



人工智能与大数据在材料科学中的融合：新范式与科学发现

杨帅¹, 刘建军³, 金帆^{4*}, 陆颖^{1,2*}

1. 中国科学院成都文献情报中心, 成都 610299;

2. 中国科学院大学经济与管理学院信息资源管理系, 北京 100049;

3. 中国科学院上海硅酸盐研究所, 高性能陶瓷和超微结构国家重点实验室, 上海 200050;

4. 中国科学院深圳先进技术研究院, 合成生物学研究所, 深圳 518055

* 联系人, E-mail: fan.jin@siat.ac.cn; luy@clas.ac.cn

2024-04-15 收稿, 2024-06-25 修回, 2024-07-04 接受, 2024-07-05 网络版发表

国家社会科学基金(22BTQ024)资助

摘要 材料科学作为一门关键学科, 在推动社会进步和科技创新中发挥着不可替代的作用. 随着人工智能(artificial intelligence, AI)和大数据技术的飞速发展, 材料科学的研究范式正在经历一场深刻的变革. 本综述探讨了AI与大数据结合, 如何重塑材料科学的研究范式(AI for materials science), 加速计算材料学的发展, 并推动实验过程的革新. 首先, 概述了大数据背景下材料数据库的基础设施建设. 这些数据库作为科研工作的基石, 为材料数据的存储、管理和分析提供了强大的支持. 接着, 讨论了AI技术在材料发现周期各阶段的应用, 并介绍了AI与实验室自动化和机器人技术融合形成的自我驱动实验室(self-driving laboratories, SDLs)的兴起. SDLs实现了材料发现的完整闭环流程, 推动了科学研究向自主科学发现模式的重要转变. 此外, 大型语言模型(large language models, LLMs)的发展更是为自然语言理解带来了革命性变化, 推动科学LLMs的兴起, 拓宽了从文本理解到科学探索的能力. 本文进一步综述了材料科学中大型语言模型的最新进展, 强调了它们在加速材料发现过程中的关键作用. 最后, 探讨了AI在材料研究中的巨大潜力, 并审视了构建材料研究智能生态系统所面临的挑战. 通过对相关研究进展的梳理, 本文旨在为科研人员提供信息参考, 并揭示AI在材料研究中的重要性.

关键词 人工智能, 材料发现, 自我驱动实验室, 大型语言模型

材料科学的进步一直是推动人类文明发展的关键因素^[1]. 从古代的陶瓷和金属工艺到现代的纳米技术和生物材料, 材料的创新一直是技术革新的基石. 材料科学的发展历程见证了人类对自然界材料的深刻理解和创新应用. 从经验法则的积累到理论模型的建立, 再到计算模拟的兴起, 材料科学的研究范式经历了多次重大转变. 每一次转变都极大地扩展了我们对材料性质的认识, 并加速了新材料的发现和应用.

在20世纪中叶, 随着计算机技术的发展, 材料科学迎来了第一波计算革命, 即通过理论计算和模拟来预测材料的性质. 这一时期, 经典热力学定律、蒙特卡罗方法、分子动力学模拟和密度泛函理论等方法的出现, 使得科学家能够在原子和分子层面上理解材料的行为.

进入21世纪, 随着实验科学和计算机模拟产生的数据量呈指数级增长, 材料科学领域迎来了第四科学研究范式, 即“数据驱动”的研究模式. 这一范式强调利

引用格式: 杨帅, 刘建军, 金帆, 等. 人工智能与大数据在材料科学中的融合: 新范式与科学发现. 科学通报, 2024, 69: 4730–4747

Yang S, Liu J J, Jin F, et al. Integration of artificial intelligence and big data in materials science: New paradigms and scientific discoveries (in Chinese). Chin Sci Bull, 2024, 69: 4730–4747, doi: 10.1360/TB-2024-0404

用大数据技术来处理和分析实验与模拟数据,以提取有价值的信息和洞察,从而加速新材料的发现和设计过程^[2]。

然而,随着材料系统的复杂性增加,传统的研究范式面临着效率和准确性的挑战。为了应对这些挑战,材料科学领域正在经历一场由人工智能(artificial intelligence, AI)和大数据驱动的革命,这一革命被称为科学智能(AI for science)。

科学智能是一种新兴的研究范式,也有学者称这一范式为第五科研范式^[3],它利用AI的强大数据处理能力和模式识别能力来加速材料的发现、设计和性能优化。在材料科学中,AI的应用不仅限于理论计算和模拟,它还涵盖了实验设计、数据采集与分析,以及材料的合成与特性表征等多个环节。通过深度学习等AI技术,研究人员现在能够处理和分析前所未有的大规模数据集,揭示材料结构与其性能之间的复杂相互作用。

科学智能的核心在于跨学科技术的整合与协同作用,通过这种综合性的方法,可以实现超越单一技术所能达成的成果总和。这种跨学科的融合性研究方法特别适合解决材料科学中的大规模、非线性问题^[4]。它为整个研究过程的各个阶段注入了新的活力,从AI加速材料研究的初期阶段,到全生命周期的持续赋能,再到自动化实验的建立,直至实现自主科学发现,科学智能的应用为材料科学研究领域带来了前所未有的机遇。

通过本文的梳理,我们旨在为科研人员提供信息,揭示AI在材料研究中的巨大潜力,并指出实现这一潜力过程中需要克服的挑战。

1 材料数据库的基础设施建设

在“人工智能+大数据”的新研究范式中,数据的角色不仅是基石,而且是驱动创新的核心力量。高质量的数据分析与挖掘依赖于优质数据源的支持,尤其是材料数据库中丰富的信息。这些信息的准确性和可靠性直接影响数据应用的有效性和研究成果的创新程度。在材料基因工程的推进中,资助机构已经采用了全面的数据管理政策。这些政策不仅强调了对数据共享和传播的重视,也为材料学领域的数据驱动研究设立了坚实的基础,并推动了其快速发展。

目前,规模庞大、内容丰富的材料数据库正在持续建设,以支持“人工智能+大数据”研究范式的发展需求。这些数据库覆盖了从结构和热力学相图到电学和

机械性能等广泛的材料属性,包括但不限于金属、陶瓷、合金、玻璃、二维材料及纳米复合材料等多种材料类别(表1)。这种资源的多样性使得这些数据库成为研究人员不可或缺资产,为他们提供了一系列符合FAIR(可发现(findable)、可访问(accessible)、可互操作(interoperable)和可重用(reusable))数据原则的高质量信息资源,确保了数据的高效利用和广泛应用。

这些数据库,如国际上知名的无机晶体结构数据库(ICSD)^[5]、Materials Project数据库^[6]、开放量子材料数据库(OQMD)^[7]等,以及国内的重要资源,包括中国科学院凝聚态物质科学数据中心、AlphaMat信息平台^[8]、Atomly数据库^[9]、电化学储能材料设计平台^[10]、MatCloud+物性数据库^[11]等,已经成为材料科学领域内的宝贵财富。它们不仅积累了海量的材料结构和性能数据,还提供了先进的数据可视化工具、搜索工具,甚至是应用程序编程接口。这些工具极大地提高了研究人员访问和利用这些数据的便捷性。

在“人工智能+大数据”研究范式的背景下,这些材料数据库的价值不可估量。它们不仅支撑了材料科学的基础研究,还加速了新材料的发现和开发进程。此外,还有如MatWeb、Total Materia和ASM等资源,提供了覆盖广泛的材料特性信息,进一步扩展了研究人员可访问的数据范围。同时,计算材料存储库^[12]、材料合成参数数据库^[13]、材料数据设施^[14]等,都是提供丰富多样化数据源的重要平台。类似的资源也已经整理在表1中,供读者参考。

2 AI加速材料科学发现

随着AI在材料科学领域的广泛应用,我们见证了其研究进展的快速发展,这或将根本性地改变材料科学研究人员在实验室中的工作方式。这一变革主要源于多种机器学习(machine learning, ML)方法的引入,这些方法将被广泛应用于现代材料科学实验室中。AI的融入不仅仅局限于材料的发现阶段,还将深入到计算和实验工作的各个环节,甚至有可能促成一个全新的、基于AI的材料智能生态系统的建立。本节将综述近年来AI驱动的材料科学领域的新进展和新兴发展趋势。

2.1 AI辅助加速理论计算

在材料科学的理论计算中,第一性原理计算,特别是密度泛函理论(density functional theory, DFT),扮演

表 1 一些重要的材料数据库或平台

Table 1 List of important material databases or platforms

名称	材料类型	数据来源
无机晶体结构数据库(FIZ ICSD)(https://icsd.products.fiz-karlsruhe.de/)	无机	实验、计算
NIST ICSD(美国)(https://icsd.nist.gov/)	无机	实验、计算
原子间势(力场)数据库(https://www.ctcms.nist.gov/potentials/)	无机	实验、计算
结构材料数据库(https://materialscommons.org/)	无机	实验、计算
Atomly数据库(https://www.atomly.net/)	无机	实验、计算
材料合成参数数据库(https://www.synthesisproject.org/)	无机	实验、计算
中国科学院凝聚态物质科学数据中心(http://condmatt.iphys.ac.cn/)	无机	实验、计算
Materials Project数据库(https://next-gen.materialsproject.org/)	无机	计算
开放量子材料数据库(https://oqmd.org/)	无机	计算
JARVIS-DFT(https://jarvis.nist.gov/)	无机	计算
AFLOW数据库(http://aflowlib.org/)	无机	计算
计算材料存储库(https://cmr.fysik.dtu.dk/)	无机	计算
f-electron 结构数据库(http://correlatedmaterials-lanl.org/)	无机	计算
CALPHAD数据平台(https://thermocalc.com/)	无机	计算
原子间势IAP平台(https://matterverse.ai/)	无机	计算
电化学储能材料设计平台(https://www.bmaterials.cn/)	无机	计算
无机薄膜材料实验数据库HTEM(https://hitem.nrel.gov/)	无机	实验
衍射数据库PDF(https://www.icdd.com/pdfsearch/)	无机	实验
Pauling file数据库(https://www.paulingfile.com/)	无机	实验
有机材料数据库(https://omdb.mathub.io/)	有机	实验、计算
高分子纳米材料设计与分析平台NanoMine(https://github.com/Duke-MatSci/nanomine)	有机	实验、计算
高分子材料设计数据库PoLyInfo(https://polymer.nims.go.jp/)	有机	实验
Springer材料数据库(https://materials.springer.com/)	有机、无机	实验、计算
材料数据设施(https://www.materialsdatafacility.org/)	有机、无机	实验、计算
晶体开放数据库(https://www.crystallography.net/)	有机、无机	实验、计算
Citrination平台(https://citrine.io/)	有机、无机	实验、计算
欧盟MatDB数据存储库(https://odin.jrc.ec.europa.eu/)	有机、无机	实验、计算
ChemSpider平台(http://www.chemspider.com)	有机、无机	实验、计算
AlphaMat信息平台(https://aimslab.cn/#/alphamat)	有机、无机	实验、计算
MatNavi材料数据库(https://mits.nims.go.jp/)	有机、无机	实验、计算
中国国家材料科学数据共享平台(https://www.materdata.cn/)	有机、无机	实验、计算
NOMAD数据库(https://nomad-coe.eu/)	有机、无机	计算
ASM数据库(https://www.asminternational.org/)	有机、无机	实验
MatWeb数据库(https://www.matweb.com/)	有机、无机	实验
Total Materia数据库(https://www.totalmateria.com/)	有机、无机	实验
CatApp平台(http://suncat.slac.stanford.edu/catapp/)	催化	计算
Catalysis Hub催化数据平台(https://www.catalysis-hub.org/)	催化	计算
OC20催化数据库(https://opencatalystproject.org/)	催化	实验
Novamag数据(http://crono.ubu.es/novamag/)	磁性材料	计算
国际玻璃数据库(www.newglass.jp/)	玻璃	实验
SciGlass数据库(https://www.akoscheminformatics.de/sciglass/sciglass.htm)	玻璃	实验
Mindata数据库(https://www.mindat.org/)	矿物质	实验
美国矿物晶体结构维护数据库(American Mineralogist CSD)	矿物质	实验
SNUMAT数据库(https://www.snumat.com/)	带隙材料	计算
可再生能源材料特性数据库NRELMatDB(https://materials.nrel.gov/)	光伏、光电材料等	计算
MatCloud+物性数据库(https://www.matcloud.com.cn/)	晶体材料	实验、计算
计算2D材料数据库(C2DB)(https://cmr.fysik.dtu.dk/c2db/c2db.html)	二维材料	计算
MDR SuperCon数据库(https://mdr.nims.go.jp/concern/datasets/mw22v8634)	超导	实验
DDSE数据库(https://ssbed-ssbed-database-gui-ntjlz.streamlit.app/)	电池电解质	实验

了关键角色. 这类计算基于量子力学原理, 如薛定谔方程, 直接模拟电子密度来预测材料性质, 不依赖于经验数据. DFT以其高效的计算能力、广泛的适用性和较高的预测精度, 在处理复杂材料系统中特别突出, 成为众多AI辅助理论计算的基础^[15]. AI技术与DFT的结合显著提高了计算模型的效率、准确性和可扩展性, 优化了计算流程 and 数据分析, 有效弥补了DFT的局限. 当前, 基于DFT的ML模型主要包括(图1): 观测值学习模型, 预测如原子力和分子能量的量子属性; 混合ML/DFT方法, 通过优化哈密顿量和密度泛函形式改进模型; ML密度泛函, 改进电子密度到能量映射的策略.

首先, AI辅助显著提升了计算效率. 以基于物理的监督量子机器学习(quantum machine learning, QML)^[16]方法为例, 它是一种高度依赖DFT的计算模型, 这种依赖主要体现在QML模型的训练过程中, 通过使用DFT计算或量子化学的数据来训练模型. 一旦训练完成, QML模型在预测速度方面显著优于传统的DFT. 具体来说, QML模型的预测主要基于高效的线性代数运算, 这些运算通常比DFT中的计算要快数个数量级. 因此, 在处理大规模数据和复杂系统时, AI与DFT的结合能够显著提高计算速度, 从而提升整体的计算效率.

当然, 效率和预测能力往往需要进行权衡. 一般而言, 训练数据较少的ML模型可能运行更高效, 但其预测准确度可能较低. 为了平衡效率和泛化能力之间的关系, 主动学习策略可以被用来更高效地抽样训练数据. 例如, Huang等人^[17]使用查询特征相似性来进行训练数据的偏向选择, 使得数据效率明显提高. 这种方法通过智能选择训练样本来提高学习效率和模型性能. 另一种提高预测准确性的方法是通过学习修正标签(Δ -ML)或利用不同量子近似之间的相关性, 如多层次网络组合的量子机器学习(combination QML, CQML)

框架^[18], 其类似于复合量子化学, 允许在基组和电子相关维度的分层数据集中进行系统误差消除, 以获得高保真的预测质量. 这些技术可以增强模型对复杂系统的理解, 从而提高预测的准确度. 因此, 通过精心设计和优化ML模型, 可以在最小化数据需求和模型复杂性的同时, 实现快速迭代, 以提高计算的效率.

在探索提高预测准确率方面, 关键策略包括使用更多的参考数据训练ML模型, 以此提升DFT预测的准确性. 在这方面, 一般策略包括强化训练集约束, 这种策略是在固定训练集大小的前提下, 通过加强模型的要求和边界条件来提高ML模型的准确性^[19]. 此外, 可以结合实验观测值. 为了达到实验精度, ML模型需要融合实验观测数据, 以有助于超越纯计算方法的局限.

在提高可扩展性方面, ML的应用使得对复杂电子系统的高计算成本DFT方法进行近似变得可行, 允许处理更大规模的系统. 例如, ONETEP线性缩放DFT中^[20], 实现了电子结构预测范围的大幅扩展; 通过ML与基于物理的方法相结合的算法和软件集合“MALA”^[21]也可以大幅拓展不同尺度的电子结构预测; 缺陷图神经网络(defect graph neural network, dGNN)^[22]等方法进一步提高了对理想晶体结构中缺陷形成焓的预测能力, 简化并增强了理论计算的适用性.

此外, ML模型在化学基础上展现出显著的泛化能力, 即在较小系统上训练的模型也能准确预测更大系统的行为^[23]. 这得益于对原子环境相似性的精确量化和局部性假设的应用, 使得模型能在降低计算负担的同时, 保持对复杂系统的适应性和精确预测能力.

因此, AI技术与DFT的结合不仅显著缩短了计算时间, 还提升了理论计算的预测性能. 未来研究将可能探索融合传统泛函与其他变量的方法以及多尺度计算, 发展出既高效又高精度、应用范围更广的ML模型.

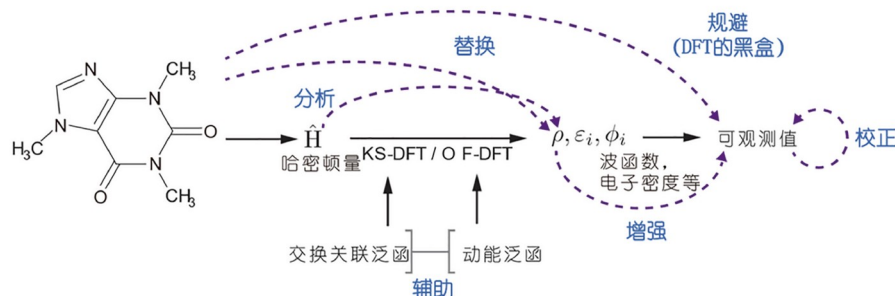


图1 (网络版彩色)机器学习方法辅助的DFT计算. 实线箭头代表传统的工作流程. 虚线箭头代表基于机器学习方法的DFT计算路线
Figure 1 (Color online) Machine learning assisted DFT calculations. The solid arrows represent the traditional workflow. The dashed arrows indicate the DFT computational route based on machine learning methods

2.2 AI辅助的材料设计与预测

在材料研究中, 一个核心目标是设计或改进材料以实现特定的性能. 面对的主要挑战之一是如何从庞大的化学化合物空间(chemical compound space, CCS)中筛选出具有所需特性的材料. 由于理论计算在处理复杂体系(如包含有机无机部分的体系、复杂固溶体等)时的效率和准确性有限, 实验数据和基于统计的预测模型变得至关重要, 可为理解材料行为提供实际信息并有效筛选设计空间. 在这个过程中, AI技术在材料空间导航和性能预测中发挥了关键作用.

ML技术在材料科学领域被广泛应用于高效地预测材料属性, 从而有助于探索、设计和筛选有前景的材料候选. 针对特定的应用场景, 可以部署一个或多个基于ML的代理模型进行深入分析. 这方面的内容在一些综述文章中已有详细讨论^[24-27], 我们在此研究策略上的讨论. 目前, 基于ML的材料筛选通常采用组合枚举、逆向设计和基于主动学习的策略, 如图2所示.

ML辅助的枚举策略结合了传统枚举法和现代ML技术, 为材料设计提供了一个高效的方法. 这种策略先界定研究范围(如材料类型、化学成分和结构), 然后在这个范围内用ML模型预测所有可能的材料组合或结构的性能, 如合金的所有元素组合. ML模型利用其数据处理和模式识别能力, 优化筛选过程, 快速识别出符合特定要求或具有最优性能的材料组合. 例如, Rhone

等人^[28]通过ML研究了大空间的单层过渡金属卤化物, 实现了一种半监督学习的ML形式, 加速了新材料的发现及其磁性的预测; Thomas等人^[29]通过建立符号回归模型, 找到了80种超绝缘材料.

逆向设计策略标志着材料科学的一次革命性转变, 它着眼于从目标性质出发来识别或发现新材料. 通过结合ML和各类优化算法(如进化算法和基于群体的优化方法), 这种策略深入探讨结构与性质之间的复杂关联. 近期研究, 例如非线性机械超材料^[30]、新型晶体材料^[31]和多环芳香族分子^[32]的逆向设计, 证实了其在应对功能性材料设计挑战中的有效性. 与传统高通量筛选相比, 逆向设计借助生成模型^[33](如强化学习、变分自编码器和生成对抗网络)发掘新材料, 具有更大潜力. 这些方法通过创建连续的材料向量空间, 能够生成具有期望属性的新材料配置, 展示了在理解和利用非线性结构-属性关系方面的先进性.

主动学习策略在材料设计中通过自适应设计流程和反馈机制创新地提升了效率, 依赖ML模型的预测来指导数据收集和模型改进. 这种方法核心在于提高模型的应用性和准确度, 通过量化预测不确定性和采用适当的获取函数来优化决策过程. 结合ML模型和贝叶斯优化等优化算法^[34], 主动学习能精细调整模型并解决如何选择下一步实验或计算对象的挑战, 采用迭代方式指导数据收集以改进模型. 此外, 主动学习策略融合现有知识的挖掘和新的可能性的探索, 有效地促进

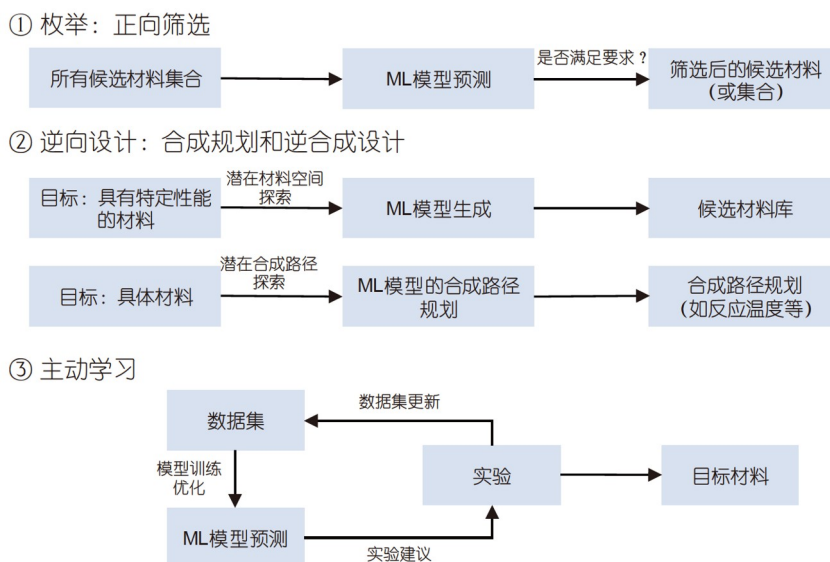


图2 (网络版彩色)材料发现和筛选的三种机器学习策略

Figure 2 (Color online) Three machine learning strategies for material discovery and screening

了模型性能的提升和对设计空间的深入探索。

近年来, 主动学习已广泛应用于材料设计和发现, 包括碳纳米材料^[35]、聚合物材料^[36]和储能材料^[34]等领域。同时, AI技术的发展促进了不同策略和过程的融合, 加速了新材料的发现并提高了预测准确性。这包括混合深度学习方法、图神经网络和基于多体交互的深度学习框架等, 展现了AI在智能化材料设计中的潜力。

2.3 AI辅助的材料表征

在材料科学领域, 深入理解材料的微观结构与其宏观性能之间的联系至关重要。得益于扫描透射电子显微镜、扫描隧道显微镜、原子力显微镜等先进的表征技术, 我们能够获得大量高分辨率的成像数据, 这些数据为揭示材料的性质提供了宝贵的信息。然而, 随着成像数据量的激增, 数据处理成为一个巨大的挑战。为了应对这一挑战, 迫切需要开发高效的算法工具来处理和分析这些数据。在这一背景下, AI技术为处理和理解材料表征数据提供了新的解决方案。AI技术在加速成像、图像分割和动态分析等方面展现出了巨大的潜力。通过AI技术, 研究人员能够更快速、更准确地处理和分析大量的成像数据, 从而深入挖掘数据中隐藏的规律和知识。

AI算法在加速材料成像方面取得了显著进展, 显著缩短成像时间的同时, 保持或提高了图像质量, 为快速、精确地表征材料特性提供了强力支持。这对于高通量实验至关重要, 允许在较短时间内处理更多样本, 加速了材料发现和研究进度。例如, 基于卷积神经网络

的方法已成功用于提高电子背散射衍射的扫描速率, 并显示出对图像噪声的高鲁棒性^[37]。此外, AI辅助扫描探针显微技术通过利用部分原始数据重建高质量图像, 大幅减少了扫描时间。这不仅加快了图像采集过程, 还提高了实验效率和可靠性^[38]。通过引入智能化表征流程^[39], AI技术正推动材料成像与表征速度和质量的进一步提升, 加速了新材料的发现和性能优化。

AI技术在图像处理领域表现出卓越的效能, 特别是在图像识别、分类、降维处理和数据增强等方面为材料结构表征和分析提供了强大支持(图3(a))。这使得从复杂成像数据中精确分离不同微观结构并进行动态分析成为可能。AI还能进行三维样品微观结构的深入表征, 提供了理解材料宏观性能所需的关键微观视角。例如, AI已被应用于光学显微镜中识别二维异质结构^[40]、晶体材料图像分类^[41], 以及在拉曼光谱分析中进行相位识别^[42], 这些应用不仅增强了我们对材料复杂结构的理解, 还显著提高了材料分析的精度和效率。

AI技术在图像处理领域的应用已经拓展到动态采样数据的收集与分析以及实验散射和光谱数据的处理, 尤其在自动图像减噪、分割、重建和微结构重组^[43-45]方面发挥关键作用。这些进展显著优化了数据处理流程, 提升了实验数据的质量和解析度, 使研究人员能够更深入地探索材料的微观结构。

进一步地, AI通过结合物理信息模型从大量的材料表征数据中提取深层科学知识, 尤其体现在建立微观结构特征与宏观性能之间更精确的相关性, 为材料设计和优化提供新视角(图3(b))。例如, 使用卷积神经

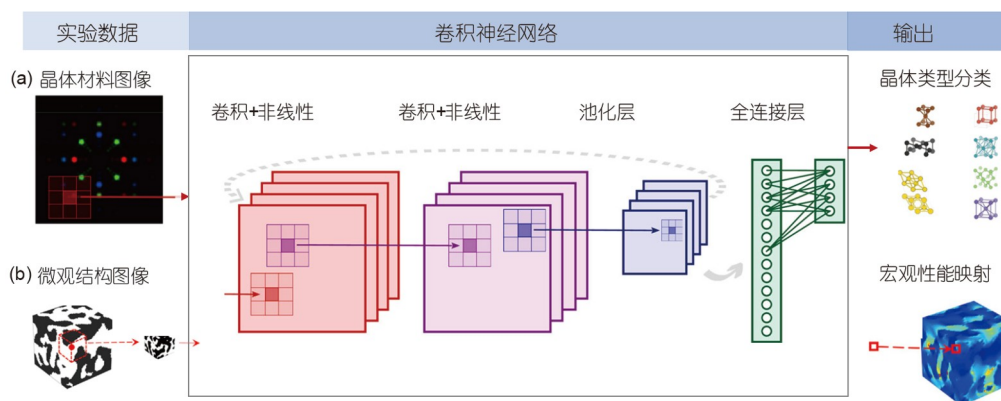


图3 (网络版彩色)人工智能技术用于材料表征。(a) 深度学习实现晶体材料图像的分类。(b) 深度学习建立“微结构-宏观性能”的映射关系

Figure 3 (Color online) Application of artificial intelligence techniques in material characterization. (a) Utilization of deep learning methods for the classification of crystalline material images. (b) Establishment of a “microstructure-macroscopic property” mapping relationship using deep learning approaches

网络将复合材料的微观结构图像直接映射到宏观刚度上,提出了一种理解和预测材料性能的新方法^[46].利用神经网络建立的图像数据与材料生长机理^[47]之间的映射模型不仅深化了对材料行为的理解,也为材料科研提供了高效的分析工具.

2.4 自动化实验室和自主科学发现

在现代材料科学领域,数据科学和AI技术的集成应用不仅促进了材料合成过程的智能化,而且推动了实验室自动化和机器人技术的深度融合,进而催生了自我驱动实验室(self-driving labs, SDLs)的发展^[48],使得材料研究的工作流程形成了一个完整的闭环.

在SDLs的架构内, AI与实验室自动化技术共同打造了一个高度模块化的研究平台. 这一平台依赖ML算法的迭代指导, 自动化地贯穿了从材料合成到表征的整个研究过程, 极大地提升了实验的效率. 如图4(a)所示, SDLs呈现出的一般架构展现了从试剂准备到材料合成,再到表征及性能评估各环节的无缝衔接,构筑了

一个从头至尾的完整实验 workflow.

SDLs的核心特点在于其卓越的自适应能力,使其能够在极小的人工干预下自主适应新信息并调整实验参数,实现了科学实验过程的自动化闭环. 这种系统不仅在实验设计阶段进行了优化,还能够根据实时反馈数据进行动态调整,充分发挥AI在加速材料发现与合成方面的潜力. 例如, SDLs在闭环合成-属性关系映射和按需合成半导体材料方面的应用展现了惊人的速度,比传统方法快超过1000倍,大幅提高了材料科学研究的效率和速度^[48]. 因此, SDLs不仅改变了操作员和研究人员在化学及材料科学工作流程中的角色,而且预示了实验室未来发展的新方向.

在过去五年里, SDLs在加速材料合成与制造方面实现了突破性进展,覆盖了从碳纳米管到复杂有机化合物、纳米材料、相变记忆材料以及薄膜材料等多个研究领域(表2). 这些进步不仅提升了实验室操作的自动化和智能化水平,而且通过其自主可重构的实验模块,极大地加速了新材料的探索. SDLs通过数字化实

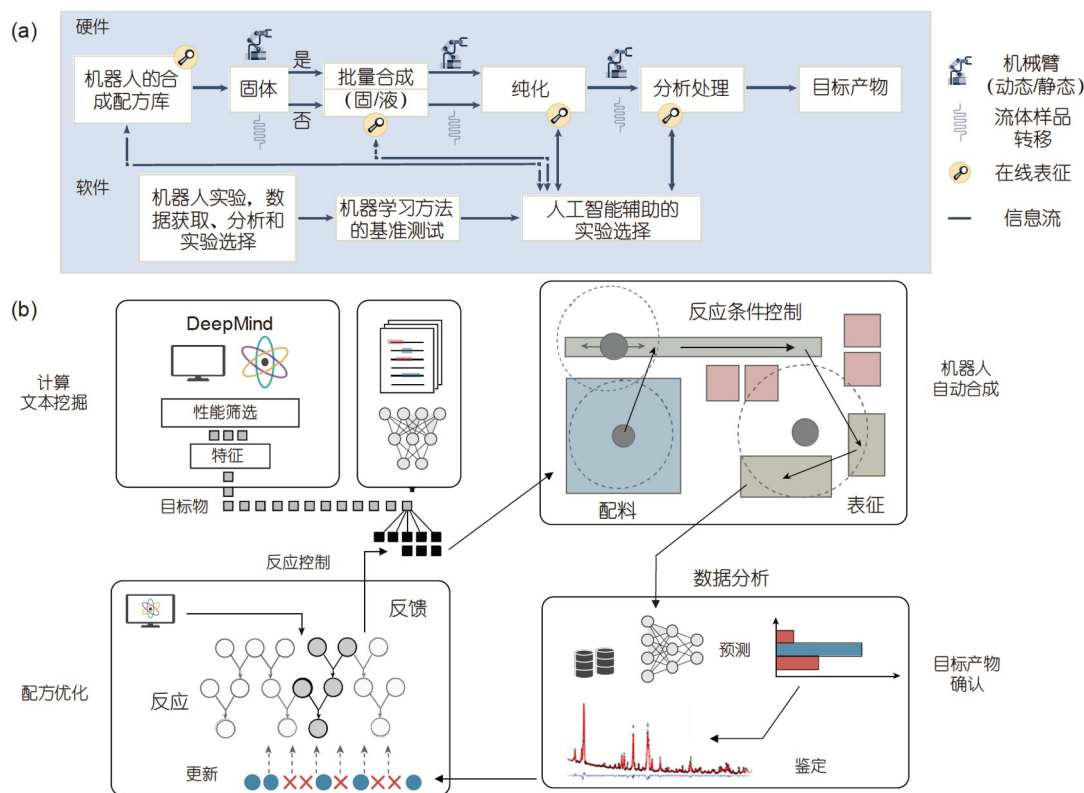


图4 (网络版彩色)基于自我驱动实验室的自主科学发现. (a) 化学和材料科学领域构建SDLs一般架构. (b) A-Lab的自主材料发现的流程示意图

Figure 4 (Color online) Autonomous scientific discovery based on self-driving laboratories (SDLs). (a) General architecture of SDLs in the field of chemistry and materials science. (b) The process and facilities for autonomous material discovery at A-Lab

表 2 近几年的SDLs研究平台

Table 2 Recent self-driving laboratories (SDLs) research platforms

研究团队	平台名称	适用性
Xi Zhu ^[49]	AIR-Chem (2018)、MAOS (2020)、MAOSIC (2020)	无机钙钛矿量子点、纳米材料
Alán Aspuru-Guzik ^[50]	ChemOS (2018)	连接研究人员、机器人和计算工具的SDL软件包
Alexei A. Lapkin ^[51]	DoE (2021)	复杂有机反应、液体配方设计
Christoph J. Brabec ^[52]	MANDA Line One (2020)、HTRobot (2021)	有机光伏材料
Keith A. Brown ^[53]	BEAR (2020, 2021)	聚乳酸材料的3D打印
Ichiro Takeuchi ^[54]	CAMEO (2020)	形状记忆材料
Andrew I. Coope ^[55]	Mobile robotic chemist (2020)	光催化剂
C. P. Berlinguette ^[56]	Ada (2020)	薄膜材料
Thomas Heumueller ^[57]	2020	薄膜材料
Emory M. Chan ^[58]	RAPID (2020)	金属卤化物钙钛矿单晶
Taro Hitosugi ^[59]	2020	无机薄膜
Paul J. A. Kenis ^[60]	2021	纳米晶体
Xiaonan Wang ^[61]	2021	银纳米粒子
Frank A. Leibfarth ^[62]	2021	聚合物
IBM	RoboRXN (2021, https://research.ibm.com/science/ibm-roborex/)	化学合成全流程平台
Eugenia Kumacheva ^[63]	MFML平台(2021)	金属纳米粒子
Yi Luo ^[64]	AI-Chemist小来(2022, 2023)	优化金属氧化物光催化剂的氢掺杂策略、合成产氧催化剂
C. Oliver Kappe ^[65]	2022	药物合成
Xue-Feng Yu ^[66]	2023	胶体纳米材料合成
Shoichi Matsuda ^[67]	NIMS-OS (2023)	实现机器人实验和AI的闭环的Python库
Samantha Y. Chong ^[68]	ARChemist (2023)	晶体生长与表征
Youn-Suk Choi ^[69]	2023	有机分子自主合成
Gerbrand Ceder ^[70]	A-Lab (2023)	无机化合物材料
Leroy Cronin ^[71]	Chemputer (2019)、MWP (2020)、CA-robot (2020)、 χ DL (2024)	过渡金属盐; 超分子簇; 金纳米颗粒、复杂配方设计; 通用化学编程语言
Milad Abolhasani ^[72]	Artificial Chemist (2020)、AlphaFlow (2023)、Smart Dope (2024)	半导体纳米材料、金属硫族化物、掺杂钙钛矿量子点、金属卤化物钙钛矿纳米晶体
Timothy Noël ^[73]	RoboChem (2024)	光化学合成

验过程, 提高了实验数据的准确性、可转移性, 同时显著减少了人为错误, 增强了数据的可重复性。

例如, 中国科学技术大学团队开发了机器人化学家“小来”^[64], 该系统集成了阅读海量文献和调用深层物理模型进行思考与模拟计算的能力, 实现了完整的科学方法闭环。它能在特殊环境下(如模拟火星环境)设计和制造复杂材料, 将原本需数千年通过“试错”方法达成的任务缩短至数周完成。劳伦斯伯克利国家实验室的A-Lab^[70](图4(b))AI机器人平台则展示了如何从大量候选物中识别并自主合成预期稳定化合物的能力, 仅用17天时间就成功合成了41种新化合物。同样, 荷兰

阿姆斯特丹大学开发的“化学机器人”(RoboChem)^[73], 通过AI驱动的ML单元自动优化分子合成过程, 利用流动化学技术和自动化设备大幅提高了分子合成和分析的效率。这种高度自动化的系统不仅极大提升了实验效率, 而且通过生成全面的数据集为化学分子的探索提供了新的维度。研究发现, RoboChem在复现选定论文实验时, 约80%的情况下能产生更佳结果。这些示例表明, AI辅助的自我驱动实验室在化学领域的应用不仅加速了科学发现的进程, 而且展现了AI在推动未来科研工作中的巨大潜能。

SDLs的兴起不仅标志着材料研究的手段向全面自

动化的转型,而且更重要的是,它象征着科学研究模式向自主科学发现的演进^[74]。这种完整闭环的研究流程,无须(或少)人类直接操作,依靠机器和软件系统的强大计算及数据分析能力进行模式识别、假设构建、实验设计及结果验证,从而在科学探索中实现创新性的发现和深入的理解。如小来^[64]、A-Lab^[70]、RoboChem^[73]等项目所示,自主科学发现的理念已经开始在实践中得到体现。这种独立于人类干预的研究模式具有显著的优势,它不仅能够提升科学研究的效率,加快新发现的步伐,而且在未来有望解决那些人类难以独立攻克的复杂科学问题。随着SDLs技术的不断发展和完善,预期将引领材料科学及整个科学界的革命性变革,推进科学探索的深度发展。然而,当前阶段,SDLs面临的财务和技术挑战仍需通过学术界与工业界的合作、开源机器人模块和云实验室资源的支持来克服,以降低技术门槛。

同时,ML方案的成熟使得自主计算实验室的概念逐渐变为现实^[75,76]。这种实验室采用主动学习以逆向方式进行模拟,通过迭代模拟生成符合预设属性和化学空间的材料,标志着新材料探索和发现效率的进一步提升。未来,结合自主计算实验室和SDLs,将实现ML不同模型的协同作用,为材料科学研究带来更加高效、精确的新方法和视角。

2.5 大型语言模型的新赋能

近年来,大型语言模型(large language models, LLMs),尤其是基于transformer架构的模型如ChatGPT,已在材料科学领域引起广泛关注,ChatGPT更是被*Nature*杂志评为2023年度十大人物之一。这不仅凸显了其在科学领域的影响力,也象征着AI工具在科学研究中的重要地位。

在材料科学中,LLMs与传统ML方法互为补充,后者擅长处理结构化数据,而LLMs则在解析复杂、非结构化的文本数据方面展现出独特优势,如科学文献和实验报告。这些模型能够理解多种数据表述格式并处理实验的复杂细节,提高数据处理的灵活性和效率。

通用型自然语言模型,如ChatGPT、GPT-4、Gemini等,在科学领域已取得显著成就,能够覆盖广泛的科学知识并处理复杂数据。在材料科学中,它们的应用逐渐扩展,已在理解材料知识、提出候选成分、生成材料结构、预测材料性能和规划合成路线等方面显示出优异的表现^[77]。此外,为了更专业地解决科学问题,

开发了专门的科学语言大模型(Sci-LLMs)^[78],这些模型专注于封装专业科学知识,解决科学领域中的复杂问题。

LLMs已成为推动材料科学研究和发现的关键力量。通过全球研究人员的广泛合作,LLMs在化学材料领域的应用潜力得到了深入探索,展示了14个典型案例^[79],表3整理了与材料研究相关的大型语言模型。在本文中,我们特别强调LLMs在化学材料科学中的两个主要应用方向。一是知识提取。LLMs具有强大的知识提取能力,能够从科学文献或数据库中提取关键信息,并将其转换为结构化数据。这是传统ML方法难以企及的。二是智能代理(agent),利用LLMs处理复杂问题的能力,可以通过简单的语言提示来进行材料设计或执行实验流程。这种方法不仅简化了科学研究过程,还提高了效率,为材料科学的创新和发展提供了新的视角和工具。

2.5.1 知识提取

在材料科学领域的自然语言处理(natural language processing, NLP)任务中,特别是知识提取,主要依靠两种模型:跨学科大型语言模型和专门针对材料科学的大型语言模型。

跨学科大型语言模型深化了多个学科领域的文本数据分析与理解,包括生物学、材料化学及其他科学学科。例如,SciBERT^[80]模型基于BERT架构,通过对广泛的科学出版物进行无监督预训练,极大地增强了处理科学领域NLP任务的能力。同理,ScholarBERT^[81]系列模型同样基于BERT,覆盖了生命科学、生物医学和物理科学等多个领域,通过针对这些领域的广泛科学语料库训练,满足了跨学科研究的需求。DARWIN-Base模型^[82]采取了纯解码器架构,利用来自各类科学文本和知识库的海量数据进行训练,并通过SciQ数据集的指令和问答数据进行精细微调,展示出在处理需广泛科学知识的任务上的出色能力。

在材料化学领域,专门针对该领域的LLMs展现了对化合物和化学反应信息理解与推理的强大能力。例如,基于BERT架构的模型如ChemBERT^[92]、ChemBERTa^[93,94]系列和ChemDFM^[97]等,在大规模化学期刊论文数据集上进行预训练,专注于自动提取化学反应信息,使用掩码语言建模技术以适应材料化学领域的NLP任务。同样,MatSciBERT^[98]、MaterialsBERT^[99]、MaterialBERT^[100]及其他针对特定材料科学分支的模型,比如电池材料的BatteryBERT^[104]和光学材料的Op-

表 3 与材料研究相关的大型语言模型^{a)}

Table 3 Large language models related to material research

时间(年.月)	模型名称	模型大小	基座结构	领域/用途
2019.09	SciBERT ^[80]	117 M	BERT	包含材料的跨学科 大型语言模型
2023.05	ScholarBERT ^[81]	340 M, 770 M	BERT	
2023.08	DARWIN-Base ^[82]	7 B	LLaMA	
2019.11	S2ORC-SciBERT ^[83]	110 M	BERT	
2020.04	SPECTRE ^[84]	110 M	BERT	
2021.03	OAG-BERT ^[85]	110 M	BERT	
2022.06	SciDeBERTa ^[86]	N/A	DeBERTa	
2022.09	CSL-T5(中文) ^[87]	220 M	T5	
2022.10	Academic RoBERTa(日语) ^[88]	125 M	RoBERTa	
2022.11	Galactica ^[89]	125 M、1.3 B、6.7 B、30 B、120 B	GPT	
2022.11	VarMAE ^[90]	110 M	RoBERTa	化学材料领域大型 语言模型
2023.05	Patton ^[91]	N/A	GNN-Nested Transformer	
2021.06	ChemBERT ^[92]	120 M	RoBERTa	
2020.10	ChemBERTa ^[93]	125 M	RoBERTa	
2022.09	ChemBERTa-2 ^[94]	125 M	RoBERTa	
2023.11	FORGE-Mat ^[95]	13 B	GPT-NeoX	
2021.10	MatBERT ^[96]	110 M	BERT	
2024.01	ChemDFM ^[97]	13 B	LLaMa-13B	
2021.09	MatSciBERT ^[98]	117 M	BERT	
2022.09	MaterialsBERT ^[99]	110 M	PubMedBERT	
2022.09	MaterialBERT ^[100]	110 M	BERT	
2024.01	MatSciRE ^[101]	110 M	MatBERT	
2023.01	BERT-PSIE ^[102]	117 M	BERT	
2021.05	NukeLM ^[103]	110 M、125 M、355 M	BERT RoBERTa	
2022.05	BatteryBERT ^[104]	110 M	BERT	
2022.10	OpticalBERT ^[105]	110 M	BERT	
2022.05	ChemGPT ^[106]	1B	GPT	
2022.08	ProcessBERT ^[107]	110 M	BERT	
2023.08	GIT-Mol ^[108]	700 M	BLIP2	
2023.09	GPT-MolBERTa ^[109]	N/A	BERT RoBERTa	
2023.10	LLM-Prop ^[110]	37 M	T5	
2023.06	MolReGPT ^[111]	N/A	ChatGPT	
2023.10	MechGPT ^[112]	1.3 B、7 B	Llama	
2023.04	ChemCrow ^[113]	N/A	GPT-4	LLMs作为智能代理 促进材料科学发现
2023.08	ChatMOF ^[114]	N/A	GPT-4、GPT-3.5-turbo	
2023.04	Coscientist ^[115]	N/A	GPT-4、Claude 1.3	
2024.01	ORGANA ^[116]	N/A	GPT-4	
2023.10	CLAIRify ^[117]	N/A	GPT-3.5	

a) “N/A”代表文献未提及或未查询到相关信息

ticalBERT^[106], 在丰富的材料科学文献语料库上训练, 证明了它们在提取材料属性信息方面的卓越性能. 此外, 基于Transformer的语言模型也被开发应用于更具体的任务, 如分子生成^[108]、分子动力学(ChemGPT)^[106]、分子属性预测(GPT-MolBERTa)^[109]以及合成流程图的优化^[118], 进一步扩展了LLMs在材料科学领域的应用范围.

同时, 通用大型语言模型, 如GPT系列和Llama, 通过提示工程、微调和上下文学习在材料科学中的应用也展现出灵活性和高效性. 这些模型通过专门的训练能够在化学和材料科学的多种任务中实现高水平的性能, 如材料逆向设计优化^[119]和催化剂合成^[120].

2.5.2 智能代理

LLMs已在文本理解和处理等NLP任务中展示出其强大能力. 除此之外, LLMs引人注目的新兴能力之一是执行链式思维提示(chain of thought prompting, CoT prompting)以实现逐步推理, 以及有效地使用外部工具. 链式思维提示^[121]是一种引导语言模型沿着一系列逻辑步骤解决问题的技术, 这种方法让模型能够像人类解决问题时那样跟踪中间步骤, 而不是直接在单个响应中提供答案.

LLMs使用外部工具的能力进一步扩展了它们的知识边界, 使它们能通过调用搜索引擎、使用计算器等方式来丰富自己的知识库. 这项能力已经证明对于提高LLMs在以往难以处理的任务上的表现至关重要.

最近的研究进展, 如Reason+Act(ReAct)^[122]系统的开发, 展示了将CoT和工具使用能力结合的巨大潜力. 通过集成外部工具进CoT框架, 这类将LLMs作为代理的系统已经展示出在解决复杂问题方面超越其他基于LLMs方法的能力.

在化学材料领域, 当LLMs作为智能代理与现有化学工具和数据库结合使用时, 可以显著提升LLMs在化学材料领域的适用性和功能性, 如图5所示. 最近, 洛桑联邦理工学院和罗切斯特大学的研究人员开发了一套以GPT-4为核心的智能助手ChemCrow^[113], 该系统整合了13种专家设计的化学工具. ChemCrow已被证明能够有效地规划和执行化学任务, 从而有效简化了药物和材料设计及合成等领域的常见化学任务的推理过程. 此外, 韩国科学技术院研究团队利用大型语言模型作为智能代理, 建立了一个能够预测和生成金属有机框架的系统ChatMOF^[114].

LLMs与实验室自动化技术的结合, 为开发自主设计和执行科学实验的系统提供了新的思路. 例如, 卡内基梅隆大学的Coscientist项目^[115], 一个基于多个LLMs的智能代理, 能够通过简单的语言提示执行整个实验过程, 自主设计、规划和执行复杂的科学实验. Coscientist可以使用工具浏览互联网和相关文档, 使用机器人应用编程接口, 并利用其他LLMs完成各种任务. 加拿大多伦多大学团队开发的ORGANA系统, 能够通过自然语言与实验室中的化学家进行交互, 稳健地执

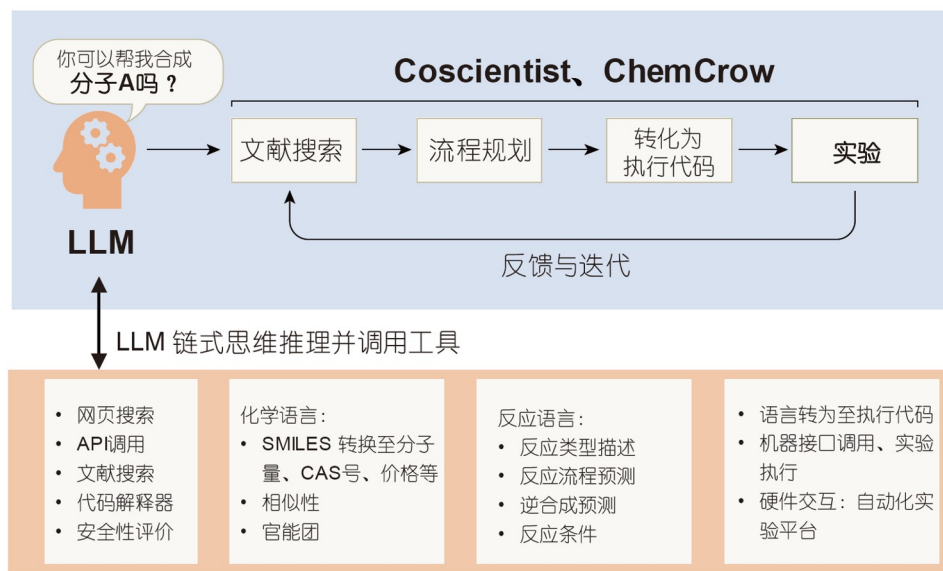


图 5 (网络版彩色)以大型语言模型作为智能代理的自动化合成实验流程

Figure 5 (Color online) Automated synthesis workflow using LLMs as intelligent agents

行由19个并行执行的步骤组成的长视距电化学实验计划^[116]。还有通过使用大型语言模型作为智能代理的CLAIRIFY方法,进行任务和运动规划,将自然语言指令翻译成机器人可执行的计划,实现自动化化学实验^[117]。

像ChemCrow、Coscientist等平台不仅可以作为独立的复杂任务解决器,更重要的是,它们可以被视为通用的化学材料研究助手,使相关工具更易于化学材料科学家访问,从而加速科学发现。此外,这些平台还突出了无缝利用工具组合性的优点,使得通过自然语言的简单请求轻松丰富不同工具之间的结果,甚至使得构建定制的工具管道成为可能。

3 展望: 构建材料研究的智能生态系统

展望未来,我们正迈向一个由AI和机器人技术驱动的智能工业革命,这将极大地扩展人类的认知边界。AI技术正在重塑材料研究的面貌,推动智能生态系统的构建。尽管如此,我们仍需面对AI在材料科学应用中的挑战。

3.1 构建高质量、标准化的数据库

在由大数据和AI驱动的研究范式下,科学数据的有效整合与利用成为创新的关键。然而,科学数据目前面临的问题包括标准不一致、质量参差不齐以及多来源数据的孤立性,这些问题共同限制了基于数据的材料科学研究的进一步发展。

为了解决这些问题,迫切需要打破数据孤岛现象,通过整合来自不同来源的理论与实验数据,构建一个融合多学科知识和多模态数据的人工智能化材料科学数据库,将成为推动材料科学领域智能化发展的重要举措。

此外,在构建数据库时,我们应重视材料数据的信息丰富性,而不仅仅是单纯追求数据量的增加。已有研究显示,不同材料属性的大型数据集之间存在显著的信息冗余^[123]。这意味着在建设数据库的过程中,应更加注重数据的质量和信

息价值。同时,材料数据通常具有高维特征空间与小样本量的特点,这限制了ML模型性能的上限。因此,建立系统、全面的材料科学数据治理流程显得尤为必要^[124]。这将有助于提高数据的质量和可用性。

3.2 复杂机器学习模型的开发与验证

在AI模型的开发和测试方面,基准数据集的应用

是推动新算法持续进步和性能测试的关键。这种做法不仅遵循了计算机科学和统计学的传统方法,也是评估算法在处理未知数据时表现的重要基础。特别是在材料科学领域,由于存在着庞大的元素类型和材料系统的高度复杂性,ML模型的泛化能力和可迁移性尤其受到挑战。在特定数据集之外,模型的预测准确度往往会显著下降,这限制了它们在新材料发现和性能预测方面的应用范围。

随着ML和深度学习技术的快速发展,对这一挑战的应对策略也在不断演进。一种途径是通过开发更加复杂的ML模型,这些模型能够处理更多层次的数据并理解更加复杂的系统。这类模型拥有更多的拟合参数,能够从更广泛的数据中学习,提高模型在新领域的适用性和预测准确性。

另一种解决方案是创建端到端的学习框架,这种框架集成了多个专门的ML模型来处理不同的子任务。这种方法的典型例子是AlphaFold,它通过多个深度学习模型的组合,成功解决了蛋白质折叠预测的问题。类似地,在材料科学中,可以开发集成了不同ML模型的框架,以有效处理从原子尺度到宏观性能的多尺度问题。

为了进一步提升复杂ML模型在材料科学中的应用,还需要着力增强其解释性和可解释性。这包括开发能够为模型预测提供直观理解的技术和工具,以及设计能够自动识别和解释数据中关键特征的算法。通过增强模型的解释性,研究人员可以更好地理解模型的工作原理,从而在模型设计和参数调整过程中做出更加明智的选择。

3.3 智能互联的自动化云实验室

随着AI技术在实验研究和材料发现领域展现出巨大潜力,赋予硅基智能系统以“手”的能力(负责材料合成、设备自动化的组装/拆卸及样品转移)和“眼”的功能(执行材料表征和多模态传感),已成为推进材料科学创新的关键趋势^[48]。这种发展不仅预示着材料发现研究向自主闭环的转变,而且随着AI系统的增强和复杂化,迫切需要引入模块化云实验室设施,以克服预算和空间的限制^[125]。这种转型将极大提高实验设备链的灵活性和可重组性,同时确保了多个自主实验室之间的高效互操作性。

实现这一目标需要建立一个综合性的AI网络系统,该系统融合实验与理论研究,旨在实现劳动力的有效

分配、经济的规模化及系统间的平衡。这种系统的建立是技术革新的体现,同时也是对当前科研工作模式的深刻改革,可以促使材料实验科学研究向更加自动化、智能化、互联化的方向发展。

智能互联的自动化云实验室将超越传统人工操作的限制,转而采用AI代理驱动的自动化实验设计、执行、数据收集与分析,以及新知识的生成和应用。在这一过程中,AI将不仅是执行者,更是决策者和创新者,不断推动材料科学领域的发展。然而,构建这样一个智能互联的自动化云实验室体系需要多学科间的合作,涉及AI技术、机器人学、云计算、材料科学等领域的共同努力。这不仅挑战现有的科研设施和流程,也要求科研人员学会如何高效地与AI系统协作完成科研任务。

3.4 科学大型语言模型的训练和评估

尽管LLMs在材料科学领域已经取得了显著进展,但必须认识到这一研究领域仍然处于初级阶段。技术层面上,LLMs面临着上下文长度限制和倾向于生成幻

觉及无效数据的问题。在针对专业科学领域大模型进行预训练上,存在着专业训练数据集规模较小、缺乏高质量和多模态数据的问题。

作为智能代理,LLMs已展示出在特定任务解决、合作和人机交互方面的卓越能力。然而,对它们的性能进行量化和客观评估依然是一项挑战,这涉及基础能力和领域专业能力的评估。基础能力评估包括规划、决策、检索和记忆等基本技能。尽管已经建立了一些评价框架,如AgentSims^[126]等,但在这一领域仍需要发展更为标准化和细化的评估指标和协议。针对领域专业能力的评估,开发基于特定执行环境和任务的基准测试,对推动领域特定LLM代理的研究和应用至关重要。为此,建立必要的数据集和评估机制是不可或缺的。

此外,还必须考虑到伴随LLM应用而来的伦理问题,包括处理敏感数据、数据安全和价值观对齐等方面。这些问题不仅关系到技术和科研的进步,也触及到LLM应用的社会责任和道德底线。

致谢 本文撰写过程中得到了深圳大学李连伟教授的指导,特此致谢。

参考文献

- Olson G B. Designing a new material world. *Science*, 2000, 288: 993–998
- Agrawal A, Choudhary A. Perspective: Materials informatics and big data: Realization of the “fourth paradigm” of science in materials science. *APL Mater*, 2016, 4: 053208
- Li G J. AI4R: The fifth scientific research paradigm (in Chinese). *Bull Chin Acad Sci*, 2024, 39: 1–10 [李国杰. 智能化科研(AI4R): 第五科研范式. 中国科学院院刊, 2024, 39: 1–10]
- Mou T, Pillai H S, Wang S, et al. Bridging the complexity gap in computational heterogeneous catalysis with machine learning. *Nat Catal*, 2023, 6: 122–136
- Bergerhoff G, Hundt R, Sievers R, et al. The inorganic crystal structure data base. *J Chem Inf Comput Sci*, 1983, 23: 66–69
- Jain A, Ong S P, Hautier G, et al. The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater*, 2013, 1: 011002
- 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). *JOM*, 2013, 65: 1501–1509
- Wang Z, Chen A, Tao K, et al. AlphaMat: A material informatics hub connecting data, features, models and applications. *npj Comput Mater*, 2023, 9: 130
- Jiang Y, Yu Z, Wang Y, et al. Screening promising CsV₃Sb₅-like kagome materials from systematic first-principles evaluation. *Chin Phys Lett*, 2022, 39: 047402
- He B, Chi S, Ye A, et al. High-throughput screening platform for solid electrolytes combining hierarchical ion-transport prediction algorithms. *Sci Data*, 2020, 7: 151
- Yang X, Wang Z, Zhao X, et al. MatCloud: A high-throughput computational infrastructure for integrated management of materials simulation, data and resources. *Comput Mater Sci*, 2018, 146: 319–333
- Landis D D, Hummelshøj J S, Nestorov S, et al. The computational materials repository. *Comput Sci Eng*, 2012, 14: 51–57
- Kim E, Huang K, Saunders A, et al. Materials synthesis insights from scientific literature via text extraction and machine learning. *Chem Mater*, 2017, 29: 9436–9444

- 14 Blaiszik B, Chard K, Pruyne J, et al. The materials data facility: Data services to advance materials science research. *JOM*, 2016, 68: 2045–2052
- 15 Huang B, von Rudorff G F, von Lilienfeld O A. The central role of density functional theory in the AI age. *Science*, 2023, 381: 170–175
- 16 von Lilienfeld O A, Müller K R, Tkatchenko A. Exploring chemical compound space with quantum-based machine learning. *Nat Rev Chem*, 2020, 4: 347–358
- 17 Huang B, von Lilienfeld O A. Quantum machine learning using atom-in-molecule-based fragments selected on the fly. *Nat Chem*, 2020, 12: 945–951
- 18 Zaspel P, Huang B, Harbrecht H, et al. Boosting quantum machine learning models with a multilevel combination technique: Pople diagrams revisited. *J Chem Theor Comput*, 2019, 15: 1546–1559
- 19 Cancès E, Dusson G, Kemlin G, et al. Practical error bounds for properties in plane-wave electronic structure calculations. *SIAM J Sci Comput*, 2023, 44: B1312–B1340
- 20 Prentice J C A, Aarons J, Womack J C, et al. The ONETEP linear-scaling density functional theory program. *J Chem Phys*, 2020, 152: 174111
- 21 Fiedler L, Modine N A, Schmerler S, et al. Predicting electronic structures at any length scale with machine learning. *npj Comput Mater*, 2023, 9: 115
- 22 Witman M D, Goyal A, Ogitsu T, et al. Defect graph neural networks for materials discovery in high-temperature clean-energy applications. *Nat Comput Sci*, 2023, 3: 675–686
- 23 Das S, Kanungo B, Subramanian V, et al. Large-scale materials modeling at quantum accuracy: *Ab initio* simulations of quasicrystals and interacting extended defects in metallic alloys. In: Proceedings of the the International Conference for High Performance Computing, Networking, Storage and Analysis, 2023. 1–12
- 24 Batra R, Song L, Ramprasad R. Emerging materials intelligence ecosystems propelled by machine learning. *Nat Rev Mater*, 2021, 6: 655–678
- 25 Pilania G. Machine learning in materials science: From explainable predictions to autonomous design. *Comput Mater Sci*, 2021, 193: 110360
- 26 Vasudevan R, Pilania G, Balachandran P V. Machine learning for materials design and discovery. *J Appl Phys*, 2021, 129: 070401
- 27 Xue Y H, He M W, Liu H, et al. Computer simulation-assisted design of the preparation of polymer chain-grafted nanoparticles (in Chinese). *Chin Sci Bull*, 2023, 68: 729–740 [薛耀红, 贺敏伟, 刘鸿, 等. 计算机模拟辅助设计聚合物接枝纳米粒子的构筑. 科学通报, 2023, 68: 729–740]
- 28 Rhone T D, Bhattarai R, Gavras H, et al. Artificial intelligence guided studies of van der Waals magnets. *Adved Theor Sims*, 2023, 6: 2300019
- 29 Purcell T A R, Scheffler M, Ghiringhelli L M, et al. Accelerating materials-space exploration for thermal insulators by mapping materials properties via artificial intelligence. *npj Comput Mater*, 2023, 9: 112
- 30 Bastek J H, Kochmann D M. Inverse design of nonlinear mechanical metamaterials via video denoising diffusion models. *Nat Mach Intell*, 2023, 5: 1466–1475
- 31 Yan D, Smith A D, Chen C C. Structure prediction and materials design with generative neural networks. *Nat Comput Sci*, 2023, 3: 572–574
- 32 Weiss T, Mayo Yanes E, Chakraborty S, et al. Guided diffusion for inverse molecular design. *Nat Comput Sci*, 2023, 3: 873–882
- 33 Liu Y, Yang Z, Yu Z, et al. Generative artificial intelligence and its applications in materials science: Current situation and future perspectives. *J Mater*, 2023, 9: 798–816
- 34 Agarwal G, Doan H A, Robertson L A, et al. Discovery of energy storage molecular materials using quantum chemistry-guided multiobjective bayesian optimization. *Chem Mater*, 2021, 33: 8133–8144
- 35 Zhang D, Yi P, Lai X, et al. Active machine learning model for the dynamic simulation and growth mechanisms of carbon on metal surface. *Nat Commun*, 2024, 15: 344
- 36 Kim C, Chandrasekaran A, Jha A, et al. Active-learning and materials design: The example of high glass transition temperature polymers. *MRS Commun*, 2019, 9: 860–866
- 37 Shen Y F, Pokharel R, Nizolek T J, et al. Convolutional neural network-based method for real-time orientation indexing of measured electron backscatter diffraction patterns. *Acta Mater*, 2019, 170: 118–131
- 38 Masubuchi S, Watanabe E, Seo Y, et al. Deep-learning-based image segmentation integrated with optical microscopy for automatically searching for two-dimensional materials. *npj 2D Mater Appl*, 2020, 4: 3
- 39 Xu Y, Wang X, Zhou H, et al. Real-time terahertz characterization for composite delamination using a lightweight CPU adaptive network. *Compos Part B-Eng*, 2022, 247: 110354
- 40 Lin X, Si Z, Fu W, et al. Intelligent identification of two-dimensional nanostructures by machine-learning optical microscopy. *Nano Res*, 2018, 11: 6316–6324
- 41 Ziletti A, Kumar D, Scheffler M, et al. Insightful classification of crystal structures using deep learning. *Nat Commun*, 2018, 9: 2775
- 42 Cui A, Jiang K, Jiang M, et al. Decoding phases of matter by machine-learning raman spectroscopy. *Phys Rev Appl*, 2019, 12: 054049
- 43 Zhao X, Yang T, Li B, et al. SwinGAN: A dual-domain Swin Transformer-based generative adversarial network for MRI reconstruction. *Comput Biol Med*, 2023, 153: 106513
- 44 Li X, Hu X, Chen X, et al. Spatial redundancy transformer for self-supervised fluorescence image denoising. *Nat Comput Sci*, 2023, 3: 1067–1080

- 45 Na J, Kim S J, Kim H, et al. A unified microstructure segmentation approach via human-in-the-loop machine learning. *Acta Mater*, 2023, 255: 119086
- 46 Yang Z, Yabansu Y C, Jha D, et al. Establishing structure-property localization linkages for elastic deformation of three-dimensional high contrast composites using deep learning approaches. *Acta Mater*, 2019, 166: 335–345
- 47 Tong Z, Wang L, Zhu G, et al. Predicting twin nucleation in a polycrystalline Mg alloy using machine learning methods. *Metall Mater Trans A*, 2019, 50: 5543–5560
- 48 Abolhasani M, Kumacheva E. The rise of self-driving labs in chemical and materials sciences. *Nat Synth*, 2023, 2: 483–492
- 49 Xu Y, Lu Y, Li J, et al. Effect of graphene quantum dot size on plant growth. *Nanoscale*, 2020, 12: 15045–15049
- 50 Roch L M, Häse F, Kreisbeck C, et al. ChemOS: Orchestrating autonomous experimentation. *Sci Robot*, 2018, 3: eaat5559
- 51 Cao L, Russo D, Felton K, et al. Optimization of formulations using robotic experiments driven by machine learning DoE. *Cell Rep Phys Sci*, 2021, 2: 100295
- 52 Du X, Lüer L, Heumueller T, et al. Elucidating the full potential of OPV materials utilizing a high-throughput robot-based platform and machine learning. *Joule*, 2021, 5: 495–506
- 53 Gongora A E, Snapp K L, Whiting E, et al. Using simulation to accelerate autonomous experimentation: A case study using mechanics. *iScience*, 2021, 24: 102262
- 54 Kusne A G, Yu H, Wu C, et al. On-the-fly closed-loop materials discovery via Bayesian active learning. *Nat Commun*, 2020, 11: 5966
- 55 Burger B, Maffettone P M, Gusev V V, et al. A mobile robotic chemist. *Nature*, 2020, 583: 237–241
- 56 MacLeod B P, Parlane F G L, Morrissey T D, et al. Self-driving laboratory for accelerated discovery of thin-film materials. *Sci Adv*, 2020, 6: eaaz8867
- 57 Langner S, Häse F, Perea J D, et al. Beyond ternary OPV: High-throughput experimentation and self-driving laboratories optimize multicomponent systems. *Adv Mater*, 2020, 32: 1907801
- 58 Li Z, Najeeb M A, Alves L, et al. Robot-accelerated perovskite investigation and discovery. *Chem Mater*, 2020, 32: 5650–5663
- 59 Shimizu R, Kobayashi S, Watanabe Y, et al. Autonomous materials synthesis by machine learning and robotics. *APL Mater*, 2020, 8: 111110
- 60 Vikram A, Brudnak K, Zahid A, et al. Accelerated screening of colloidal nanocrystals using artificial neural network-assisted autonomous flow reactor technology. *Nanoscale*, 2021, 13: 17028–17039
- 61 Mekki-Berrada F, Ren Z, Huang T, et al. Two-step machine learning enables optimized nanoparticle synthesis. *npj Comput Mater*, 2021, 7: 55
- 62 Reis M, Gusev F, Taylor N G, et al. Machine-learning-guided discovery of ^{19}F MRI agents enabled by automated copolymer synthesis. *J Am Chem Soc*, 2021, 143: 17677–17689
- 63 Tao H, Wu T, Kheiri S, et al. Self-driving platform for metal nanoparticle synthesis: Combining microfluidics and machine learning. *Adv Funct Mater*, 2021, 31: 2106725
- 64 Zhu Q, Huang Y, Zhou D, et al. Automated synthesis of oxygen-producing catalysts from Martian meteorites by a robotic AI chemist. *Nat Synth*, 2023, 3: 319–328
- 65 Sagmeister P, Ort F F, Jusner C E, et al. Autonomous multi-step and multi-objective optimization facilitated by real-time process analytics. *Adv Sci*, 2022, 9: 2105547
- 66 Zhao H, Chen W, Huang H, et al. A robotic platform for the synthesis of colloidal nanocrystals. *Nat Synth*, 2023, 2: 505–514
- 67 Tamura R, Tsuda K, Matsuda S. NIMS-OS: An automation software to implement a closed loop between artificial intelligence and robotic experiments in materials science. *Sci Technol Adv Mater*, 2023, 3: 2232297
- 68 Lunt A M, Fakhruddin H, Pizzuto G, et al. Modular, multi-robot integration of laboratories: An autonomous workflow for solid-state chemistry. *Chem Sci*, 2024, 15: 2456–2463
- 69 Ha T, Lee D, Kwon Y, et al. AI-driven robotic chemist for autonomous synthesis of organic molecules. *Sci Adv*, 2023, 9: eadj0461
- 70 Szymanski N J, Rendy B, Fei Y, et al. An autonomous laboratory for the accelerated synthesis of novel materials. *Nature*, 2023, 624: 86–91
- 71 Rauschen R, Guy M, Hein J E, et al. Universal chemical programming language for robotic synthesis repeatability. *Nat Synth*, 2024, 3: 488–496
- 72 Bateni F, Sadeghi S, Orouji N, et al. Smart dope: A self-driving fluidic lab for accelerated development of doped perovskite quantum dots. *Adv Energy Mater*, 2024, 14: 2302303
- 73 Slattey A, Wen Z, Tenblad P, et al. Automated self-optimization, intensification, and scale-up of photocatalysis in flow. *Science*, 2024, 383: eadj1817
- 74 Maffettone P M, Friederich P, Baird S G, et al. What is missing in autonomous discovery: Open challenges for the community. *Digital Discov*, 2023, 2: 1644–1659
- 75 Takahashi A, Terayama K, Kumagai Y, et al. Fully autonomous materials screening methodology combining first-principles calculations, machine learning and high-performance computing system. *Sci Technol Adv Mater*, 2023, 3: 2261834
- 76 Tan G M, Jia W L, Wang Z, et al. Computing system for simulation intelligence (in Chinese). *Bull Chin Acad Sci*, 2024, 39: 17–26 [谭光明, 贾伟

乐, 王展, 等. 面向模拟智能的计算系统. 中国科学院院刊, 2024, 39: 17–26]

- 77 AI4Science M R, Quantum M A. The impact of large language models on scientific discovery: A preliminary study using GPT-4. 2023, arXiv: [2311.07361](#)
- 78 Zhang Q, Ding K, Lyv T, et al. Scientific large language models: A survey on biological & chemical domains. 2024, arXiv: [2401.14656](#)
- 79 Jablonka K M, Ai Q, Al-Feghali A, et al. 14 examples of how LLMs can transform materials science and chemistry: A reflection on a large language model hackathon. *Digital Discov*, 2023, 2: 1233–1250
- 80 Beltagy I, Lo K, Cohan A. SciBERT: A pretrained language model for scientific text. 2019, arXiv: [1903.10676](#)
- 81 Hong Z, Ajith A, Pauloski J, et al. The diminishing returns of masked language models to science. In: Proceedings of the the Findings of the Association for Computational Linguistics: ACL 2023, 2023. 1270–1283
- 82 Xie T, Wan Y, Huang W, et al. DARWIN series: Domain specific large language models for natural science. 2023, arXiv: [2308.13565](#)
- 83 Lo K, Wang L L, Neumann M, et al. S2ORC: The semantic scholar open research corpus. 2019, arXiv: [1911.02782](#)
- 84 Cohan A, Feldman S, Beltagy I, et al. Specter: Document-level representation learning using citation-informed transformers. 2020, arXiv: [2004.07180](#)
- 85 Liu X, Yin D, Zheng J, et al. OAG-BERT: Towards a unified backbone language model for academic knowledge services. In: Proceedings of the the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2022. 3418–3428
- 86 Jeong Y, Kim E. SciDeBERTa: Learning DeBERTa for science technology documents and fine-tuning information extraction tasks. *IEEE Access*, 2022, 10: 60805–60813
- 87 Li Y, Zhang Y, Zhao Z, et al. CSL: A large-scale Chinese scientific literature dataset. 2022, arXiv: [2209.05034](#)
- 88 Yamauchi H, Kajiura T, Katsurai M, et al. A Japanese masked language model for academic domain. In: Proceedings of the the Third Workshop on Scholarly Document Processing, 2022. 152–157
- 89 Taylor R, Kardas M, Cucurull G, et al. Galactica: A large language model for science. 2022, arXiv: [2211.09085](#)
- 90 Hu D, Hou X, Du X, et al. VarMAE: Pre-training of variational masked autoencoder for domain-adaptive language understanding. 2022, arXiv: [2211.00430](#)
- 91 Jin B, Zhang W, Zhang Y, et al. Patton: Language model pretraining on text-rich networks. 2023, arXiv: [2305.12268](#)
- 92 Guo J, Ibanez-Lopez A S, Gao H, et al. Automated chemical reaction extraction from scientific literature. *J Chem Inf Model*, 2022, 62: 2035–2045
- 93 Chithrananda S, Grand G, Ramsundar B. ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction. 2020, arXiv: [2010.09885](#)
- 94 Ahmad W, Simon E, Chithrananda S, et al. Chemberta-2: Towards chemical foundation models. 2022, arXiv: [2209.01712](#)
- 95 Yin J, Dash S, Wang F, et al. FORGE: Pre-training open foundation models for science. In: Proceedings of the the International Conference for High Performance Computing, Networking, Storage and Analysis, 2023. 1–13
- 96 Trewartha A, Walker N, Huo H, et al. Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science. *Patterns*, 2022, 3: 100488
- 97 Zhao Z, Ma D, Chen L, et al. ChemDFM: Dialogue foundation model for chemistry. 2024, arXiv: [2401.14818](#)
- 98 Gupta T, Zaki M, Krishnan N M A, et al. MatSciBERT: A materials domain language model for text mining and information extraction. *npj Comput Mater*, 2022, 8: 102
- 99 Shetty P, Rajan A C, Kuenneth C, et al. A general-purpose material property data extraction pipeline from large polymer corpora using natural language processing. *npj Comput Mater*, 2023, 9: 52
- 100 Yoshitake M, Sato F, Kawano H, et al. MaterialBERT for natural language processing of materials science texts. *Sci Technol Adv Mater*, 2022, 2: 372–380
- 101 Mullick A, Ghosh A, Chaitanya G S, et al. MatSciRE: Leveraging pointer networks to automate entity and relation extraction for material science knowledge-base construction. *Comput Mater Sci*, 2024, 233: 112659
- 102 Gilligan L P J, Cobelli M, Taufour V, et al. A rule-free workflow for the automated generation of databases from scientific literature. *npj Comput Mater*, 2023, 9: 222
- 103 Burke L, Pazdernik K, Fortin D, et al. NukeLM: Pre-trained and fine-tuned language models for the nuclear and energy domains. 2021, arXiv: [2105.12192](#)
- 104 Huang S, Cole J M. BatteryBERT: A pretrained language model for battery database enhancement. *J Chem Inf Model*, 2022, 62: 6365–6377
- 105 Zhao J, Huang S, Cole J M. OpticalBERT and OpticalTable-SQA: Text- and table-based language models for the optical-materials domain. *J Chem Inf Model*, 2023, 63: 1961–1981
- 106 Frey N C, Soklaski R, Axelrod S, et al. Neural scaling of deep chemical models. *Nat Mach Intell*, 2023, 5: 1297–1305
- 107 Kato S, Kanegami K, Kano M. ProcessBERT: A pre-trained language model for judging equivalence of variable definitions in process models. *IFAC-PapersOnLine*, 2022, 55: 957–962

- 108 Liu P, Ren Y, Tao J, et al. GIT-Mol: A multi-modal large language model for molecular science with graph, image, and text. *Comput Biol Med*, 2024, 171: 108073
- 109 Balaji S, Magar R, Jadhav Y. GPT-MolBERTa: GPT molecular features language model for molecular property prediction. 2023, arXiv: [2310.03030](#)
- 110 Rubungo A N, Arnold C, Rand B P, et al. Llm-prop: Predicting physical and electronic properties of crystalline solids from their text descriptions. 2023, arXiv: [2310.14029](#)
- 111 Li J, Liu Y, Fan W, et al. Empowering molecule discovery for molecule-caption translation with large language models: A ChatGPT perspective. 2023, arXiv: [2306.06615](#)
- 112 Buehler M J. MechGPT, a language-based strategy for mechanics and materials modeling that connects knowledge across scales, disciplines, and modalities. *Appl Mech Rev*, 2023, 76: 021001
- 113 Bran A M, Cox S, White A D, et al. ChemCrow: Augmenting large-language models with chemistry tools. 2023, arXiv: [2304.05376](#)
- 114 Kang Y, Kim J. Chatmof: An autonomous ai system for predicting and generating metal-organic frameworks. 2023, arXiv: [2308.01423](#)
- 115 Boiko D A, MacKnight R, Kline B, et al. Autonomous chemical research with large language models. *Nature*, 2023, 624: 570–578
- 116 Darvish K, Skreta M, Zhao Y, et al. ORGANA: A robotic assistant for automated chemistry experimentation and characterization. 2024, arXiv: [2401.06949](#)
- 117 Yoshikawa N, Skreta M, Darvish K, et al. Large language models for chemistry robotics. *Auton Robot*, 2023, 47: 1057–1086
- 118 Balhorn L S, Caballero M, Schweidtmann A M. Toward autocorrection of chemical process flowsheets using large language models. 2023, arXiv: [2312.02873](#)
- 119 Jablonka K M, Schwaller P, Ortega-Guerrero A, et al. Leveraging large language models for predictive chemistry. *Nat Mach Intell*, 2023, 6: 161–169
- 120 Ramos M C, Michtav S S, Porosoff M D, et al. Bayesian optimization of catalysts with in-context learning. 2023, arXiv: [2304.05341](#)
- 121 Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. *Adv Neural Inf Process Syst*, 2022, 35: 24824–24837
- 122 Yao S, Zhao J, Yu D, et al. React: Synergizing reasoning and acting in language models. 2022, arXiv: [2210.03629](#)
- 123 Li K, Persaud D, Choudhary K, et al. Exploiting redundancy in large materials datasets for efficient machine learning with less data. *Nat Commun*, 2023, 14: 7283
- 124 Liu Y, Yang Z, Zou X, et al. Data quantity governance for machine learning in materials science. *Natl Sci Rev*, 2023, 10: nwad125
- 125 Hysmith H, Foadian E, Padhy S P, et al. The future of self-driving laboratories: From human in the loop interactive AI to gamification. *Digital Discov*, 2024, 3: 621–636
- 126 Lin J, Zhao H, Zhang A, et al. Agentsims: An open-source sandbox for large language model evaluation. 2023, arXiv: [2308.04026](#)

Summary for “人工智能与大数据在材料科学中的融合: 新范式与科学发现”

Integration of artificial intelligence and big data in materials science: New paradigms and scientific discoveries

Shuai Yang¹, Jianjun Liu³, Fan Jin^{4*} & Ying Lu^{1,2*}

¹ National Science Library (Chengdu), Chinese Academy of Sciences, Chengdu 610299, China;

² Department of Information Resources Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100049, China;

³ State Key Laboratory of High Performance Ceramics and Superfine Microstructure, Shanghai Institute of Ceramics, Chinese Academy of Sciences, Shanghai 200050, China;

⁴ Institute of Synthetic Biology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China

* Corresponding authors, E-mail: fan.jin@siat.ac.cn; luy@clas.ac.cn

Materials science plays an irreplaceable role in driving societal progress and technological innovation. With the rapid development of artificial intelligence (AI) and big data technologies, the research paradigm in materials science is undergoing profound transformations. This review discusses how the integration of AI and big data is reshaping the research paradigm in materials science (AI for materials science), accelerating the advancement of computational materials science, and innovating the experimental process.

It begins by outlining the infrastructure development of material databases in the context of big data, which serve as the cornerstone of scientific work by providing robust support for the storage, management, and analysis of material data. These databases facilitate efficient data handling, enabling researchers to extract valuable insights from vast amounts of experimental and simulation data. Subsequently, the review explores the application of AI technologies across different stages of the material discovery cycle, including theoretical calculations, experimental design, data collection, and synthesis. AI algorithms, particularly deep learning, have revolutionized these stages by enhancing the ability to process and analyze complex datasets, revealing intricate relationships between material structures and their properties.

A significant highlight of this review is the introduction of self-driving laboratories (SDLs). Resulting from the integration of AI with laboratory automation and robotics technology, SDLs have realized a complete closed-loop process for material discovery, promoting a significant shift towards autonomous scientific discovery models. These laboratories can independently design and execute experiments, analyze results, and iteratively refine hypotheses, greatly increasing the efficiency and accuracy of material discovery.

Furthermore, the development of large language models (LLMs) has brought about revolutionary changes in natural language understanding, leading to the emergence of scientific LLMs, thus expanding the capabilities from text understanding to scientific exploration. The review provides an overview of the latest advancements in LLMs within materials science, emphasizing their critical role in expediting the material discovery process. These models can parse and understand vast amounts of scientific literature, enabling researchers to stay abreast of the latest developments and identify novel research directions.

The review concludes by evaluating the challenges involved in building an intelligent ecosystem for material research. These challenges include the need for high-quality, standardized data, the integration of diverse AI tools, and the development of robust methodologies for cross-disciplinary collaboration. Despite these challenges, the substantial potential of AI in materials science is evident. AI technologies promise to transform material research, enabling the discovery of new materials with unprecedented speed and precision.

In summary, this review aims to inform researchers about the significance of AI in materials science, highlighting the transformative impact of AI and big data on the research paradigm. It underscores the importance of developing intelligent systems and methodologies to harness the full potential of AI, thereby advancing the field of materials science and contributing to technological innovation and societal progress.

artificial intelligence, material discovery, self-driving laboratory, large language model

doi: [10.1360/TB-2024-0404](https://doi.org/10.1360/TB-2024-0404)