储能新材料设计与先进表征专刊

数据驱动的机器学习在电化学储能材料研究中的应用

施思齐[1,2,5]，涂章伟[1]，邹欣欣[3]，孙拾雨[2]，杨正伟[3]，刘 悦[3,4]
（[1]上海大学材料科学与工程学院；[2]上海大学材料基因组工程研究院；[3]上海大学计算机工程与科学学院；[4]上海市智能计算系统工程技术研究中心，上海 200444；[5]之江实验室，浙江 杭州 311100）

摘 要：储能电池的关键是材料。继实验观测、理论研究和计算模拟之后，数据驱动的机器学习具有快速捕捉材料成分-结构-工艺-性能间复杂构效关系的优势，有望为电化学储能材料的研发提供新的范式。本文从结构化和非结构化数据驱动两方面，系统评述了机器学习在电化学储能材料研究中的最新进展。全面概括了可用于电化学储能材料机器学习的国内外材料数据库，分析了其数据的收集、共享和质量检测存在的问题；重点阐述了电化学储能材料中机器学习的工作流程和应用，包括结构化数据驱动下数据收集、特征工程和机器学习建模以及图形、表征图像和文献文本这类非结构化数据驱动下的模型构建和应用。进一步，厘清电化学储能材料领域机器学习面临的三大矛盾且给出对策，即高维度与小样本数据的矛盾与协调、模型准确性与易用性的矛盾与统一、模型学习结果与专家经验的矛盾与融合，并提出构建“领域知识嵌入的机器学习方法”有望调和这些矛盾。本文将为机器学习在电化学储能材料设计和性能优化中的应用提供参考。

关键词：电化学储能材料；机器学习；材料数据库；领域知识
doi：10.19799/j.cnki.2095-4239.2022.0051
中图分类号：TP 181 文献标志码：A 文章编号：2095-4239（2022）03-739-21

Applying data-driven machine learning to studying electrochemical energy storage materials

SHI Siqi[1,2,5], *TU Zhangwei*[1], *ZOU Xinxin*[3], *SUN Shiyu*[2], *YANG Zhengwei*[3], *LIU Yue*[3,4]
([1]School of Materials Science and Engineering, Shanghai University; [2]Materials Genome Institute, Shanghai University; [3]School of Computer Engineering and Science, Shanghai University; [4]Shanghai Engineering Research Center of Intelligent Computing System, Shanghai, 200444, China; [5]Zhejiang Laboratory, Hangzhou 311100, Zhejiang, China)

Abstract: Materials are key to energy storage batteries. With experimental observations, theoretical research, and computational simulations, data-driven machine learning should provide a new paradigm for electrochemical energy storage material research and development. Its advantages include rapid capture of the complex structure-activity relationship between material composition, structure, process, and performance. In this study, the latest developments in employing machine learning in electrochemical energy storage materials are reviewed systematically from structured and unstructured data-driven perspectives. The material databases from China and abroad are summarized for electrochemical energy storage material use, and data collection and quality inspection problems are analyzed. Data-driven machine learning workflows and applications in

收稿日期：2022-01-31；**修改稿日期：**2022-02-10。
基金项目：国家重点研发计划项目（2021YFB3802100），国家自然科学基金面上项目（52073169），之江实验室科研攻关项目（2021PE0AC02）。

第一作者：施思齐（1978—），男，教授，研究方向为电化学储能材料基础科学问题解析、计算方法发展和新材料设计，E-mail：sqshi@shu.edu.cn；**通讯作者：**刘悦，博士，教授，研究方向为机器学习、数据挖掘及其在材料领域的应用，E-mail：yueliu@shu.edu.cn。

electrochemical energy storage materials are demonstrated. They contain data collection, feature engineering, and machine learning modeling under structured data, and the model construction and application under unstructured data of graphics, representation images, and literature. Three principal contradictions and countermeasures faced by machine learning in electrochemical energy storage materials are highlighted: the contradiction and coordination of high-dimensional and small sample data, the contradiction and unity of model accuracy and ease of use, and the contradiction and contradiction fusion between model learning results and expert experience. We highlighted that “Machine Learning Embedded with Domain Knowledge” construction should reconcile these contradictions. This review provides a reference for applying machine learning in electrochemical energy storage materials’ design and performance optimization.

Key words: electrochemical energy storage materials; machine learning; materials database; domain knowledge

作为能源互联网的关键环节，以锂/钠离子电池为代表的储能电池正处在与信息产业深度融合的阶段，目前的发展目标是突破储能电池能量密度低、电池安全性差、大电流充放电能力不足以及使用寿命短等方面的瓶颈并进一步拓宽其应用场景[1]。然而，储能电池的综合性能受各类材料的复杂构效关系共同影响，如电极脱嵌锂过程的结构演化[2]、电解质的离子输运机制[3]和电极与电解质间的界面性质[4]等，这为储能电池的研发和性能提升带来了挑战。

早期的储能电池研发是基于经验主义的实验方法，涉及人工合成、材料表征和性能分析等步骤，耗时长且成本高。于是，研究人员进一步发展了基于物理化学定律的理论研究方法[5]，但该方法在解决许多电化学储能材料科学问题时往往过于复杂，难以求解。后来，随着材料科学、物理学和计算机科学的交叉与融合，微观−介观−宏观尺度的计算模拟方法逐渐兴起，包括第一性原理计算、分子动力学模拟、蒙特卡罗模拟、CALPHAD 方法、相场模拟和有限元模拟等[6]。这些方法涵盖了不同的空间和时间尺度范围，在可充电电池领域得到了广泛的应用[7-8]。然而，计算模拟方法依赖于材料的微观结构和高性能计算设备，其计算速度和准确性仍然受到限制，且该方法每次往往只能对材料的单一性能进行研究与优化，很难同时筛选或设计出综合性能优异的电化学储能材料。近几年，随着实验、理论和计算数据的大量积累以及高效、准确的人工智能技术的迅速发展，材料科学研究进入了第四科学范式[9-10]，即数据驱动的材料科学研究，有望实现储能电池的高效研发。

如图 1 所示，数据驱动的材料科学利用传统实验、理论和计算模拟方法积累的大量数据，借助数据驱动的人工智能方法对电化学储能材料的性能驱动机制进行建模和分析，以加速新型高性能电化学储能材料的研发与设计。目前，作为数据驱动的人工智能方法的典型代表之一，机器学习已经被广泛应用于材料的性能预测和新材料发现[11-18]。机器学习在电池领域的应用可以追溯到 1999 年 Salkind 等[19]使用模糊逻辑方法来确定电池的充电状态和健康状态。随后，Ceder 等[20-22]利用机器学习技术预测材料晶体结构并用于汽车电池锂基材料的发现。2011 年，美国政府提出了“材料基因组计划”(materials genome initiative，MGI)[23]，其目标之一便是通过机器学习方法将“实验”、“计算”和“数据”相结合，以快速开发出清洁能源系统的相关材料[24-25]。自此，以数据驱动的机器学习方法助力电化学储能材料研发的工作不断涌现出来。已有一些优秀综述从不同的角度介绍了电化学储能材料领域中机器学习的研究现状。例如，Guo 等[26]从材料原子建模的角度，介绍了机器学习在固态电池材料的势能函数构建、性能预测和逆向设计中的应用；陈翔等[27]从多尺度电池应用的角度，评述了机器学习与微观、介观和宏观尺度的理论或实验融合的方法在电池材料的研究现状；Lombardo 等[28]从材料研发到电池实际应用的角度，总结了机器学习在电池制造、材料表征和电池诊断等方面的研究进

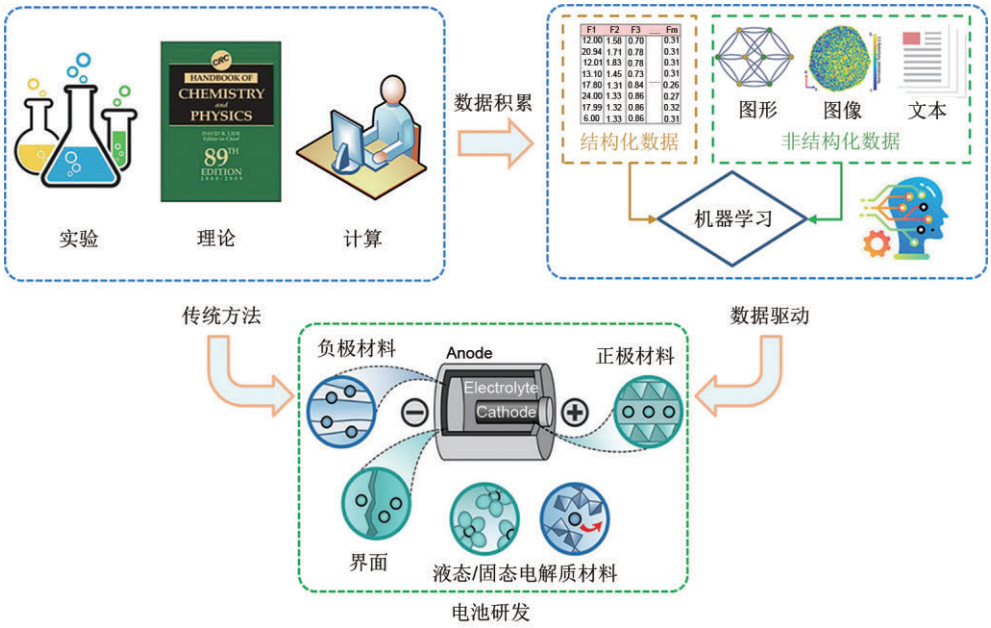

图1　电池研发四大范式：实验、理论、计算和数据驱动

Fig. 1　Four paradigms of battery research and development: Experimentation, theory, computation and data driven

展；刘悦等[29]从机器学习工作流程的角度，综述了机器学习在充电电池材料领域的应用现状，并分析和总结了机器学习方法在材料领域应用普遍面临的三大挑战问题和相应的解决策略。

在MGI的推动下，电化学储能材料数据被不断产生和积累，包括结构化数据和非结构化数据。其中，结构化数据一般能够形式化存储在数据表格中，且每列都有具体的含义；非结构化数据则通常指结构化数据之外的一切数据，包括节点和边组成的图形数据、像素点组成的图像数据和字符组成的文本数据。然而，利用这些异构数据来驱动电化学储能材料的研发，对机器学习建模过程中的数据表示、模型选择、评估与应用提出了新的挑战。本文以不同类型数据驱动的机器学习在电化学储能材料研发中的应用为主线，全面介绍了可用于电化学储能材料研究的材料数据资源，并指出了其未来发展方向；重点总结了结构化数据驱动下机器学习的工作流程及其在电极和电解质材料的性能预测与成分优化、电池健康状态评估的应用现状，以及非结构化数据驱动下机器学习在材料性能预测、表征图像分析和文献文本挖掘等方面的相关工作；系统厘清了机器学习在电化学储能材料领域应用所面临的三大矛盾，并结合机器学习的最新发展提出了相应的调和策略；最后，对全文内容进行了总结。

1　电化学储能材料的数据资源

过去若干年里，全世界范围内材料研究学者们通过实验测量和计算模拟积累了大量的材料数据，由此建立了大量可用于电化学储能材料研究的涵盖材料结构与性能的数据库(表1)。电化学储能材料中重要的性能如脱/嵌锂电位、热力学稳定性和化学稳定性等均可从密度泛函计算得到的能量、电子结构等信息中获得，因此包含这些信息的通用型材料数据库都可用于电化学储能材料本征性质的研究[30]。从这些材料数据库中能够得到电化学储能材料的实验或计算的原始数据，为数据驱动的机器学习提供样本。

实验测量作为沿用至今的材料科学研究关键手段之一，对材料的研发起着至关重要的作用。科学工作者们通过对文献中实验测量数据的收集，建立了一些材料数据库，其中包含了化学组成、材料结构、文献引用等基本信息。剑桥结构数据库(cambridge structural database，CSD)由英国剑桥大学Kennard等在1965年创建，从文献中收录了115万种小分子有机物和金属有机化合物晶体结构数据，其中包含了晶胞参数、原子坐标和引用文献等[31-33]。德国波恩大学Bergerhoff等[34]在1983年创建了无机晶体结构数据库(inorganic crystal structure database，ICSD)来作为剑桥结构数据库

表1 主要的材料数据库及其数据特点

Table 1 Main material databases and data characteristics

数据库名称	数据库特点	数据来源	网址
CSD	小分子有机物和金属有机化合物晶体结构数据	实验测量	https://www.ccdc.cam.ac.uk/
ICSD	无机晶体结构数据	实验测量	https://icsd.products.fiz-karlsruhe.de/
Pauling file	无机晶体材料、相图和物理性能	实验测量	https://www.paulingfile.com/
Materials Project	无机晶体材料、分子、纳米孔隙材料、嵌入型电极材料和转化型电极材料以及材料性能	ICSD/计算模拟	https://materialsproject.org/
AFLOWlib	无机晶体材料、二元合金、多元合金以及材料性能	ICSD/计算模拟	http://aflowlib.org/
OQMD	无机晶体材料以及热力学和结构特性	ICSD/计算模拟	http://www.oqmd.org/
材料学科领域基础科学数据库	金属材料和无机非金属材料	实验测量	http://www.matsci.csdb.cn/
国家材料科学数据共享网	各类材料体系数据	实验测量/计算模拟	http://www.materdata.cn/
材料基因工程数据库	各类材料体系数据	实验测量/计算模拟	https://www.mgedata.cn/
Atomly	无机晶体结构以及材料性能	ICSD/计算模拟	https://atomly.net/#/matdata
电池材料离子输运数据库	无机晶体材料以及离子输运性能	ICSD/计算模拟	http://e01.iphy.ac.cn/bmd/
电化学储能材料高通量计算平台	无机晶体材料以及离子输运性能和机器学习描述符	ICSD/计算模拟	https://matgen.nscc-gz.cn/solidElectrolyte/

的补充，收录了1913年以来出版的21万多条实验表征的无机晶体结构详细信息，包含化学名称、化学式、矿物名、晶胞参数、空间群、原子坐标、原子占位及文献引用等[35]。1995年，日本科学技术厅等[36]单位合作组建了Paulina Film项目，收集了从1900年至今超过35000种出版物中的无机材料数据，包含了35万个晶体结构、5万个相图和15万条物理性能。为了有效地应用和积累科学数据，我国在1987年由中国科学院牵头正式启动科学数据资源建设。其中，中国科学院金属研究所承建的“材料学科领域基础科学数据库”，(http://www.matsci.csdb.cn/)拥有金属材料数据6万余条和无机非金属材料数据1万余条，涵盖了材料的热学、力学和电学等各种性能。2001年我国开始逐步启动科学数据共享工程，其中北京科技大学建设的“国家材料科学数据共享网”(http://www.materdata.cn/)汇集了全国30余家科研单位包括有色金属材料、有机高分子材料和能源材料等超过60万条材料科学数据。虽然这些基于实验测量的材料数据库记录的数据可靠且直观，但是获得这些数据的成本高昂。

随着计算机算力的提升，材料研究模式开始以“经验试错法”到基于“材料基因”设计方法转变，期间催生了许多高通量材料计算平台和数据库。劳伦斯伯克利国家实验室Ceder等[37]在2011年创立Materials Project数据库，存储了75万多种材料，涉及无机化合物、分子、纳米孔隙材料、嵌入型电极材料和转化型电极材料以及包括9万多条能带结构、弹性张量、压电张量等性能的第一性原理计算数据。2012年，杜克大学Curtarolo等[38]发布了AFLOWlib计算材料数据库，存储了包括无机化合物、二元合金与多元合金等超过356万种材料结构和7亿条第一性原理计算的材料性能数据，是诸多数据库中数据量最大的一个。2013年，西北大学Wolverton等[39]推出了开放量子材料数据库(open quantum materials database，OQMD)，通过DFT计算了102万种材料的热力学和结构特性，其中以钙钛矿数据居多。以上三个数据库的数据都是从无机晶体结构数据库衍生而来，不同之处在于其所包含的虚拟材料的数量[40]。相比于国外，国内的材料计算数据库发展较晚。2016年，北京科技大学牵头建立的“材料基因工程专用数据库”(http://www.mgedata.cn/)，包含超过76万条催化材料、特种合金及其材料热力学和动力学等数据；2020年，中国科学院物理研究所等单位创建的Atomly数据库(http://atomly.net/#matdata)，包含从ICSD数据库和DFT计算得到的18万个无机晶体结构并计算其详细的电子结构信息以及热力学相图。这些基于计算的数据库拥有着庞大的数据量，使得数据驱动的材料研究得到迅速的发展。

然而，电化学储能材料的研发需要考虑离子输运性质、能量密度、充放电速率等特定的材料性能，上述通用数据库往往不能满足这些需求。因此，

专门为电化学储能材料建立的数据库开始被研究与使用。例如，中国科学院物理研究所在2018年推出了电池材料离子输运数据库(http://eol.iphy.ac.cn/bmd/)，采用键价方法计算得到了2万多条无机晶体化合物离子迁移势垒数据，可快速筛选已知结构化合物中离子迁移势垒较低的潜在快离子导体。本课题组于2020年发布了电化学储能材料高通量计算平台(https://matgen.nscc-gz.cn/solidElectrolyte/)，集成了晶体结构几何分析(CAVD)[41]、键价和计算(BVSE)、多精度融合算法[42]和相稳定性计算等程序，并基于CAVD和BVSE构建了包含2.9万条数据的离子输运特性数据库[43]，能够为下游的机器学习任务提供相应的学习样本，如图2所示。为构建全面系统的电化学储能材料数据库，本团队正在引入相图计算、蒙特卡洛、相场模拟和连续介质等模块以进一步为该领域的研究提供技术支持。

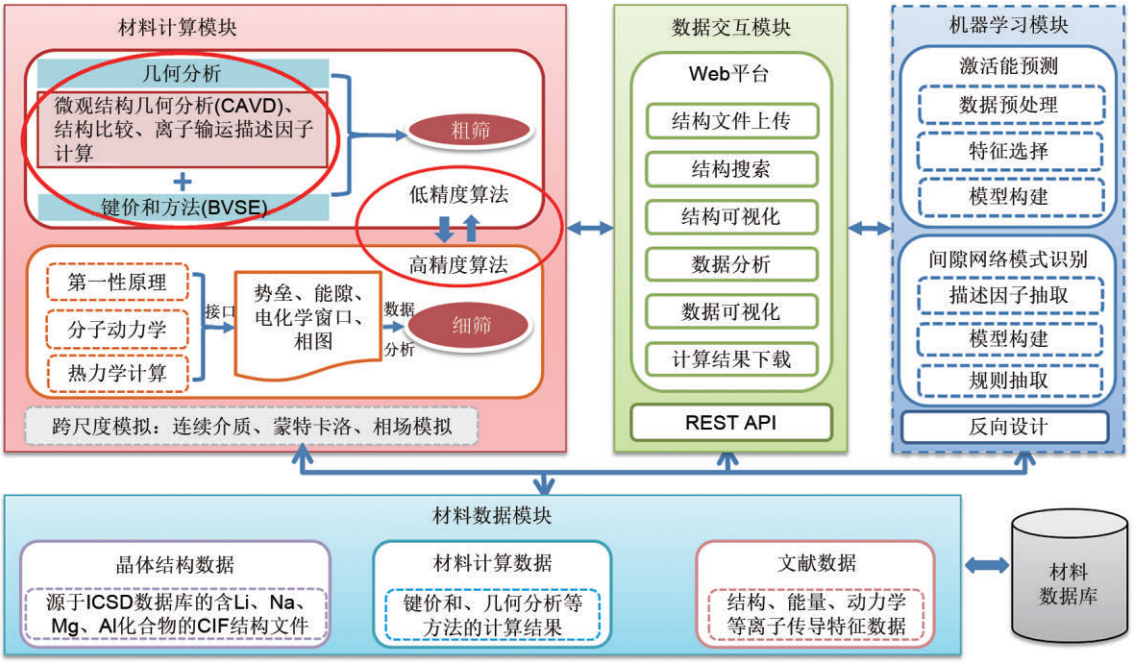

图2　电化学储能材料高通量计算平台总览

Fig. 2　Overview of high-throughput computing platform for electrochemical energy storage materials

综上所述，国内外各研究机构和团队建立了各种各样的通用和专用材料数据库，为数据驱动的电化学储能材料研发提供了丰富的数据资源。进一步，为支持数据驱动的电化学储能材料研发，还可以在以下三个方面对数据库建设进行完善。

第一，目前建立的电化学储能材料相关数据库收集的数据资源不够系统全面，无法满足储能电池的研发需求。一方面，在研究过程中只有小部分结果理想的数据被发表了出来，还存在大量失败实验数据并未公开，这些反例数据已经被证明能够辅助机器学习发现新材料[44]。因此，在搜集成功数据的同时，可以鼓励研究人员有针对性地将失败的实验结果保留。另一方面，现有电化学储能材料数据库中的材料数据尺度单一，储能电池的综合性能不仅与材料的本征性质相关，也与材料的微观形貌、外界环境场及器件的宏观构造等因素相互耦合[30]。因而可以建立电化学储能材料DFT计算参数库、分子动力学模拟参数库、相场模拟参数库、组分表征数据库、表界面数据库和结构表征数据库，为机器学习在电化学储能材料的应用提供多尺度数据。

第二，上述数据库主要包含了材料的结构和性能数据，通常由材料专家从中提取结构化数据或者把材料结构表示为非结构化图形数据作为机器学习模型的数据集。对于图像和文本类型的非结构化数据还无法从已有的材料数据库中获取。图像数据主要储存在材料测试机构中，通常无法公开获取。文本数据分散在各大材料科学出版物中，从海量文献中标记集成可用于机器学习的数据非常困难，且几乎没有开源具有标注信息的材料文本数据集。因此，有必要建立开源的材料图像数据库和文本数据库，推

动非结构化数据驱动的电化学储能材料研发应用。

第三，对于数据的使用者来说，数据的质量决定着机器学习模型的上限。实验测量的数据质量主要受材料缺陷、污染物和实验条件以及实验设备的不确定性影响；计算模拟的数据质量主要与计算模拟方法本身的精度相关。在数据集成过程中，不同来源数据的误差相结合，使得材料数据的质量更加难以确定[45]，如晶体的形成能，其计算值和实验值显著不同[46]。此外，研究人员在实验或计算过程中关注的参量具有差异性，收集材料数据时可能存在数据记录不一致的问题，造成了数据集的稀疏性。因而急需设计电化学储能材料数据质量检测方法，以提升机器学习模型的性能。

总之，通过上述方案能够优化完善材料数据资源，为材料数据和领域知识创建可持续的生态系统，从而促进数据驱动下的电化学储能新材料发现。

2 电化学储能材料数据驱动的机器学习

本节将对电化学储能材料中结构化和非结构化数据驱动的机器学习建模和应用分别进行系统地介绍，重点分析其存在的困难和挑战。

2.1 结构化电化学储能材料数据驱动的机器学习

目前机器学习在电化学储能材料领域的应用大多数是基于结构化数据驱动的，这需要针对特定的目标属性选取合适的描述符，并对其进行结构化表示，构建学习样本，进行机器学习建模和应用。具体工作流程如图3所示。首先，可以使用实验测量、计算模拟或者直接从现有的材料数据库中收集材料原始数据，并从中提取合适的描述符，这些描述符一般包括材料结构、化学成分和材料性能等；其次，经过特征选择或者特征转换将描述符数据集转换为学习样本；然后，通过选择合适的机器学习算法并调整最优超参数，模拟条件属性与目标属性之间的映射关系；最后，研究人员可以利用这些模型来预测材料的性质或指导新材料的发现，如液态/固态电解质和电极材料的性能预测与成分优化以及电池健康状态评估。

2.1.1 数据收集

电化学储能材料内部的微观结构与材料性能之间的关系纷繁复杂，任何一种性能都与多种因素耦合相关。从实验或者计算中收集到与目标属性相关的材料原始数据之后，还需要从中选取合适的描述符构建数据集。一般来说，相似的材料对应的描述符也要相似且数量和获取成本尽可能低[47]。然而，

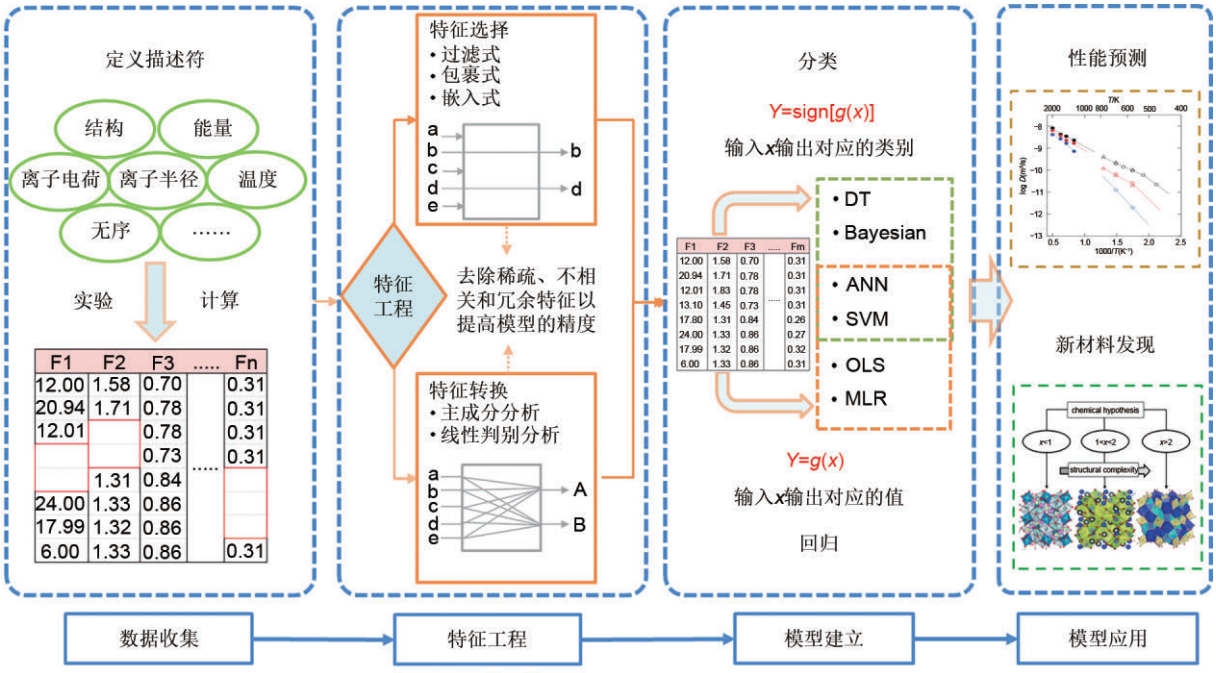

图3 结构化数据驱动的机器学习在电化学储能材料应用的工作流程[29]

Fig. 3 Workflow of structured data-driven machine learning in energy storage material application [29]

目前还没有普遍认可的描述符选择方法，其很大程度上依赖于研究者的领域知识。

针对特定的性能选取合适的描述符有助于建立更精确的模型，从而实现对电化学储能材料性能的精准预测。Sendek 等[48]根据原子的位置、质量、电负性和半径计算了与离子导电性相关的 20 个表征晶体局域原子排列和化学环境的描述符，进而利用逻辑回归算法对锂离子电池固体电解质离子电导率的高低进行分类；赵倩等[49]基于离子传导相关因素的分析，通过整合全局及局域离子传导环境对离子传导快慢的影响，构建了一套分层编码晶体结构基描述符框架，包含组成、结构、传导通道、离子分布和特殊离子 5 个部分共 32 个描述符，并采用偏最小二乘分析(PLS)方法成功地预测了立方相 Li-Argyrodites 的激活能；王爱平等[50]提取了有机溶剂小分子性质、最高占据分子轨道、最低未占据分子轨道和偶极矩以及官能团的原子性质共 13 个描述符，使用梯度提升决策树(GBDT)预测了溶剂与 LiOH 分子的结合能，发现磷酸酯溶剂能够显著加快 Li-O 电池的反应动力学。这些工作都是以目标属性为导向，依靠材料专家对材料体系的认知来选取的描述符。

材料专家针对不同材料性质所选取的描述符往往不能完全通用，这导致描述符的可扩展性差。为了将无机材料原始数据转换为机器学习算法所需的学习样本，Ward 等[51]根据材料的物理和化学性质提出了一套通用的描述符计算框架，包括化学计量属性、元素属性统计、电子结构属性和离子化合物属性共 145 个描述符。这些描述符在电化学储能材料性能预测研究中已经得到了成功的应用[52-55]。例如，Rajendra 等[52]通过上述框架得到 273 个描述符，开发了预测电极电压的机器学习模型，为钠/钾离子电池筛选了近 5000 种候选电极材料；Jo 等[53]和 Choi 等[54]利用上述框架和 Voronoi 镶嵌方法[56]分别提取了 145 个化学描述符和 126 个结构描述符并构建机器学习模型来预测钠离子固态电解质的力学性能；Verduzco 等[55]通过选取元素属性、元素分数、化学计量属性、价轨道和实验温度共 105 个描述符设计了基于随机森林的主动学习方法，用于预测石榴石型固态电解质离子电导率。上述工作证明了该描述符计算框架在无机材料性能预测的适用性。

为了提高描述符的计算效率，一些研究人员开发了计算工具包对现有的描述符计算方法进行集成。如 Ward 等[57]结合前期的工作基础[51]开发了基于 Python 的特征生成方法库 Matminer，其中包含了 47 个不同的特征提取模块，能够生成数千个物理相关的描述符，大大降低了描述符计算的难度。Himanen 等[58]创建了一个对原子结构进行编码的描述符库 DScribe，包含库仑矩阵[59]、Ewald 和矩阵[60]、正弦矩阵[60]、多体张量表示(MBTR)[61]、原子中心对称函数(ACSF)[62]和原子位置平滑重叠(SOAP)[63]等结构描述符，并通过周期性晶体的形成能和有机分子的离子电荷预测来说明其适用性。

总的来说，上述工作的推出加速了结构化描述符的构建，为后续的机器学习模型提供了可靠的数据集。但是，目前材料样本量少且描述符的选取存在稀疏性、不相关性和冗余性导致小样本高维度问题，从而影响模型的性能。此外，尽管目前已经开发了一些集成式的描述符计算工具，但是储能材料性能影响因素的复杂性导致能够适用于任意目标属性的通用描述符提取方案还未实现。

2.1.2　特征工程

由于描述符的选择往往取决于材料专家知识，这些描述符通常存在稀疏性、不相关性和冗余性，导致模型性能较差。因此，特征工程是机器学习模型构建中的一个重要步骤，包括特征转换和特征选择。特征转换是把高维特征空间映射到低维特征空间的方法，在降低特征维度的同时特征数值也会改变。特征选择是从全部特征中选择一个特征子集，以降低样本维度，进而提高机器学习模型的预测精度和泛化性能。目前，已有学者从数据的角度利用现有的统计或机器学习方法进行纯数据驱动的特征转换或选择，试图从电化学储能材料众多描述符中挑选出材料可解释、预测精度高的描述符。

2.1.2.1　特征转换

特征转换方法主要有主成分分析[64]和线性判别分析[65]。主成分分析通过线性投影并使得所投影的维度上数据的方差最大，以降低数据集的维数、提高可解释性的同时最大限度地减少信息丢失[64]。线性判别分析是将一个高维空间中的数据投影到一个较低维的空间中，且投影后要保证各个类别的类内方差小而类间均值差别大[65]。这两种方法一般用于储能电池的系统诊断。如 Banguero 等[66]将主成分分析模型应用于与电池储能系统的容量、内阻和开

路电压相关的参数集处理；Wang 等[67]利用主成分分析对电动汽车动力电池一致性多参数评价；Chen 等[68]基于线性判别分析的分类模型识别锂离子电池故障。

2.1.2.2 特征选择

特征选择方法可以分为过滤式、包裹式和嵌入式三大类[69]。过滤式特征选择方法使用基于统计理论和信息论的评分标准(例如距离函数、统计相关系数和互信息等)评估相关特征的重要性并进行排序，然后在机器学习模型中使用得分高的特征子集[70]，如图 4(a)所示。该方法具有简单和高效的优点，然而，其特征选择过程与机器学习模型分离，忽略了所选特征子集对模型性能的影响，这通常会导致模型的预测精度较低[71]。包裹式特征选择方法首先根据预先定义的搜索策略(如穷举法，遗传算法等)生成若干初始候选特征子集，其次训练一个特定的机器学习模型来评估每个候选特征子集，保留一些候选特征子集并用于生成下一组特征子集，该过程反复进行，直到选定的特征子集满足迭代停止条件[72](模型预测精度或循环次数)，如图 4(b)所示。该方法能够选择出具备高精度预测性能的最优特征子集，但往往以计算时间和复杂度为代价[73]。与包裹式方法类似，嵌入式方法同样与特定的机器学习模型绑定。但不同的是，该方法通过在目标函数和建模过程中引入正则化系数或随机因素实现模型构建和特征选择的协同(例如偏最小二乘分析、LASSO 和随机森林)，简化了特征选择的过程，但受限于特定的机器学习模型，普适性有待提高[74]。

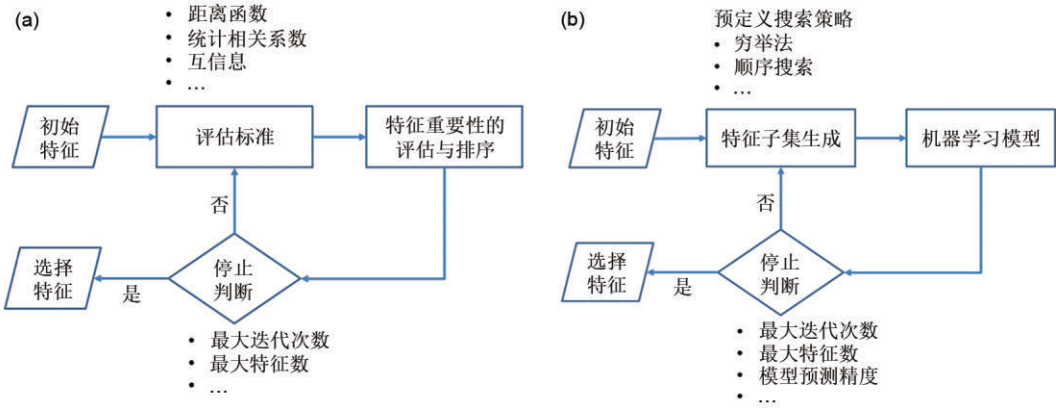

图 4 特征选择方法工作流程[75]：(a) 过滤式；(b) 包裹式

Fig. 4 Workflow of feature selection method[75]: (a) filter; (b) wrapper

在电化学储能材料性能预测研究中，包裹式方法由于考虑了特征对模型性能的影响已被广泛地应用。例如，Sendek 等[48]采用穷举策略从 20 个结构化描述符中选择了 5 个描述符，利用逻辑回归对锂离子电导率的高低进行分类；Gharagheizi 等[76]采用顺序搜索策略成功筛选出 10 个关键描述符，并建立最小二乘支持向量机(LSSVM)模型预测离子液体电导率；Wu 等[77]利用顺序搜索方法从 111 个描述符中选择了 23 个关键描述符，采用高斯核岭回归模型预测 FCC 溶质扩散势垒。嵌入式方法在选择特征的同时可以根据特征的重要性进行排序，使得专家可以更有针对性地进行材料设计，对于电化学储能材料的研究有着重要意义。例如，Shandiz 等[78]为 339 条硅酸盐阴极材料样本构建了 9 个描述符，利用极大随机化树(ERT)预测其晶系结构，发现晶胞体积是最重要的特征。赵倩等[49, 79]通过分层编码晶体结构描述符为 50 条立方相 Li-Argyrodites 样本构建了 32 个描述符，并借助偏最小二乘分析(PLS)方法推断各描述符与激活能之间的因果关系。

另外，过滤式和包裹式方法组合也是一种有效的特征选择方法，这种方法可以从数据的不同角度对特征进行处理[80]。例如 Hsu 等[81]先通过计算效率高的过滤器从原始数据集中选择候选描述符，然后通过更准确的包裹器进一步优化得到训练样本。在电化学储能材料研究领域，刘悦等[75]首次提出了一种融合加权评分领域专家知识的多层级特征选择方法，其方法框架如图 5 所示。该方法将过滤式和包裹式方法相结合自动去除稀疏、不相关和冗余特征，在特征选择过程中引入领域专家知识，消除了关键特征被删除的风险，并在四个电池材料数据集

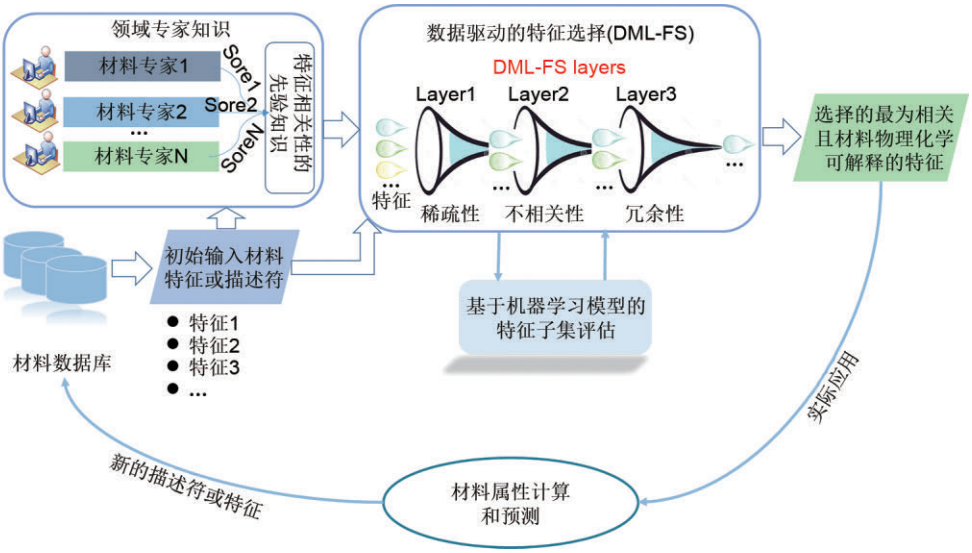

图5　融合加权评分领域专家知识的多层级特征选择方法框架[75]

Fig. 5　Multi-level feature selection method framework combining weighted scoring domain expert knowledge[75]

上进行了实验，显示出比其他方法更好的预测性能。

总之，许多结构化数据不仅维数高且样本量小，导致机器学习模型的过拟合，降低了模型的泛化能力。这也是电化学储能材料科学中需要特征工程的重要原因。然而，由于特征选择方法复杂多样，且涉及的超参数和策略也需要手动设置和调整。例如过滤式方法需要设置所选特征的数量和过滤阈值；包裹式方法需要指定子集搜索策略以生成候选特征子集；嵌入式方法需要优化机器学习算法的超参数以获得更好的性能。这将导致没有相关经验的材料专家不易使用这些方法。另外，特征工程仅仅通过特征空间的分布来选择描述符，这可能使得一些关键描述符重要度被弱化，导致学习结果与领域知识不一致。

2.1.3　机器学习建模与应用

目前，机器学习在储能电池领域得到了广泛的应用，其优越性在时间效率和预测精度上都得到了证明。其中，各种算法具有不同的特点和适应范围，选择合适的机器学习算法是构建机器学习模型的关键步骤，这极大地影响了其预测的准确性和泛化能力[82]。当前常用于储能电池研发的机器学习方法如表2所示。下面介绍这些方法在储能电池应用中的最新进展。

2.1.3.1　液态电解质研究中的应用

液态电解质是电池的重要组成部分，它在正负极之间传输离子的同时也起着阻碍电子传导的作用，对电池的性能至关重要[96]。机器学习已经被成功用于液态电解质化学稳定性、离子与溶剂的配位能预测以及溶剂成分优化。化学成分之间的稳定性和兼容性是在配置电解液时需要考虑的基本参数，Lee 等[93]将机器学习方法与传统反应指数相结合开发了神经网络回归模型并准确预测了93种电解质溶剂和氧化还原介质之间化学稳定性。离子与溶剂的配位能是影响离子输运的重要因素之一[97]，Ishikawa 等[89]计算了70种溶剂分别与5种碱族元素的配位能，选择了13个离子和溶剂相关描述符并采用高斯过程回归预测元素配位能。电解质添加剂及其成分的优化组合是实现高压电池长循环性能的有效方法，Duong 等[94]选择电解质添加剂比例、负极和正极的容量比及循环次数作为输入参数，使用人工神经网络模型预测电池容量并成功地找到性能优异的电解质成分。

2.1.3.2　固态电解质研究中的应用

近年来，固态电解质因比液态电解质具有更好的安全性、更高的能量密度和更长的寿命备受关注[1]。利用机器学习对其离子电导率、化学组成和带隙进行预测是一个研究热点。离子电导率是衡量一种材料是否可以用作固体电解质的重要指标之一，Xu 等[83]收集127种实验合成的 NASICON 和 LISICON 材料并利用包裹式特征选择方法选取7个简单描述符，然后通过逻辑回归模型预测离子电导

表2 常用于储能电池研发的机器学习模型对比

Table 2 Comparison of machine learning models commonly used in energy storage battery research and development

方法	简介	优点	缺点	适用范围	相关文献
逻辑回归	面向分类问题，建立代价函数，然后通过优化方法迭代求解出最优的模型参数	简单高效；可解释性强	容易欠拟合；对于异常值和缺失值敏感	线性可分数据	[47, 82]
偏最小二乘分析	通过最小化误差的平方和找到一组数据的最佳函数匹配	计算简单；预测精度高；易于定性解释	降维导致信息损失	小样本数据	[48]
决策树	一种树形结构，其中每个内部节点表示一个属性上的判断，每个分支代表一个判断结果的输出，最后每个叶节点代表一种分类结果	计算简单，易于理解，可解释性强	容易过拟合	稀疏性数据、不相关性数据	[49]
随机森林	以决策树作为基学习器，通过构建和组合多个弱学习器来完成学习任务	抗过拟合能力强；对缺失数据不敏感	噪音敏感	高维度、小样本、非均衡数据	[54]
支持向量机	通过最小化寻求结构化风险以提高学习器的泛化能力，实现经验风险和置信范围的最小化	使用核函数可以解决非线性的分类回归	计算耗时；对参数和核函数的选择比较敏感	高维数据	[75, 83-86]
岭回归	一种改良的最小二乘法，在MLR基础上加了L2正则项	稳定性较好	特征之间为稀疏的线性关系时效果差	多重共线性数据、病态数据	[76, 83]
朴素贝叶斯	基于贝叶斯定理和特征条件独立假设的分类方法，属于生成式模型	能个处理多分类任务；算法简单	需要计算先验概率；对输入数据的表达形式敏感	小样本数据；稀疏性数据	[87]
高斯过程回归	使用高斯过程先验对数据进行回归分析的非参数模型	预测值是观察值的插值；预测值具有概率评估	在高维空间效果差	时间序列数据	[85, 88-90]
XGBoost	经过优化的分布式梯度提升库，旨在高效、灵活且可移植	收敛速度快；内置交叉验证	在高维空间效果差	稀疏性数据	[91]
人工神经网络	通过调整内部大量节点之间相互连接的关系，以达到处理信息的目的	鲁棒性和容错能力强；能充分逼近复杂的非线性关系	参数复杂；不易解释	复杂非线性数据	[92-94]

率。提高材料离子电导率的常见策略是掺杂添加剂或混合异质材料，Homma等[90]通过实验制备了15个多相三元Li_3PO_4-Li_3BO_3-Li_2SO_4混合物样本，使用高斯过程回归的贝叶斯优化成功找到离子电导率性能优异的三元相化学组成比例。此外，带隙也是影响固体电解质性能的关键因素，Wang等[92]从Materials Project数据库中收集了286个具有计算带隙的石榴石结构并选取28个描述符来训练XGBoost模型，最后筛选出12个潜在的石榴石型固态电解质材料。

2.1.3.3 电极材料研究中的应用

电极材料的平均电压、体积变化、界面反应能、初始放电容量、库仑效率和电极制造参数对电池综合性能有着重要的影响，通过实验和计算来获得这些性质总是困难和昂贵的，因此有必要通过机器学习对其进行预测。电极材料的平均电压和充放电时的体积变化分别影响着电池的能量密度和安全性能，Moses等[95]从Materials Project数据库收集了4860个材料，通过电极的化学计量以及Matminer工具包生成了306个描述符，使用神经网络模型预测电极材料充放电时的平均电压和体积变化。为了探寻锂金属负极的枝晶生长和高反应性导致电池循环效率低和安全性差的原因，刘波等[84]计算了100种LLZOM化合物的界面反应能，将掺杂元素的15个相关特性视为描述符，通过支持向量集和核岭回归模型对界面稳定性和反应能进行准确预测。富锂层状氧化物正极材料在充放电过程中由于不可逆相变导致其结构稳定性降低、容量衰减和电压下降，Kireeva等[85]收集了99种富锂层状氧化物正极材料，选择化学成分、原子性质、合成方法和实验条件作为描述符，使用支持向量机模型成功预测了富锂层状氧化物的初始放电容量和库仑效率。

此外，由于电极的制造过程、电极浆料特性和涂层参数强烈影响电池的性能和寿命，Duquesnoy 等[88]制备了 144 个涂层电极，通过主成分分析、K 均值聚类和高斯朴素贝叶斯分类器方法相结合，从而预测了与特定制造参数相关的电极异质性。

2.1.3.4　电池健康状态评估中的应用

电池健康状态的评估对于电池系统的平稳可靠运行至关重要。而电池老化是一个复杂的过程，涉及阳极、阴极和电解质/电极界面的许多电化学反应。另外，温度和负载等操作条件也会影响电池老化过程[98]。目前机器学习技术已被成功应用于预测电池的容量和健康状态，以确保设备可靠运行和及时维护。Nagulapati 等[86]使用 18650 个电池充放电数据，利用高斯过程回归和支持向量机模型将放电循环过程中的电压、电流和温度数据相关联预测电池容量，并通过组合多电池数据集的方法提高了模型的预测精度。与常用的电流−电压数据相比，电化学阻抗谱通过测量电流对电压扰动的响应来获得宽频率范围内的阻抗，包含有关材料特性、界面现象和电化学反应的丰富信息。Zhang 等[91]收集了 2 万多个商业锂离子电池的电化学阻抗谱，使用高斯过程回归模型将整个阻抗谱作为输入，准确预测了不同温度下电池健康状态和剩余使用寿命。大幅度延长电池使用寿命的关键挑战是减少所需实验的数量和持续时间，Attia 等[99]通过弹性网络早期模型结合贝叶斯优化算法，从前 100 个循环的电化学测量数据预测了最终循环寿命并有效地探测充电协议的参数空间。

综上所述，通过提取材料的结构特征、元素属性和实验环境等结构化描述符建立机器学习模型，能够指导研究人员设计和优化液态/固态电解质和电极材料以及评估电池的健康状态。然而，由于其内部复杂的电化学行为，电化学储能材料的微观结构和材料性能之间常常存在复杂的非线性关系，导致线性模型性能较差，而神经网络等非线性模型的复杂性高可解释性差，且一般需要大量数据进行训练。因此，还需要研究人员对结构化数据驱动的电化学储能材料研发进行进一步的探索。

2.2　非结构化电化学储能材料数据驱动的机器学习

随着对电化学储能材料的深入研究，能够通过材料结构、表征技术和科学文献中得到大量的图形、图像和文本等非结构数据。这些数据含有非常丰富的材料信息，但传统的机器学习方法难以对其处理。作为机器学习的重要组成部分，深度学习能够从非常原始的非结构化图形和图像数据中提取信息，实现自动模型参数估计(即“端到端”学习)，从而避免繁琐但重要的描述符设计[100-101]。此外，通过文本挖掘和自然语言处理技术能够从非结构化的文本中提取材料的数据和知识，为进一步数据挖掘和分析提供数据集。

2.2.1　基于图形的材料性能预测

近年来，图深度学习(graph deep learning)因能够对任意大小和形状的图形进行“端到端”学习，无需研究人员构建繁琐而重要的描述符受到了材料领域越来越多的关注。

起初，图深度学习被应用在分子体系的性能预测中[102-103]，随后被用于周期性晶体体系[104-108]。工作流程如图 6 所示，首先需要获取材料结构数据集，一般来源于 ICSD 数据库或 CSD 数据库；然后对材料结构进行图形表示，这个过程需要选择合适的材料信息进行编码，如原子轨道相互作用、原子属性、键属性、全局状态和阴离子配位多面体基序等；最后通过图深度学习模型预测材料性能。这里总结了不同材料信息嵌入下图深度学习在材料性能预测中的应用。

2.2.1.1　原子信息嵌入

最近大多数材料图深度学习研究是基于原子级别的图形数据作为深度学习模型的输入[104, 107-108]。在谢天等[104]提出的晶体图卷积神经网络(CGCNN)中，每个晶体由一个晶体图形表示，并且满足原子索引置换不变性和晶胞选择不变性，该模型准确地预测了晶体结构的形成能、带隙、费米能和弹性特性等性能，最后通过钙钛矿材料说明了模型的可解释性。Ahmad 等[109]应用 CGCNN 模型筛选能够抑制锂金属阳极枝晶形成的无机固体电解质。Zhou 等[110] 基于 CGCNN 模型从 Materials Project 和 AFLOW 材料数据库中筛选了 13 万多种无机材料，成功预测了 80 种可用于锌离子电池的高压正极材料。作为 CGCNN 的改进模型，Park 等[108]提出了 iCGCNN 模型，该模型的晶体图包括 Voronoi 镶嵌晶体结构的信息、相邻组成原子的显式三体关联以及化学键的表示，在预测热力学稳定性方面，iCGCNN 的精度相较于 CGCNN 提高了 20%。

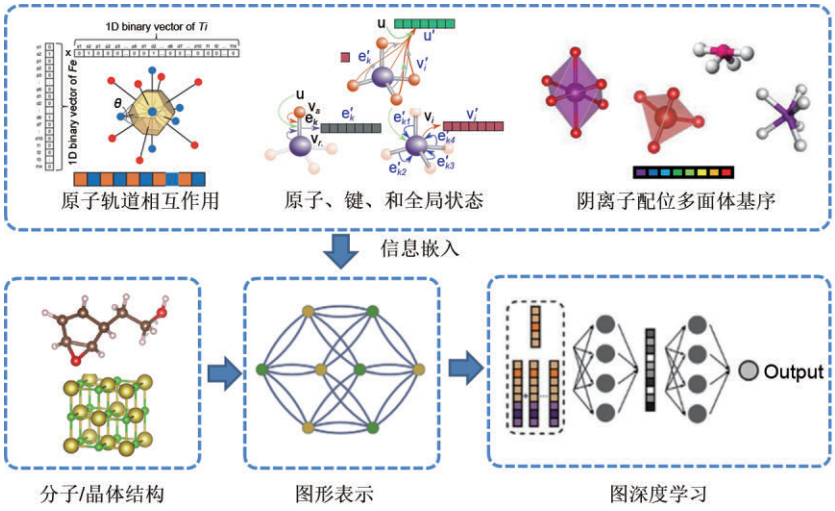

图6 图深度学习在分子/晶体结构应用的工作流程

Fig. 6 Workflow of application of graph deep learning in crystal and molecular structure

2.2.1.2 原子和全局状态信息嵌入

由于之前绝大多数模型是单独基于分子或晶体数据集开发的，且缺乏对温度和压力等全局状态的描述，致使模型缺少必要信息影响其预测性能。因此，Chen 等[107]提出了一个基于图形的深度学习框架(MEGNet)并将其应用于分子和晶体结构。该模型通过将原子属性、键属性和全局状态属性嵌入图神经网络模块中，然后通过信息传递过程反复更新，最终利用多层感知器预测材料性能。MEGNet 在预测晶体的形成能、带隙和弹性模量方面显著优于现有的机器学习模型。在此基础上，该团队还开发了一个能够处理多保真度数据和无序材料的图神经网络模型[111]。该模型将数据保真度级别编码为整数并传递给可训练的保真度嵌入矩阵，并通过元素嵌入的线性组合表示无序位点。实验结果表明该模型对于实验带隙预测的平均绝对误差降低了 22%～45%，但潜在限制是它依赖于大型低保真数据集来学习有效的结构表示，导致只能对少数目标属性进行高精度预测。

2.2.1.3 多尺度材料信息嵌入

与弹性模量等力学性能相比，预测晶体材料的电子结构性质需要更详细的材料信息，因此仅嵌入原子信息的图深度学习模型可能性能不佳。Banjade 等[106]提出了一种 Atom-Motif 双图网络模型(AMDNet)以增强对电子结构相关材料性能的预测。该模型利用阴离子配位多面体构建结构基序图，及基于原子的图形一起输入神经网络模型。与已有的模型相比，AMDNet 预测金属氧化物带隙等性能更加准确。此外，原子轨道也是一种与电子结构高度相关的材料特征。Karamad 等[105]提出了一种轨道图卷积神经网络(OGCNN)，以考虑晶体材料的原子轨道信息。该模型将原子轨道之间的键合信息编码为轨道场矩阵(OFM)表示，然后将改进的节点和边特征传递给 CGCNN 框架以进行性能预测，最后在形成能和带隙预测方面都具有比 CGCNN 更好的性能。

综上所述，本文对图深度学习的图形表示方法和模型框架进行了介绍，这些框架能够对材料的性能进行高效准确地预测，加快材料的研发速度。然而其中还存在一个基本问题：通过简单地设置截断距离可能导致原子间距离的微小变化使邻居原子数量的突然变化，晶体结构中原子的连接性难以判断。因此，图形表示更适合于具有共价键的分子材料，而具有离子键和金属键的晶体材料需要特别注意识别节点的连接性[112]。

2.2.2 材料表征图像分析

材料内部的显微组织结构决定着材料的性能，通过现代材料分析技术对其进行表征，可得到图像类型的非结构化数据。这些数据通常需要依赖材料专家对其进行分析，从中提取出显微化学成分、晶体结构和微观形貌等材料信息。但是仅仅依赖材料专家自身经验分析容易遗漏其中的隐藏信息，且耗时费力。计算机视觉领域的深度学习方法可以自动提取图像中的特征，与材料图像数据分析的强烈需

求相吻合，有助于提高材料表征的速度和准确性。这里主要针对不同的材料表征技术，介绍深度学习在电化学储能材料图像数据分析的研究进展。

2.2.2.1　X射线断层扫描图像分割

X射线断层扫描是一种强有力的表征方法，可以对材料的微观结构和化学成分进行动态无损成像，提供电池运行和退化的定量或定性分析[113]。量化锂电极中微结构的形态转变需要严格和一致的分割程序，Dixit等[114]实现了一个基于ResNet-34的深度卷积神经网络对锂金属X射线断层扫描低对比度图像中的锂金属和孔隙进行分割，以定量跟踪锂金属电极和固态电解质固固界面的形态变化。与传统的二值化过程相比，机器学习识别锂金属孔隙特征的保真度和准确性明显提高。复合电极的微观结构决定了电极颗粒在充放电过程中的行为，颗粒与碳/黏合剂分离的程度与容量损失相关。为了对严重破碎的颗粒进行识别，Jiang等[115]使用高分辨率硬X射线纳米断层扫描对复合正极材料可视化，开发了一个掩模区域卷积神经网络模型并自动识别和分割了650多个正极颗粒，消除了使用传统图像技术报告中表征结果存在的偏差。

2.2.2.2　拉曼高光谱成像特征提取

拉曼高光谱成像具有同时对多种化学特征进行成像的能力。同步识别锂离子电池电极中多个光谱特征有助于将分析技术用于在线质量控制和产品开发。Baliyan等[116]提出了一个神经网络分析框架来自动从锂离子电池电极拉曼高光谱数据集中识别光谱特征并分配类别标签，从而计算容量保留系数来定量评估锂离子电池的容量退化。该方法有效地避免了宇宙噪声带来的错误定量分析，且实现了对高光谱分析整个生命周期的自动化处理。

2.2.2.3　电子背散射衍射图像晶界增强

电子背散射衍射通过分析晶粒两侧像素之间的取向来检测多晶样品中的晶界，可以在晶粒尺度上改善正极材料的评估和量化，这对理解锂离子电池的锂传输、速率限制和降解机制至关重要[116]。Furat等[117]使用电子背散射衍射技术对正极材料颗粒进行成像，通过卷积神经网络对标记的图像进行训练并应用于整个图像数据，从而产生具有增强晶界的新图像。该方法避免了常规图像处理方法繁琐的处理步骤和参数校正过程，实现了晶粒结构的有效形态表征。

总之，深度学习技术能够从复杂的电化学储能材料图像数据中识别特定的特征，从而有效应用于X射线断层扫描图像分割、拉曼高光谱成像特征提取和电子背散射衍射图像晶界增强。然而，深度学习模型强烈依赖于大量标记的图像数据，以及非专业研究者对深度学习模型使用的复杂性等问题还阻碍着其在材料图像领域的应用。此外，上述例子仅仅是对电化学储能材料图像本身进行了建模应用，通过深度学习技术还可以进一步地探索化学成分-介观尺度显微组织结构-材料性能之间的构效关系，加速材料性能预测[118]。

2.2.3　材料文本挖掘

文本挖掘是指从文本语料库中提取有价值信息和知识的方法。近年来，材料科学的文本挖掘主要依靠自然语言处理技术和机器学习方法，从数量庞大且不断增长的科学出版物中快速获取非结构化科学知识，进而指导材料相关领域的研究。文本挖掘的工作流程可以概括为文本收集与解析、文本预处理、文本分析、信息提取、数据挖掘，如图7所示[119]。随着文本挖掘技术的逐渐成熟，已有学者将其应用到电化学储能材料领域，从而追踪材料研究动态、指导材料合成和建立材料数据库等。

2.2.3.1　追踪研究动态

文本挖掘可帮助读者找到某个领域的突破性论文并跟踪最新技术的进展。Torayev等[120]使用基于机器学习的文本挖掘技术从1800多篇文献中识别Li-O_2电池研究领域的全球趋势。结果显示，该领域的电解质研究已从碳酸盐转向了甘醇二甲醚和二甲基亚砜，且大部分文献都关注电池的循环稳定性、容量和倍率性能。El-Bousiydy等[121]使用基于

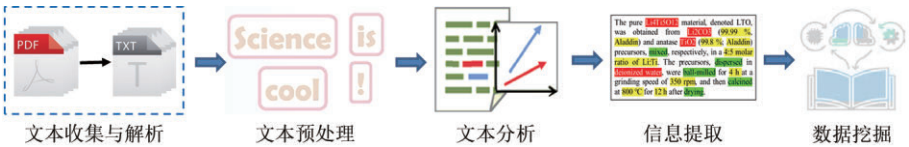

图7　文本挖掘的工作流程[119]

Fig. 7　Workflow of text mining[119]

关键字搜索文本挖掘算法，分析了1.3万份锂和钠离子电池科学文献中研究人员的习惯，发现大多文献缺乏对某些关键特征的系统报告，例如厚度、孔隙率、电解质体积、表面积和质量载荷。通过文本挖掘技术构建材料知识图谱，能够从海量材料科学文献中进行信息抽取，建立实体之间的对应关系，从而自动化地提供材料科学领域信息。Nie等[122]收集了超过290万篇材料领域的文章及其作者信息，结合机器学习和依赖匹配算法对材料知识图谱中的主体进行高精度消歧，并使用剪枝策略实现高效信息匹配和搜索，从而构建了材料知识图谱(MatKG)框架。利用该框架对$LiFePO_4$进行自动化分析，关联相关学者及其研究信息，建立了用于锂离子电池的$LiFePO_4$材料发展里程碑图。

2.2.3.2 指导材料合成

优化电解质低温处理协议能够最大程度地减少电池界面的不兼容性[123]。Mahbub等[124]使用基于规则和机器学习方法自动提取硫化物和氧化物的锂固态电解质文本中实验合成部分，然后通过神经网络模型对每个段落中的单词进行标记和分类，以预测句子中每个单词的重要合成关键词(例如材料名称、操作名称、数量、条件等)，将这些分类的标记组合成一个数据库对象并对其进一步数据挖掘以提取合成趋势。该团队从中识别出高电位氧化物基锂石榴石电解质的低温合成方法，降低了固态电解质组装到电池过程中的界面复杂性。

2.2.3.3 建立材料数据库

化学感知自然语言处理工具包ChemDataExtractor[125]是化学信息提取和文本处理的常用工具，在文本处理、标记化和词性标注方面灵活而准确，能够用于识别化学物质实体、相关属性及其相互依赖关系。大型电池材料数据库对于数据驱动的新材料发现至关重要，Huang等[126]使用ChemDataExtractor，通过文章检索、数据提取、数据清理、数据后处理和评估过程，从22万余篇电池研究论文中自动提取数据，然后创建了一个大型电池材料同源属性数据库，包括1.7万种化合物和对应的21万多条电池材料属性(容量、电压、电导率、库仑效率和能量)。

综上所述，目前只有少数基于文本挖掘的工作专注于电化学储能材料领域，其限制主要有以下三点：一是材料文本标注数据稀缺性，大多数现有的标注数据集都是以特定的材料领域而创建的，难以直接应用于其他材料体系；二是材料命名方法差异性，材料文本中存在各种专业术语，缺乏标准的命名方法容易导致歧义的产生；三是材料文本的复杂性，材料科学文本的专业性强可读性差，使得文本处理异常困难。即使如此，随着大型材料文本数据库的建立和自然语言处理技术的发展，相信文本挖掘技术会对电化学储能材料的发展起到重要的作用。

3 电化学储能材料机器学习面临的挑战与对策

如前所述，结构化和非结构化数据驱动的机器学习模型已经在电化学储能材料领域得到了广泛应用，但仍存在一些问题制约着机器学习的进一步发展。本节对这些问题进行了系统性分析，并将其归结为机器学习在电化学储能材料领域应用面临的三大矛盾，包括高维度和小样本数据的矛盾、模型准确性和易用性的矛盾以及学习结果与领域知识的矛盾。调和这些矛盾以提升机器学习模型在电化学储能材料领域应用的准确性、易用性和可解释性，将有助于进一步加速电化学储能材料的研发与设计。

3.1 高维度与小样本数据的矛盾与协调

电化学储能材料数据通常是多源(如实验数据、计算数据、生产数据和文献数据)且异构的(如结构化和非结构化数据)，不同来源数据的外部一致性很难得到保证，导致最终用于机器学习建模的数据集往往是小样本的。尤其是文献数据的标记难，小样本问题更显突出。此外，电化学储能材料性能受多种物理/化学因素影响，研究人员总是定义大量描述符来表示复杂的材料性能驱动机制，这又导致机器学习建模使用的数据集通常是高维度的。较小的数据量和较高的维度容易使得机器学习模型过度拟合现有数据，从而影响机器学习模型的泛化性能和可解释性，故电化学储能材料领域机器学习面临着高维度和小样本数据的矛盾。

正如第2节中指出的，从数据的源头抓起，注重反例数据的收集、多尺度数据的收集，积累更加丰富多样的结构化和非结构化数据。同时，提高各种数据的集成与共享，克服其质量参差不齐、数据标注不足、缺乏有效融合以及大型标准数据集建设不足等问题，是调和上述矛盾的可行和必经途径。

从技术上来说，目前普遍的做法是通过特征工程或选择适用于小样本高维度数据的学习器来解决上述问题。其中，如3.1.2节所述，特征工程方法旨在通过降低特征维度或构建“更好”的描述符来削弱高维度数据集对机器学习模型性能的负面影响。支持向量机是通过核函数将低维数据投影到高维空间中进行划分，在形式上更容易处理高维特征的数据集[127]。例如，Fujimura等[87]利用支持向量机对LISICON型固态电解质材料的离子电导率进行预测，平均绝对误差为0.373 S/cm。然而，在处理一些复杂的小样本高维度数据问题时，这些方法的效果并不理想。因此，研究人员提出数据增强、主动学习和迁移学习等方法解决电化学储能材料数据的小样本问题来调和这一矛盾。数据增强技术基于现有数据，通过物理增强、统计学和无监督生成模型等方法生成更多新数据[128]。例如，Naaz等[129]设计了一种基于生成对抗网络的数据增强方法用于预测锂离子电池的充电状态和健康状态；Hsu等[130]利用生成对抗网络框架来学习和生成固体氧化物燃料电池电极的微观结构。主动学习利用预构建的机器学习预测模型迭代地对候选化学空间进行自适应采样，从而为代价高昂的模拟计算或实验验证提供最有价值的候选样本，以加速新型高性能材料的筛选[131]。例如，Verduzco等[55]利用主动学习方法指导高离子电导率的石榴石的合成，从而减少了30%的实验次数。迁移学习旨在通过迁移包含在相关领域中的知识来提高目标学习器在目标领域上的性能，以减少目标学习器对目标数据集大数据量的依赖[132]。例如，Wang等[133]基于CGCNN的迁移学习模型在低精度PBE数据集预训练，到高精度HSE06数据集进行参数微调预测晶体结构的高精度带隙，从而克服了小数据集导致的精度低和过拟合的问题。

3.2　模型准确性与易用性的矛盾与统一

机器学习的最初目标是从数据中提取可解释的知识，并在追求算法准确性的同时强调其可解释性[127]。以线性回归、偏最小二乘法等多元线性模型为主的机器学习算法可以构建多个因素与目标属性之间的线性关系，模型简单、易于实现且学习结果容易理解。但是，电化学储能材料内部复杂的电化学行为导致线性模型的预测精度通常较低，而神经网络(NN)和支持向量机(SVM)等模型由于能够建立影响因素与目标性能间的复杂非线性关系，在电化学储能材料研究中得到了更广泛的应用。然而，这些非线性模型大都是“黑箱”模型，其内部原理复杂、结果难以解释，且通常需要大量繁琐的调参工作才能获得最优性能。因此，机器学习在电化学储能材料领域的应用中存在着模型准确性和易用性的矛盾。

降低材料专家使用模型的复杂性和提高复杂模型的可理解性是提高模型易用性的有效途径。自动机器学习(auto machine learning，AutoML)是在有限的计算资源内全部或部分配置适用于机器学习方法的参数，主要通过随机搜索、进化优化、贝叶斯优化、元学习等方法减少模型的选择、优化以及实现过程中的人工参与，从而自动构建最佳机器学习模型[134-135]。例如，Dunn等[136]提出自动机器学习模型Automatminer以预测无机固体材料的性能。该模型利用Matminer[57]生成特定材料的描述符，自动地执行数据预处理和特征工程，并通过广泛的内部数据测试来确定最佳机器学习模型。而机器学习的可解释性是机器学习模型以人类可理解术语向人类提供解释的能力。在计算机领域，机器学习的可解释性可使得机器学习模型的结构和预测结果两方面都易于理解，从而提高机器学习模型的易用性。例如：规则提取方法采用“如果输入特征x则被归类为y类”的解释方式，将机器学习中隐含的知识以一种易于理解的方式表达，以提高机器学习方法的可解释性[137]。目前，可解释性方法还没有在电化学储能材料领域得到应用。未来可以引入对模型决策过程解释的内部解释和对模型决策结果解释的外部解释来共同提高机器学习应用方法的可解释性。内部解释一般把已训练好的机器学习/深度学习等模型的内部决策结构(如网络结构、参数权重、特征向量等)映射成易于理解的If-Then-Else规则或内部决策过程的权重可视化展示[138]。外部解释可将不同的特征组合输入到已训练好的模型，来找到输入特征与模型决策结果之间的关系，以发现模型的决策规则来提高模型的可解释性；也可将更复杂的模型转换为易于理解的模型，再从中抽取规则，使得抽取的规则对模型有重现能力[139]。此外，根据提取的规则构建概念嵌入表达，再将规则表达嵌入其他模型中，也有望在保证预测精度的前提下，提高模型的可解释性。

3.3 模型学习结果与领域专家知识的矛盾与融合

目前，广泛应用于电化学储能材料发现和性能预测的机器学习模型大都是纯数据驱动的，严重依赖于样本数据进行学习，对材料领域知识的重视度不够，导致在实际应用中仍然会出现机器学习结果与领域专家知识相矛盾的现象。针对该问题，一方面，可以通过描述符定义和选择过程[75]将材料专家的领域知识融入到问题定义中，从而指导模型学习领域知识。例如，Li 等[140]提出了“中心-环境”(center-environment，CE)特征构建模型，通过将基本属性集合映射到由组分和结构信息组成的基集中来构建特征，用于预测尖晶石氧化物的形成能、晶格参数和带隙；Weng 等[141]利用符号回归得到了描述符 μ/t (μ 为八面体因子，t 为容忍因子)，并在该描述符的指导下成功地合成了五种新的氧化物钙钛矿；Gong 等[142]通过机器学习和理论模型相结合来预测二维金属材料上的锂吸附能，提高了模型的泛化能力。另一方面，在机器学习模型构建过程中嵌入领域知识是一个有效的解决方案，典型的算法有贝叶斯网络和模糊学习。其中，贝叶斯网络通过在训练过程中结合先验知识来确定网络拓扑结构[143]，而模糊学习则使用隶属函数来整合专家经验[144]。例如，Ren 等[145]通过施加基于物理场的约束来创建参数化过程模型，从而将过程优化变量与所得材料的体积和界面特性耦合起来；再添加额外的推理层将过程变量和材料属性之间的联系扩展到器件性能，并利用神经网络代理模型预测电流电压曲线；最后，通过贝叶斯网络推断结果优化太阳能电池工艺参数。此外，可通过机器学习结果建立知识库与电池材料专家先验知识共同指导材料开发。例如，Martin 等[146]提出一种将领域专家与机器学习相结合构建知识库的方法来实现两者的相互补充，从而提高专家系统的推理能力。

目前，绝大部分的机器学习方法是纯数据驱动的，机器学习的全过程仅围绕着提升模型精度这一单一目标展开，往往忽略了领域知识的重要性，这是导致上述三大矛盾的主要原因之一。基于此，为充分发挥材料领域知识在机器学习建模中的作用，本团队提出的融合加权评分领域专家知识的多层级特征选择方法[75]和分而治之的自适应机器学习建模方法[147]，已经初步证明了领域知识嵌入在改善机器学习模型预测精度和可解释性方面的有效性。进一步地，将领域知识符号化表示为机器学习模型的前处理条件、建模约束或目标函数、后解释规则等并嵌入到机器学习全生命周期过程中，同时结合知识、数据、算法和算力四大要素，构建具有一定可解释性的领域知识嵌入的机器学习新模型，将有望系统性地解决上述三大矛盾。

4 结 语

数据驱动材料科学的最新研究表明，机器学习技术的应用可以极大地促进电化学储能材料的设计和发现。本文首先介绍了可用于电化学储能材料研究的数据资源，并对电化学储能材料专用数据库发展方向提出建议，如收集实验反例数据和材料多尺度数据、共享图像和文本等非结构化数据、设计数据质量检测方案；随后，详细阐述了结构化数据驱动下的机器学习工作流程及其在电化学储能材料领域的最新应用，以及基于图形、图像、文本的非结构化数据驱动下的机器学习在电化学储能材料领域的研究进展；最后，总结了机器学习在电化学储能材料领域应用所面临的三大矛盾和相关的解决策略，并提出进一步构建面向机器学习全流程的“领域知识嵌入的机器学习方法”，将有望系统地调和上述三大矛盾。本文对机器学习在电化学储能领域应用的总结和未来发展策略的提出，将为实现高性能电化学储能材料的精准、高效研发指明方向。

参 考 文 献

[1] 张舒, 王少飞, 凌仕刚, 等. 锂离子电池基础科学问题(X)——全固态锂离子电池[J]. 储能科学与技术, 2014, 3(4): 376-394.
ZHANG S, WANG S F, LING S G, et al. Fundamental scientific aspects of lithium ion batteries(X) —All-solid-state lithium-ion batteries[J]. Energy Storage Science and Technology, 2014, 3(4): 376-394.

[2] 李泓. 锂离子电池基础科学问题(XV)——总结和展望[J]. 储能科学与技术, 2015, 4(3): 306-318.
LI H. Fundamental scientific aspects of lithium ion batteries(XV)—Summary and outlook[J]. Energy Storage Science and Technology, 2015, 4(3): 306-318.

[3] 任元, 邹喆乂, 赵倩, 等. 浅析电解质中离子输运的微观物理图像[J]. 物理学报, 2020, 69(22): 46-62.
REN Y, ZOU Z Y, ZHAO Q, et al. Brief overview of microscopic physical image of ion transport in electrolytes[J]. Acta Physica Sinica, 2020, 69(22): 46-62.

[4] 郑杰允, 李泓. 锂电池基础科学问题(Ⅴ)——电池界面[J]. 储能科学与技术, 2013, 2(5): 503-513.

ZHENG J Y, LI H. Fundamental scientific aspects of lithium batteries (Ⅴ)—Interfaces[J]. Energy Storage Science and Technology, 2013, 2(5): 503-513.

[5] 彭佳悦, 祖晨曦, 李泓. 锂电池基础科学问题(Ⅰ)——化学储能电池理论能量密度的估算[J]. 储能科学与技术, 2013, 2(1): 55-62.
PENG J Y, ZU C X, LI H. Fundamental scientific aspects of lithium batteries(Ⅰ)—Thermodynamic calculations of theoretical energy densities of chemical energy storage systems[J]. Energy Storage Science and Technology, 2013, 2(1): 55-62.

[6] 施思齐, 徐积维, 崔艳华, 等. 多尺度材料计算方法[J]. 科技导报, 2015, 33(10): 20-30.
SHI S Q, XU J W, CUI Y H, et al. Multiscale materials computational methods[J]. Science & Technology Review, 2015, 33(10): 20-30.

[7] SHI S Q, GAO J, LIU Y, et al. Multi-scale computation methods: Their applications in lithium-ion battery research and development[J]. Chinese Physics B, 2016, 25(1): doi: 10.1088/1674-1056/25/1/018212.

[8] FRANCO A A, RUCCI A, BRANDELL D, et al. Boosting rechargeable batteries R&D by multiscale modeling: Myth or reality?[J]. Chemical Reviews, 2019, 119(7): 4569-4627.

[9] AGRAWAL A, CHOUDHARY A. Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science[J]. APL Materials, 2016, 4(5): doi: 10.1063/1.4946894.

[10] BELL G, HEY T, SZALAY A. Beyond the data deluge[J]. Science, 2009, 323(5919): 1297-1298.

[11] 谢建新, 宿彦京, 薛德祯, 等. 机器学习在材料研发中的应用[J]. 金属学报, 2021, 57(11): 1343-1361.
XIE J X, SU Y J, XUE D Z, et al. Machine learning for materials research and development[J]. Acta Metallurgica Sinica, 2021, 57(11): 1343-1361.

[12] BUTLER K T, DAVIES D W, CARTWRIGHT H, et al. Machine learning for molecular and materials science[J]. Nature, 2018, 559(7715): 547-555.

[13] WANG H S, JI Y J, LI Y Y. Simulation and design of energy materials accelerated by machine learning[J]. WIREs Computational Molecular Science, 2020, 10(1): doi: 10.1002/wcms.1421.

[14] GAO T H, LU W. Machine learning toward advanced energy storage devices and systems[J]. iScience, 2021, 24(1): doi: 10.1016/j.isci.2020.101936.

[15] LYU C, ZHOU X, ZHONG L, et al. Machine learning: An advanced platform for materials development and state prediction in lithium-ion batteries[J]. Advanced Materials, 2021, doi: 10.1002/adma.202101474.

[16] LIU Y, ZHAO T L, JU W W, et al. Materials discovery and design using machine learning[J]. Journal of Materiomics, 2017, 3(3): 159-177.

[17] LIU Y, WU J M, YANG G, et al. Predicting the onset temperature (Tg) of Ge_xSe_{1-x} glass transition: A feature selection based two-stage support vector regression method[J]. Science Bulletin, 2019, 64(16): 1195-1203.

[18] LIU Y, ZHAO T L, YANG G, et al. The onset temperature (Tg) of As_xSe_{1-x} glasses transition prediction: A comparison of topological and regression analysis methods[J]. Computational Materials Science, 2017, 140: 315-321.

[19] SALKIND A J, FENNIE C, SINGH P, et al. Determination of state-of-charge and state-of-health of batteries by fuzzy logic methodology[J]. Journal of Power Sources, 1999, 80(1/2): 293-300.

[20] MORGAN D, CEDER G, CURTAROLO S. Data mining approach to ab-initio prediction of crystal structure[J]. MRS Proceedings, 2003, 804: 305-310.

[21] CURTAROLO S, MORGAN D, PERSSON K, et al. Predicting crystal structures with data mining of quantum calculations[J]. Physical Review Letters, 2003, 91(13): doi: 10.1103/PhysRevLett.91.135503.

[22] MORGAN D, CEDER G, CURTAROLO S. High-throughput and data mining with *ab initio* methods[J]. Measurement Science and Technology, 2005, 16(1): 296-301.

[23] The materials genome initiative at the national science foundation: A status report after the first year of funded research[J]. JOM, 2014, 66(3): 336-344.

[24] WIDENER A. Materials genome initiative[J]. Chemical & Engineering News Archive, 2013, 91(31): 25-27.

[25] 汪洪, 项晓东, 张澜庭. 数据+人工智能是材料基因工程的核心[J]. 科技导报, 2018, 36(14): 15-21.
WANG H, XIANG X D, ZHANG L T. Data+AI: The core of materials genomic engineering[J]. Science & Technology Review, 2018, 36(14): 15-21.

[26] GUO H Y, WANG Q, STUKE A, et al. Accelerated atomistic modeling of solid-state battery materials with machine learning[J]. Frontiers in Energy Research, 2021, 9: doi: 10.3389/fenrg.2021.695902.

[27] CHEN X, LIU X Y, SHEN X, et al. Applying machine learning to rechargeable batteries: From the microscale to the macroscale[J]. Angewandte Chemie, 2021, 60(46): 24354-24366.

[28] LOMBARDO T, DUQUESNOY M, EL-BOUYSIDY H, et al. Artificial intelligence applied to battery research: Hype or reality?[J]. Chemical Reviews, 2021, doi: 10.1021/acs.chemrev.1c00108.

[29] LIU Y, GUO B R, ZOU X X, et al. Machine learning assisted materials design and discovery for rechargeable batteries[J]. Energy Storage Materials, 2020, 31: 434-450.

[30] 吴思远, 王宇琦, 肖睿娟, 等. 电池材料数据库的发展与应用[J]. 物理学报, 2020, 69(22): 9-16.
WU S Y, WANG Y Q, XIAO R J, et al. Development and application of battery materials database[J]. Acta Physica Sinica, 2020, 69(22): 9-16.

[31] ALLEN F H. The cambridge structural database: A quarter of a million crystal structures and rising[J]. Acta Crystallographica Section B, Structural Science, 2002, 58(Pt 3 Pt 1): 380-388.

[32] GROOM C R, ALLEN F H. The cambridge structural database in retrospect and prospect[J]. Angewandte Chemie, 2014, 53(3): 662-671.

[33] KENNARD O, ALLEN F, BELLARD S, et al. Current developments in the cambridge structural database[J]. Acta Crystallographica Section A

Foundations of Crystallography, 1984, 40: C445.

[34] BERGERHOFF G, HUNDT R, SIEVERS R, et al. The inorganic crystal structure data base[J]. Journal of Chemical Information and Computer Sciences, 1983, 23(2): 66-69.

[35] BELSKY A, HELLENBRANDT M, KAREN V L, et al. New developments in the inorganic crystal structure database (ICSD): Accessibility in support of materials research and design[J]. Acta Crystallographica Section B, Structural Science, 2002, 58(Pt 3 Pt 1): 364-369.

[36] VILLARS P, BERNDT M, BRANDENBURG K, et al. The Pauling file, binaries edition[J]. Journal of Alloys and Compounds, 2004, 367(1/2): 293-297.

[37] 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): doi: 10.1063/1.4812323.

[38] CURTAROLO S, SETYAWAN W, WANG S D, et al. AFLOWLIB. ORG: A distributed materials properties repository from high-throughput *ab initio* calculations[J]. Computational Materials Science, 2012, 58: 227-235.

[39] 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.

[40] NOSENGO N, CEDER G. Can artificial intelligence create the next wonder material?[J]. Nature, 2016, 533(7601): 22-25.

[41] HE B, YE A, CHI S, et al. CAVD, towards better characterization of void space for ionic transport analysis[J]. Scientific Data, 2020, 7: doi: 10.1038/s41597-020-0491-x.

[42] HE B, MI P H, YE A J, et al. A highly efficient and informative method to identify ion transport networks in fast ion conductors[J]. Acta Materialia, 2021, 203: doi: 10.1016/j.actamat.2020.116490.

[43] ZHANG L W, HE B, ZHAO Q, et al. A database of ionic transport characteristics for over 29 000 inorganic compounds[J]. Advanced Functional Materials, 2020, 30(35): doi: 10.1002/adfm.202003087.

[44] RACCUGLIA P, ELBERT K C, ADLER P D, et al. Machine-learning-assisted materials discovery using failed experiments[J]. Nature, 2016, 533(7601): 73-76.

[45] HIMANEN L, GEURTS A, FOSTER A S, et al. Data-driven materials science: Status, challenges, and perspectives[J]. Advanced Science, 2019, 6(21): doi: 10.1002/advs.201900808.

[46] KIRKLIN S, SAAL J E, MEREDIG B, et al. The open quantum materials database (OQMD): Assessing the accuracy of DFT formation energies[J]. npj Computational Materials, 2015, 1: doi: 10.1038/npjcompumats.2015.10.

[47] GHIRINGHELLI L M, VYBIRAL J, LEVCHENKO S V, et al. Big data of materials science: Critical role of the descriptor[J]. Physical Review Letters, 2015, 114(10): doi: 10.1103/PhysRevLett.114.105503.

[48] SENDEK A D, YANG Q, CUBUK E D, et al. Holistic computational structure screening of more than 12 000 candidates for solid lithium-ion conductor materials[J]. Energy & Environmental Science, 2017, 10(1): 306-320.

[49] ZHAO Q, AVDEEV M, CHEN L Q, et al. Machine learning prediction of activation energy in cubic Li-argyrodites with hierarchically encoding crystal structure-based (HECS) descriptors[J]. Science Bulletin, 2021, 66(14): 1401-1408.

[50] WANG A P, ZOU Z Y, WANG D, et al. Identifying chemical factors affecting reaction kinetics in Li-air battery via *ab initio* calculations and machine learning[J]. Energy Storage Materials, 2021, 35: 595-601.

[51] WARD L, AGRAWAL A, CHOUDHARY A, et al. A general-purpose machine learning framework for predicting properties of inorganic materials[J]. npj Computational Materials, 2016, 2: doi: 10.1038/npjcompumats.2016.28.

[52] JOSHI R P, EICKHOLT J, LI L L, et al. Machine learning the voltage of electrode materials in metal-ion batteries[J]. ACS Applied Materials & Interfaces, 2019, 11(20): 18494-18503.

[53] JO J, CHOI E, KIM M, et al. Machine learning-aided materials design platform for predicting the mechanical properties of Na-ion solid-state electrolytes[J]. ACS Applied Energy Materials, 2021, 4(8): 7862-7869.

[54] CHOI E, JO J, KIM W, et al. Searching for mechanically superior solid-state electrolytes in Li-ion batteries *via* data-driven approaches[J]. ACS Applied Materials & Interfaces, 2021, 13(36): 42590-42597.

[55] VERDUZCO J C, MARINERO E E, STRACHAN A. An active learning approach for the design of doped LLZO ceramic garnets for battery applications[J]. Integrating Materials and Manufacturing Innovation, 2021, 10(2): 299-310.

[56] WARD L, LIU R Q, KRISHNA A, et al. Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations[J]. Physical Review B, 2017, 96(2): doi: 10.1103/PhysRevB.96.024104.

[57] 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.

[58] 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: doi: 10.1016/j.cpc.2019.106949.

[59] RUPP M, TKATCHENKO A, MÜLLER K R, et al. Fast and accurate modeling of molecular atomization energies with machine learning[J]. Physical Review Letters, 2012, 108(5): doi: 10.1103/PhysRevLett.108.058301.

[60] FABER F, LINDMAA A, VON LILIENFELD O A, et al. Crystal structure representations for machine learning models of formation energies[J]. International Journal of Quantum Chemistry, 2015, 115(16): 1094-1101.

[61] RUPP M. Many-body tensor representation for machine learning of materials[C]//APS March Meeting Abstracts, 2017.

[62] BEHLER J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials[J]. The Journal of Chemical Physics, 2011, 134(7): doi: 10.1063/1.3553717.

[63] BARTÓK A P, KONDOR R, CSÁNYI G. On representing chemical environments[J]. Physical Review B, 2013, 87(8): doi: 10.1103/

physRevB.87.184115.

[64] JOLLIFFE I T, CADIMA J. Principal component analysis: A review and recent developments[J]. Philosophical Transactions Series A, Mathematical, Physical, and Engineering Sciences, 2016, 374(2065): doi: 10.1098/rsta.2015.0202.

[65] THARWAT A, GABER T, IBRAHIM A, et al. Linear discriminant analysis: A detailed tutorial[J]. AI Communications, 2017, 30(2): 169-190.

[66] BANGUERO E, CORRECHER A, PÉREZ-NAVARRO Á, et al. Diagnosis of a battery energy storage system based on principal component analysis[J]. Renewable Energy, 2020, 146: 2438-2449.

[67] WANG L Y, WANG L F, LIAO C L, et al. Research on multi-parameter evaluation of electric vehicle power battery consistency based on principal component analysis[J]. Journal of Shanghai Jiaotong University (Science), 2018, 23(5): 711-720.

[68] CHEN K L, ZHENG F D, JIANG J C, et al. Practical failure recognition model of lithium-ion batteries based on partial charging process[J]. Energy, 2017, 138: 1199-1208.

[69] CHANDRASHEKAR G, SAHIN F. A survey on feature selection methods[J]. Computers & Electrical Engineering, 2014, 40(1): 16-28.

[70] PENG H C, LONG F H, DING C. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and Min-redundancy[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(8): 1226-1238.

[71] LI W, JACOBS R, MORGAN D. Predicting the thermodynamic stability of perovskite oxides using machine learning models[J]. Computational Materials Science, 2018, 150: 454-463.

[72] LIN F Y, LIANG D, YEH C C, et al. Novel feature selection methods to financial distress prediction[J]. Expert Systems With Applications, 2014, 41(5): 2472-2483.

[73] TEKIN ERGUZEL T, TAS C, CEBI M. A wrapper-based approach for feature selection and classification of major depressive disorder-bipolar disorders[J]. Computers in Biology and Medicine, 2015, 64: 127-137.

[74] GENUER R, POGGI J M, TULEAU-MALOT C. Variable selection using random forests[J]. Pattern Recognition Letters, 2010, 31(14): 2225-2236.

[75] LIU Y, WU J M, AVDEEV M, et al. Multi-layer feature selection incorporating weighted score-based expert knowledge toward modeling materials with targeted properties[J]. Advanced Theory and Simulations, 2020, 3(2): doi: 10.1002/adts.201900215.

[76] GHARAGHEIZI F, SATTARI M, ILANI-KASHKOULI P, et al. A “non-linear” quantitative structure-property relationship for the prediction of electrical conductivity of ionic liquids[J]. Chemical Engineering Science, 2013, 101: 478-485.

[77] WU H, LORENSON A, ANDERSON B, et al. Robust FCC solute diffusion predictions from ab-initio machine learning methods[J]. Computational Materials Science, 2017, 134: 160-165.

[78] SHANDIZ M A, GAUVIN R. Application of machine learning methods for the prediction of crystal system of cathode materials in lithium-ion batteries[J]. Computational Materials Science, 2016, 117: 270-278.

[79] ZHAO Q, ZHANG L W, HE B, et al. Identifying descriptors for Li^+ conduction in cubic Li-argyrodites via hierarchically encoding crystal structure and inferring causality[J]. Energy Storage Materials, 2021, 40: 386-393.

[80] UNCU Ö, TÜRKŞEN I B. A novel feature selection approach: Combining feature wrappers and filters[J]. Information Sciences, 2007, 177(2): 449-466.

[81] HSU H H, HSIEH C W, LU M D. Hybrid feature selection by combining filters and wrappers[J]. Expert Systems With Applications, 2011, 38(7): 8144-8150.

[82] NASRABADI N M. Book review: Pattern recognition and machine learning[J]. Journal of Electronic Imaging, 2007, 16: doi: 10.1117/1.2819119.

[83] XU Y J, ZONG Y, HIPPALGAONKAR K. Machine learning-assisted cross-domain prediction of ionic conductivity in sodium and lithium-based superionic conductors using facile descriptors[J]. Journal of Physics Communications, 2020, 4(5): doi: 10.1088/2399-6528/ab92d8.

[84] LIU B, YANG J, YANG H L, et al. Rationalizing the interphase stability of Li|doped-$Li_7La_3Zr_2O_{12}$via automated reaction screening and machine learning[J]. Journal of Materials Chemistry A, 2019, 7(34): 19961-19969.

[85] KIREEVA N, PERVOV V S. Materials informatics screening of Li-rich layered oxide cathode materials with enhanced characteristics using synthesis data[J]. Batteries & Supercaps, 2020, 3(5): 427-438.

[86] NAGULAPATI V M, LEE H, JUNG D, et al. A novel combined multi-battery dataset based approach for enhanced prediction accuracy of data driven prognostic models in capacity estimation of lithium ion batteries[J]. Energy and AI, 2021, 5: doi: 10.1088/2399-6528/ab92018.

[87] FUJIMURA K, SEKO A, KOYAMA Y, et al. Accelerated materials design of lithium superionic conductors based on first-principles calculations and machine learning algorithms[J]. Advanced Energy Materials, 2013, 3(8): 980-985.

[88] DUQUESNOY M, BOYANO I, GANBORENA L, et al. Machine learning-based assessment of the impact of the manufacturing process on battery electrode heterogeneity[J]. Energy and AI, 2021, 5: doi: 10.1016/j.egyai.2021.100090.

[89] ISHIKAWA A, SODEYAMA K, IGARASHI Y, et al. Machine learning prediction of coordination energies for alkali group elements in battery electrolyte solvents[J]. Physical Chemistry Chemical Physics: PCCP, 2019, 21(48): 26399-26405.

[90] HOMMA K, LIU Y, SUMITA M, et al. Optimization of a heterogeneous ternary Li_3PO_4-Li_3BO_3-Li_2SO_4 mixture for Li-ion conductivity by machine learning[J]. The Journal of Physical Chemistry C, 2020, 124(24): 12865-12870.

[91] ZHANG Y, TANG Q, ZHANG Y, et al. Identifying degradation patterns of lithium ion batteries from impedance spectroscopy using machine learning[J]. Nature Communications, 2020, 11: doi: 10.1038/s41467-020-15235-7.

[92] WANG Z L, ZHANG H K, LI J J. Accelerated discovery of stable spinels in energy systems via machine learning[J]. Nano Energy,

2021, 81: doi: 10.1016/j.nanoen.2020.105665.

[93] LEE B, YOO J, KANG K. Predicting the chemical reactivity of organic materials using a machine-learning approach[J]. Chemical Science, 2020, 11(30): 7813-7822.

[94] DUONG V M, TRAN T N, GARG A, et al. Machine learning technique-based data-driven model of exploring effects of electrolyte additives on $LiNi_{0.6}Mn_{0.2}Co_{0.2}O_2$/graphite cell[J]. Journal of Energy Storage, 2021, 42: doi: 10.1016/j.est.2021.103012.

[95] MOSES I A, JOSHI R P, OZDEMIR B, et al. Machine learning screening of metal-ion battery electrode materials[J]. ACS Applied Materials & Interfaces, 2021, 13(45): 53355-53362.

[96] 刘亚利, 吴娇杨, 李泓. 锂离子电池基础科学问题(Ⅸ)——非水液体电解质材料[J]. 储能科学与技术, 2014, 3(3): 262-282.
LIU Y L, WU J Y, LI H. Fundamental scientific aspects of lithium ion batteries (Ⅸ)—Nonaqueous electrolyte materials[J]. Energy Storage Science and Technology, 2014, 3(3): 262-282.

[97] ZHANG J G, XU W, XIAO J, et al. Lithium metal anodes with nonaqueous electrolytes[J]. Chemical Reviews, 2020, 120(24): 13312-13348.

[98] LI Y Y, STROE D I, CHENG Y H, et al. On the feature selection for battery state of health estimation based on charging-discharging profiles[J]. Journal of Energy Storage, 2021, 33: doi: 10.1016/j.est.2020.102122.

[99] ATTIA P M, GROVER A, JIN N, et al. Closed-loop optimization of fast-charging protocols for batteries with machine learning[J]. Nature, 2020, 578(7795): 397-402.

[100] GONG W Y, YAN Q M. Graph-based deep learning frameworks for molecules and solid-state materials[J]. Computational Materials Science, 2021, 195: doi: 10.1016/j.commatsci.2021.110332.

[101] TRÉMEAU A, XU S X, MUSELET D. Deep Learning for Material recognition: Most recent advances and open challenges[C]// International Conference on “Big Data, Machine Learning and Applications” (BIGDML), Silchar, India, 2019.

[102] DUVENAUD D, MACLAURIN D, AGUILERA-IPARRAGUIRRE J, et al. Convolutional networks on graphs for learning molecular fingerprints[C]// Proceedings of the 28th International Conference on Neural Information Processing Systems, 2015: 2224-2232.

[103] WU Z Q, RAMSUNDAR B, FEINBERG E N, et al. MoleculeNet: a benchmark for molecular machine learning[J]. Chemical Science, 2017, 9(2): 513-530.

[104] XIE T, GROSSMAN J C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties[J]. Physical Review Letters, 2018, 120(14): doi: 10.1103/PhysRevLett.120.145301.

[105] KARAMAD M, MAGAR R, SHI Y T, et al. Orbital graph convolutional neural network for material property prediction[J]. Physical Review Materials, 2020, 4(9): doi: 10.1103/PhysRevMaterials.4.093801.

[106] BANJADE H R, HAURI S, ZHANG S S, et al. Structure motif-centric learning framework for inorganic crystalline systems[J]. Science Advances, 2021, 7(17): doi: 10.1126/sciadv.abf1754.

[107] CHEN C, YE W K, ZUO Y X, et al. Graph networks as a universal machine learning framework for molecules and crystals[J]. Chemistry of Materials, 2019, 31(9): 3564-3572.

[108] PARK C W, WOLVERTON C. Developing an improved crystal graph convolutional neural network framework for accelerated materials discovery[J]. Physical Review Materials, 2020, 4(6): doi: 10.1103/PhysRevMaterials.4.063801.

[109] AHMAD Z, XIE T, MAHESHWARI C, et al. Machine learning enabled computational screening of inorganic solid electrolytes for suppression of dendrite formation in lithium metal anodes[J]. ACS Central Science, 2018, 4(8): 996-1006.

[110] ZHOU L M, YAO A M, WU Y J, et al. Machine learning assisted prediction of cathode materials for Zn-ion batteries[J]. Advanced Theory and Simulations, 2021, 4(9): doi: 10.1002/adts.202100196.

[111] CHEN C, ZUO Y, YE W, et al. Learning properties of ordered and disordered materials from multi-fidelity data[J]. Nature Computational Science, 2021, 1(1): 46-53.

[112] LI S N, LIU Y J, CHEN D, et al. Encoding the atomic structure for machine learning in materials science[J]. WIREs Computational Molecular Science, 2022, 12(1): doi: 10.1002/wcms.1558.

[113] PIETSCH P, WOOD V. X-ray tomography for lithium ion battery research: A practical guide[J]. Annual Review of Materials Research, 2017, 47: 451-479.

[114] DIXIT M B, VERMA A, ZAMAN W, et al. Synchrotron imaging of pore formation in Li metal solid-state batteries aided by machine learning[J]. ACS Applied Energy Materials, 2020, 3(10): 9534-9542.

[115] JIANG Z, LI J, YANG Y, et al. Machine-learning-revealed statistics of the particle-carbon/binder detachment in lithium-ion battery cathodes[J]. Nature Communications, 2020, 11: doi: 10.1038/s41467-020-16233-5.

[116] BALIYAN A, IMAI H. Machine learning based analytical framework for automatic hyperspectral Raman analysis of lithium-ion battery electrodes[J]. Scientific Reports, 2019, 9: doi: 10.1038/s41598-019-54770-2.

[117] FURAT O, FINEGAN D P, DIERCKS D, et al. Mapping the architecture of single lithium ion electrode particles in 3D, using electron backscatter diffraction and machine learning segmentation[J]. Journal of Power Sources, 2021, 483: doi: 10.1016/j. jpowsour. 2020.229148.

[118] KHATAVKAR N, SWETLANA S, SINGH A K. Accelerated prediction of Vickers hardness of Co- and Ni-based superalloys from microstructure and composition using advanced image processing techniques and machine learning[J]. Acta Materialia, 2020, 196: 295-303.

[119] KONONOVA O, HE T J, HUO H Y, et al. Opportunities and challenges of text mining in materials research[J]. iScience, 2021, 24(3): doi: 10.1016/j.isci.2021.102155.

[120] TORAYEV A, MAGUSIN P C M M, GREY C P, et al. Text mining assisted review of the literature on Li-O_2 batteries[J]. Journal of Physics: Materials, 2019, 2(4): doi: 10.1088/2515-7639/ab3611.

[121] EL-BOUSIYDY H, LOMBARDO T, PRIMO E N, et al. What can text mining tell us about lithium-ion battery researchers’ habits? [J].

Batteries & Supercaps, 2021, 4(5): doi: /10.1002/batt.202000288.

[122] NIE Z W, LIU Y J, YANG L Y, et al. Construction and application of materials knowledge graph based on author disambiguation: Revisiting the evolution of LiFePO 4[J]. Advanced Energy Materials, 2021, 11(16): doi: 10.1002/aenm.202003580.

[123] ZHU Y Z, HE X F, MO Y F. Origin of outstanding stability in the lithium solid electrolyte materials: Insights from thermodynamic analyses based on first-principles calculations[J]. ACS Applied Materials & Interfaces, 2015, 7(42): 23685-23693.

[124] MAHBUB R, HUANG K, JENSEN Z, et al. Text mining for processing conditions of solid-state battery electrolytes[J]. Electrochemistry Communications, 2020, 121: doi: 10.1016/j.elecom.2020.106860.

[125] SWAIN M C, COLE J M. ChemDataExtractor: A toolkit for automated extraction of chemical information from the scientific literature[J]. Journal of Chemical Information and Modeling, 2016, 56(10): 1894-1904.

[126] HUANG S, COLE J M. A database of battery materials auto-generated using ChemDataExtractor[J]. Scientific Data, 2020, 7: doi: 10.1038/s41597-020-00602-2.

[127] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.
ZHOU Z H. Machine learning[M]. Beijing: Tsinghua University Press, 2016.

[128] VAN DYK D A, MENG X L. The art of data augmentation[J]. Journal of Computational and Graphical Statistics, 2001, 10(1): 1-50.

[129] NAAZ F, HERLE A, CHANNEGOWDA J, et al. A generative adversarial network-based synthetic data augmentation technique for battery condition evaluation[J]. International Journal of Energy Research, 2021, 45(13): 19120-19135.

[130] HSU T, EPTING W K, KIM H, et al. Microstructure generation via generative adversarial network for heterogeneous, topologically complex 3D materials[J]. JOM, 2021, 73(1): 90-102.

[131] LOOKMAN T, BALACHANDRAN P V, XUE D, et al. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design[J]. Npj Computational Materials, 2019, 5: doi: 10.1038/s41524-019-0153-8.

[132] ZHUANG F Z, QI Z Y, DUAN K Y, et al. A comprehensive survey on transfer learning[J]. Proceedings of the IEEE, 2021, 109(1): 43-76.

[133] WANG Z L, WANG Q X, HAN Y Q, et al. Deep learning for ultra-fast and high precision screening of energy materials[J]. Energy Storage Materials, 2021, 39: 45-53.

[134] THORNTON C, HUTTER F, HOOS H H, et al. Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms[C]//KDD '13: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2013: 847-855.

[135] FEURER M, KLEIN A, EGGENSPERGER K, et al. Auto-sklearn: Efficient and Robust Automated Machine LearningAutomated Machine Learning[M]. USA: Springer, 2019: 113-134.

[136] DUNN A, WANG Q, GANOSE A, et al. Benchmarking materials property prediction methods: The Matbench test set and Automatminer reference algorithm[J]. Npj Computational Materials, 2020, 6: doi: 10.1038/s41524-020-00406-3.

[137] DENKER J, SCHWARTZ D, WITTNER B, et al. Large automatic learning, rule extraction, and generalization[J]. Complex Systems, 1987, 1: 877-922.

[138] KOH P W, LIANG P. Understanding black-box predictions via influence functions[C]//34th International Conference on Machine Learning, 2017: 1885-1894.

[139] FRYE C, FEIGE I, ROWAT C. Asymmetric shapley values: Incorporating causal knowledge into model-agnostic explainability[J]. Advances in Neural Information Processing Systems, 2020, 33: 1229-1239.

[140] LI Y H, XIAO B, TANG Y C, et al. Center-environment feature model for machine learning study of spinel oxides based on first-principles computations[J]. The Journal of Physical Chemistry C, 2020, 124(52): 28458-28468.

[141] WENG B, SONG Z, ZHU R, et al. Simple descriptor derived from symbolic regression accelerating the discovery of new perovskite catalysts[J]. Nature Communications, 2020, 11: doi: 10.1038/s41467-020-17263-9.

[142] GONG S, WANG S, ZHU T S, et al. Screening and understanding Li adsorption on two-dimensional metallic materials by learning physics and physics-simplified learning[J]. JACS, 2021, 1(11): 1904-1914.

[143] FLORES M J, NICHOLSON A E, BRUNSKILL A, et al. Incorporating expert knowledge when learning Bayesian network structure: A medical case study[J]. Artificial Intelligence in Medicine, 2011, 53(3): 181-204.

[144] TANG W Y, MAO K Z, MAK L O, et al. Adaptive fuzzy rule-based classification system integrating both expert knowledge and data[C]//2012 IEEE 24th International Conference on Tools with Artificial Intelligence, Athens, Greece. IEEE, 2012: 814-821.

[145] REN Z, OVIEDO F, THWAY M, et al. Embedding physics domain knowledge into a Bayesian network enables layer-by-layer process innovation for photovoltaics[J]. npj Computational Materials, 2020, 6: doi: 10.1038/s41524-020-0277-x

[146] GHALLAB M, SPYROPOULOS C D, FAKOTAKIS N, et al. Fighting knowledge acquisition bottleneck with argument based machine learning[J]. Frontiers in Artificial Intelligence and Applications, 2008, 178: 234-238.

[147] LIU Y, WU J M, WANG Z C, et al. Predicting creep rupture life of Ni-based single crystal superalloys using divide-and-conquer approach based machine learning[J]. Acta Materialia, 2020, 195: 454-467.