第 51 卷第 2 期　　硅　酸　盐　学　报　　Vol. 51，No. 2
2023 年 2 月　　JOURNAL OF THE CHINESE CERAMIC SOCIETY　　February，2023
http://www.gxyb.cbpt.cnki.net　　DOI：10.14062/j.issn.0454-5648.20220991

综 合 评 述

面向材料领域机器学习的数据质量治理

刘　悦[1,4]，马舒畅[1]，杨正伟[1]，邹欣欣[1]，施思齐[2,3]

(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)

摘　要：数据驱动的机器学习凭借其准确高效的预测能力广泛应用于材料的性能预测和构效关系研究。数据决定了机器学习的上限。然而，目前材料领域的数据存在来源广、噪音大、样本少、维度高等数据质量问题，阻碍了机器学习在材料领域更广泛的应用。本文从数据品质和数据数量 2 个视角系统梳理并全面剖析了材料领域数据质量问题及其相关治理工作，发现数据品质与数据数量共同决定数据质量。基于此，提出了面向材料领域机器学习全过程的领域知识嵌入的数据质量治理框架。该框架定义了 12 种维度用于解析材料数据质量的内涵；构建了数据质量治理的生命周期模型以确保数据质量治理活动有序进行；建立了一系列数据质量治理处理模型，从领域知识与数据驱动 2 个方面对数据质量进行精准全面治理，为生命周期模型的具体实施提供技术支持。该框架实现了材料数据质量的综合评估与提升，为高质量数据获取提供理论指导与候选方案，加速机器学习在材料研发中的深入应用。

关键词：材料科学；机器学习；数据质量；领域知识
中图分类号：TP181；TB3　　**文献标志码：**A　　**文章编号：**0454–5648(2023)02–0427–11
网络出版时间：2023–01–17

A Data Quality and Quantity Governance for Machine Learning in Materials Science

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

(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)

Abstract: Data-driven machine learning is widely used in materials property prediction and structure-activity relationship research due to its accurate and efficient predictive ability. Data determines the upper limit of machine learning. However, materials data often have various quality and quantity problems (*i.e.*, multiple sources, large noise, small samples, and high dimensionality), affecting the application of machine learning in the materials field. In this paper, by analyzing the data quality and quantity problems and their related governance work, we find that data quality and data quantity jointly determine this problem. Following this, a data quality and quantity governance framework embedded by materials domain knowledge in the whole process of materials machine learning is proposed. We define twelve dimensions to analyze the connotation of materials data quality and quantity. A life cycle model of data quality and quantity governance is constructed to ensure that data quality and quantity governance activities are carried out in an orderly manner. To manage data quality and quantity accurately and comprehensively, a series of corresponding governance processing models are established from domain knowledge and data-driven aspects, which provides technical support for the specific implementation of the life cycle model. This framework realizes the overall evaluation and improvement of materials data quality and quantity, providing theoretical guidance and candidate solutions for high-quality and appropriate-quantity data acquisition and accelerating the in-depth application of machine learning in materials research and development.

Keywords: materials science; machine learning; data quality and quantity; domain knowledge

收稿日期：2022–11–18。　修订日期：2022–12–12。
基金项目：国家重点研发计划项目(2021YFB3802101)；国家自然科学基金面上项目(52073169)；国家自然科学基金重大研究计划培育项目(92270124)。
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: 2022–11–18.　**Revised date:** 2022–12–12.
First author: LIU Yue (1975–), female, Ph.D., Professor.
E-mail: yueliu@shu.edu.cn
Correspondent author: SHI Siqi (1978–), male, Ph.D., Professor.
E-mail: sqshi@shu.edu.cn

近年来，数据驱动的机器学习正被成功地应用于材料的性能预测、新材料发现和过程优化中[1–4]。然而，数据决定了机器学习的上限[5–6]。在数据收集过程中，由于实验误差、环境差异、计算缺陷等各种不确定性因素，数据的品质往往受到影响，导致其具有稀疏性、高噪音、多源异构等特点[7]。此外，数据的数量也是影响数据质量的关键因素之一：一方面，由于材料样本的获取依赖于复杂的实验或劳动密集型的采集工作，这导致其数量普遍较小(样本量不足)；另一方面，材料专家在获取样本的过程中，通常定义多个描述符(特征)来描述材料性能复杂的驱动机制，使得材料数据通常具有较高的维度[8] (特征量偏大)。

为了进一步提升机器学习的准确性和可信性，使其结果更容易被材料科学领域的研究人员所接受和推广，在应用机器学习解决材料科学问题前，对材料数据进行充分的数据质量治理具有十分重要的意义。值得注意的是，并非所有数据都存在质量问题。当样本量适中且特征较少时，数据治理会引入噪声，降低模型精度。因此，对数据质量进行治理前需要定义合理指标评估数据是否存在质量问题。

目前，材料领域已有一些工作针对数据质量问题展开研究。针对数据品质问题，一些研究将统计分析方法和机器学习算法相结合，对异常数据进行识别并剔除[9–10]，通过对比剔除异常数据前后机器学习模型的预测精度，发现进行数据品质治理能够在某种程度上提升模型精度；针对数据数量问题，一些研究在机器学习建模前使用特征选择事先删除冗余特征，通过对比删除冗余特征前后机器学习模型的预测精度，发现进行数据数量治理能够有效降低模型复杂度，从而构建具有良好泛化性能和较高预测精度的机器学习模型[11]。

综上所述，材料领域已有研究从单一角度探索了数据质量问题，表明提升数据质量能够在一定程度上提高机器学习模型的预测精度。但仍有以下 2 方面需要重视：1) 需要数据质量治理统一框架的指导，以对数据质量进行全面系统地评估与提升；2) 材料领域知识对于数据治理具有非常重要的作用。例如：Yuan 等[12]在领域知识的指导下对数据库中数据进行预处理，高效优化了合金成分，提高了机器学习模型预测精度；刘悦等[13]全面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。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，面向材料领域机器学习全流程，在领域知识的指导下，探索数据在机器学习每一阶段可能存在的质量问题，实现数据质量的动态监测和全面控制，从而指导研究人员进行更准确的数据分析和更可靠的科学决策。

本文综述了面向材料领域机器学习的数据品质治理和数量治理的研究现状，并对材料数据在机器学习应用全过程中各阶段面临的质量问题进行了阐述。基于此，提出一种面向材料领域机器学习应用全过程的领域知识嵌入的数据质量治理框架，旨在为机器学习任务提供更高质量的学习样本，提升机器学习在材料领域应用的可靠性；最后，展望了材料数据质量治理的未来发展。

1 材料领域机器学习中的数据品质治理

为提升材料数据品质以获得更好的机器学习模型预测性能，研究人员开始采用数据统计或机器学习方法对数据品质问题进行定量分析，例如，Gharagheizi[14]和 Hemmati-Sarapardeh[15]等使用最小二乘支持向量机对离子液体的电子电导率进行预测时，通过人工校验、统计分析成功剔除由于文献资料错误报道或实验测量误差造成的异常样本，保证了机器学习模型的准确性和可靠性；Li 等[16]探讨了数据分布的不均衡性对机器学习模型的影响，通过在不同组分空间中建立不同的分类或回归模型，再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；Xu 等[17]开发了一种程序来识别材料工程数据库(MP)中化学式组成为 ABX_3 和 $(A'A'')(B'B'')X_6$ 的化合物的钙钛矿的形成性，发现了 11 个 ABO_3 化合物的形成能数据存在异常并进行了合理校正；此外，也有研究从定性分析的角度出发，提出多种数据品质评估指标或管理原则对数据品质进行治理。例如，针对材料科学领域实验数据往往不能够充分描述元数据，并且经常以不连贯的文档形式显示等问题，Thorsten Wuest 等[18]提出了数据品质管理指南，该指南支持材料科学实验研究数据的完整性、可用性和可重用性检查，在一定程度上保证了实验数据的品质；Madison Wenzlick 等[19]提出了一种评估合金拉伸、蠕变–应变松弛和疲劳性能数据品质的方法，从数据完整性、准确性、可用性和标准化 4 个指标出发，制定相应指标的品质评定等级对数据进行品质评估，有效解决了合金设计和寿命预测中数据来源不同的问题。FAIR 原则[20–21]

是在大数据环境中提出的一套数据管理准则，倡导数据在开放共享过程中应尽量实现可发现、可访问、可互操作和可重用，旨在保证多源数据的可溯源性、格式一致性和完整性。2018 年，Claudia 等[22]将 FAIR 原则引入到大数据驱动的材料科学研究中，主张通过构建原始数据仓库、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的品质，但并未探讨具体数据品质问题的解决方案。因此，目前材料科学领域仍缺乏全面的数据品质评估策略，以提升用于机器学习建模的材料数据的品质，进而提升机器学习模型的性能。

综上所述，材料科学领域的研究人员已经认识到了数据品质的重要性，并针对某些特定材料体系的具体品质问题进行了探索性研究。然而，现有研究均未对机器学习建模各阶段所涉及的多种材料数据品质问题进行系统探讨，由于数据品质差而导致的机器学习模型泛化能力差、结果无法解释等问题依旧无法避免。此外，现有的数据驱动方法仅根据数据的分布情况对数据的品质进行评估与提升，往往忽略了数据领域背景的重要性，从而易产生与材料领域知识不一致甚至是相悖的结果。因此，实现领域知识与数据品质治理方法的有机融合以及全面的数据品质治理对于机器学习建模至关重要。

2　材料领域机器学习中的数据数量治理

目前，对于数据数量的治理往往通过减少特征量或者扩充样本量的方式实现，本节将对这些方法进行综述。

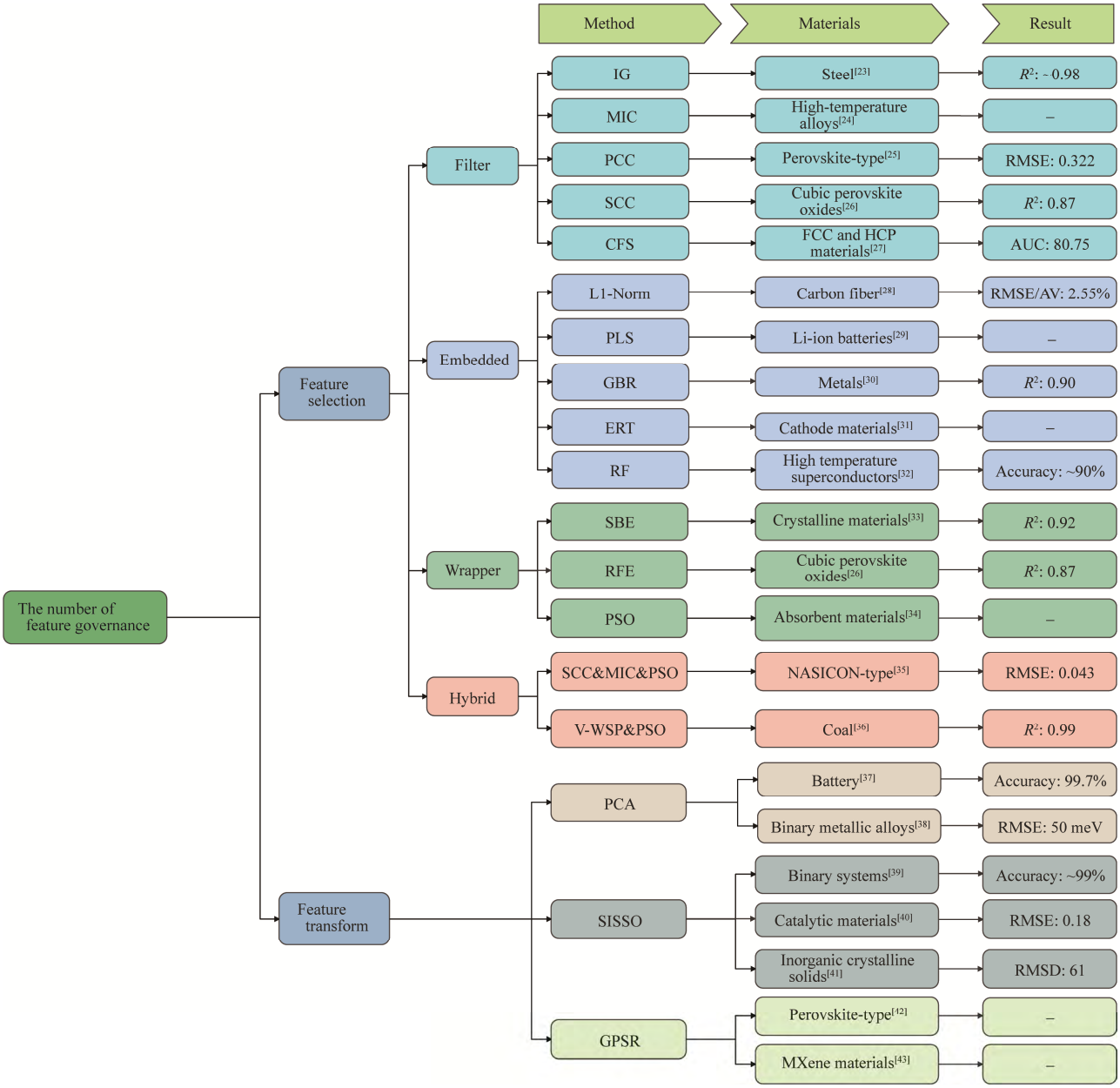
2.1　特征量治理

机器学习模型建立在描述特定材料任务的特征组成的数据集之上。然而，用于机器学习建模的特征并不是越多越好。在高维特征空间中，相关性高的冗余特征会对机器学习模型的预测性能和可解释性产生负面影响。因此，如何有效地控制特征数量，以帮助研究人员以一种减少特征数量的方式进行机器学习建模，同时使模型能够探索隐藏在材料数据中的通用模式，是一个亟待解决的问题。

目前，常用的特征量治理方法有特征选择(Feature Selection)和特征转换(Feature Transform)。图 1 总结了这些方法在不同材料体系中的应用[23–43]。

特征选择能够通过去除冗余特征实现快速捕捉数据之间隐含的潜在模式[44]，已成为材料特征量治理最重要的方法。按照所使用的特征选择方法不同，可以将材料特征量治理的工作分为过滤式(Filter)、嵌入式(Embedded)、包裹式(Wrapper)和混合式(Hybrid)等 4 类。过滤式特征选择采用基于统计理论和信息论的度量指标评估相关特征的重要性并进行排序，然后选择得分高的特征子集用于机器学习[45]。目前，材料研究人员已成功将这些度量指标应用在了多种材料体系中。例如，将信息增益(IG)用于钢的疲劳强度预测[23]；将最大互信息系数(MIC)用于高温合金蠕变性能分析[24]；将 Pearson 相关系数(PCC)用于钙钛矿型材料识别[25]；将 Spearman 相关系数(SCC)用于立方钙钛矿氧化物的结构–成分关系研究[26]；将基于相关性的特征选择(CFS)用于 FCC 和 HCP 材料的应力热点分类问题[27]。与过滤式方法相比，嵌入式特征选择与特定的机器学习模型绑定，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，以实现模型构建和特征选择的协同，简化了特征选择过程[46]。目前，材料研究人员也将嵌入式方法应用在了多种材料问题中。例如，将 L1 正则化(L1-Norm)用于碳纤维性能分析[28]；将偏最小二乘(PLS)用于挖掘影响锂离子电池阴极体积变化的相关变量研究[29]；将梯度提升回归(GBR)用于间隙扩散活化能预测[30]；将极端随机树(ERT)用于锂离子电池正极材料的晶系预测[31]以及将随机森林(RF)用于超导体临界温度建模研究[32]。包裹式特征选择也依赖于模型，通常与机器学习模型和元启发式算法(如遗传算法、粒子群优化算法等)结合使用，反复构建模型评估候选特征子集，直至选定的特征子集满足迭代停止条件[44]。逐步向后消除(SBE)和递归特征消除(RFE)通过评估机器学习模型按顺序添加或删除特征时预测精度的变化来选择最佳特征，已用于晶体材料的特征筛选[33]和立方钙钛矿氧化物的结构–成分关系研究[26]，粒子群优化算法(PSO)通过反复迭代以从现有特征子集中产生更好的子集，也被用于吸波材料的厚度优化中[34]。混合式特征选择将过滤式和包裹式方法相结合，先通过过滤式方法过滤掉不相关特征，缩小寻优空间，再借助包裹式方法寻找重要特征，减少算法执行时间的同时提升了模型精度[47]。目前，已有研究人员将混合式特征选择用于 NASICON 型固态电解质激活能预测[35]、煤样光谱数据分析[36]等机器学习建模任务，在花费较少时间的前提下获得最优特征子集和较高预测精度。

特征转换通过将原始特征映射到低维空间来生成新的特征。一般来说，它通过构造预测能力更强



"R^2" represents the conformity between predicted value and actual value; "–" represents this term is not mentioned in this study; "RMSE" represents the root mean square error between the predicted value and actual value; "AUC (Area Under Curve)" represents the area under the receiver operating characteristic curve (ROC); "RMSE/AV" is a ratio that shows the overall deviation degree of the predicted sample; "Accuracy" represents the ratio of the number of correctly classified samples to the total number of samples; "RMSD" means the root mean square deviation.

图 1 材料领域中常用的样本量治理方法

Fig. 1 Number of features governance methods commonly used in materials field

但数量更少的特征来实现有效的降维。主成分分析(PCA)能够从原始数据中提取最重要特征，有效降低特征维度，被用于电池容量优化[37]和二元合金结构预测[38]。确定独立筛选和稀疏操作符(SISSO)方法在确定独立筛选之后通过稀疏化算子自主寻找最优N维描述符，目前已被用于二元体系[39]、催化剂[40]和无机晶体固体[41]等材料的发现。此外，基于遗传编程的符号回归(GPSR)方法已应用于材料科学中，用于设计新的氧化物钙钛矿催化剂[42]，并通过描述稳定性[43]对 MXene 材料进行分类。

2.2 样本量治理

近年来，一系列高通量计算平台的搭建为开展广泛的材料研发提供了充足的数据积累[48–50]。然而，在一些材料研究中，特别是对于新材料，仍然面临缺乏足够数据来构建可靠机器学习模型的困境[51]。上述问题可以转化为如何通过增加原始样本或修改学

习算法实现样本量的扩充。

数据增强通常是指借助辅助数据或信息[52]对原始小样本数据集进行数据扩充。随着深度学习的发展，基于神经网络的方法如生成对抗网络(Generative Adversarial Networks)[53–54]，变分自编码器(Variational Autoencoder)[55–56]等数据增强方法已成功应用于材料中，均带来了预测精度的提升。此外，主动学习(Active Learning)[57]利用预先建立的机器学习预测模型来迭代和自适应地对候选化学空间进行采样，为昂贵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[58–59]。因此，它也常被用来解决材料领域小样本学习问题。表 1 列出了上述方法在材料领域的具体应用以及取得的效果。

表 1　样本量治理方法在材料领域的应用

Table 1　Applications of sample governance method in materials science

ML algorithm	Materials	# of initial sample	# of governed sample	Result	Reference
Generative adversarial networks	2D materials	291 840	2 650 264	AUC: 0.96	[53]
	Inorganic materials	251 368 (OQMD) 57 530 (MP) 25 323 (ICSD)	1 831 648 (OQMD) 1 969 633(MP) 1 983 231 (ICSD)		[54]
Variational autoencoder	Inorganic materials	10 981 (MP)	>19 000 (MP)		[55]
	3-D molecules	46 744			[56]
Active learning	Inorganic perovskite	5 218	79		[58]
	NiTi-based shape memory alloy	256	15	R^2: 0.85	[59]

OQMD: Open Quantum Materials Database; MP: Materials Project; ICSD: Inorganic Crystal Structure Database; "#" represents a number; "AUC (Area Under Curve)" represents the area under the receiver operating characteristic curve (ROC); "R^2" represents the conformity between predicted value and actual value.

3　领域知识嵌入的数据质量治理框架

虽然针对数据品质和数量治理的研究取得了一定的成果，但这些研究仅面向特定材料问题中个别数据质量维度。此外，材料领域在长期探索中已经积累了大量的专家经验、计算公式、理论推导等领域知识。然而，鲜有研究在机器学习建模时将材料领域知识考虑在内，可能导致模型预测结果与材料专家认知不一致甚至相悖，这极大地限制了机器学习在材料领域更大范围的应用。因此，材料领域知识对于机器学习的指导作用也至关重要。为了全面评估影响机器学习模型性能的多种关键数据质量问题，本节提出了面向材料领域机器学习全过程的领域知识嵌入的数据质量治理框架(a Data Quality & Quantity GoveRnanCe framework embedded by materials domain knowledge in the whole process of materials machine learning, DQQ-GRC)。该框架通过明确机器学习在材料领域应用全过程中各阶段面临的数据质量问题，提出了数据驱动方法与材料领域知识协同的数据质量评估和提升策略，有望帮助研究人员为广泛的机器学习任务提供更高质量的学习样本，进而提升机器学习在材料领域应用的可信度。

本节提出的 DQQ-GRC 可表示为一个三元组，如式(1)所示：

$$\text{DQQ}-\text{GRC}=<\text{DQQDs},\text{LCM},\text{PMs}> \tag{1}$$

式中：DQQDs 是一系列数据质量维度的集合，LCM 是数据质量治理的生命周期模型，PMs 是针对不同质量问题的数据质量处理模型的集合。如图 2 所示，DQQDs 规定了 DQQ-GRC 的边界，用于理解材料数据质量的内涵，并为进一步的数据质量治理活动的设计和组织提供理论支持。LCM 规划了材料数据质量治理的行动路线。在 DQQDs 的范围内，LCM 可以通过明确材料数据质量治理的全生命周期，以确保各质量治理活动可以按需有序地进行。PMs 是材料数据质量治理的实际执行者，针对每个材料数据质量维度，都提供了一系列数据驱动方法(包括已经应用于材料领域的和尚未应用于材料领域的方法)与可用于评估和提升特定数据质量维度的材料领域知识的协同策略。

3.1　材料数据质量维度 DQQDs

材料数据质量是一个复杂的多维概念，为全方面多维度评估材料领域用于机器学习的数据质量，本文定义了两类数据品质维度和一类数据数量维度，并解释了定义每一维度的必要性，如表 2 所示。其中，通用品质维度(Inherent Quality Dimensions)对数据所属的领域类型不敏感，是不同领域的研究人员普遍关注的品质维度，包括数据的可溯源性(Traceability)、准确性(Accuracy)、冗余性(Redundancy)、

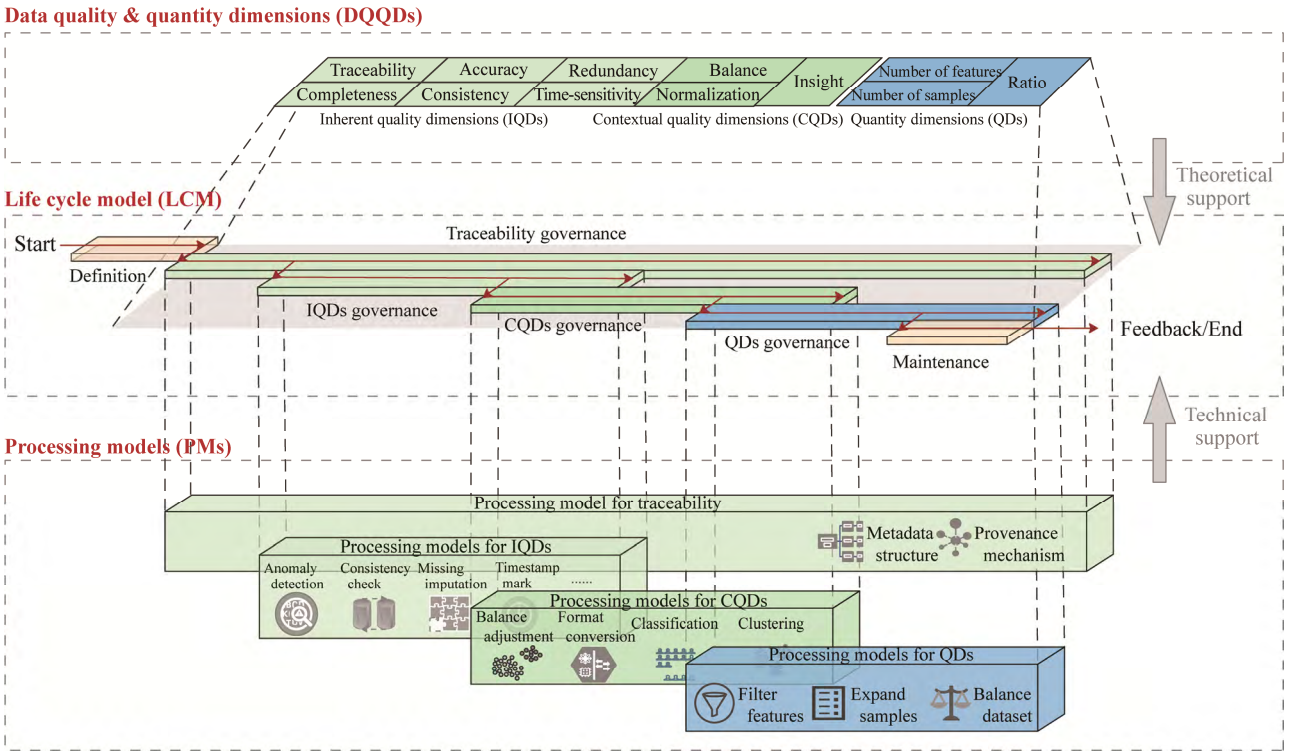


图 2 面向材料领域机器学习全流程的领域知识嵌入的数据质量治理框架

Fig. 2 A data quality & quantity governance framework embedded by materials domain knowledge in the whole process of materials machine learning

完整性(Completeness)、一致性(Consistency)和时间敏感性(Time-sensitivity)；上下文品质维度(Contextual Quality Dimensions)对数据所属的领域类型敏感，是材料数据用于机器学习建模所要考虑的特殊品质维度，包括均衡性(Balance)、规范性(Normalization)和洞察力(Insight)；数量维度(Quantity Dimensions)是用于机器学习任务的数据集均需考虑的另一类维度，包括样本量(Number of samples)、特征量(Number of features)以及二者的平衡(Ratio)。

3.2 材料数据质量治理的生命周期模型 LCM

LCM 制定了材料数据质量治理的行动路线，以实现机器学习过程、材料数据生命周期和数据质量治理生命周期的协同，有序组织机器学习全过程中的各种数据质量治理活动。

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、模型应用等 6 个阶段[13]。由于不同阶段对数据进行不同操作，因此每一阶段涉及不同的数据质量问题，但主要集中在数据准备、数据预处理和特征工程 3 个阶段。具体来说，数据准备阶段存在一致性、时间敏感性问题；数据预处理阶段存在准确性、完整性、均衡性、规范性、样本量不足等问题；特征工程阶段存在冗余性、特征量、洞察力等问题。

因此，针对机器学习全过程中的各种数据质量问题，LCM 明确了在机器学习哪个阶段执行哪种数据质量治理活动。如图 2 所示，LCM 由定义阶段、可溯源性治理阶段、通用品质治理阶段、上下文品质治理阶段、数量治理阶段和维护阶段组成。定义阶段明确了各阶段需要治理的对象及其相关的质量问题。由表 2 可知，可溯源性记录数据从收集到应用过程中的所有操作信息。因此，可溯源性治理在定义阶段结束后启动并贯穿 LCM 全过程。随后，通用品质治理在机器学习的数据准备阶段启动，解决数据准确性、一致性、完整性、冗余性、时间敏感性等品质问题。上下文品质和数量治理在数据预处理阶段同时启动，但二者的执行顺序是不固定的，因为它们解决的数据品质和数量问题不同，使用的数据质量治理方法也互不影响。为了保证机器学习模型预测结果的可靠性，这些数据品质和数量问题都需要在模型构建之前进行有效治理。因此，上述 3 个治理阶段均在特征工程阶段结束。此外，以上 3 个质量治理阶段产生的所有数据都反馈给可溯源性治理，以保证数据质量治理过程的可溯源性。最后，维护阶段关注机器学习应用过程中数据、模型和结果的更新，针对不同的更新对象，LCM 的某些阶段可能需要被重新执行。例如，如果用于机器学

表 2　材料数据质量维度

Table 2　Materials data quality & quantity dimensions

Category	Dimensions	Definition	Description
Inherent quality dimensions	Traceability	Measures whether the acquisition and ML-based processing of materials data is traceable	In order to make the machine learning results repeatable and trustworthy, traceability is required to record all operation information of the data.
	Accuracy	Measures whether data was recorded correctly and reflects realistic values	Materials data derived from industrial production, experiments, and calculations usually have uncertainties or are subject to errors, thus accuracy assessment is required.
	Redundancy	Measures whether data contains redundant information	Due to the complexity of materials performance driving mechanism and the multi-source of materials data, there are duplicate samples or redundant descriptors in data, thus redundancy elimination is required.
	Completeness	Measures whether all data used for training ML models and other critical additional information are recorded with no missing entries	Equipment defects or errors during the integration of multi-source materials data may lead to missing values in the data, thus completeness governance is required.
	Consistency	Measures whether all samples in the multi-source materials dataset are represented in the same way	Different data sources generally have different standards for data generation, storage and characterization, which may lead to the inconsistency of the final integrated dataset and consistency detection is required.
	Time-sensitivity	Captures the different characteristics of materials data according to certain time-related factors	For temporal materials data or materials properties related to time factors, time sensitivity analysis is required.
Contextual quality dimensions	Balance	Measures whether the population of different classes in the entire dataset is balanced	Materials data with good performance and poor performance are often difficult to be adequately collected, so balance governance is required.
	Normalization	Measures whether materials data has been converted into the representation or organization forms suitable for ML modeling as required	Data may have dimensional inconsistencies and uneven data division, so normalization governance is required before machine learning modeling.
	Insight	Measures the learnability of materials data quickly before carrying out more time-consuming analyses	In order to exploring whether the data used for machine learning modeling contains useful information, insight exploration is required.
Quantity dimensions	Number of features	Measures whether the number of features is redundancy in the materials data	Materials experts usually define multiple features to describe the complex driving mechanism of materials properties, so the features may be redundant and need to be controlled by the number of features governance.
	Number of samples	Measures whether the number of samples are appropriate in the materials data	The acquisition of materials samples depends on complex experiments with high cost, so the number of samples may be insufficient, which requires number of samples governance.
	Ratio	Measures whether the number of features and samples are balance in the materials data	Unbalanced number of features and samples often lead to poor machine learning modeling result, so it is necessary to consider the balance between them.

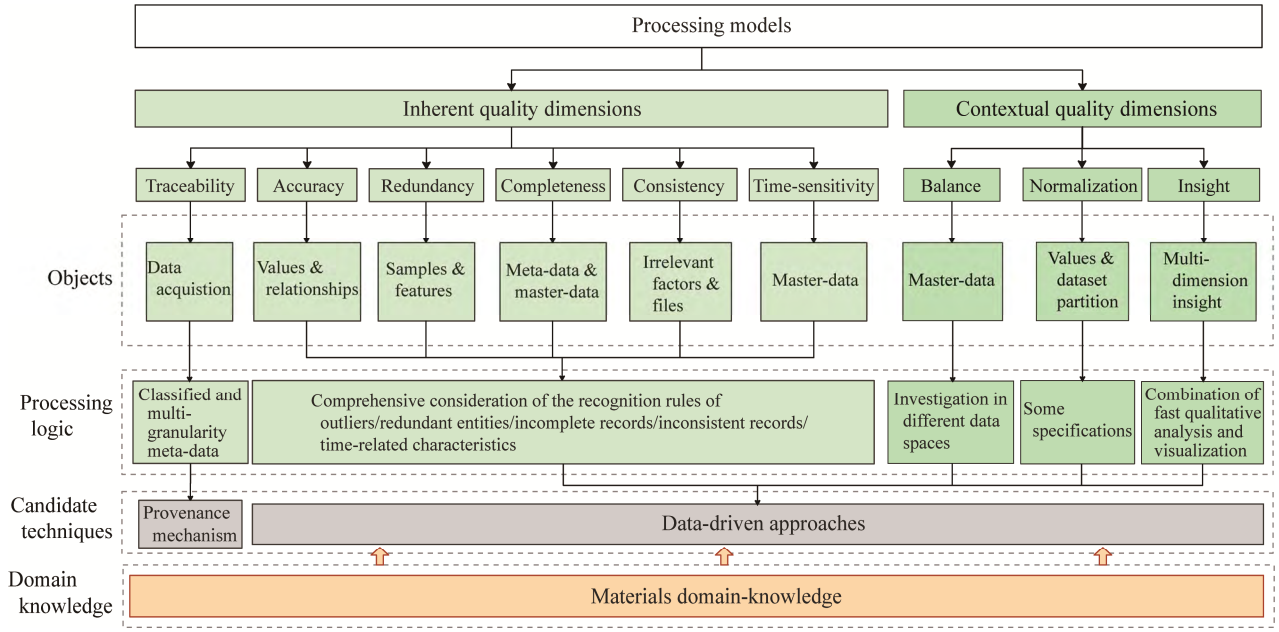
习的材料数据被修改或更新，那么所有的通用和上下文品质治理都需要被重新执行。

3.3　材料数据质量治理的处理模型 PMs

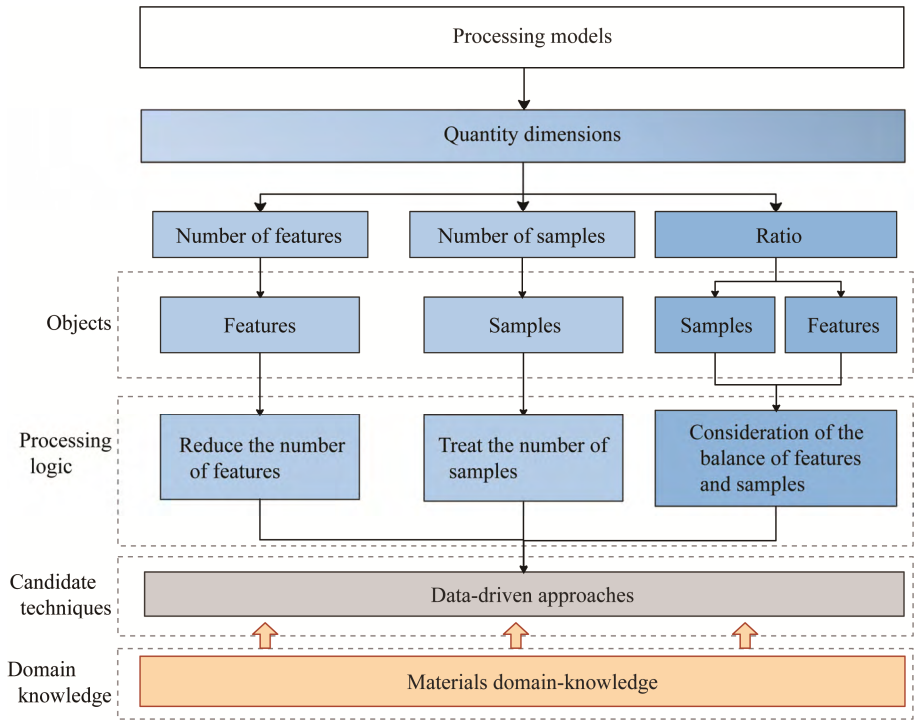
虽然许多数据驱动方法能够解决材料数据的某些质量问题，但如何合理地选择和应用这些模型对材料专家来说仍然是一个难题。此外，在不考虑领域知识的情况下，这些模型的处理结果可能并不可靠。为了解决这一问题，本节探索了材料领域知识在数据质量治理中的应用。如图 3 所示，针对 3.1 节定义的 12 种数据质量维度，分别构建了 12 个领域知识嵌入的数据质量处理模型，每个处理模型均定义了相关质量治理活动的治理对象，提出了如何选择合适处理方法的处理逻辑，提供了大量可用于评估和提升相关质量问题的候选技术，并分析了材料领域知识在每一质量问题中的协同参与方式，从而为材料专家提供数据质量治理实际实施的指导原则和候选方案，以在领域知识指导下帮助材料专家选择合适方法进行数据质量治理。具体如下：

1) 通用品质维度

通用品质维度综合考虑数据的异常取值、冗余特征、不完整记录、时间相关特性，目前普遍使用的数据驱动方法有数据一致性检查、缺失值补全、



(a) Data quality governance model



(b) Data quantity governance model

图 3 数据质量治理的处理模型总览

Fig. 3 Data quality & quantity governance processing models overview

异常值检测等，这些方法均能在一定程度上实现通用数据品质的治理。然而，在材料知识的背景下，描述符或目标属性的值总是具有物理限制的，材料数据类型不同，取值规范也会不同。另外，材料领域知识在剔除冗余特征和筛选关键特征方面也发挥着不可替代的作用。因此，结合材料领域知识开发新的材料数据品质治理方法具有重要意义[60]。针对准确性治理，本研究团队提出了融合材料领域知识的数据准确性检测方法[61]，将材料领域知识和数据驱动方法相协同，依次对描述符单个维度的正确性、描述符间相关关系的相关性、样本间的可靠性进行检测，确保数据集从机器学习初始阶段就有较高准

确性。该方法在 NASICON 型固态电解质激活能预测数据集上有效识别了潜在的异常数据并进行了合理修正，最终使得最优模型的 R^2 提升了 33%。针对冗余性治理，本研究团队分别将材料领域专家对“描述符对性能的影响程度”和“描述符之间的相关关系”的认知分别数值化表示为特征的重要度和符号化表示为不共现规则，并融入到数据驱动的特征选择过程中，提出了融合加权评分领域专家知识的多层级特征选择方法[62]和嵌入领域知识以降低特征相关性的特征选择方法[35]。2 种方法均成功遴选出符合领域认知且内部相关性更低的描述符，在此基础上构建的预测模型取得了更好的泛化性能和更高的稳定性。此外，还可将领域知识嵌入到机器学习的目标函数或约束条件中，协同领域知识和数据驱动方法开展其他品质维度治理活动，进一步提高机器学习模型的可解释性及预测结果的可靠性。

2) 上下文品质维度

上下文品质维度旨在探究数据空间分布的均衡性、数据取值的规范性及数据的潜在价值，通用的数据驱动方法有很多，比如通过聚类查看样本分布，统一数据规范化方法以及数据划分方式，利用主成分分析、*t* 分布–随机近邻嵌入等可视化手段帮助材料专家分析数据中蕴含的潜在模式等等，从而提升机器学习模型的准确性和泛化性。此外，材料领域知识在选择合适处理方法、洞察数据价值方面也有一定的指导作用。例如，针对均衡性治理，可结合材料领域知识对材料类别进行划分，引入基于生成对抗网络或变分自编码器的数据增强方法扩充少数类样本；对于文本数据，可以利用现有材料的语料库对预训练的语言模型 DistilRoBERTa 进行微调并学习上下文语义信息，在掩码语言模型的基础上对一句话中被随机掩盖掉的单词或短语结合其语境和标签进行预测，从而得到与原始文本语义相似但内容不同的生成样本；针对规范性治理，使用最小–最大规范化、零均值规范化方法，并结合材料专家对数据的理解划分训练集–验证集–测试集是一种有效的方案。

3) 数量维度

数量维度通常采用减少特征数量或扩充样本数量的方式寻求二者之间平衡，以达到数据数量治理的目的。目前普遍使用的数据驱动方法有数据增强、主动学习、特征选择、特征转换等。为进一步提高机器学习建模结果的准确性，可将描述特征间相关关系或约束样本取值的领域知识符号化或规则化表示，嵌入到数据驱动方法执行过程中。例如，针对样本量治理，可将材料领域知识以公式形式嵌入到数据增强的训练过程，以捕捉最相关的特征向量，保留数据关键信息；此外，还可将材料领域知识转化为材料数据分布规则，干预主动学习中贝叶斯优化的抽样过程，以从大量数据中挑选出和机器学习任务最相关的样本。为保持样本量和特征量的平衡，自动机器学习通过将数据预处理、特征工程、模型构建等过程自动化，是一种可行的方案。材料领域知识有望嵌入自动机器学习模型选择、超参数寻优等过程实现样本量和特征量的协同治理。

4　总结与展望

虽然材料领域的研究人员已经意识到了用于机器学习建模的材料数据质量的重要性，但仍然缺乏对材料数据质量内涵的深入理解和材料数据质量治理的有效策略。首先综述了材料领域数据品质治理和数量治理的研究现状；随后，提出了面向材料领域机器学习全过程的领域知识嵌入的数据质量治理框架，针对材料科学领域机器学习研究中缺乏全面系统的数据质量治理方法指导问题，定义了多种数据质量维度、构建了数据质量治理的生命周期模型、建立了一系列数据质量问题处理模型以综合实现材料数据质量的评估和提升，为材料研究人员进行数据质量治理提供理论指导与可行方案。

随着文本挖掘(TM)和自然语言处理(NLP)的发展，材料领域已有大量研究表明 TM 和 NLP 技术能够从科学文献中大规模地提取数据[63–64]。继续发展 TM 和 NLP 技术，从海量科学文献中挖掘的材料知识或许可为领域知识嵌入的数据质量治理框架提供材料领域知识。此外，将数据质量处理模型和现有数据分析工具相结合，开发一系列自动化的质量治理工具，将帮助材料研究人员在高质量数据驱动下构建高性能机器学习模型以加速新材料研发。

参考文献:

[1] ROBERT C. Machine learning, a probabilistic perspective[J]. Chance, 2014, 27(2): 62–63.

[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(1): 1–36.

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

[5] CHEN H H, CHEN J P, DING J H. Data evaluation and enhancement for quality improvement of machine learning[J]. IEEE Trans Reliab, 2021, 70(2): 831–847.

[6] MEHRABI N, MORSTATTER F, SAXENA N, et al. A survey on bias and fairness in machine learning[J]. Acm Comput Surveys, 2021, 54(6): 1–32.

[7] OAKI Y, IGARASHI Y. Materials informatics for 2d materials combined with sparse modeling and chemical perspective: Toward small-data-driven chemistry and materials science[J]. Bull Chem Soc Jpn, 2021, 94(10): 2410–2422.

[8] 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.

[9] BEAL M S, HAYDEN B E, LE GALL T, et al. High throughput methodology for synthesis, screening, and optimization of solid state lithium ion electrolytes[J]. ACS Comb Sci, 2011, 13(4): 375–381.

[10] RAJAN A C, MISHRA A, SATSANGI S, et al. Machine-learning-assisted accurate band gap predictions of functionalized mxene[J]. Chem Mater, 2018, 30(12): 4031–4038.

[11] LU P, ZHUO Z, ZHANG W H, et al. A hybrid feature selection combining wavelet transform for quantitative analysis of heat value of coal using laser-induced breakdown spectroscopy[J]. APPL Phys B-Lasers O, 2021, 127(19): 1–11.

[12] YUAN J, WANG Q, LI Z, et al. Domain-knowledge-oriented data pre-processing and machine learning of corrosion-resistant γ-u alloys with a small database[J]. Comput Mater Sci, 2021, 194: 110472.

[13] 刘悦, 邹欣欣, 杨正伟, 等. 材料领域知识嵌入的机器学习[J]. 硅酸盐学报, 2022, 50(3): 863–876.
LIU Yue, ZOU Xinxin, YANG Zhengwei, et al. J Chin Ceram Soc, 2022, 50(3): 863–876.

[14] 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]. Chem Eng Sci, 2013, 101: 478–485.

[15] HEMMATI-SARAPARDEH A, TASHAKKORI M, HOSSEINZADEH M, et al. On the evaluation of density of ionic liquid binary mixtures: Modeling and data assessment[J]. J Mol Liq, 2016, 222: 745–751.

[16] 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.

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

[18] WUEST T, MAK-DADANSKI J, THOBEN K-D. Data quality in materials science: A quality management manual approach[C]//IFIP International conference on advances in production management systems, Springer, 2014: 42–49.

[19] WENZLICK M, MAMUN O, DEVANATHAN R, et al. Assessment of outliers in alloy datasets using unsupervised techniques[J]. J Materiomics, 2022, 74(7): 2846–2859.

[20] WILKINSON M D, DUMONTIER M, AALBERSBERG I J, et al. The fair guiding principles for scientific data management and stewardship[J]. Sci Data, 2016, 3: 160018.

[21] 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. Inform Document Services (in Chinese), 2021, 42(1): 57–68.

[22] 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): 1–6.

[23] AGRAWAL A, DESHPANDE P D, CECEN A, et al. Exploration of data science techniques to predict fatigue strength of steel from composition and processing parameters[J]. Integr Mater Manuf I, 2014, 3: 90–108.

[24] SHIN D, YAMAMOTO Y, BRADY M P, et al. Modern data analytics approach to predict creep of high-temperature alloys[J]. Acta Mater, 2019, 168: 321–330.

[25] 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): 1–8.

[26] DENG Q, LIN B. Exploring structure-composition relationships of cubic perovskite oxides *via* extreme feature engineering and automated machine learning[J]. Mater Today Commun, 2021, 28: 102590.

[27] MANGAL A, HOLM E A. A comparative study of feature selection methods for stress hotspot classification in materials[J]. Integr Mater Manuf I, 2018, 7(3): 87–95.

[28] QI Z C, ZHANG N X, YONG L, et al. Prediction of mechanical properties of carbon fiber based on cross-scale fem and machine learning[J]. Compos Struct, 2019, 212: 199–206.

[29] 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.

[30] ZENG Y Z, LI Q X, BAI K W. Prediction of interstitial diffusion activation energies of nitrogen, oxygen, boron and carbon in bcc, fcc, and hcp metals using machine learning[J]. Comput Mater Sci, 2018, 144: 232–247.

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

[32] STANEV V, OSES C, KUSNE A G, et al. Machine learning modeling of superconducting critical temperature[J]. NPJ Comput Mater, 2018, 4(29): 1–14.

[33] FURMANCHUK A, SAAL J E, DOAK J W, et al. Prediction of Seebeck coefficient for compounds without restriction to fixed stoichiometry: A machine learning approach[J]. J Comput Chem, 2018, 39(4): 191–202.

[34] FENG H Q, WU B H, LIU Y Y, et al. The application of particle swarm optimization algorithm on absorbent materials[J]. Appl Mech Mater, 2014, 446–447: 1541–1545.

[35] LIU Y, ZOU X X, MA S C, et al. Feature selection method reducing correlations among features by embedding domain knowledge[J]. Acta Mater, 2022, 238: 118195.

[36] YAN C, LIANG J, ZHAO M, et al. A novel hybrid feature selection strategy in quantitative analysis of laser-induced breakdown spectroscopy[J]. Anal Chim Acta, 2019, 1080: 35–42.

[37] STURLAUGSON L E, SHEPPARD J W. Principal component analysis preprocessing with bayesian networks for battery capacity estimation[C]//2013 IEEE International Instrumentation and Measurement Technology Conference (I2MTC), Minneapolis, MN, USA, 2013: 98–101.

[38] CURTAROLO S, MORGAN D, PERSSON K, et al. Predicting crystal structures with data mining of quantum calculations[J]. Phys Rev Lett, 2003, 91(13): 135503.

[39] 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: 083802.

[40] ANDERSEN M, LEVCHENKO S V, SCHEFFLER M, et al. Beyond scaling relations for the description of catalytic materials[J]. Acs Catalysis, 2019, 9(4): 2752–2759.

[41] BARTEL C J, MILLICAN S L, DEML A M, et al. Physical descriptor for the Gibbs energy of inorganic crystalline solids and temperature-dependent materials chemistry[J]. Nat Commun, 2018, 9: 4168–4177.

[42] WENG B, SONG Z, ZHU R, et al. Simple descriptor derived from symbolic regression accelerating the discovery of new perovskite catalysts[J]. Nat Commun, 2020, 11: 3513–3520.

[43] HE M, ZHANG L. Machine learning and symbolic regression investigation on stability of mxene materials[J]. Comput Mater Sci, 2021, 196: 110578.

[44] TRAN B, XUE B, ZHANG M, et al. A new representation in pso for discretization-based feature selection[J]. IEEE Trans Cybern, 2018, 48(6): 1733–1746.

[45] HANCHUAN P, FUHUI L, DING C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy[J]. IEEE Trans Pattern Anal Mach Intell, 2005, 27(8): 1226–1238.

[46] GENUER R, POGGI J M, TULEAU-MALOT C. Variable selection using random forests[J]. Pattern Recognit Lett, 2010, 31(14): 2225–2236.

[47] BALAKRISHNAN K, DHANALAKSHMI R. Feature selection techniques for microarray datasets: A comprehensive review, taxonomy, and future directions[J]. Front Inf Technol Electron Eng, 2022, 23(10): 1451–1478.

[48] JAIN A, ONG S P, HAUTIER G, et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation[J]. Apl Materials, 2013, 1(1): 011002.

[49] KIRKLIN S, SAAL J E, MEREDIG B, et al. The open quantum materials database (oqmd): Assessing the accuracy of dft formation energies[J]. NPJ Comput Mater, 2015, 1(1): 15010.

[50] HE B, CHI S, YE A, et al. High-throughput screening platform for solid electrolytes combining hierarchical ion-transport prediction algorithms[J]. Sci Data, 2020, 7(1): 151.

[51] WU Y J, FANG L, XU Y B. Predicting interfacial thermal resistance by machine learning[J]. NPJ Comput Mater, 2019, 5(1): 56.

[52] WANG Y Q, YAO Q M, KWOK J T, et al. Generalizing from a few examples: A survey on few-shot learning[J]. Acm Comput Surv, 2020, 53(3): 1–34.

[53] SONG Y, SIRIWARDANE E M D, ZHAO Y, et al. Computational discovery of new 2D materials using deep learning generative models[J]. ACS Appl Mater Interfaces, 2021, 13(45): 53303–53313.

[54] DAN Y, 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.

[55] NOH J, KIM J, STEIN H S, et al. Inverse design of solid-state materials *via* a continuous representation[J]. Matter, 2019, 1(5): 1370–1384.

[56] HOFFMANN J, MAESTRATI L, SAWADA Y, et al. Data-driven approach to encoding and decoding 3-d crystal structures[J]. Arxiv, 2019. Doi: 10.48550/arXiv.1909.00949.

[57] 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(1): 21.

[58] MIN K, CHO E. Accelerated discovery of potential ferroelectric perovskite *via* active learning[J]. J Mater Chem C, 2020, 8(23): 7866–7872.

[59] 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 Mater, 2019, 20(1): 1010–1021.

[60] JEONG M H, SULLIVAN C J, GAO Y Z, et al. Robust abnormality detection methods for spatial search of radioactive materials[J]. Trans GIS, 2019, 23(4): 860–877.

[61] 施思齐, 孙拾雨, 马舒畅, 等. 融合材料领域知识的数据准确性检测方法[J]. 无机材料学报, 2022, 37(12): 1311–1320.
SHI Siqi, SUN Shiyu, MA Shuchang, et al. J Inorg Mater (in Chinese), 2022, 37(12): 1311–1320.

[62] LIU Y, WU J, AVDEEV M, et al. Multi-layer feature selection incorporating weighted score-based expert knowledge toward modeling materials with targeted properties[J]. Adv Theory Simul, 2020, 3: 1900215.

[63] LIU Y, GE X Y, YANG Z W, et al. An automatic descriptors recognizer customized for materials science literature[J]. J Power Sources, 2022, 545: 231946.

[64] 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.