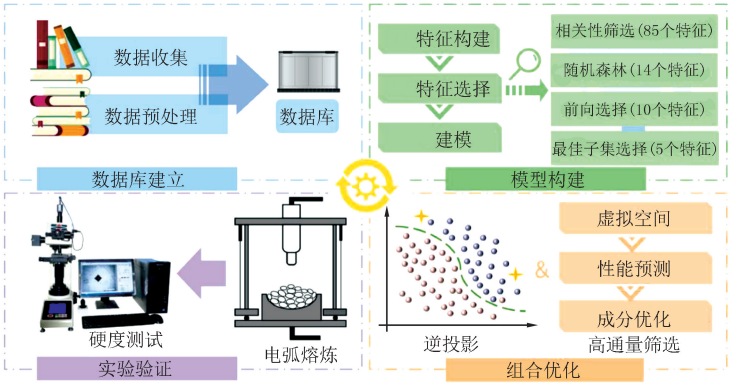


图4 机器学习结合模式识别突破高熵合金硬度示意图[35]

Fig. 4 Breakthrough in high-entropy alloy hardness through integration of machine learning and pattern recognition-diagram[35]

本课题组[45]提出了一种基于领域知识的支持向量机模型,用于预测不同系列铝合金的疲劳寿命.通过引入领域知识描述符,显著提升了支持向量机模型的性能.所设计的特征与目标性能之间具有高相关系数,相较于未使用领域知识的模型,该模型的预测能力显著提高.利用该模型成功预测了 7 种铝合金的疲劳寿命,突显了特征选择在材料机器学习建模中的重要性.同时,FATRIANSYAH 等[46]使用 344 条数据建立了两个人工神经网络模型,以探索铝合金疲劳强度与其他性能之间的相互影响.第一个模型以原子成分参数和性能参数为描述符,预测疲劳寿命;第二个模型以性能参数和疲劳寿命为输入,预测原子成分参数,实现了第一个模型的逆过程.这两个模型经过训练后,相关系数 R^2 均达到了 0.9 以上,成功地设计了基于疲劳性能预测原子成分的反演模型,可用于指导特定疲劳寿命下的铝合金合成.此外,探索铝合金缓蚀剂的结构-性能关系对于腐蚀防护技术具有重要意义.这些研究为铝合金材料设计和性能优化提供了有力的工具和方法.

GALVÄO 等[47]运用了 KNN、DT、ANN、SVM 和 RF 等多种机器学习模型,探索有机缓蚀剂的潜在影响.在这些模型的比较中,他们发现 RF 模型表现最佳,为有机缓蚀剂的研究提供了有力的指导.通过深入研究,他们鉴定出了二聚化熵是保护机制中的关键因素,为有机缓蚀剂的研究提供了指导性的认识.这项研究对于深入理解有机缓蚀剂的影响机制以及优化其性能具有重要的意义.

RIVERA 等[48]针对激光焊接中铝合金产生的气孔等问题进行了研究.为解决这一问题,他们利用了 SVM、RF、CatBoost 和 ANN 等机器学习算法构建模型,并通过高速 X 射线分析进行特征提取和选择,同时进行了样本不平衡处理.通过对模型的准确性、AUC 和 F1 等指标的评估,研究确定了在基于高速摄像机监测的情况下解决气孔问题的最佳机器学习算法为随机森林,其准确度达到了 75%,AUC 值为 0.83.无气孔的 F1 得分为 0.75,有气孔的 F1 得分为 0.76.该研究结果表明,所提出的模型和方法在工业应用中具有实施的可行性.这些方法有望提升最终产品的质量,减少工艺浪费和产品质量分析的时间,同时增加操作性能.这对于解决激光焊接过程中的孔隙问题,提高生产效率和产品质量具有实际应用的潜力.

除了高熵合金和铝合金,机器学习方法在预测其他合金性能方面也取得了广泛应用.本课题组[49]利用岭回归(RR)、支持向量机回归(SVR)和 XGBoost 模型的集成方法成功设计了熔点分别为 16 ℃和 90 ℃的低熔点合金,并通过实验验证了这些设计.同时,提出了一种机器学习结合主动渐进式搜索方法的逆向设计策略,从海量的虚拟空间中成功设计出了熔点在 11 ℃可用于低温润滑剂、冷却剂和熔点在 70 ℃可用于消防设施启动零件的低熔点合金.后通过实验验证,6 个实验样本与预报值平均相对误差小于 8%[50].此外,HOU 等[51]运用了 RR、SVR、梯度提升决策树(GBDT)、RF、CatBoost 和高斯过程回归(GPR)等 6 个模型,构建了针对镁合金抗拉强度、屈服强度和延伸率的预测模型.通过模型的集成,他们进一步提升了模型的准确性,为预测生物医用镁合金的力学性能开辟了新途径.WANG 等[52]提出了一种机器学习设计系统,该系统建立了两种反向传播神经网络模型,分别实现了从组成到性能以及从性能到组成的映射关系.这一系统成功地用于高性能复杂铜合金的成分设计.这些研究展示了机器学习在合金材料领域的应用潜力,为合金设计和优化提供了有力的支持.

3.2　机器学习在钙钛矿方面的应用

钙钛矿材料因其高光吸收系数、大载流子迁移率和简便的合成方法备受瞩目,被认为是未来最有潜力的光电材料之一[53].这种材料在太阳能电池、光电检测器、发光二极管、光催化等众多领域展现出广泛的应用前景,还有望应用于柔性光电子器件等领域[54].

ZHAO 等[55]通过使用高通量机器人系统和机器学习方法,致力于解决钙钛矿材料稳定性的问题.他们研究了有机铅碘钙钛矿电池的最佳稳定性条件.通过机器人系统在不同的材料成分、实验条件和测试条件下合成了 1 400 多个钙钛矿电池样本,以电池性能衰减时长作为衡量电池稳定性的标准.将材料成分、实验条件和测试条件作为自变量,电池性能衰减时长作为目标值,他们构建了梯度提升树模型.该模型在测试集上的 RMSE 为 169,明显优于其他模型的 RMSE 值.通过变量分析,他们得出了最佳实验条件和最佳有机铅碘钙钛矿材料配比成分(MA0.1Cs0.05FA0.85PbI3).通过这一研究,其成功地合成了性能衰减时长超过 4 000 h 的钙钛矿电池,远超过大多数已报道的电池装置.这项工作通过高通量实验和机器学习分析,揭示了最佳实验条件对电池性能衰减的影响,有效地推动了钙钛矿电池稳定性研究的进展.

在有机无机杂化钙钛矿(HOIP)材料的后处理胺应用方面,本课题组[56]利用 4 种机器学习方法构建了 HOIPs 后处理胺类别("反应性胺"和"非反应性胺")的预测模型.课题组收集了 55 个后处理胺的数据样本,并利用 Dragon7 生成了 5 270 个 HOIPs 分子描述符.通过特征变量筛选,选出了 7 个关键的特征变量,然后通过 4 种模型的性能比较,表明支持向量分类(SVC)模型预测 HOIPs 后处理胺类别的正确率和稳定性较好.利用所建 SVC 模型结合采用 SHAP 方法可以对模型所用主要特征变量进行解释分析,SVC 模型为加快 HOIPs 后处理胺的筛选提供了有效参考.

本课题组[57]还构建了 HOIPs 材料的形成性和带隙的机器学习模型,预测了具有合适带隙的无铅 HOIPs 新材料.利用类别提升算法建立了 HOIPs 材料形成性判别模型,独立测试集的预报准确率达到了 95.5%.预测出 3 个新的 Sn-Ge 混合体系的 HOIPs 材料,与实验结果相符.构建了基于 4 个树算法的 HOIPs 带隙预测的加权投票回归模型,结合自主研发的主动渐进搜索方法,预测出指定带隙值且无铅的 HOIPs 新材料,为 HOIPs 材料的逆向设计提供了快速搜索的新方法和相应的 Python 开源软件.

本课题组[58]还采用了基于机器学习的逐步设计策略,研究了多目标下的钙钛矿在催化产氢速率方面的性能.首先,分别建立了比表面积(SSA)、带隙(Eg)、微晶尺寸(CS)3 个机器学习模型.然后,在容忍因子为 0.8～1.0 的条件下,通过高通量筛选,筛选出了 5 368 个候选样本.接着,利用 Eg 模型筛选出了在 1.4～2.6 eV范围内的样本.最后,通过比表面积模型和微晶尺寸模型,筛选出具有高 SSA 和小 CS 的样本,最终选择出35 种钙钛矿候选物.随后收集了 80 个具有催化产氢速率样本的数据,建立了催化产氢回归模型,并从 8 种模型中筛选出 Gradient Boosting Regression(GBR)模型.利用 GBR 模型对之前筛选出的 35 个候选物进行预测,结果显示这些候选物具有高的催化产氢速率,表明筛选出的候选物在光催化方面可能具有令人满意的性能.此外,将相关模型开发成在线预测程序,以方便研究人员的使用.这种逐步设计策略不仅为钙钛矿材料在催化应用中提供了指导,还提供了方便的工具用于在线预测.

本课题组[59]还构建了 ABO_3 型钙钛矿的铁电性分类模型,以及分别用于预测钙钛矿材料的比表面积 SSA、Eg、居里温度 Tc 和介电损耗(tan δ)的机器学习模型.根据 ABO_3 型钙钛矿 3 种不同的应用场景,设计了适用于催化领域的钙钛矿材料,具有特定 Eg(0.9 eV≤Eg≤1.7 eV)和 SSA(>18.35 m^2/g).同时还设计了适用于铁电半导体的材料,具有特定 Eg(0.5 eV≤Eg≤2.5 eV)、居里温度(>278 K)和介电损耗(tan δ<0.04).此外还设计了适用于水分解的材料,具有特定 Eg(1.6 eV≤Eg≤2.4 eV)和 SSA(>18.35 m^2/g).通过这种多目标设计方法,成功筛选出了候选样本,为定制合成具有目标性质的特定材料提供了新思路.这一研究方法对于材料科学领域的定制合成和设计具有一定的推动作用.

综上所述,机器学习在钙钛矿材料的不同性质研究中展现出强大的应用潜力,无论是催化性能、铁电性质还是其他特性,都为材料科学研究提供了新的解决方案.

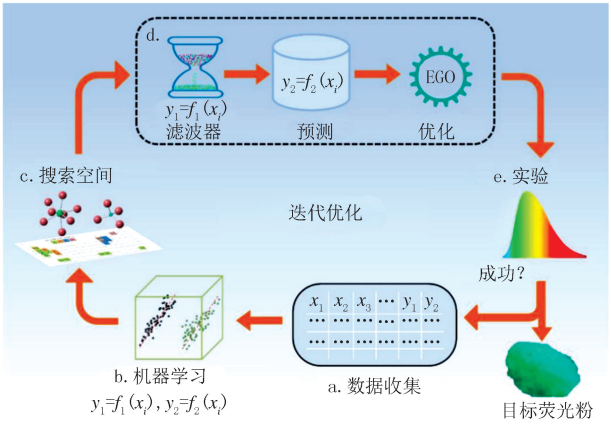
3.3　机器学习在其他材料方面的应用

WU 等[60]通过迁移学习和贝叶斯分子设计算法成功建立了高精度的聚合物热导率预测模型.其利用 PolyInfo 和 QM9 中的聚合物样本数据建立了预训练模型,并通过对比模型性能选择了具有最高预测精度

的热容 C_v 预训练模型.然后,利用 28 个实验测得的聚合物热导率数据调整预训练模型参数,使其成功迁移至热导率的预测,实现了较高的预测精度,MAE 为 0.024 W/mK,比直接使用 28 个数据点进行训练的模型低 40%.作者采用贝叶斯算法设计了大量具有重复单元结构的分子,并通过筛选最终确定了24 个分子结构.其中,经过实验验证的 3 个聚合物样本表现出较高的热导率,且高于已发表论文中的聚合物材料.这项研究证实了迁移学习和贝叶斯分子设计在聚合物材料设计和发现中的成功应用.

文献[61]报道采用机器学习、密度泛函和实验相结合的方法,成功研发了一种高效的电催化 CO_2 还原制乙烯的 Cu—Al 催化剂.首先,通过对 Materials Project 数据库中 244 种铜基金属晶体可能吸附位点的子集进行密度泛函理论模拟,计算 CO 吸附能,以评估催化活性.采用原子序数、电负性、原子数和单金属吸附能等特征,利用随机森林回归算法和时间序列划分方法建立了机器学习模型,预测各位点的 CO 吸附能,模型平均绝对值误差为 0.18 eV 和 0.29 eV.将机器学习模型与密度泛函相结合,通过主动学习找到接近理想吸附能的吸附位点.研究发现,在 Cu(100)表面,Al 含量为 4%～12%时具有更多接近理想吸附能的吸附位点.最终,实验合成了 Al 含量为 10%的 Cu—Al 电催化剂,其在 400 mA/cm^2 电流密度下的法拉第效率超过 80%,电流密度为 150 mA/cm^2 时阴极乙烯的能量转换效率达到 55%.该研究展示了材料基因理念在多金属电催化剂探索中的重要性.

北京科技大学宿彦京教授课题组[62]通过主动学习的多目标机器学习方法,经过 5 次理论指导与实验验证的循环迭代,从 171 636 种石榴石结构的化合物中筛选出了 25 个具有目标波长(480～510 nm)和优异热稳定性的候选材料.经制备和表征,$Lu_{1.5}Sr_{1.5}Al_{3.5}Si_{1.5}O_{12}$ ∶Ce 表现最佳,具有出色的热稳定性(在 640 K 时保持≥60%的发射强度),发射峰约为 505 nm,是非常有应用前景的荧光粉.研究流程如图 5 所示,首先从文献中筛选得到 70 个 Ce^{3+} 掺杂石榴石荧光粉样本,构建数据集,并通过特征筛选得到 47 个特征.利用 XGBoost、SVR 等算法和前进后退特征筛选方法将特征空间缩减至 10 维以内,其中 KRR(核岭回归)模型在波长和 $T60$ 的预测中表现最佳.作者根据已知样本的化学式构造了包含 171 636 种化合物的未知搜索空间,并在模型的指导下通过 5 次迭代成功探索得到了 $T60$ 超过 640 K、发光波长合适的绿色荧光粉材料 $Lu_{1.5}Sr_{1.5}Al_{3.5}Si_{1.5}O_{12}$ ∶Ce,其性能超过了已知数据集中的最大 $T60$(633 K).该研究充分体现了机器学习在合理设计石榴石荧光粉方面的实际应用价值.



(a)从文献和特征构建中收集数据;(b)特征选择和ML建模;(c)未知的石榴石结构的样本集,定义为待搜索空间;(d)通过自适应实验设计需要合成和表征的最佳.

图5 ML辅助的荧光粉设计示意图

Fig.5 ML-assisted phosphor design illustration

西安交通大学开发了一种快速预测未知多组分材料相图的方法[63],并成功应用于 $BaTiO_3$ 基铁电陶瓷和 NiTi 基形状记忆合金体系.流程包括建立陶瓷和合金数据集,使用泛克里金算法构建机器学习模型,模拟化合物的相图.模型在陶瓷体系中表现良好,相变温度的误差较小.在合金体系中,误差比率分别为 5.67%和 20.60%.通过该模型成功预测了未知体系的相图,经过少于 3 轮的理论预测与实验验证的循环迭代后,模型预测的不确定性大幅降低.该研究为快速预测未知多组分材料相图提供了基础.

4 展望

本文详细探讨了机器学习技术在材料科学中的应用.材料科学对于推动人类文明进步至关重要,但传统研究方法在复杂材料设计和性能预测上有较大的局限性.机器学习的崛起为材料科学带来了新的机遇,通过数据驱动的机器学习模型,实现了高通量筛选、性能预测、晶体结构预测和材料配方优化等.机器学习与第一

性原理计算和实验验证相结合,必将在加速材料创新、降低试错成本、定制化材料设计方面发挥越来越重要的作用.

随着机器学习技术不断发展,其在材料科学领域的前景十分广阔.可以预见机器学习在材料设计方面的进一步深化,尤其是在预测新材料性能、结构和稳定性方面的应用将得到加强.借助机器学习,科研人员将能够更快速地发现具有特定性能的材料,并为不同领域的需求定制化设计材料,从而加速技术发展和应用创新.然而,机器学习在材料科学中也面临挑战.数据的质量和数量、模型的解释性以及可靠性等问题仍然需要深入研究和解决.此外,机器学习模型的泛化能力以及对于稀有事件的处理也需要进一步优化.在应对这些挑战的过程中,跨学科的合作将会变得更加重要,更好地融合材料科学、计算机科学和统计学等领域的知识,将推动机器学习在材料科学中的应用取得更大突破.

总之,机器学习为材料科学带来了前所未有的机遇,不仅加速了材料创新,还为解决全球性问题提供了新的思路.随着技术不断发展和深入,我们有理由期待机器学习将在材料科学中持续发挥重要作用,为科技和人类社会的发展做出更大贡献.

参　考　文　献

[1] 侯腾跃,孙炎辉,孙舒鹏,等.机器学习在材料结构与性能预测中的应用综述[J].材料导报,2022,36(6):165-176.
HOU T Y,SUN Y H,SUN S P,et al.A review of the application of machine learning in material structure and performance prediction[J].Materials Reports,2022,36(6):165-176.

[2] KUMAR J N,LI Q X,TANG K Y T,et al.Machine learning enables polymer cloud-point engineering via inverse design[J].NPJ Computational Materials,2019,5:73.

[3] XU P C,JI X B,LI M J,et al.Small data machine learning in materials science[J].NPJ Computational Materials,2023,9:42.

[4] HOLLINGSWORTH S A,DROR R O.Molecular dynamics simulation for all[J].Neuron,2018,99(6):1129-1143.

[5] SCHLEDER G R,PADILHA A C M,ACOSTA C M,et al.From DFT to machine learning:recent approaches to materials science-a review[J].Journal of Physics:Materials,2019,2(3):032001.

[6] 王海伟,叶波,冯晶,等.机器学习在钢铁材料研究中的应用综述[J].中国材料进展,2023,42(10):806-813.
WANG H W,YE B,FENG J,et al.Application of machine learning in steel materials:a survey[J].Materials China,2023,42(10):806-813.

[7] FABER F A,HUTCHISON L,HUANG B,et al.Prediction errors of molecular machine learning models lower than hybrid DFT error[J].Journal of Chemical Theory and Computation,2017,13(11):5255-5264.

[8] WEI J,CHU X,SUN X Y,et al.Machine learning in materials science[J].InfoMat,2019,1(3):338-358.

[9] MAHESH B.Machine learning algorithms-a review[J].International Journal of Science and Research(IJSR)[Internet],2020,9(1):381-386.

[10] KIM P.Matlab deep learning[M].CA:Apress Berkeley,2017.

[11] BRAMLEY G N.A small predator removal experiment to protect north island weka (Gallirallus australis greyi) and the case for single-subject approaches in determining agents of decline[J].New Zealand Journal of Ecology,1996,20(1):37-43.

[12] ZHANG Q,CHANG D P,ZHAI X Y,et al.OCPMDM:online computation platform for materials data mining[J].Chemometrics and Intelligent Laboratory Systems,2018,177:26-34.

[13] PANG B,NIJKAMP E,WU Y N.Deep learning with TensorFlow:a review[J].Journal of Educational and Behavioral Statistics,2020,45(2):227-248.

[14] PEDREGOSA F,VAROQUAUX G,GRAMFORT A,et al.Scikit-learn:Machine learning in Python[J].The Journal of Machine Learning Research,2011,12:2825-2830.

[15] JAIN A,ONG S P,HAUTIER G,et al.Commentary:the Materials Project:a materials genome approach to accelerating materials innovation[J].APL Materials,2013,1(1):011002.

[16] HELLENBRANDT M.The inorganic crystal structure database (ICSD):present and future[J].Crystallography Reviews,2004,10(1):17-22.

[17] GROOM C R,BRUNO I J,LIGHTFOOT M P,et al.The Cambridge structural database[J].Acta Crystallographica Section B,Structural Science,Crystal Engineering and Materials,2016,72(Pt 2):171-179.

[18] CURTAROLO S,SETYAWAN W,HART G L W,et al.AFLOW:an automatic framework for high-throughput materials discovery[J].Computational Materials Science,2012,58:218-226.

[19] GRAŽULIS S,CHATEIGNER D,DOWNS R T,et al.Crystallography Open Database-an open-access collection of crystal structures[J].Journal of Applied Crystallography,2009,42(Pt 4):726-729.

[20] SAAL J E,KIRKLIN S,AYKOL M,et al.Materials design and discovery with high-throughput density functional theory:the open quantum materials database (OQMD)[J].JOM,2013,65(11):1501-1509.

[21] KIM S,CHEN J,CHENG T J,et al.PubChem in 2021:new data content and improved web interfaces[J].Nucleic Acids Research,2021,49(D1):D1388-D1395.

[22] RUDDIGKEIT L,VAN DEURSEN R,BLUM L C,et al.Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17[J].Journal of Chemical Information and Modeling,2012,52(11):2864-2875.

[23] IRWIN J J,SHOICHET B K.ZINC:a free database of commercially available compounds for virtual screening[J].Journal of Chemical Information and Modeling,2005,45(1):177-182.

[24] TALIRZ L,KUMBHAR S,PASSARO E,et al.Materials Cloud,a platform for open computational science[J].Scientific Data,2020,7:299.

[25] CHENG Z Y,LIN J.Layered organic-inorganic hybrid perovskites:structure,optical properties,film preparation,patterning and templating engineering[J].CrystEngComm,2010,12(10):2646-2662.

[26] WARD L,DUNN A,FAGHANINIA A,et al.Matminer:an open source toolkit for materials data mining[J].Computational Materials Science,2018,152:60-69.

[27] HIMANEN L,JÄGER M O J,MOROOKA E V,et al.DScribe:library of descriptors for machine learning in materials science[J].Computer Physics Communications,2020,247:106949.

[28] HJORTH LARSEN A,JØRGEN MORTENSEN J,BLOMQVIST J,et al.The atomic simulation environment-a Python library for working with atoms[J].Journal of Physics Condensed Matter,2017,29(27):273002.

[29] LANDRUM G.RDKit:A software suite for cheminformatics,computational chemistry,and predictive modeling[J].Greg Landrum,2013,8:31.

[30] 赵娟娟,叶顺,徐可,等.基于提取不同中红外光谱特征信息的烟叶部位判别研究[J].河南师范大学学报(自然科学版),2021,49(1):45-49.
ZHAO J J,YE S,XU K,et al.Research on discrimination of tobacco leaf parts based on extracting different information of MIR[J].Journal of Henan Normal University(Natural Science Edition),2021,49(1):45-49.

[31] HE J B,FAN X T.Evaluating the performance of the K-fold cross-validation approach for model selection in growth mixture modeling[J].Structural Equation Modeling:A Multidisciplinary Journal,2019,26(1):66-79.

[32] RASCHKA S.Model evaluation,model selection,and algorithm selection in machine learning[J].ArXiv e-Prints,2018:arXiv:1811.12808.

[33] 徐荣幸,赵鸿.机器学习在座逾渗相变问题中的应用[J].河南师范大学学报(自然科学版),2019,47(1):45-51.
XU R X,ZHAO H.Study site percolation phase transitions based on machine learning[J].Journal of Henan Normal University(Natural Science Edition),2019,47(1):45-51.

[34] NANDY A,DUAN C R,TAYLOR M G,et al.Computational discovery of transition-metal complexes:from high-throughput screening to machine learning[J].Chemical Reviews,2021,121(16):9927-10000.

[35] YANG C,REN C,JIA Y F,et al.A machine learning-based alloy design system to facilitate the rational design of high entropy alloys with enhanced hardness[J].Acta Materialia,2022,222:117431.

[36] ZHOU X,ZHENG Z,LU T,et al.Interpretable machine learning assisted multi-objective optimization design for small molecule hole transport materials[J].Journal of Alloys and Compounds,2023:171440.

[37] LU T,LI H Y,LI M J,et al.Inverse design of hybrid organic-inorganic perovskites with suitable bandgaps via proactive searching progress[J].ACS Omega,2022,7(25):21583-21594.

[38] 申俊丽,余堃.通过 PET-CT 图像纹理特征预测软组织肉瘤转移性[J].河南师范大学学报(自然科学版),2021,49(2):25-30.
SHEN J L,YU K.Prediction of soft tissue sarcoma metastasis by PET-CT image texture features[J].Journal of Henan Normal University(Natural Science Edition),2021,49(2):25-30.

[39] GAO C,MIN X,FANG M,et al.Innovative materials science via machine learning[J]. Advanced Functional Materials,2021,32(1):2108044.

[40] FU Z Y,LIU W Y,HUANG C,et al.A review of performance prediction based on machine learning in materials science[J].Nanomaterials,2022,12(17):2957.

[41] HART G L W,MUELLER T,TOHER C,et al.Machine learning for alloys[J].Nature Reviews Materials,2021,6:730-755.

[42] LIU X J,XU P C,ZHAO J J,et al.Material machine learning for alloys:applications,challenges and perspectives[J].Journal of Alloys and Compounds,2022,921:165984.

[43] GEORGE E P,RAABE D,RITCHIE R O.High-entropy alloys[J].Nature Reviews Materials,2019,4:515-534.

[44] RAO Z Y,TUNG P Y,XIE R W,et al.Machine learning-enabled high-entropy alloy discovery[J].Science,2022,378(6615):78-85.

[45] LIAN Z H,LI M J,LU W C.Fatigue life prediction of aluminum alloy via knowledge-based machine learning[J].International Journal of Fatigue,2022,157:106716.

[46] FATRIANSYAH J F,RIZQILLAH R K,SUHARIADI I,et al.Composition-based aluminum alloy selection using an artificial neural network[J].Modelling and Simulation in Materials Science and Engineering,2023,31(5):055011.

[47] GALVÄO T L P,NOVELL-LERUTH G,KUZNETSOVA A,et al.Elucidating structure-property relationships in aluminum alloy corrosion inhibitors by machine learning[J].The Journal of Physical Chemistry C,2020,124(10):5624-5635.

[48] RIVERA J S，GAGNé M O，TU S Y，et al.Quality classification model with machine learning for porosity prediction in laser welding aluminum alloys[J].Journal of Laser Applications，2023，35(2)：022011.

[49] CHEN H M，SHANG Z W，LU W C，et al.A property-driven stepwise design strategy for multiple low-melting alloys via machine learning[J].Advanced Engineering Materials，2021，23(12)：2100612.

[50] WU Y M，SHANG Z W，LU T，et al.Target-directed discovery for low melting point alloys via inverse design strategy[J].Journal of Alloys and Compounds，2024，971：172664.

[51] HOU H B，WANG J F，YE L，et al.Prediction of mechanical properties of biomedical magnesium alloys based on ensemble machine learning[J].Materials Letters，2023，348：134605.

[52] WANG C S，FU H D，JIANG L，et al.A property-oriented design strategy for high performance copper alloys via machine learning[J].NPJ Computational Materials，2019，5：87.

[53] 胡扬，张胜利，周文瀚，等.基于机器学习探索钙钛矿材料及其应用[J].硅酸盐学报，2023，51(2)：452-468.
HU Y，ZHANG S L，ZHOU W H，et al.Studies on perovskite material and its applications via machine learning[J].Journal of the Chinese Ceramic Society，2023，51(2)：452-468.

[54] 冯顺.基于机器学习的无机钙钛矿材料形成能预测[J].无线互联科技，2023，20(16)：47-51.
FENG S.Prediction of formation energy of inorganic perovskite materials based on machine learning[J].Wireless Internet Technology，2023，20(16)：47-51.

[55] ZHAO Y C，ZHANG J Y，XU Z W，et al.Discovery of temperature-induced stability reversal in perovskites using high-throughput robotic learning[J].Nature Communications，2021，12：2191.

[56] ZHENG J，LU T，LIAN Z H，et al.Machine learning assisted classification of post-treatment amines for increasing the stability of organic-inorganic hybrid perovskites[J].Materials Today Communications，2023，35：105902.

[57] LU T，LI H Y，LI M J，et al.Predicting experimental formability of hybrid organic-inorganic perovskites via imbalanced learning[J].The Journal of Physical Chemistry Letters，2022，13(13)：3032-3038.

[58] TAO Q L，CHANG D P，LU T，et al.Multiobjective stepwise design strategy-assisted design of high-performance perovskite oxide photocatalysts[J].The Journal of Physical Chemistry C，2021，125(38)：21141-21150.

[59] XU P C，CHANG D P，LU T，et al.Search for ABO(3) type ferroelectric perovskites with targeted multi-properties by machine learning strategies[J].Journal of Chemical Information and Modeling，2022，62(21)：5038-5049.

[60] WU S，KONDO Y，KAKIMOTO M A，et al.Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm[J].NPJ Computational Materials，2019，5：66.

[61] ZHONG M，TRAN K，MIN Y M，et al.Accelerated discovery of CO_2 electrocatalysts using active machine learning[J].Nature，2020，581：178-183.

[62] JIANG L P，JIANG X，ZHANG Y，et al.Multiobjective machine learning-assisted discovery of a novel cyan-green garnet：Ce phosphors with excellent thermal stability[J].ACS Applied Materials & Interfaces，2022，14(13)：15426-15436.

[63] TIAN Y，YUAN R H，XUE D Z，et al.Determining multi-component phase diagrams with desired characteristics using active learning[J].Advanced Science，2020，8(1)：2003165.

Machine learning-based materials design

Lu Wencong[1]，Wu Yanmiao[1]，Liu Taiang[2]，Lu Tian[2]，Ji Xiaobo[1]，Xing Leikai[1]

(1. Department of Chemistry，College of Sciences，Shanghai University，Shanghai 200444，China；
2. Shanghai Shuzhiwei Information Technology Co.，LTD.，Shanghai 200444，China)

Abstract：Material innovation has always been a crucial driver for human civilization，and with the growing demand for high-performance materials in modern technology，the importance of materials science is becoming increasingly evident. However，relying solely on traditional trial-and-error methods and first-principles approaches for complex materials design has significant limitations. Machine learning has evolved into a new paradigm for materials science research，enabling high-throughput screening，performance prediction，crystal structure prediction，and material formulation optimization through the analysis of large datasets. The combination of machine learning with first-principles methods in materials design has introduced innovative approaches to material research. This review looks back on the applications of machine learning in materials design and explores its potential applications in accelerating material innovation，reducing trial-and-error costs，and customizing material design，providing an outlook on the opportunities and challenges in the field of materials science.

Keywords：machine learning；materials design；recipe optimization

[责任编校　赵晓华　陈留院]