第 50 卷第 3 期
2022 年 3 月

硅 酸 盐 学 报
JOURNAL OF THE CHINESE CERAMIC SOCIETY

Vol. 50，No. 3
March，2022

http://www.gxyb.cbpt.cnki.net

DOI：10.14062/j.issn.0454-5648.20220093

综 合 评 述

材料领域知识嵌入的机器学习

刘 悦[1,4,5]，邹欣欣[1]，杨正伟[1]，施思齐[2,3,5]

(1. 上海大学计算机工程与科学学院，上海 200444；2. 上海大学材料科学与工程学院，上海 200444；
3. 上海大学材料基因组工程研究院，上海 200444；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，杭州 311100)

摘 要：数据驱动的机器学习因其能够快速拟合历史数据中的潜在模式并实现材料性能的精准预测，已被广泛应用于材料性能优化和新材料设计。然而，由于缺乏描述符间关联关系、材料性能驱动机制等材料领域知识的指导，数据驱动的机器学习在实际应用中常常出现与材料基础理论认知或原理不一致的结果。本工作通过分析材料数据的特点和数据驱动的机器学习建模原理，厘清了数据驱动的机器学习应用于材料领域面临的三大矛盾：高维度与小样本数据的矛盾、模型准确性与易用性的矛盾、模型学习结果与领域专家知识的矛盾。藉此提出材料领域知识嵌入的机器学习作为上述矛盾的调和策略。进一步，面向“目标定义–数据准备–数据预处理–特征工程–模型构建–模型应用”的机器学习全流程，通过剖析相关的基础性和探索性工作，探讨了在机器学习各阶段实现材料领域知识嵌入的关键技术。最后，展望了材料领域知识嵌入机器学习的发展机遇和挑战。

关键词：材料设计；机器学习；材料数据
中图分类号：TB3; TP3　　**文献标志码**：A　　**文章编号**：0454–5648(2022)03–0863–14
网络出版时间：2022–03–01

Machine Learning Embedded with Materials Domain Knowledge

LIU Yue[1,4,5], *ZOU Xinxin*[1], *YANG Zhengwei*[1], *SHI Siqi*[2,3,5]

(1. School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China;
2. School of Materials Science and Engineering, Shanghai University, Shanghai 200444, China;
3. Materials Genome Institute, Shanghai University, Shanghai 200444, China;
4. Shanghai Engineering Research Center of Intelligent Computing System, Shanghai 200444, China;
5. Zhejiang Laboratory, Hangzhou 311100, China)

Abstract: Data-driven Machine Learning (ML) has been widely used in materials performance optimization and novel materials design due to its ability to quickly fit potential data patterns and achieve accurate prediction. However, the results of data-driven ML are often inconsistent with the materials basic theory or principle, which results mainly from the lack of the guidance of materials domain knowledge, e.g., the correlation among descriptors and the driving mechanism associated with the properties. Herein, by analyzing the characteristics of materials data and the modeling principle of data-driven ML methods, we clarify the three main contradictions occurring to the application of ML in materials science, i.e., the contradictions between high dimension and small sample, accuracy and usability of models, learning results and domain knowledge. Following this, we propose the ML method embedded with materials domain knowledge to reconcile these three contradictions. Further, surrounding the whole ML process including target definition, data collection and preprocessing, feature engineering, model construction and application, we explore some key techniques to realize domain knowledge embedding by summarizing the related basic and exploratory efforts. Finally, opportunities and challenges facing the ML method embedded with domain knowledge are also discussed.

Keywords: materials design; machine learning; materials data

收稿日期：2022–01–30。　修订日期：2022–02–15。
基金项目：国家自然科学基金面上项目(52073169)；国家重点研发计划(2021YFB3802100)；之江实验室科研攻关项目(2021PE0AC02)。
第一作者：刘 悦(1975—)，女，博士，教授。
通信作者：施思齐(1978—)，男，博士，教授。

Received date: 2022–01–30.　**Revised date:** 2022–02–15.
First author: LIU Yue (1975–), female, Ph.D., Professor.
E-mail: yliu@staff.shu.edu.cn
Correspondent author: SHI Siqi (1978–), male, Ph.D., Professor.
E-mail: sqshi@shu.edu.cn

机器学习是研究计算机如何模拟人类学习行为以自动获取知识，从而不断改善自身性能的一门学科[1]。机器学习方法能够更快速、准确地从历史数据中挖掘出有效的关联信息。相对于基于实验测量和模拟计算的传统材料研究模式，机器学习方法具有更快、更准、更省地获得“成分–结构–工艺–性能”间相互关联的特点，近年来被广泛地应用于材料领域的科学研究[2–4]。

根据学习方式的不同，机器学习可分为有监督学习[3]、无监督学习[5]、自监督学习[6]、半监督学习[7]、主动学习[8]和强化学习[9]等。如图 1 所示，这些机器学习方法均已在材料领域得到了应用，涉及高温合金[8–12]、高熵合金[13–15]、特种玻璃[16]、快离子导体[17–20]、热电[21]、铁电[22]、光伏[23]等多种材料体系。有监督学习方法在材料领域的应用最为广泛，其通过训练材料标注数据来实现精准的性能预测[3]；无监督和自监督学习方法以无标注的材料数据为监督信息，通过为材料数据构造伪标签或学习材料数据特征来区分材料的性能优劣[5, 20]，并分析导致不同材料性能差异的关键因素[23, 25]；半监督学习方法在标记样本数量较少的情况下，通过在模型训练过程中引入无标记样本来避免传统监督学习在训练样本不足(训练不充分)时出现模型性能退化的问题，在材料的逆向设计、性能预测、材料文本挖掘等方面都有初步应用[26–29]；主动学习方法以预先构建的机器学习预测模型为基础，自动建立平衡材料性能预测值与模型不确定性的效能方程，主要应用于材料的优化设计[22, 26–32]；强化学习方法通过智能体与环境的交互不断更新算法策略和获取数据，在一定程度上摆脱了数据量的束缚，主要用于解决从超大参数组合空间中寻找最优设计方案的问题[33–35]。

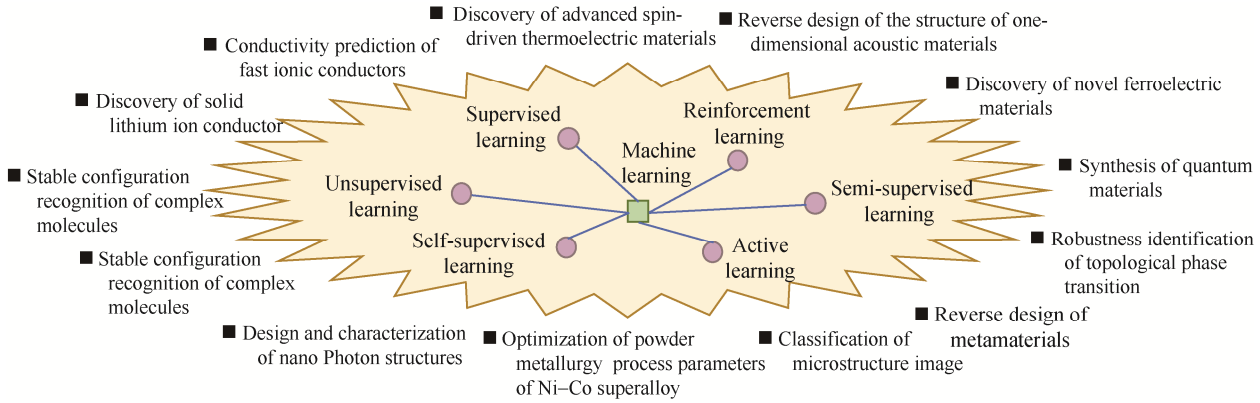


图 1 各类机器学习方法在材料领域科学研究中的应用举例

Fig. 1 Application cases of various machine learning methods in materials science

综上所述，材料领域的科学研究中已经涌现出了诸多机器学习应用的成功案例。然而，传统的机器学习方法仅以数据为驱动力，通常假设学习样本符合某种特定的数据分布，但材料数据往往是高维度、小样本、高噪音的，难以服从某种特定的数据分布。这就导致在实际应用机器学习方法进行材料研究时，经常会出现学习结果与领域知识或专家经验不一致，甚至是相悖的现象，这一问题严重制约着机器学习方法在材料领域更大范围地应用，尤其是用于大规模的材料工业设计和制备。因此，研究人员在利用数据驱动的机器学习方法的同时，还应该注重材料领域知识的重要性。构建材料领域知识嵌入的机器学习新方法，削弱机器学习模型对数据分布的敏感性，实现机器学习建模过程的领域知识指导，将有利于提升机器学习在材料领域应用的准确性、普适性和可解释性。

本综述旨在阐明构建材料领域知识嵌入的机器学习新方法的必要性、可行性和潜在发展方向。首先，厘清了数据驱动的机器学习在材料领域应用的全流程及其面临的三大关键矛盾；然后，为调和三大关键矛盾，提出了面向机器学习全流程的材料领域知识嵌入的机器学习方法，并总结了材料领域中调和三大矛盾的基础性和探索性工作；最后，展望了材料领域知识嵌入机器学习方法的未来发展。

1 材料领域机器学习全流程

如图 2 所示，机器学习在材料领域应用的全流程可分为目标定义、数据准备、数据预处理、特征工程、模型构建和模型应用等 6 个阶段。整个过程伴随着材料数据和相关材料领域知识在各阶段间的

流动(对应图 2 中的材料数据流和领域知识流)。其中，材料数据是每个阶段的基本操作对象，而材料领域知识则负责帮助研究人员针对特定的机器学习任务建立最优的数据分析方案。

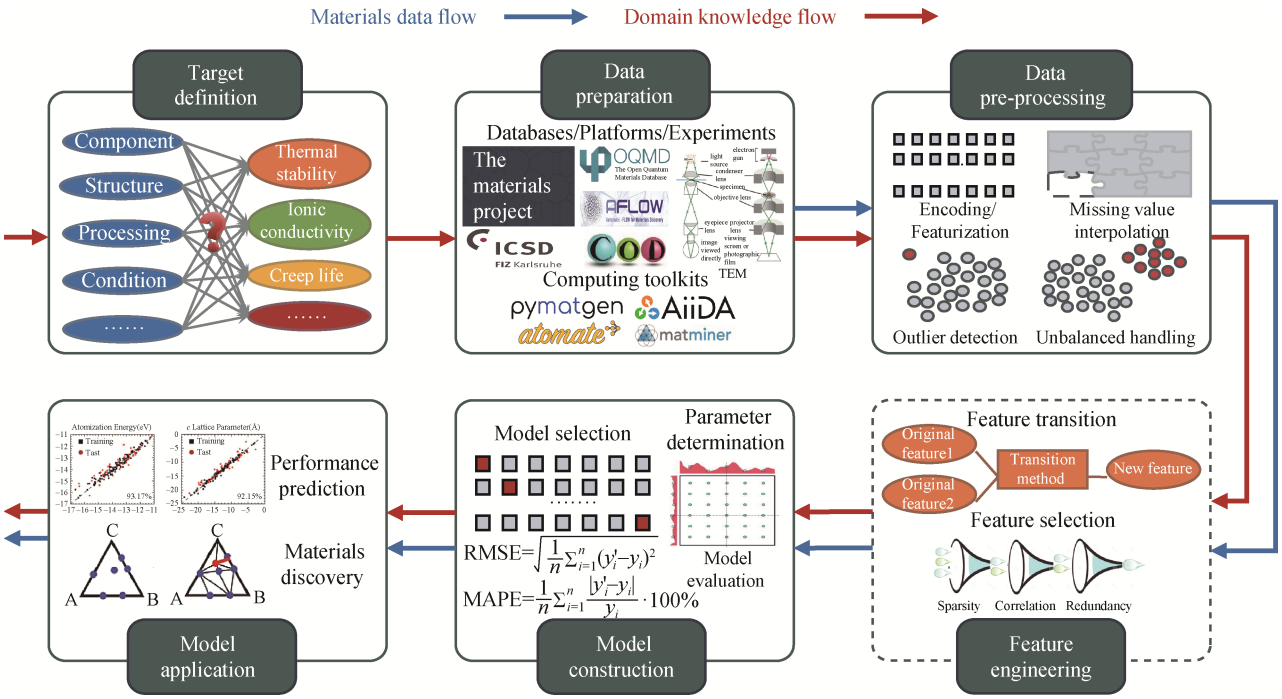


图 2　机器学习在材料领域应用的全过程

Fig. 2　Whole process of machine learning application in materials science

目标定义和数据准备阶段主要由材料领域知识驱动。前者的目的是将特定材料研究问题转换成机器学习建模任务。通常情况下，研究人员需要明确具体的材料体系以及拟学习的“构效关系”，后者则要求研究人员根据预定义的目标，做出一系列的材料数据获取决策，包括明确材料结构文件的获取途径、确定材料性能的评价指标及其获取方式、定义影响材料性能的材料/物理/化学参数(即描述符)及其表征方式等，进而根据这些决策构建用于机器学习建模的原始材料数据集。

基于上述材料数据集，数据预处理、特征工程、模型构建及其应用阶段同时拥有材料数据流和领域知识流。数据预处理阶段为机器学习模型提供可学习的高质量样本，包括数据的编码和特征化、缺失值插补[36]、异常值检测[37]和对数据不均衡分布的处理[38]等，在该阶段，材料领域知识常被用来校对数据驱动的通用预处理方法[39–40]的学习结果。特征工程阶段旨在通过特征转换以扩充稀疏的材料数据集，或对高维的材料数据集进行特征约简，以识别出更合适的特征用于建模，从而提高机器学习模型的预测性能[41]。然而，数据驱动的特征工程方法有时并不能直接给出最优特征，需要材料专家凭借自身掌握的领域知识进行最终确定，如单变量的相关性分析方法仅能识别彼此高度相关的特征，却难以决定这些特征的去留，因为材料“构效关系”总是多个特征耦合的结果。

模型构建阶段的任务是通过模型评估来选择模型并确定其超参数，以找到具有最佳预测性能的机器学习模型，在该阶段，由于现有机器学习模型种类繁多，研究人员常常凭借自身对材料潜在“构效关系”的理解或猜测，预先对候选机器学习模型进行选择，如线性模型[17, 42–43]、树形模型[44–45]、集成模型[46–48]等，进而通过评估候选模型的性能来选择最佳模型，材料领域主要以模型的预测精度为评估指标，构建机器学习模型的最终目的是指导新型高性能材料的设计与发现。因此，在模型应用阶段，除了模型的预测精度之外，领域专家还需要凭借材料领域知识对机器学习模型的结果进行解释，通过明确材料性能的驱动机制来指导新型材料的设计与合成。

综上所述，机器学习在材料领域应用的每个过程都离不开领域知识的参与。然而，目前材料领域知识虽然在目标定义和数据准备阶段起主导作用，但在机器学习建模过程中却仅起到校验数据驱动算法结果的作用，并没有对其学习过程进行实质性干预，仍然无法消除数据驱动方法受制于高维度、小样本、高噪音的材料数据，易产生错误结论的现象。

2 材料领域机器学习的三大矛盾

为了解决材料领域知识不充分参与导致的数据驱动机器学习方法易产生不可解释结论的问题，需要明确数据驱动的机器学习方法与材料领域应用不适配的基本情景。在此，将其归结为机器学习在材料领域应用的三大矛盾[49]，并在以下小节中进行介绍。

2.1 高维度与小样本数据的矛盾

相对于其他领域，材料数据通常是小样本的且有时更加多样化[50]。图 3 统计了 103 篇材料领域科研文献(明确说明数据集规模)中的 105 份机器学习建模用材料数据集所包含的样本量。其中，约 56%的数据集样本量不足 500，约 65%的数据集的样本量不足 1 000，而仅有约 35%的数据集样本量超过 1 000。小样本材料数据必然会影响机器学习模型的构建[51]。如 Faber 等[52]发现，随着训练集的增大，机器学习模型对钾冰晶石形成能的预测精度有了明显提高。Schmidt 等[53]发现当训练集增加 1 倍时，钙钛矿化合物的形成能预测模型的误差减少了约 20%。

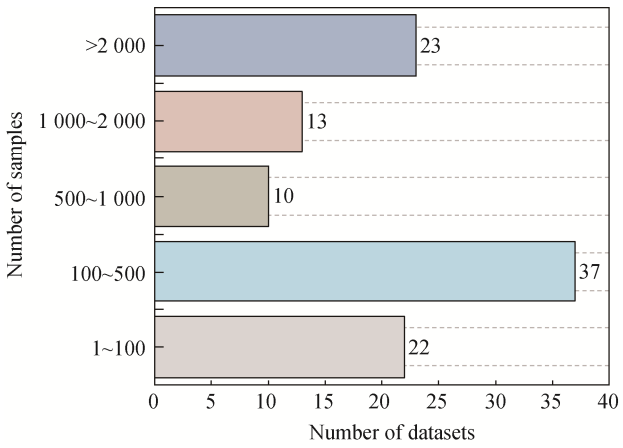


图 3 105 份用于机器学习的材料数据集的样本数量

Fig. 3 Number of samples in 105 materials datasets for machine learning

由于材料性能具有复杂的驱动机制，常常受多种物理/化学因素的影响，故材料数据集通常具有较高的维度，即材料研究人员通常会定义较多的特征来描述一种不确定的材料性能驱动机制。如图 4 所示，在图 3 基础上，进一步统计了样本量不超过 500 的材料数据集(共 59 份数据集)特征数量与样本数量的比率。其中，约 34%的数据集中特征数量超过样本量的 1/4，甚至有约 8%的数据集中特征数量超过了样本总量。一方面，高维特征中容易存在冗余特征，会降低机器学习模型的预测性能和效率[3, 54]；另一方面，高维特征会导致机器学习模型过于复杂，从而不利于材料专家对结果进行解释与分析[3]。

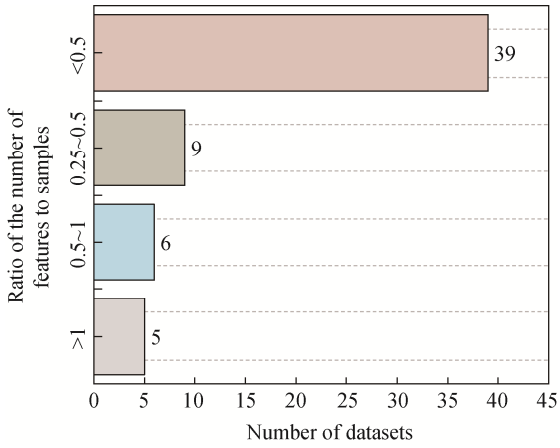


图 4 59 份小样本材料数据集的特征量与样本量比率

Fig. 4 Ratio of number of features to sample size in 59 small-sample materials datasets

2.2 模型准确性与易用性的矛盾

应用机器学习解决材料问题的目的是通过构建精准的机器学习模型，并从中抽取可解释的规则，来帮助材料领域专家更深层次地理解材料的性能驱动机制。线性模型可以为材料专家提供多个影响因素与目标性能之间的线性关系，从而能够直观地表达影响因素对目标性能的正负相关性和重要程度。Sendek 等[17]通过建立逻辑回归(LR)模型，准确地识别出平均子晶格化学键离子性特征与固态锂离子导体材料离子电导率的负相关关系；Wang 等[55]利用偏最小二乘回归(PLS)方法发现了 X_4^+ 半径和 X 八面体相关描述符对锂离子电池阴极体积变化有着显著影响。

事实上，大部分材料的性能驱动机制难以通过简单的线性模型进行拟合，故复杂的非线性模型得到更为广泛的应用。如 Zhao 等[56]利用支持向量回归模型(SVR)准确预测了沥青层的孔隙率，测试集上的拟合优度(R^2)为 0.8865；Jiang 等[57]建立了用于陶瓷耐火材料接触角预测的 Gaussian 过程回归模型(GPR)，预测准确率高达 96%；Ye 等[58]利用深度神经网络(DNNs)模型成功预测了混合石榴石和钙钛矿材料的稳定性。然而，这些非线性模型大多是“黑盒”模型，其建模过程和结果往往难以解释，且得到预测性能优异的非线性模型通常需要经过繁琐耗时的参数确定过程。因此，机器学习在用于解决材料问题时面临着模型准确性与易用性的矛盾。

2.3 模型学习结果与领域专家知识的矛盾

目前应用于材料领域的机器学习方法绝大多数

是纯数据驱动的，其准确性严重依赖于可收集的材料样本数据，无法从根本上保证机器学习结果与领域知识的一致性，从而导致机器学习结果与领域专家知识的矛盾。如 Iwasaki 等[21]在应用机器学习识别先进自旋驱动热电材料时发现，通用目标的随机森林模型(RF)、套索方法(LASSO)和多元线性回归(MLR)虽然具备较高的预测精度，但均无法找到影响自旋驱动热电材料热能的、与领域知识一致的关键属性。Hu 等[59]发现通用的 K 最近邻(KNN)算法一般假设特征具有相等权重，但实际上不同因素对锂离子电池容量的影响程度不同，于是，利用粒子群优化算法为每个特征寻找合理的权重，结果发现 KNN 方法的预测精度得到了明显提升，且得到了可解释的特征重要度。

3　材料领域知识嵌入的机器学习

长期以来，材料领域在经验、理论和计算科学的研究范式中，已经积累了丰富的领域知识。如今，材料领域迎来了数据驱动科学的研究范式[60]。目前的工作更多地使用数据驱动的机器学习方法来辅助材料科学研究，期望其能够从历史数据中挖掘出有价值的信息，以缩短材料科学的研究周期。

人类的认知过程表明，知识的来源不仅是对实例的分析，还有历史经验的指导。因此，机器学习不仅要从历史数据中挖掘潜在模式，材料领域知识的指导也至关重要。基于这一点，本研究认为材料领域知识的作用不能仅停留在定义目标、准备数据和校验数据驱动算法的结果上，而应该贯穿机器学习全流程的各个阶段。为此，本研究提出了一种新型机器学习方法——材料领域知识嵌入的机器学习，将材料领域知识进行符号化表示，并嵌入到机器学习方法的模型(Model)、策略(Strategy)和算法(Algorithm)中，建立作用于机器学习各阶段的领域知识嵌入方法，实现材料领域知识在机器学习全流程的有机融入，从而通过数据和知识的双向驱动来构建高精度且具有一定可解释性的机器学习新模型，以系统地调和材料领域机器学习三大关键矛盾。具体如下：

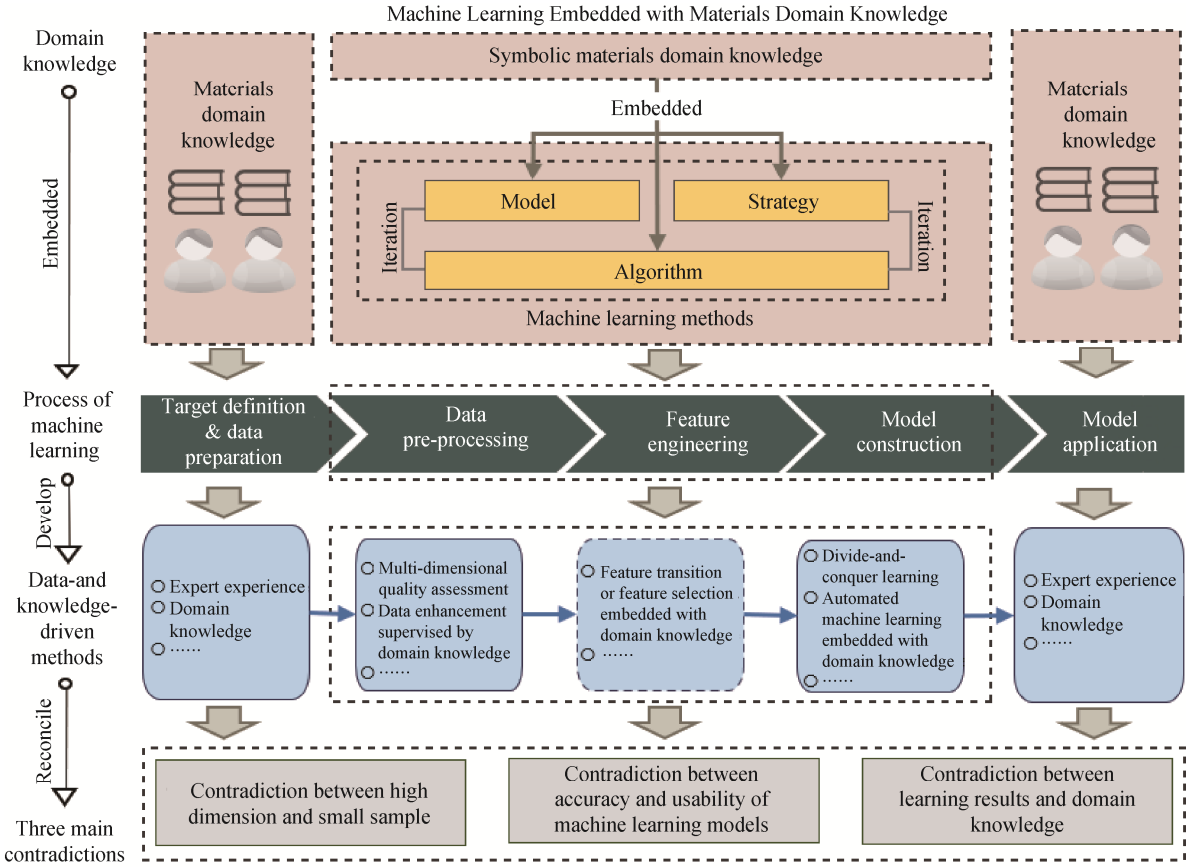


图 5　材料领域知识嵌入的机器学习

Fig. 5　Machine learning with materials domain knowledge embedding

3.1　数据预处理阶段

机器学习模型性能的上限由输入数据的质量决定[60]。然而，材料数据往往具有多源、异构、不确定性强、样本少、维度高等特点。因此，在应用机

器学习解决材料科学问题前，对材料数据进行质量评估和提升具有十分重要的意义。融合材料领域知识对材料数据进行品质提升和样本增强，有利于提高材料数据的“质”和扩增建模所用材料数据的“量”，从而调和材料数据小样本与高维度的矛盾。

对于材料数据的“质”提升，领域知识可用于材料数据的修正。收集描述符取值的经验范围、描述符取值与其他物理现象的关联关系、相似或不相似数据的产生条件等材料领域知识，并将其转化为数值不等式、二元关系和 IF–THEN–ELSE 规则等定量表示形式，再结合数据驱动的置信度评估、相似性度量和异常点检测等方法，可识别不符合物理化学规律的异常数据。领域知识也可用于将原始材料数据转换为适合机器学习的表示形式。如考虑不同描述符的经验取值范围或样本的潜在类别关系，对不同描述符或不同样本执行不同的归一化操作，从而加速模型收敛、提升模型预测准确度。基于生成式对抗网络(GAN)的数据增强方法[63]可用于扩增材料数据的“量”。通过 GAN 的生成模型学习材料数据的分布并生成合成数据，再利用物理模型和融合领域知识的数据质量检测方法辅助 GAN 的判别模型对生成数据的可靠性进行判断，并将结果反馈给生成模型以用于其自身优化，从而实现虚拟材料数据的可靠生成。

3.2 特征工程阶段

建立材料领域知识嵌入的特征选择和转换方法，合理地筛选或构造有物理意义的描述符，将有利于降低特征维度，进而构建更简单但准确的机器学习模型。因此，该阶段构建的材料领域知识嵌入的机器学习模型将有望同时调节高维度和小样本的矛盾，以及模型准确性和易用性的矛盾。

材料领域知识可用于冗余特征的消除或关键描述符的遴选。如将材料领域专家对描述符重要度的评估进行粗粒度定量化表示[54]，并联合相关性系数[64]、LASSO 或 RF 模型、Shapley 值[65]等属性重要度分析技术，构建综合的特征重要度评估指标作为数据驱动算法的特征选择依据，能够避免数据驱动的算法对关键描述符的误删或误判，高度相关的描述符同时用于训练易导致机器学习模型结果的不可解释，利用描述符间关联关系的领域知识判别多元描述符之间关联关系，构建数学函数统计高度相关的描述符同时出现在训练数据集中的情况，并将其与评估描述符集合预测能力的适应度函数联合作为特征选择的优化目标，能够有效地遴选出内部相关性较低但预测能力较高的描述符子集[66]。

3.3 模型构建阶段

一方面，直接利用半监督学习、主动学习、迁移学习等建模方法能在一定程度上解决小样本数据的问题。其中，半监督[29]和主动学习[67]方法通过有策略地利用未标注数据来优化模型或参数空间，迁移学习则通过迁移相关领域/模型中的知识/经验来提高目标学习器的性能，从而减少目标学习器对数据量的依赖[68]。

另一方面，模型、策略和算法是机器学习方法的三要素[61]，实现材料领域知识到三要素的嵌入对提升机器学习的准确性和可解释性具有十分重要的意义。其中，模型主要是指模型的假设空间，决定了所要学习的条件概率分布或决策函数，如线性模型的假设空间是所有线性函数构成的函数集合。策略是从假设空间中选择最优模型的学习准则，可形式化为机器学习模型的目标函数，一般由损失函数或风险函数和正则项构成。在某些情况下，目标函数还伴有约束条件。如支持向量机模型(SVM)的策略就是一个凸二次规划问题。算法则是求解最优模型的具体计算方法，如梯度下降法、Newton's 法和共轭梯度法等。具体来说：

1) 在“模型”上的材料领域知识嵌入可借鉴材料领域专家对材料数据潜在模式的理解，将材料问题“化繁为简”，建立可解释的集成模型，从而起到调和模型学习结果与材料领域专家知识矛盾的作用。如根据不同组分空间、实验和处理工艺等条件下的材料性能驱动机制可能存在的不同，可基于“分而治之”理念，构建能够自动对数据集进行分组，并在不同分组下构建不同最优模型的集成模型[11]。

2) 从材料领域知识中抽取理论或经验准则，并将其进行定量化表示以嵌入到机器学习的目标函数或约束条件中，是材料领域知识到“策略”嵌入的主要途径。如可通过结合材料领域知识为 SVM 定义专属核函数或者为 KNN 分配特征权重等途径来实现领域知识到约束条件的嵌入；可通过增加正则化项来实现领域知识到目标函数的嵌入。

3) 对于机器学习方法的“算法”，可引入求解材料问题的理论推导模型，通过改进梯度下降法、牛顿法等传统优化方法的求解过程来实现材料领域知识的嵌入。

此外，借鉴领域专家的建模经验或相关领域知识，实现机器学习模型的自动化或半自动化建模，

也有望调和模型准确性与易用性的矛盾。

4　材料领域知识嵌入机器学习的基础与探索

材料领域知识嵌入的机器学习方法以传统数据驱动的算法为基础，故现阶段应用于材料领域的数据驱动机器学习方法能够为构建材料领域知识嵌入的机器学习方法提供研究基础。本研究团队也在构建材料领域知识嵌入机器学习方面展开了一些探索性的工作，能够为构建面向机器学习全流程的领域知识嵌入的机器学习提供实证和借鉴。

4.1　数据预处理阶段的基础与探索

伴随着机器学习在材料领域的广泛应用，决定机器学习模型上限的材料数据质量越来越被重视。材料领域的研究者们分别围绕数据的“质”和“量”展开了研究，期望通过提升数据的品质和数量来进一步提升机器学习结果的可信性。

对于材料数据的“质”提升问题，Xu 等[69]和 Hafiz 等[70]分别开发了一个编译程序和多层感知机(MLP)模型以提升钙钛矿材料形成能数据和 f-电子体系结构数据的准确性；Li 等[71]探讨了数据分布的不均衡性对 ML 模型的影响，通过在不同组分空间中建立不同的分类或回归模型，进而对预测结果进行集成以提升钙钛矿氧化物凸包能量的预测准确性；Wang 等[72]强调了在许可、法律和知识产权保护允许的情况下材料数据可溯源性的重要性，并规定了一系列科研手稿中必要的数据溯源信息；Ghiringhelli 等[73]通过定义层次化的元数据实现了先进电子结构计算过程与结果的溯源。FAIR 原则[74–75]是大数据环境下的一项科学数据共享准则，倡导努力实现科研活动产出数据可发现、可访问、可互操作和可重用。2018 年，Claudia 等[76]首次将 FAIR 原则引入到大数据驱动的材料科学研究中，主张通过构建数据仓库、规范化数据存档、数据知识百科全书、大数据分析和可视化工具以确保材料数据品质[77]。

聚焦材料数据的“量”提升问题，许多材料数据库和高通量计算平台被建立，目的是为数据驱动方法提供数据基础，如剑桥结构数据库(CSD)[78]、无机晶体结构数据库(ICSD)[79]、开放量子材料数据库(QQMD)[80]、Paulina 文件库[81]、材料工程数据库(MP)[82]和 Aflowlib[83]等。此外，基于 GAN 的数据增强方法也已经被用于实现样本量扩增。如 Sanchez-Lengeling 等[84]回顾了材料领域使用机器学习生成模型进行逆向设计的主要过程和若干主流方法；DAN 等[85]基于 GAN 提出了 1 种生成式机器学习模型(MatGAN)，以高效地生成假想的新型无机材料。

综上所述，材料科学领域的研究人员已经意识到了数据质量的重要性，并针对某些特定的质量问题进行了针对性研究，但仍缺乏对机器学习各阶段涉及的通用数据质量问题的系统探讨和指导框架。如何结合材料领域知识，开发材料数据的专有质量评估和提升方法是一个有待深入研究的问题。目前，其他领域已有少量工作通过融合领域知识来评估和提升数据质量，初步证明了领域知识在数据预处理阶段嵌入的有效性。如通过分析机器状态检测中缺失段和异常段的特点，Xu 等[86]提出了一种基于改进局部异常因子(LOF)方法的错误数据检测方法；根据“导致 SuperCOSMOS 巡天观测中 4 种异常记录的事件通常较短而难以被传统的异常值探测方法识别”的领域知识，Storkey 等[87]提出了一种更新字符串(Renewal string)方法以清理巡天观测数据库。

4.2　特征工程阶段的基础与探索

寻找和定量化合适的影响因子(即描述符或特征)来表征材料的性能是材料领域机器学习建模的一个关键问题[42]，已有大量的工作为不同材料体系的广泛性能提供了描述符的构造和选择方法。

一些工作致力于为材料领域的机器学习模型构建“更好”的描述符。Ward 等[88]定义了一个化学意义上不同的属性列表来帮助研究人员快速建立高精度的机器学习预测模型，包括化学计量属性、元素属性的统计量、电子结构属性和离子化合物属性等 4 大类，适用于广泛的无机材料体系。Li 等[89]提出了“中心–环境”(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。该方法成功应用于 5329 种尖晶石氧化物的形成能、晶格参数和能带隙预测。符号回归(SR)[90–91]是一种可解释的机器学习方法，它通过同时搜索参数集和数学公式集来生成表征特定关联的最优数学表达式。Weng 等[92]利用 SR 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 5 种新的氧化物钙钛矿。Wang 等[47]基于 24 种元素描述符和 8 种典型函数，利用特征交叉方法生成了 23 200 个特征，并从中筛选出 52 个特征用于类金刚石结构硫族化合物的能带预测，预测精度为 90.48%。

然而，描述符的累积也为机器学习带来了困扰，即高维描述符空间中可能存在的冗余描述符影响了

机器学习模型的预测精度和可解释性。于是，遴选出更符合材料领域知识和客观事实的描述符是材料领域科学研究在特征工程阶段的又一目标。如 Wen 等[93]开发了一种融合相关性分析和梯度提升算法的混合式特征选择方法，从 59 个描述符中成功筛选出 3 个关键描述符，以预测单相高熵合金的固溶强化效果，预测精度远高于现有的物理模型；IM 等[94]利用 Pearson 相关系数法分析了 32 个初始描述符之间的相关性，并剔除了 12 个冗余描述符，藉此建立了卤化物双钙钛矿的电子结构数据集生成热和带隙预测的梯度提升树(GBRT)模型，在测试集上的平均均方根误差(RMSE)分别为 0.021 eV/atom 和 0.022 3 eV。本研究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[54]，是材料领域知识在特征工程阶段嵌入的一个初步探索。如图 6 所示，该方法联合材料专家对描述符重要度的评估值和数据驱动方法的测量结果，建立了描述符重要度的综合评价指标。在 8 个材料数据集上，该方法被证明能够有效遴选出预测精度较高且符合专家经验的描述符集合。基于该方法，本研究进一步提出了领域规则约束的混合式特征选择方法[66]，将表征描述符间关联的材料领域知识表示成“不共现”规则，并将其作为约束条件来限制和引导基于粒子群优化算法的特征选择方法的进化方向，从而找到稀疏性尽可能小、相关性尽可能大、冗余性尽可能低且符合材料领域知识的描述符集合。在包含 85 个三方相 NASICON 型固态电解质材料的数据集上，该方法成功构建了测试集上 R^2 为 0.95 的激活能预测模型。

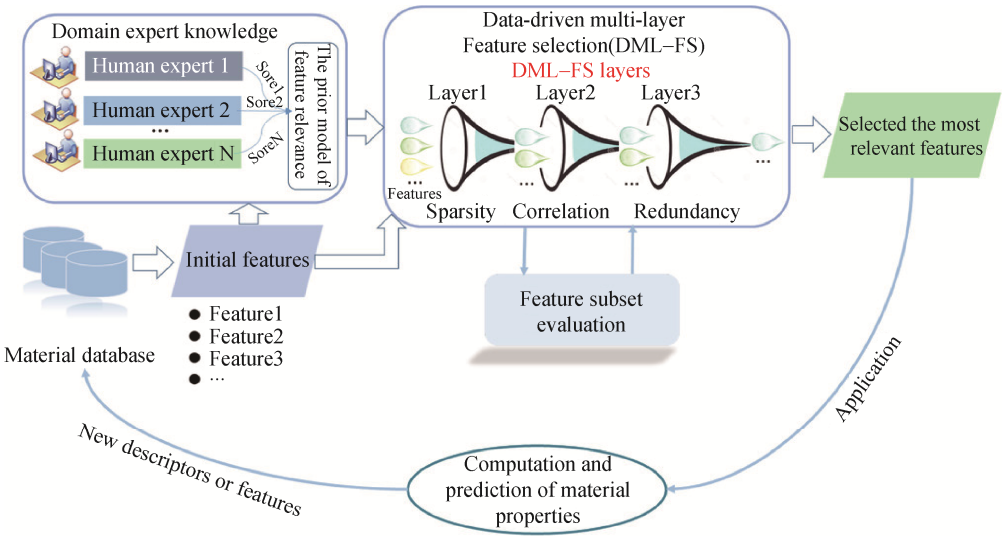


图 6 融合加权评分领域专家知识的多层级特征选择方法框架图[54]

Fig. 6 Framework of multi-level feature selection method incorporating weighted score-based expert knowledge[54]

描述符的定义和选取体现了材料领域专家对材料性能驱动机制的理解或猜测，其背后蕴含着丰富的材料领域知识或见解，但这些非结构化知识尚未得到充分利用。因此，挖掘和探索这些材料领域知识的多元表示方式及其到数据驱动的特征选择或转换方法中的融合方式，可能会成为在特征选择阶段发展材料领域知识嵌入的研究重点。

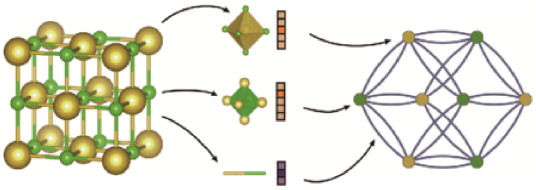
4.3 模型构建阶段的基础与探索

随着机器学习技术的发展，计算机领域已经积累了大量算法，可用于解决材料领域的分类[95–96]、回归[97–99]、聚类[5, 20]、优化[31, 84, 100–101]等问题。而材料领域的研究人员为了解决更广泛或特殊的材料问题，也在不断引入更多的机器学习新方法或尝试开发新的机器学习方法。

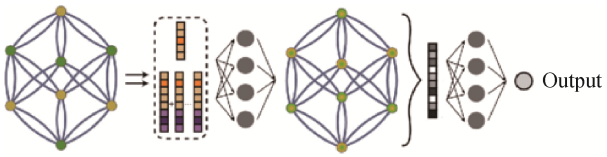
近几年，由于主动学习和迁移学习技术能够有效解决材料数据的小样本问题，在材料领域得到了广泛应用。如 Doan 等[32]通过构建基于 Bayesian 优化(BO)的主动学习框架，实现了均苯醚分子的有效识别。该模型在训练一组包含 112 000 个均苯醚分子数据集时，仅评估 100 个分子就成功地确定了 42 个最佳 HBE 候选分子。Pruksawan 等[102]将主动学习和 BO 结合，利用主动学习进行黏合剂材料的预测和优化，提出了一种最佳黏合剂材料制备方法。在该工作中，初始数据集仅包含 32 个黏合剂样本的黏合剂强度，经过 3 个主动学习周期后，成功获得了具有极高黏合剂连接强度(35.8±1.1) MPa 的材料。Sheng 等[103]利用迁移学习，针对基类电池数据(含 25 338 条样本)的预训练模型，通过参数共享对小样

本锂电池数据(含 525 条样本)进行迁移建模，在可接受误差范围内成功预测了其电池容量。Cubuk[17]提出了一种小数据迁移学习方法，基于不同材料之间物理知识的相似性，以关系迁移的方式对 12 716 个来自 MP 数据库的含锂材料的锂离子电导率进行建模，并成功用于数十亿固体锂离子导体的筛选以辅助材料设计。由于难以获得橡胶材料退化区域的训练数据，Togo 等[104]提出了一种基于迁移学习的深度自编码高斯混合模型(TL-DAGMM)，该模型利用从预先训练的深度学习模型中提取代表性特征，再通过特征迁移实现普通橡胶材料区域特征的自动学习。最后，通过基于 4 张图像数据计算得到的异常分数成功估计了橡胶材料的 12 个劣化区域，在真实橡胶材料的电子显微镜图像上的实验也证明了该模型的有效性。

针对非结构化材料图形或图像数据的机器学习建模研究，是应用领域知识开发材料领域专用模型的典型代表。Isayev 等[105]提出构建基于图形的属性标记材料片段(PLMFs)来表征无机晶体材料的晶体结构。该方法将晶体无向图划分为子图片段，再基于子图片段中节点的原子属性，利用平方距离矩阵的倒数和邻接矩阵的乘积生成片段描述符。如图 7 所示，Xie 等[106]提出了一种晶体图卷积神经网络(CGCNN)，通过将晶体结构表示成满足原子索引置换位置和晶胞选择不变性的晶体图来构建卷积神经网络，已经被成功应用于多种晶体材料的形成能、带隙、费米能等性能的预测[106–108]。



(a) Construction of the crystal graph



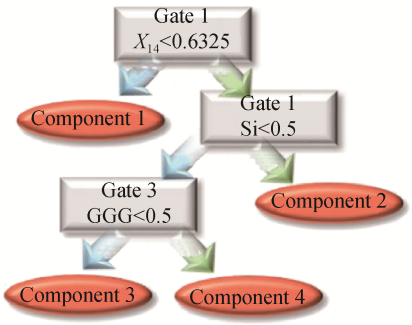
(b) Structure of the convolutional neural network on top of the crystal graph

图 7　晶体图神经网络说明[106]

Fig. 7　Illustration of the crystal graph convolutional neural networks[106]

此外，还有工作通过实现机器学习模型和物理模型或实验的协同，来搭建材料领域知识和数据驱动模型之间的桥梁。如以发现具有低热迟滞 Ni–Ti 形状记忆合金为目的，西安交通大学的薛德祯等[31]展示了与实验紧密结合的自适应设计策略如何通过顺序识别下一个实验或计算来加速发现过程，从而有效地探索复杂的搜索候选化合物空间，并利用该方法发现了具有极低热迟滞(1.84 K)的 Ni–Ti 形状记忆合金 $Ti_{50.0}Ni_{46.7}Cu_{0.8}Fe_{2.3}Pd_{0.2}$；美国西北大学的 Meredig 等[109]提出了一种物理启发式模型与旋转森林算法相结合的方法来预测 160 万个新三元化合物的生成能，该模型最终实现了 0.9 的 R^2，借此通过 DFT 计算初步验证了 8 种可能的新型稳定材料。

根据已知材料领域知识对未知材料性能驱动机制的理解或直觉，制定可解释的机器学习方案是材料领域知识在模型构建阶段嵌入的另一种表现。Iwasaki 等[21]利用专家层次混合的分解渐进 Bayesian 推理方法(FAB/HMRs)精准识别出影响自旋驱动热电材料热能的关键属性，藉此发现了具有目前为止最大热能的新型热电材料 $Co_{48.9}Pt_{51.1}N_{7.2}$。如图 8 所示，FAB/HMRs 通过构建树结构自适应地分组数据集，并为每组数据构建不同的线性预测模型来得到整个数据集上的最优预测结果，而这种建模方式正与相关材料专家的研究思路及其掌握的领域知识相匹配。基于镍基单晶高温合金在不同组分、环



(a) Tree structure

Component 1(Non-magnetic materials) $S_{STE}=0$
Component 2(Magnetic materials on Si substrate) $S_{STE}=-0.754X_2X_3+0.166X_7^2+0.173X_1X_8+0.348X_2X_8$ $+0.491X_1X_{13}-1.52X_6X_{13}+0.63X_{11}X_{14}+0.543$
Component 3(Magnetic materials on AIN substrate) $S_{STE}=0.763X_2X_3+0.554X_1X_{13}+2.41X_6X_{13}$ $-1.52X_{14}X_2+0.961X_{11}X_{14}+2.19$
Component 4(Magnetic materials on GGG substrate) $S_{STE}=0.053\ 9X_2X_8-0.011\ 1X_6X_{13}-0.654$

(b) Linear regression models

图 8　层次专家混合的分解渐进贝叶斯推理方法说明[21]

Fig. 8　Illustration of the factorized asymptotic Bayesian inference hierarchical mixture of experts[21]

境和工艺条件下具有不同蠕变机制的专家经验，曾提出了结合材料多尺度属性的分而治之的自适应学习方法(DCSA)[11]。该方法通过预先聚类数据集以区分具有不同蠕变机制的镍基单晶高温合金材料，再对不同类别的材料进行自适应建模。最终，DCSA 在包含 266 条样本的镍基单晶高温合金蠕变断裂寿命数据集上的 10 折交叉验证 R^2 高达 0.917 6。

综上所述，在材料领域知识的指导下，一些材料领域专用的机器学习方法已经被开发，但仍然限制于某些特定的机器学习模型，如卷积神经网络和 FAB/HMRs。因此，获取和符号化表示更充分和多元的材料领域知识，并研究其嵌入到通用机器学习模型的方式，将是进一步拓展材料领域知识嵌入机器学习方法的关键。

5 总结与展望

随着模型多样性和研究人员对算法原理关注的增加，机器学习建模过程和学习结果的可解释性成为除预测精度之外，能够决定机器学习模型是否可用于实际材料开发和制造的重要指标。然而，尽管许多机器学习模型已经成功地解决了多种材料问题，但材料专家对其建模过程的合理性依然持怀疑态度，其结果的可解释性也存在偶然发生的可能性。本研究认为导致这一现象的原因是：大部分机器学习模型是纯数据驱动的，忽略了材料领域知识的重要性。为此，通过厘清数据驱动机器学习在材料领域应用的全流程及其面临的三大关键矛盾，提出面向机器学习全流程的领域知识嵌入的机器学习方法，以提升机器学习结果与材料领域知识的一致性，并总结了相关的基础性和探索性工作。

构建材料领域知识嵌入的机器学习模型需要解决以下三个关键问题：1) 获取哪些材料领域知识以及如何获取? 2) 如何将非结构化的材料领域知识转化为计算机可读或易读的形式? 3) 如何将计算机可读或易读的材料领域知识嵌入到机器学习模型的学习过程中? 上述问题有望通过以下思路解决：

1) 大量材料领域知识以非结构化文本的形式贮存在数以百计的已发表科研文献中。文本挖掘(TM)和自然语言处理技术(NLP)旨在建立算法以解释字符序列并从中获得逻辑信息[110]。材料领域的最新进展[110−112]已经显示，TM 和 NLP 技术能够大规模地从科学文本中提取数据。继续发展 TM 和 NLP 技术，以自动化地从科学文本中抽取更多元的材料领域知识，或可成为构建领域知识嵌入的机器学习方法的主要领域知识来源。

2) 将材料领域知识表示为逻辑谓词可能是材料领域知识嵌入到机器学习模型的有效手段。在计算机科学领域，已经有一些工作通过将特定的领域知识转换为一阶逻辑规则的方式来优化数据驱动的神经网络的建模过程，并应用于情感分类[113]、服装搭配推荐[114]、手写体识别[115]等领域。明确材料领域知识的类别和特点，构建材料领域知识的符号化表示体系，可为材料领域更广泛地构建领域知识嵌入的机器学习新模型提供规范化标准。

3) 目前，大部分的材料领域知识用于材料数据的前处理和机器学习结果的后解释。掌握机器学习模型的建模原理，将符号化的领域知识嵌入到机器学习方法的三要素中(Model、Strategy 和 Algorithm)，从而优化模型的求解过程，是提升机器学习模型准确性和可解释性的重要途径。南方科技大学的张东晓团队在工业领域的测井曲线补全[116]、电网负荷预测[117]和地下渗流场求解[118]等应用中，利用领域算法和控制方程等领域知识优化了神经网络模型的模型结构设计和模型效果评估；本课题组开发了融合加权评分领域知识的特征选择方法[54]、材料领域规则约束的混合式特征选择方法[66]、分而治之的自适应建模方法[11]等。这些工作都有力地证明了领域知识嵌入到机器学习模型三要素中可有效地提升模型的预测精度和效率，同时也为构建面向机器学习全流程的领域知识嵌入机器学习积累了经验。

未来，实现材料领域知识的自动化获取和统一符号化表示，发展符号化领域知识到各类机器学习方法三要素中的嵌入方式，从而构建领域知识嵌入的机器学习方法来协调材料领域机器学习的三大矛盾，对进一步提升机器学习的普适性、准确性和可解释性具有十分重要的意义。

参考文献:

[1] PARK H, JUNG K, NEZAFATI M, et al. Sodium ion diffusion in Nasicon ($Na_3Zr_2Si_2PO_{12}$) solid electrolytes: effects of excess sodium[J]. ACS Appl Mater Inter, 2016, 8(41): 27814–27824.

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

[3] SCHMIDT J, MARQUES M R G, BOTTI S, et al. Recent advances and applications of machine learning in solid-state materials science[J]. NPJ Comput Mater, 2019, 5(83): 1–36.

[4] CHEN C, ZUO Y, YE W, et al. A critical review of machine learning of energy materials[J]. Adv Energy Mater, 2020, 10(8): 1903242.

[5] CERIOTTI M. Unsupervised machine learning in atomistic

simulations, between predictions and understanding[J]. J Chem Phys, 2019, 150(15): 150901.

[6] JING L L, TIAN Y L. Self-supervised visual feature learning with deep neural networks: a survey[J]. IEEE T Pattern Anal, 2021, 43(11): 4037–4058.

[7] 屠恩美, 杨杰. 半监督学习理论及其研究进展概述[J]. 上海交通大学学报, 2018, 52(10): 1280–1291.
TU Enmei, YANG Jie. J Shanghai Jiao Tong University (in Chinese), 2018, 52(10): 1208–1291.

[8] 袁睿豪, 廖玮杰, 唐斌, 等. 数据驱动的航空发动机材料设计研究进展[J]. 航空制造技术, 2021, 64(18): 22–30.
YUAN Ruihao, LIAO Weijie, TANG Bin, et al. Aeron Manuf Technol (in Chinese), 2021, 64(18): 22–30.

[9] RAJAK P, WANG B B, NOMURA K, et al. Autonomous reinforcement learning agent for stretchable kirigami design of 2D materials[J]. NPJ Comput Mater, 2021, 7(1): 102.

[10] DEL VECCHIO C, FENU G, PELLEGRINO F A, et al. Support vector representation machine for superalloy investment casting optimization[J]. Appl Math Model, 2019, 72: 324–336.

[11] 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 Mater, 2020, 195: 454–467.

[12] TAYLOR P L, CONDUIT G. Machine learning predictions of superalloy microstructure[J]. Comp Mater Sci, 2022, 201: 110916.

[13] WEN C, ZHANG Y, WANG C X, et al. Machine learning assisted design of high entropy alloys with desired property[J]. Acta Mater, 2019, 170: 109–117.

[14] BATCHELOR T A A, PEDERSEN J K, WINTHER S H, et al. High-entropy alloys as a discovery platform for electrocatalysis[J]. Joule, 2019, 3(3): 834–845.

[15] ZHANG Y, WEN C, WANG C X, et al. Phase prediction in high entropy alloys with a rational selection of materials descriptors and machine learning models[J]. Acta Mater, 2020, 185: 528–539.

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

[17] SENDEK A D, YANG Q, CUBUK E D, et al. Holistic computational structure screening of more than 12000 candidates for solid lithium-ion conductor materials[J]. Energ Environ Sci, 2017, 10(2): 306–320.

[18] SENDEK A D, CUBUK E D, ANTONIUK E R, et al. Machine learning-assisted discovery of solid Li-ion conducting materials[J]. Chem Mater, 2019, 31, 2: 342–352.

[19] CUBUK E D, SENDEK A D, REED E J. Screening billions of candidates for solid lithium-ion conductors: A transfer learning approach for small data[J]. J Chem Phys, 2019, 150(21): 214701.

[20] ZHANG Y, HE X F, CHEN Z Q, et al. Unsupervised discovery of solid-state lithium ion conductors[J]. Nat Commun, 2019, 10: 5260.

[21] IWASAKI Y, SAWADA R, STANEV V, et al. Identification of advanced spin-driven thermoelectric materials *via* interpretable machine learning[J]. NPJ Comput Mater, 2019, 5: 103.

[22] MIN K, CHO E. Accelerated discovery of potential ferroelectric perovskite *via* active learning[J]. J Mater Chem C, 2020, 8: 7866–7872.

[23] MA W, LIU Y M. A data-efficient self-supervised deep learning model for design and characterization of nanophotonic structures[J]. Sci China Phys Mech, 2020, 63(8): 284212.

[24] HO C T, WANG D W. Robust identification of topological phase transition by self-supervised machine learning approach[J]. New J Phys, 2021, 23(8): 083021.

[25] CHEN D, ZHENG J X, WEI G W, et al. Extracting predictive representations from hundreds of millions of molecules[J]. J Phys Chem Lett, 2021, 12(44): 10793–10801.

[26] MA W, CHENG F, XU Y, et al. Probabilistic representation and inverse design of metamaterials based on a deep generative model with semi-supervised learning strategy[J]. Adv Mater, 2019, 31(35): 1901111.

[27] SAHOO P, ROY I, WANG Z, et al. MultiCon: a semi-supervised approach for predicting drug function from chemical structure analysis[J]. J Chem Inf Model, 2020, 60(12): 5995–6006.

[28] KUNSELMAN C, ATTARI V, MCCLENNY L, et al. Semi-supervised learning approaches to class assignment in ambiguous microstructures[J]. Acta Mater, 2020, 188: 49–62.

[29] CHEN D, SUN D, FU J, et al. Semi-supervised learning framework for aluminum alloy metallographic image segmentation[J]. IEEE Access, 2021, 9: 30858–30867.

[30] 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 Jianxin, SU Yanjing, XUE Dezhen, et al. Acta Metall SIN (in Chinese), 2021, 57(11): 1343–1361.

[31] XUE D Z, BALACHANDRAN P V, HOGDEN J, et al. Accelerated search for materials with targeted properties by adaptive design[J]. Nat Commun, 2016, 7: 11241.

[32] DOAN H A, AGARWAL G, QIAN H, et al. Quantum chemistry-informed active learning to accelerate the design and discovery of sustainable energy storage materials[J]. Chem Mater, 2020, 32: 6338–6346.

[33] SAIEDIAN I, BADLOE T, LEE H, et al. Deep Q-network to produce polarization-independent perfect solar absorbers：a statistical report[J]. Nano Converg, 2020, 7(1): 26.

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

[35] 狄少丞, 冯云田, 瞿同明, 等. 基于深度强化学习算法的颗粒材料应力–应变关系数据驱动模拟研究[J]. 力学学报, 2021, 53(10): 2712–2723.
DI Shaoceng, FENG Yuntian, QU Tongming, et al. Chin J Theor App Mech-pol (in Chinese), 2021, 53(10): 2712–2723.

[36] WOHLRAB L, FURNKRANZ J. A review and comparison of strategies for handling missing values in separate-and-conquer rule learning[J]. J Intell Inf Syst, 2011, 36(1): 73–98.

[37] XU X D, LIU H W, YAO M H. Recent progress of anomaly detection[J]. Complexity, 2019: 2686378.

[38] GUO H X, LI Y J, SHANG J, et al. Learning from class-imbalanced data: Review of methods and applications[J]. Expert Syst Appl, 2017, 73: 220–239.

[39] BERTI-QUILLE L. Measuring and modelling data quality for quality-awareness in data mining[M]. Quality Measures in Data Mining. Springer Berlin Heidelberg, 2007.

[40] WANG Y D, PAN Z B, PAN Y W, et al. A training data set cleaning method by classification ability ranking for the k-nearest neighbor classifier[J]. IEEE T Neur Net Lear, 2020, 31(5): 1544–1556.

[41] ROY K, KAR S, DAS R N. A primer on QSAR/QSPR modeling: fundamental concepts[M]. Springer, 2015.

[42] GHIRINGHELLI L M, VYBIRAL J, LEVCHENKO S V, et al. Big data of materials science: critical role of the descriptor[J]. Phys Rev Lett, 2015, 114(10): 105503.

[43] SHANDIZ M A, GAUYIN R. Application of machine learning methods for the prediction of crystal system of cathode materials in lithium-ion batteries[J]. Comp Mater Sci, 2016, 117: 270–278.

[44] Li Y, ZOU C F, BERECIBAR M, et al. Random forest regression for online capacity estimation of lithium-ion batteries[J]. Appl Energ, 2018, 232: 197–210.

[45] CHELGANI S C, MATIN S S, HOWER J C. Explaining relationships between coke quality index and coal properties by random forest method[J]. Fuel, 2016, 182: 754–760.

[46] IM J, LEE S, KO T W, et al. Identifying Pb-free perovskites for solar cells by machine learning[J]. NPJ Comput Mater, 2019, 5: 37.

[47] WANG X M, XU Y L, YANG J, et al. ThermoEPred-EL: Robust bandgap predictions of chalcogenides with diamond-like structure via feature cross-based stacked ensemble learning[J]. Comp Mater Sci, 2019, 169: 109117.

[48] WEN C, WANG C X, ZHANG Y, et al. Modeling solid solution strengthening in high entropy alloys using machine learning[J]. Acta Mater, 2021, 212: 116917.

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

[50] DE JONG M, CHEN M, NOTESTINE R, et al. A statistical learning framework for materials science: application to elastic moduli of k-nary inorganic polycrystalline compounds[J]. Sci Rep, 2016, 6: 34256.

[51] ZHANG Y, LING C. A strategy to apply machine learning to small datasets in materials science[J]. NPJ Comput Mater, 2019, 3(5): 71–78.

[52] FABER F A, LINDMAA A, VON Lilienfeld O A, et al. Machine learning energies of 2million Elpasolite (ABC_2D_6) crystals[J]. Phys Rev Lett, 2016, 117(13): 135502.

[53] SCHMIDT J, SHI J M, BORLIDO P, et al. Predicting the thermodynamic stability of solids combining density functional theory and machine learning[J]. Chem Mater, 2017, 29(12): 5090–5103.

[54] 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]. Adv Theor Simul, 2020, 3(2): 1900215.

[55] WANG X L, XIAO R J, LI H, et al. Quantitative structure-property relationship study of cathode volume changes in lithium ion batteries using ab-initio and partial least squares analysis[J]. J Materiomics, 2017, 3(3): 178–183.

[56] ZHAO Y L, ZHANG K, ZHANG Y, et al. Prediction of air voids of asphalt layers by intelligent algorithm[J]. Constr Build Mater, 2022, 317: 125908.

[57] JIANG D W, WANG Z Y, ZHANG J L, et al. Predictive modelling for contact angle of liquid metals and oxide ceramics by comparing Gaussian process regression with other machine learning methods[J]. Ceram Int, 2022, 48(1): 665–673.

[58] YE W K, CHEN C, WANG Z B, et al. Deep neural networks for accurate predictions of crystal stability[J]. Nat Commun, 2018, 9: 3800.

[59] HU C, JAIN G, ZHANG P Q, et al. Data-driven method based on particle swarm optimization and k-nearest neighbor regression for estimating capacity of lithium-ion battery[J]. Appl Energ, 2014, 129: 49–55.

[60] HALEVY A, NORVIG P, PEREIRA F. The unreasonable effectiveness of data[J]. IEEE Intell Syst, 2009, 24(2): 8–12.

[61] 李航. 统计学习方法[M]. 北京: 清华大学出版社, 2019: 15–19.

[62] AGRAWAL A, CHOUDHARY A. Perspective: materials informatics and big data: realization of the “fourth paradigm” of science in materials science[J]. Apl Mater, 2016, 4(5): 053208.

[63] MA B Y, WEI X Y, LIU C N, et al. Data augmentation in microscopic images for material data mining[J]. NPJ Comput Mater, 2020, 6(1): 125.

[64] RESHEF D N, RESHEF Y A, FINUCANE H K, et al. Detecting novel associations in large data sets[J]. Science, 2011, 334(6062): 1518–1524.

[65] LUNDBERG S M, LEE S I. A unified approach to interpreting model predictions[C]//31st Annual Conference on Neural Information Processing Systems, Long Beach, CA, 2017.

[66] 郭碧茹. 基于机器学习的 NASICON 型固态电解质激活能预测方法研究[D]. 上海: 上海大学, 2020.
GUO Biru. Prediction of activation energy of NASICON solid electrolyte based on machine learning (in Chinese, dissertation). Shanghai: Shanghai University, 2020.

[67] LOOKMAN T, BALACHANDRAN P V, XUE D Z, et al. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design[J]. NPJ Comput Mater, 2019, 5: 21.

[68] ZHUANG F Z, QI Z Y, DUAN K Y, et al. A comprehensive survey on transfer learning[J]. P IEEE, 2021, 109(1): 43–76.

[69] XU Q C, LI Z Z, LIU M, et al. Rationalizing perovskite data for machine learning and materials design[J]. J Phys Chem Lett, 2018, 9(24): 6948–6954.

[70] HAFIZ H, KHAIR A I, CHOI H, et al. A high-throughput data

analysis and materials discovery tool for strongly correlated materials[J]. NPJ Comput Mater, 2018, 4: 63.

[71] LI W, JACOBS R, MORGAN D. Predicting the thermodynamic stability of perovskite oxides using machine learning models[J]. Comput Mater Sci, 2018, 150: 454–463.

[72] WANG A Y T, MURDOCK R J, KAUWE S K, et al. Machine learning for materials scientists: an introductory guide toward best practices[J]. Chem Mater, 2020, 32(12): 4954–4965.

[73] GHIRINGHCLLI LM, CARBOGNO C, LEVCHCNKO S, et al. Towards a common format for computational material science data[J]. arXiv Materials Science: 1607.04738v1.

[74] WILKINSON M D, DUMONTIER M, AALBERSBERG I J, et al. Comment: the FAIR guiding principles for scientific data management and stewardship[J]. Sci Data, 2016, 3: 160018.

[75] 宋佳, 温亮明, 李洋. 科学数据共享 FAIR 原则：背景、内容及实践[J]. 情报资料工作, 2021, 42(1): 57–68.
SONG Jia, WEN Liangming, LI Yang. Inf Doc Serv(in Chinese), 2021, 42(1): 57–68.

[76] DRAXL C, SCHEFFLER M. NOMAD: The FAIR concept for big data-driven materials science[J]. MRS Bull, 2018, 43(9): 676–682.

[77] The NOMAD (Novel Materials Discovery) Center of Excellence (CoE): https://nomad-coe.eu.

[78] ALLEN F H. The Cambridge Structural Database: a quarter of a million crystal structures and rising [J]. Acta Crystallogr B, 2002, 58: 380–388.

[79] BERGERHOFF G, HUNDT R, SIEVERS R, et al. The inorganic crystal structure data base[J]. J Chem Info Comput Sci, 1983, 23(2): 66–69.

[80] 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: 1501–1509.

[81] VILLARS P, BERNDT M, BRANDENBURG K, et al. The Pauling File, binaries edition[J]. J Alloy Compd, 2004, 367: 293–297.

[82] JAIN A, ONG S P, HAUTIER G, et al. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation [J]. APL Mater, 2013, 1: 011002.

[83] CURTAROLO S, SETYAWAN W, WANG S, et al. AFLOWLIB.ORG: A distributed materials properties repository from high-throughput ab initio calculations[J]. Comp Mater Sci, 2012, 58: 227–235.

[84] SANCHEZ-LENGELING B, ASPURU-GUZIK A. Inverse molecular design using machine learning: Generative models for matter engineering[J]. Science, 2018, 361(6400): 360–365.

[85] DAN Y B, ZHAO Y, LI X, et al. Generative adversarial networks (GAN) based efficient sampling of chemical composition space for inverse design of inorganic materials[J]. NPJ Comput Mater, 2020, 6(1): 84.

[86] XU X F, LEI Y G, LI Z D. An incorrect data detection method for big data cleaning of machinery condition monitoring[J]. IEEE T Ind Electron, 2020, 67(3): 2326–2336.

[87] STORKEY A J, HAMBLY N C, WILLIAMS C K I, et al. Cleaning sky survey data bases using Hough transform and renewal string approaches[J]. Mon Not R Astron Soc, 2004, 347(1): 36–51.

[88] WARD L, AGRAWAL A, CHOUDHARY A, WOLVERTON C. A general-purpose machine learning framework for predicting properties of inorganic materials[J]. NPJ Comput Mater, 2016, 2: 16028.

[89] 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]. J Phys Chem C, 2020, 124(52): 28458–28468.

[90] OUYANG R H, CURTAROLO S, AHMETCIK E, et al. SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates[J]. Phys Rev Mater, 2018, 2(8): 083802.

[91] WANG Y, WAGNER N, RONDINELLI J M. Symbolic regression in materials science[J]. MRS Commun, 2019, 9(3): 793–805.

[92] WENG B C, SONG Z L, ZHU R L, et al. Simple descriptor derived from symbolic regression accelerating the discovery of new perovskite catalysts[J]. Nat Commun, 2020, 11(1): 3513.

[93] WEN C, WANG C X, ZHANG Y, et al. Modeling solid solution strengthening in high entropy alloys using machine learning[J]. Acta Mater, 2021, 212: 116917.

[94] IM J, LEE S, KO T W, et al. Identifying Pb-free perovskites for solar cells by machine learning[J]. NPJ Comput Mater, 2019, 5: 37.

[95] TONG Z N, WANG L Y, ZHU G M, et al. Predicting twin nucleation in a polycrystalline Mg alloy using machine learning methods[J]. Metall Mater Trans A, 2019, 50(12): 5543–5560.

[96] KHARKOV Y A, SOTSKOV V E, KARAZEEV A A, et al. Revealing quantum chaos with machine learning[J]. Phys Rev B, 2020, 101(6): 064406.

[97] 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 Mater, 2021, 35: 595–601.

[98] AHMAD A, AHMAD W, ASLAM F, et al. Compressive strength prediction of fly ash-based geopolymer concrete *via* advanced machine learning techniques[J]. CASE Stud Constr Mat, 2022, 16: e00840.

[99] SARKER S, TANG-KONG R, SCHOEPPNER R, et al. Discovering exceptionally hard and wear-resistant metallic glasses by combining machine-learning with high throughput experimentation[J]. Appl Phys Rev, 2022, 9(1): 011403.

[100] 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.

[101] LAMBARD G, SASAKI T T, SODEYAMA K, et al. Optimization of direct extrusion process for Nd–Fe–B magnets using active learning assisted by machine learning and Bayesian optimization[J]. Scripta Mater, 2022, 209: 114341.

[102] PRUKSAWAN S, LAMBARD G, SAMITSU S, et al. Prediction and optimization of epoxy adhesive strength from a small dataset through active learning[J]. Sci Technol Adv Mat, 2020, 20(1): 1010–1021.

[103] SHEN S, SADOUGHI M, LI M, et al. Deep convolutional neural networks with ensemble learning and transfer learning for capacity

estimation of lithium-ion batteries[J]. Appl Energ, 2020, 260: 114296.

[104] TOGO R, SAITO N, OGAWA T, et al. Estimating regions of deterioration in electron microscope images of rubber materials *via* a transfer learning-based anomaly detection model[J]. IEEE Access, 2019, 7: 162395–162404.

[105] ISAYEV O, OSES C, TOHER C, et al. Universal fragment descriptors for predicting properties of inorganic crystals[J]. Nat Commun, 2017, 8: 15679.

[106] XIE T, GROSSMAN J C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties[J]. Phys Rev Lett, 2018, 120(14): 145301.

[107] AHAMD 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 Sci, 2018, 4(8): 996–1006.

[108] ZHOU L M, YAO A M Z, WU Y J, et al. Machine learning assisted prediction of cathode materials for Zn-ion batteries[J]. Adv Theor Simul, 2021, 4(9): 2100196.

[109] MEREDIG B, AGRAWAL A, KIRKLIN S, et al. Combinatorial screening for new materials in unconstrained composition space with machine learning[J]. Phys Rev B, 2014, 89(9): 094104.

[110] KONONOVA O, HE T J, HUO H Y, et al. Opportunities and challenges of text mining in materials research[J]. ISCIENCE, 2021, 24(3): 102155.

[111] OLIVETTI E A, COLE J M, KIM E, et al. Data-driven materials research enabled by natural language processing and information extraction[J]. Appl Phys Rev, 2021, 7(4): 041317.

[112] TSHITOYAN V, DAGDELEN J, WESTON L, et al. Unsupervised word embeddings capture latent knowledge from materials science literature[J]. Nature, 2019, 571(7763): 95–98.

[113] HU Z T, MA X Z, LIU Z Z, et al. Harnessing deep neural networks with logic rules[C]//54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 2016: 2410–2420.

[114] SONG X M, FENG F L, HAN X J. Neural compatibility modeling with attentive knowledge distillation[C]//41st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Univ Michigan, Ann Arbor, MI, 2018: 5–14.

[115] DAI W Z, XU Q L, YU Y, et al. Tunneling neural perception and logic reasoning through abductive Learning. arXiv: 1802.01173, 2018.

[116] CHEN Y T, ZHANG D X. Physics constrained deep learning of geomechanical logs[J]. IEEE T Geosci Remote, 2020, 58(8), 5932–5943.

[117] CHEN Y T, ZHANG D X. Theory guided deep-learning for load forecasting (TgDLF) *via* ensemble long short-term memory[J]. Adv Appl Energ, 2020, 1: 1–15.

[118] CHEN Y T, HUANG D, ZHANG D X. Theory-guided hard constraint projection (HCP): a knowledge-based data-driven scientific machine learning method[J]. J Comput Phys, 2021, 445: 110624.