

科技评论

面向 AI for Materials 的数据基座重塑：全生命周期数据质量管控框架与可信生态构建

施思齐^{1,2}

摘要 高质量数据是驱动材料科学研究向智能化转型的核心要素。然而,当前材料数据生产体系的结构性缺陷,不仅造成数据碎片化、孤岛化、格式异构、记录残缺,更因生产者责任意识薄弱和系统性治理框架缺位,使得质量管控责任被过度转嫁给下游使用者。尽管数据使用者尝试通过在诸多数据质量治理方法中引入领域知识以提升治理成效,但由于对数据生产过程理解有限,其治理策略的普适性及结果的可靠性均难以保障。围绕如何持续生产高质量材料数据这一核心问题,系统分析数据全生命周期中各阶段数据质量问题产生的原因,并据此提出综合性材料数据质量管控框架的构建思路,旨在引导数据生产与使用的各方参与者重新审视从既有数据获取、管理到应用的完整流程,为促进高质量材料数据高效交付的实施路径提供参考,进而为构建材料数据可信生态奠定基础。

关键词 材料科学;数据质量管控框架;可信生态;全生命周期

1 AI 时代材料数据质量面临的系统性挑战

数据是指通过观察、实验或计算得出的结果。随着数据从一般性资源向核心生产要素与可复用资产跃升,高质量数据的生产与治理能力日益成为竞争制高点。近年来,国家持续推进数据资源化、资产化和要素化进程,先后出

台《中共中央 国务院关于构建数据基础制度更好发挥数据要素作用的意见》和《“数据要素×”三年行动计划(2024—2026 年)》,并组建国家数据局,持续推进高质量数据集建设和行业数据资源体系建设,为数据要素价值释放和数字经济高质量发展提供制度支撑。2025 年 8 月出台的《国务院关于深入实施“人工智能+”行动的意见》更是将“打造开放共享的高质量科学数据集”写入“人工智能+”科学技术的第一条。这些制度安排与政策导向,不仅为数据要素价值释放提供了顶层设计,也为各垂直领域

的数据治理与应用明确了方向。

就材料领域而言,长期积淀的实验、计算、表征与工程数据中蕴藏着丰富的科学规律与工程知识,是国家基础性数据资源的重要组成部分。与此同时,以机器学习(machine learning, ML)为代表的人工智能(artificial intelligence, AI)技术,正在重塑材料科学的方法论,推动 AI for Materials 这一前沿交叉领域快速兴起^[1-8]。这一新范式的核心优势在于,其能够从高维数据集中高效地挖掘潜在的复杂关联规律^[9],揭示传统理论与方程难以描述的新现象^[10],并可在无

1. 上海大学核电关键材料全国重点实验室,材料科学与工程学院,上海 200444
2. 上海大学材料基因组工程研究院,上海 200444

收稿日期: 2025-12-16; 修回日期: 2026-05-11

基金项目: 重点新材料研发及应用国家科技重大专项(2026ZD0623700); 国家自然科学基金项目(92472207)

作者简介: 施思齐, 教授, 研究方向为算法-数据-知识互利共生且与实验融合的材料设计新范式构建, 致力于推动人工智能赋能新材料创制与评价, 电子信箱: sqshi@shu.edu.cn

引用格式: 施思齐. 面向 AI for Materials 的数据基座重塑: 全生命周期数据质量管控框架与可信生态构建[J]. 科技导报, 2026, 44(16): 17-28;
doi:10.3981/j.issn.1000-7857.2025.12.00082

需预设算法指令的前提下实现有效外推与概率预测^[11-12]。然而,在 AI 快速发展的当下,模型效能仍受高质量数据供给不足的制约^[13-14]。因此,如何系统提升材料数据的质量,促进其有效开发利用,进而支撑 AI 驱动的材料研发新范式,已成为材料学科向数字化、智能化转型过程中亟须解答的命题。

事实上,数据质量作为一个跨学科议题,已在信息科学^[15-17]、生物医学^[18-19]等领域得到广泛而深入的探讨。在材料领域,多项研究从概念层面总结了高质量材料数据应具备的关键属性^[20-21],为现有数据的评估与治理提供了宝贵的理论参考。例如,路勇超等^[21]基于“AI-Ready 型数据”理念,从单样本与整体数据集双重视角出发,系统梳理了支撑 AI 应用的核心数据属性。为规范 ML 建模全过程中的数据处理,施思齐团队提出了一种融合领域知识的数据质量治理框架^[22],并引入材料领域启发式规则以提升数据异常检测效能^[23]。尽管这些努力已奠定一定基础,然而,从概念共识向可重复、可推广的实践规范转化,仍存在结构性断层,主要体现在 3 个方面。

1) 质量治理责任倒置。上游数据生产环节中薄弱的质量管控机制,常致使本应在前端解决的质量问题被过度转移至下游的数据使用者。他们不得不投入大量资源与精力开展数据清洗、对齐与验证,甚至需要进行“逆向工程”以推断数据的原始语境,不仅负担沉重,而且效率低下。

2) 过程可追溯性缺失。在

Gate-stage 工序交接模式下,各数据生产环节存在“交差”式交付心态与规避溯源动机,致使过程记录普遍缺失,数据谱系模糊,质量问题难以溯源。由此引发的非结构化的治理操作(如过度清洗等),不仅无法从根本上解决质量缺陷问题,还可能造成关键物理特征丢失,削弱数据的科学价值。

3) 解决方案碎片化。现有的数据质量提升多为“项目式”或“救火式”的处理,即针对特定数据集或具体建模任务,开发定制化的清洗脚本与验证规则。这些解决方案缺乏普适性与可扩展性,治理成果难以迁移复用,造成大量低水平重复劳动。

当前,高质量数据已然超越单纯的技术诉求,成为突破硬件算力瓶颈与算法设计固有限制的核心杠杆,更是推动 AI 模型性能实现质的飞跃的根本路径。放眼全局,材料数据质量问题并非局限于某一环节,而是深刻贯穿数据生产、流通直至消费的全生命周期之中,环环相扣、相互渗透^[24]。若仍将数据质量提升仅视为模型训练前的“事后修补”或局部优化,无疑会低估或忽视其系统性风险。这种碎片化思维不仅难以从根本上解决问题,反而容易酿成数据偏差累积、模型泛化能力下降乃至科研结论失真等连锁隐患,最终阻滞 AI for Materials 从数据驱动向知识发现的范式跃迁。

因此,材料数据质量的保障与优化绝非可有可无的辅助措施,也不应被视为数据产生后的修补性任务,而是一项亟需多方主体协同参与、分阶段逐层推进的系统性

数据质量工程(Data Quality Engineering)。只有通过科学识别材料数据生命周期各阶段的潜在质量风险,并据此构建起覆盖全链条、动态可迭代的数据质量管控框架,方能实现高质量数据资源的持续稳定供给。这一框架的缺失,会严重制约 AI 在材料领域的应用潜力;反之,其系统建立与高效运转,已从理想化的“优化选项”演变为推动学科革新的“必然抉择”,直接关乎该前沿领域能否真正突破数据瓶颈,实现从经验试错向智能预测的深刻转型。

2 解构材料数据全生命周期的质量风险链

要系统性地破解上述挑战,关键在于对材料数据全生命周期链条的深入剖析,厘清各环节如何悄然催生、传递并层层放大质量风险。本文将材料数据生命周期划分为 3 个关键阶段:数据获取、管理和应用。每个阶段虽各有其独特困境,却并非孤立存在。前一阶段遗留的质量隐疾往往会作为“技术债务”积累并传递至后续阶段,形成误差逐级放大的效应。

2.1 数据获取阶段

数据获取是数据质量管理的起点,承担着为后续质量分析与提升提供原始信息的根本任务。高质量的数据,有赖于在获取时广泛覆盖多类别材料,并进行系统化记录,这是支撑深度挖掘与可靠应用的前提。遗憾的是,当前的材料数据获取实践普遍存在三重风险:分布失衡、记录失范与采集失真,致使其从源头处便制约了数据质量

的提升空间。

学术资源与关注度的不均衡分配,是造成材料数据生态呈现结构性分化的直接原因^[25-26]。在材料科学计算领域, Materials Project 数据库已收录逾 20 万种材料及超过 57 万种物质的物性数据^[27], AFLOWLIB 数据库更是涵盖了 350 万种化合物的模拟计算性质记录^[28-29];而全球规模最大的材料实验数据库 HTEM,仅包含约 8 万条实验数据^[30]。并且,在数据高度富集的成熟或热点材料体系内,冗余且同质化的数据积累问题尤为突出。以钙钛矿太阳能电池材料为例,相关文献计量研究对 11092 篇科学出版物进行分析后发现,此类热点体系的数据分布呈现显著的长尾分布特征,即少量经典组分占据绝大部分数据量,而多数新型组分数据极为稀缺^[31]。这种同质化积累可能掩盖潜在的系统偏差,致使局部质量评估标准趋于固化,难以有效甄别隐匿于一致性表象之下的低质数据。与之相对,在研究门槛高、关注度低或尚处于探索阶段的新兴或复杂材料体系中,其样本稀缺性不仅阻碍了体系专属质量控制模型的构建,更导致数据波动范围与异常阈值难以合理界定,后续质量评估往往被迫沿用其他体系的通用标准,进而引入系统性误判风险。这一结构性失衡,实质上在数据生态两端分别埋下了“过拟合”与“欠识别”的质量隐患。化解这一结构性失衡问题,迫切需要领域知识的深度介入:基于物理化学原理识别数据采集的“关键空白区域”,判断哪些体系虽当前冷门却具有潜在科学价值,

从而指导数据资源的战略性布局。这种知识引导的采样策略,能从源头上缓解数据生态的“马太效应”,避免因学术热点漂移造成结构性数据缺失。

其次,材料研发人员或工程师偏重于结果数据的记录,而忽视对数据生成全流程上下文的系统捕获,致使数据原始记录难以形成完整、连贯的知识载体。这不仅直接削弱了结果数据本身的可解释性与可验证性,更在数据源头处制造深层障碍,导致后续治理过程面临溯源断裂、纠偏效率低下等严峻困境。在面对数据异常或来源不一致时,治理工作往往只能依赖统计方法剔除离群值,而非基于物理机制的精准修正。这不仅难以彻底消除偏差隐患,还可能在去除噪声的同时,丢失蕴含新现象或边界条件的宝贵信号,使质量优化过程本身潜藏信息减损的风险。而社区层面共识性元数据框架的缺失,进一步加剧了记录质量的参差不齐与格式异构问题,导致多源数据难以有效兼容与合并,间接固化了材料数据的碎片化状态。

此外,数据获取阶段亦普遍存在受发表导向与资源约束影响而衍生的主观筛选倾向:研发人员仅记录实验成功的“正面”结果,而将严格受控条件下未能达成目标的“负面”数据点(如合成失败的路径)予以弃置。然而,此类看似“无用”的数据实则蕴含独特的学术价值。在后续的数据使用阶段,它们不仅是 AI 模型辨识成功-失败边界、辅助探明材料失效机制的重要依据,更是构建稳健模型、提升预测可靠性的宝贵资源^[32-34]。

例如, Raccuglia 等^[32]基于系统整理的“暗反应”数据集(包含大量失败的合成实验)训练支持向量机模型,其对模板化硒化钒合成反应的预测准确率可达 89%,超越了多数资深专家的直觉判断。此外, Moosavi 等^[33]利用 ML 方法从失败的金属有机框架合成数据中重建实验轨迹,成功解析出最佳合成条件。上述案例充分表明,领域知识的介入能够帮助研究者建立更科学的“数据价值”评判标准:基于对材料形成机理、反应路径与失效模式的理解,科研人员可以识别哪些“失败实验”实则揭示了有意义的相空间边界,哪些看似异常的结果可能指向未被发现的新现象。这种知识引导的价值判断,能够有效克服“唯成功论”的数据采集偏见,推动建立更为客观、全面的数据记录文化。偏离客观中立的数据采集行为,虽有利于聚焦核心发现,却也导致材料数据集呈现人为净化的非自然分布状态。这不仅削弱了数据集表征真实材料行为的统计效力,也致使后续数据清洗与优化过程中误差难以根除,甚至在去噪过程中无意剔除了蕴含潜在高价值的数据信号,从而在质量优化的名义下造成不可逆的信息损失。

在数据获取方式日益多元化的当下,基于自然语言处理(natural language processing, NLP)技术,尤其是大语言模型(large language model, LLM)技术从海量科学文献中自动化抽取关键信息并转化为结构化数据,正逐步发展为一种高效、低成本的数据获取新范式。然而,此类方法在加速材料

数据聚拢、提升现有材料数据规模的同时,亦引入一系列不容忽视的质量风险:跨文献术语、单位与表述的模糊性与歧义性易引发抽取偏差;通用 LLM 在材料数据整理中可能引入命名实体识别不准、材料合成路径信息缺失、由 LLM“幻觉”(hallucination)现象导致的虚假实体或关系生成等问题;多模态材料领域数据(文本-表格-图像)常难以实现精准对齐与信息协同提取;高质量标注数据的获取成本高昂,成为制约专用模型性能的主要瓶颈。若缺乏领域本体对齐约束、人工抽样复核等配套的质量校验机制,自动化抽取不仅难以保障数据获取的全面性,更可能隐含系统性偏差,影响后续分析与建模可靠性。

更为严峻的是,占据材料科学文献与研究过程核心地位的图像数据(如微观形貌扫描电子显微镜/透射电子显微镜显微图像、晶体衍射斑点、光谱图以及各种复杂数据图等),长期处于“多模态数据治理”的盲区。相较于文本数据的抽取,图像数据的标注质量更令人担忧:现有的图像存档往往仅有简单的图题,缺乏对图像中关键物理特征(如位错类型、析出相分布、界面晶格畸变程度)的像素级语义标注与矢量对齐。对于以视觉语言模型(vision-language models, VLM)为代表的多模态大模型(multimodal large language models, MLLM)技术而言,未经过精细化语义标注的图像数据无异于“哑数据”,难以参与模型训练或导致模型产生严重的特征幻觉。因此,建立面向材料显微学与谱学

的图像标注元数据规范,是释放 MLLM 在新材料研发中潜力的前置条件。

2.2 数据管理阶段

在数据驱动的研究范式中,数据获取如同原油开采,是资源获取的起点;而数据管理则承担着炼油厂的核心职能,旨在将原始、混杂的“数据原油”精炼为可信、可用、可复用的“高质量数据燃料”,支撑材料研究社区实现高效协同的创新。

现实中,材料数据管理仍主要服务于孤立的课题研究,而非面向开放、可互操作的社区知识体系。尽管数据经初步整理,但格式与语义异构、数据关联薄弱等结构性缺陷仍普遍存在,致使材料数据生态中的质量提升进程陷入低效循环,优化努力屡屡受挫,暴露出数据治理认知滞后于数据规模化生产能力的现实困境。

当前,尽管部分材料子领域已出现局部标准与本体构建的尝试^[35-37],但由于社区层面长期缺乏统一的数据表达框架,使得材料数据在格式与语义层面仍处于双重异构状态。这种异构并非技术实现上的偶然疏忽,而是源于材料学界对标准化、结构化数据表达规范的集体性低估与投入不足。不同课题组、平台乃至期刊在描述相同或类似材料体系时,常使用各自定义的术语、单位、结构模板与元数据体系,形成大量无法直接互通的数据“方言”。在数据管理阶段,这种双重异构性迫使治理工作陷入高成本、低效率的手工对齐与转化,大量资源消耗在重复性的清洗、映射与合并操作上,难以实

现规模化、自动化的质量提升。在数据使用阶段,语义模糊与结构各异也大幅度削弱了数据的可发现性、可理解性与可复用性,不仅阻碍跨研究、跨机构的数据集成与联合分析,也使得基于此类数据训练的模型,其可靠性与泛化能力存疑,最终制约了材料研发从数据驱动中获取整体效能的可能。

材料领域内的数据关联薄弱现状,源于数据管理实践中对复杂科学关系(如“工艺—结构—性能”链)的捕捉不足。尽管知识图谱等语义技术已在部分场景中得到探索^[38-40],但面向材料全生命周期、机器可理解且可扩展的关联框架,仍处于早期构建阶段。多数数据关联仍依赖于隐式或表浅链接,缺乏结构化、语义化的关系表达。这既是技术路径上的惯性延续,也反映出社区对深度关联建模的投入有限。由此产生的一个直接后果便是,材料数据在跨域整合、因果溯源与知识迁移过程中频繁发生语义断裂,上游的信息模糊被逐级放大为下游的知识断层与推断误差,不仅严重制约了质量优化工作的系统性与可解释性,也使得数据驱动材料的发现难以突破局部关联的局限,进而无法实现真正意义上的跨尺度、跨体系知识融合。

在多模态数据治理层面,这种断裂体现得尤为突出:数值型数据(如元素组成、硬度值)与对应的图像表征数据(如断口形貌、晶粒尺寸分布图)往往被割裂存储于不同的非结构化文档或文件夹中。由于缺乏将图像特征向量化并与结构性能参数进行显式知识图谱

关联的治理机制,大量蕴含“形貌—性能”隐性映射规律的图像数据沦为无法被机器理解的视觉背景。因此,要构建高质量材料数据生态,必须将图像数据的语义解析与关联治理纳入质量管控的核心议程,而非将其视为数值计算的附属插图。

2.3 数据应用阶段

在材料数据的全生命周期中,数据应用是释放数据价值、驱动科学认知的核心阶段。研发人员通过融合领域知识与物理化学规则,从多维数据中识别潜在模式,并将其形式化为可验证的原理与模型,为同类问题提供方法论参考。在这一过程中,领域知识的系统嵌入是提升数据治理智能化水平的关键环节。唯有将材料科学的先验规则与数据驱动方法深度结合,才能有效提升异常检测的准确性、增强模型的物理可解释性,并在数据稀疏场景下实现可靠的外推推断。而数据的准确可靠与全面覆盖,是确保所得研究结论与模型具备科学有效性与泛化能力的根本前提,也直接决定了数据驱动研究能否从源头实现质量闭环。

在数据驱动的材料科研范式下,数据准确度是衡量其能否成为可持续复用研究资产的关键。真正“准确”的数据应同时具备3个维度上的可信性:(1)真实性,即数据如实反映实验观测、测量结果以及统计行为;(2)逻辑自洽性,即数据内部及与相关领域知识之间不存在矛盾;(3)时效性,即数据与其所依赖的方法、条件与技术背景在时间维度上保持一致。

满足这些条件的数据,才能在长期引用、跨项目复用和持续集成中不断积累信用,形成可被社区持续信任与挖掘的数据资产;反之,若数据失真,其影响会在后续使用与衍生过程中被逐级放大。初始的测量偏差或逻辑矛盾,往往会在模型训练、数据融合与知识推演中被非线性传递,不仅导致当前结论失效,更会污染基于此衍生的新数据集,使误差在迭代研究中被固化甚至加剧,最终造成“劣质数据驱逐优质数据”的长期风险。

材料科学发现的深度与广度则由数据的综合性所决定。其内涵通常包含2个关键维度:(1)数据特征的全面性,即覆盖与研究对

象相关的关键变量及其相互关系,形成可解释的“成分—结构—工艺—性能”知识链路;(2)数据类别的多样性,即样本应在类别、条件与体系间具有充分且均衡的分布。若特征覆盖不全,数据集便如同一副残缺的拼图,即便每一个“拼图碎片”都具备高度的准确性,也无法支撑研究者构建材料科学的完整认知框架。若类别分布失衡,则会导致ML模型或科学假设易对优势类别过拟合,在外推至稀缺或未知体系时表现失准,且这种偏差会在后续数据生成与模型迭代中被进一步固化。

3 贯穿材料数据全生命周期的质量管控框架

为构建可靠、可扩展、可持续发展的高质量材料数据生态,需从源头到终端建立贯穿全生命周期的递进式质量管控框架(图1)。

3.1 数据获取阶段:以“规范生成”奠定质量基石

建立规范化的数据生产和记录流程,是从源头上确保数据“原生质量”的关键。针对当前材料数

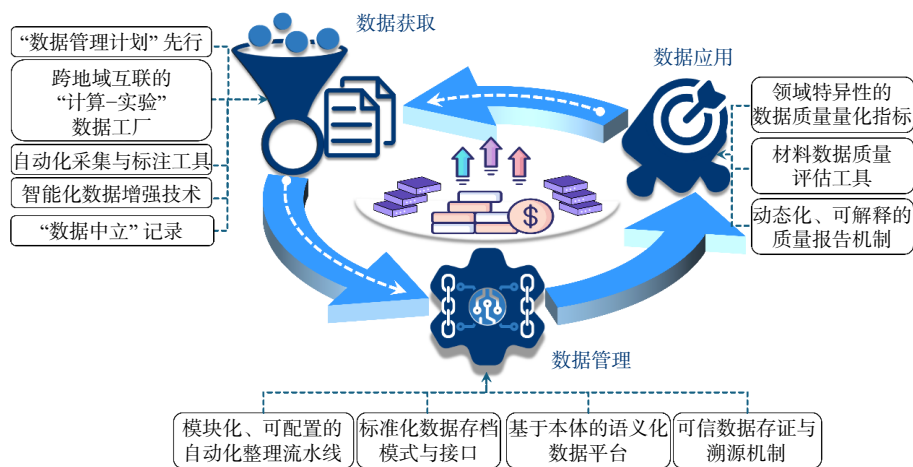


图1 面向材料数据全生命周期的质量管控框架

据获取阶段存在的分布失衡、记录失范与采集失真风险,建议实施以下措施。

1) 推广“数据管理计划”先行机制:在项目启动时,即要求制定详细的数据管理计划,明确数据格式、元数据标准、存储方案和共享政策,从源头对数据质量进行统筹。

2) 构建跨地域互联的“计算-实验”数据工厂:通过标准化接口与统一调度系统,将计算模拟与自动化实验平台通过智能体(Agent)技术整合为闭环流水线,以“标准化”的机器流程替代人工干预,实现高通量、规范化的数据生产,从根源上缓解数据分布不均问题,同时提升数据产出的可靠性。

3) 部署自动化采集与标注工具:开发并集成面向实验设备、计算软件和科研流程的数据采集工具,实现原始数据、全过程参数及标准化元数据的自动捕获。针对非结构化文本抽取,进一步结合检索增强生成(retrieval-augmented generation, RAG)、领域适应的LLM(如 MatSciBERT^[41])以及科研智能体技术,使数据采集、语义补全、质量检测 and 标准映射由传统人工驱动模式逐步转向智能辅助模式,并结合物理一致性校验以提升鲁棒性,最大程度减少人工干预带来的错误与模型幻觉风险。

4) 发展智能化数据增强技术:针对数据稀缺领域,特别警惕通过生成式模型(如生成对抗网络、扩散模型)产生违反物理规律的“幻觉样本”而造成的材料数据污染。数据生成与增强过程必须嵌入物理约束机制,如将热力学稳

定性、晶体对称性等作为生成条件的硬性约束,或利用物理信息神经网络(physics-informed neural networks, PINN)等架构,将控制方程直接融入损失函数,确保合成数据在补充真实数据集的同时,在统计特征与物理规律层面均具科学合理。

5) 倡导并制度化“数据中立”记录原则:在实验室信息管理系统和计算平台中,通过强制或激励机制,推动科研人员记录所有实验/计算尝试(包括失败案例),并将这类数据作为有价值的数字资产进行规范化管理。此外,可以针对“中立”数字资产建立贡献评价体系,例如成立专门的数据协会收录规范化、系统化的高质量数据报告,设置量化指标作为数据生产人员的贡献评价依据和激励途径。

3.2 数据管理阶段:以“智能治理”实现质量赋能

要实现材料数据管理阶段从“风险放大”向“质量赋能”的范式转变,亟需通过“智能治理”路径,将原始混杂数据精炼为高可信、可复用的优质资产。为应对当前格式与语义双重异构、关联浅层化的结构性缺陷,推荐以下实施策略。

1) 构建模块化的自动化整理流水线:设计涵盖数据探查、规则清洗、格式转换、一致性验证及关联关系自动构建的标准化处理流程。该流水线应支持领域规则与ML模型相结合的混合质检模式。其中,领域知识可以形式化载体嵌入流程,例如将物相命名规范、元素价态规则等转换为可执行的符号规则库,或封装为预训练的特征

提取模块,实现对“合成条件-结构-性能”关联的有效约束。同时,该流水线应允许根据具体材料体系灵活配置处理模块,实现高效、可复用的质量提升效果。

2) 推行标准化数据存档模式与接口:在社区内推广统一或可互操作的数据存档模式(如 JSON-LD),并提供标准的应用程序编程接口(application programming interface, API),降低数据汇交与集成的技术门槛,促进跨平台、跨团队数据的无缝流动与质量共同治理。

3) 构建基于本体的语义化数据平台:采用领域本体(如 Materials Design Ontology)作为上层语义模型,将异构数据映射到统一的概念框架下,从根本上解决语义异质性,同时赋予数据明确的科学内涵。

4) 实施基于智能合约的可信数据存证与溯源机制:在保障基础安全(加密、备份、权限控制)的前提下,针对原创实验结果、基准数据集等关键数据,引入区块链等防篡改技术,确保数据的完整性、保密性与可用性,为数据质量的可信度与权威性提供底层保障,赋能数据的可靠流通与复用。

3.3 数据应用阶段:以“量化评估”保障质量兑现

要在数据应用中实现质量闭环,必须在数据使用前建立系统化的量化评估机制,通过客观、多维度的质量检测,保障数据价值的可靠兑现。这需从指标定义、工具开发到报告生成的全流程建设,具体策略如下。

1) 开发领域特异性的数据质

量化指标:围绕材料数据的真实性、逻辑自洽性与时效性等核心维度,构建可计算的质量指标体系。除通用指标外,既需涵盖材料特有的关键属性,如相图一致性、结构-性能关联合理性、合成路径的可重复性等,也应将可发现、可访问、可互操作、可重用(findable, accessible, interoperable and reusable, FAIR)数据原则^[42]中的可查找性、可访问性等元数据质量维度纳入指标体系,为数据质量评估提供科学、细粒度且兼具技术与管理的度量基准。需要简要说明的是,FAIR原则与材料特有指标之间存在交叉验证关系:一方面,FAIR原则中的可查找性与可访问性确保材料特有指标(如相图一致性、结构-性能关联合理性)所依赖的原始数据及元数据易于定位、获取与复用,从而支持指标的可重复计算与跨研究对比;另一方面,材料特有指标的验证结果(如合成路径可重复性检验失败)可反向定位元数据中可能存在的缺失、不准确或不完整问题,进而驱动FAIR原则的持续改进。这种双向互验机制既增强了材料数据质量评估的科学性与可操作性,也为通用数据管理规范在材料学科的具体落地提供了领域适配的检验手段。

2) 开发材料数据质量评估、多参数矩阵分析及AI比对工具:开发集成规则引擎与ML模型的质量评估系统,实现数据完整性、一致性、异常值、特征覆盖度及分布均衡性等指标的全自动计算与可视化诊断。领域知识的形式化表达可通过将相图热力学稳定性

判据、元素周期律约束等编码为逻辑规则,集成至规则引擎,或将材料科学中成熟的描述符(如晶体场稳定能、配位数等)作为先验特征融入ML模型,提升异常识别的物理可解释性。同时,开发数据多参数矩阵分析软件,支持多维质量参数的交叉关联与模式挖掘,并构建基于AI的数据系统比对软件,实现不同来源、不同批次材料数据的一致性对比与差异诊断。工具需支持批量处理与交互式探查功能,快速定位质量薄弱环节,并生成结构化质量概要,为后续治理提供明确指引。

3) 推广动态化、可解释的质量报告机制:建立与数据发布同步的标准质量报告模板,不仅提供整体质量评分与多维质量画像(如完整性热图、领域规则违例清单),还需明确标注数据局限性、潜在应用场景与改进建议。报告应遵循动态更新原则,随数据集的迭代演进同步优化,既帮助用户快速预判数据适用性,也为深度数据遴选与针对性质量提升提供结构化依据。

需要特别指出的是,领域知识在材料数据质量管控中的嵌入并非单一维度的技术操作,而是一个贯穿数据全生命周期的分层耦合过程。依据知识发挥作用的机制与位置,可将其解构为3个核心层级。

1) 数据层的“特征化嵌入”:将物理和化学原理(如元素周期律、晶体场理论)转化为可计算的描述符或衍生特征,作为额外维度注入原始数据集。其核心目的是丰富数据的物理语义密度,使模型

无需从零归纳基础规律,从而降低对样本规模的依赖并加速收敛。例如,在智能化数据增强中,热力学稳定性判据即作为生成条件的硬性约束,确保合成数据在统计与物理层面的双重合理性。

2) 算法层的“结构化嵌入”:将守恒律、对称性等先验知识编码为模型的归纳偏置(Inductive Bias),直接融入网络架构或损失函数。其核心作用是约束模型的假设空间,强制其在符合物理规律的解域内搜索最优解,有助于抑制模型幻觉并提升外推可靠性。例如,PINN将控制方程嵌入损失函数,即是此层级的典型实践。

3) 输出层的“规则化嵌入”:将相图一致性、电荷中性等确定性规则形式化为可执行的符号逻辑库,置于模型推理后端作为校验与过滤机制。其核心价值在于提供结果可信度的硬性边界,对违反基本科学原理的预测进行拦截或修正,保障数据应用阶段的质量闭环。例如,质量评估工具中的规则引擎,正是这一层级的具体落地。

4 高质量材料数据生态的长远愿景

高质量数据是驱动材料科学研究从经验试错迈向数据智能的核心要素,其价值在于构建可积累、可复用、可演进的数字基座。构建贯穿全生命周期的综合性质量管控框架,有助于系统性地破除“劣质数据陷阱”,将分散、异构、低可信的原始材料数据,精炼为语义清晰、关联丰富、质量可度量、应用可追溯的战略性数据资产。

同时,未来材料数据生态不应停留于静态数据库建设,而应向动态演化的数据生态(Living Materials Data Ecosystem)发展。在该模式下,数据、知识和模型形成持续反馈闭环;模型应用过程中暴露的数据缺陷能够反向驱动数据补充与优化,新实验和新计算结果能够持续丰富数据资产,使材料数据体系具备自我完善和持续演进能力。

需要进一步注意的是,动态材料数据生态中的反馈闭环并不必然意味着自我纠错。随着 AI 逐步进入实验选择、信号重建和证据组织等数据生产环节,模型先验、主动采样策略及生成算法开始反向影响后续数据获取与知识形成过程。当不同平台共享相近的基础模型、知识库和分析流程时,局部偏差可能被持续放大并跨平台传播,而学习型信号重建和生成数据的广泛应用,也进一步模糊了原始观测与算法推断之间的边界。由此,材料数据质量风险将由单点误差演化为沿“选择—观测—解释—更新”闭环持续累积的系统性偏差。因此,未来动态材料数据生态不仅需要具备持续演进能力,更应具备主动识别、诊断并纠正系统性偏差的能力。

为适应上述变化,材料数据质量管控的基本对象应由孤立的数据记录进一步拓展为可追溯的证据事件。除样品信息、实验条件和测量结果外,证据事件还应同步记录候选空间、实验选择策略、模型与算法版本、探测器原始信号及其重建过程、主动干预关系以及独立验证路径,并将多个证据事件连接为覆盖问题设定、实验选

择、仪器执行、算法观测到模型更新全过程的认知行动链。其中,全生命周期管理回答“数据如何被获取、管理和应用”,认知行动链则进一步回答“该数据为何产生、观测如何形成以及结论由何种独立证据支持”,二者共同构成面向 AI 时代的数据质量治理框架。在此基础上,未来材料数据基座的核心价值将不仅体现在数据汇聚和复用上,更在于知识形成过程的可追溯、观测过程的可重放、科学主张的可证性和科研系统的持续纠错能力。

上述质量治理框架的有效实施,最终仍需依托统一、开放且持续演进的数据标准体系。作为确保全球材料数据一致性、准确性、可交换性与可互操作性的结构化规范框架,数据标准不仅能够保障数据生产过程中关键信息的完整捕获,更通过统一的数据模型与接口规范,确保跨地域、跨机构数据的互操作与可信共享能力。为推动材料数据标准的可持续发展与有效实施,国内外先后组建了跨学科标准委员会,包括美国国家标准与技术研究院(National Institute of Standards and Technology, NIST)、欧洲标准化委员会(European Committee for Standardization, CEN)、中国材料与试验团体标准委员会(Chinese Society for Testing & Materials, CSTM)。其中, CSTM 于 2019 年发布全球首个材料基因工程通则标准,并陆续推出高通量计算数据、高通量实验数据、材料数据标识等一系列配套标准,为材料数据标准化体系的系统化建设提供了参考方案。但任何旨在成

为标准并付诸实际应用的新规范,必须解决“激励”这一关键性问题。只有当材料数据工作者切实认识并体会到数据标准对研究效能的提升价值,甚至为研究进程带来可量化的加速效应,同时还能获取可量化、可流通的贡献评价指标为职业发展赋能时,才能真正激发其对应用数据标准的重视,逐步改善既有数据记录缺陷,并着力将其宝贵的科研数据进行重组,形成规范化、结构化的标准数据。通常而言,这种认知重构需要经历“工具赋能—价值显化—行为转化”的三阶段跃迁过程:首先通过智能化数据采集、记录工具,降低标准遵循门槛(如智能体驱动的实验记录助手),继而建立数据质量与科研产出的因果关联模型(如引用标准化数据集对论文影响力的增益分析),最终形成“标准遵循—效率提升—成果涌现”的正向反馈闭环。在此逻辑框架下,开发用户友好的数据基础设施成为关键突破方向,其不仅能持续促进高质量材料数据的产生,还可以收获科研人员对数据标准应用的实践反馈,推动数据标准的迭代更新,可持续地满足数据用户需求。

当前,在工具赋能层面,一系列智能化工具将数据标准转化为可执行流程,有效驱动多源异构数据持续积累转化为高质量资产。其中,刘悦等^[43]提出一种材料科学文本挖掘高质量数据集的构建方法,通过可溯源的文献自动获取、下游任务驱动的预处理、基于材料四面体准则的标签注释,以及融合领域知识的有条件数据增强模型(cDA-DK),有效实现非结构化

科学文献向高质量结构化数据集的转化。该方法不仅为材料数据标准化体系提供了高质量的数据供给,更为数据标准的实际落地与迭代优化奠定了坚实基础。同样, Roy 等^[44]开发的多智能体框架 Com-ProScanner, 能够从科学文献中精准提取材料化学组分与性质等关键数据, 并将其转化为遵循标准化格式、可互操作的结构化数据资产, 直接服务于材料数据标准化体系的建设与完善。此外, 为便于材料数据的标准化查询, OPTIMADE 联盟已通过 OPTIMADE API^[45-46] 定义了跨数据库统一查询规范, 并已被 Materials Project、AFLOW 等主流材料数据库采纳。而在国内, 国家新材料大数据中心正统筹推进材料数据的汇聚、共享与标准化应用, 并积极引导广大材料研发人员参与标准化材料数据的反馈与迭代过程。

从标准制定流程角度看, 应遵循“受控词表—领域本体—标准注册”的规范流程。以领域专业术语为基础单元, 整合同义词簇、厘清层级关系, 凝练为体系化的受控词表, 并最终将其注册存档至社区标准库中, 是数据标准赖以确立的核心前提。何冰等^[47]开发了面向材料数据与 workflow 全生命周期管理的 IMat-DWM 语义工具包, 以材料领域受控词汇表为基础, 规范术语表达, 并进一步构建领域本体以描述材料概念及其层级关系, 实现数据、工具和工作流等资源的统一语义建模, 为材料资源的标准化管理和智能协同提供了支撑。需特别指出的是, 领域数据标准的构建不应一味地追求绝对统一性

与一致性。由于材料领域内研究对象与学科文化存在显著差异, 强行进行数据统一易引发数据标准兼容性不足、冲突与冗余等问题。因此, 数据标准构建建议遵循“核心约束+领域扩展”的设计原则。对于基础性材料数据, 其核心要素, 包括关键数据字段定义、测量协议、质量阈值等, 应当优先在顶层确立, 从根源上杜绝跨实验室、跨研究项目间数据记录的差异, 增强数据集可比性与可整合性。例如, 材料的元素成分、原子排布、电子结构及其本征性质等信息是材料性能的基础性决定因素, 应当予以重点化标准保障, 确保基础信息表征与解读的一致性。在领域扩展层面, 各材料子学科可基于研究对象的特性、物理化学属性、合成路径参数及服役性能指标, 自定义、构建并拓展具备领域适配性的数据标准拓展模块。这种弹性的标准设计与构建框架既有利于维护学科核心知识的完整性, 又避免了过度标准化对领域知识表达的束缚, 从而提升现有标准化高价值数据资产的复用潜力。

构建高质量材料数据衍生资产库, 实现高质量数据长效、可持续、可衍生、资产化运营。当前形势下, 高质量数据是驱动 AI 在材料上实现创新的核心动力。围绕打造材料数据生态, 将建设高质量材料数据衍生资产库作为数据增值的核心支撑。该理念区别于传统存储系统, 以全生命周期质量管控、标准化、可信溯源、数据资产化和安全流通为建设核心。科研人员可依托原始数据, 在数据资产库中获取数据激励和成果增益。

其中衍生的数据既包含平台技术加工形成的增值数据, 也包括运用原始数据开展拓展研究得到的新数据及平台运营过程中的数据交易, 以此形成长期正向循环。将合规核验后的材料数据打造为可管控、可共享、可增值的数字资产, 系统性解决数据治理现存难题。

最后, 建成高质量材料数据生态不仅是材料科学自身发展的内在需求, 更是国家战略层面的重要布局。这一目标的实现, 不仅依赖技术平台的支撑, 更需要科研机构、企业、数据平台等多元主体的协同参与和深度协作。在国家持续推进高质量数据集建设的制度背景下, 企业作为材料数据的核心生产者与应用主体, 既是高质量数据生态的受益者, 更是其建设者与维护者, 理应积极参与数据治理、标准制定与生态共建, 推动高价值材料数据的持续积累与价值有序释放。作为材料智能研发的核心基础设施, 成熟的数据生态将有效支撑关键材料的快速自主研发, 保障产业链安全稳定, 并持续释放材料数据的资产价值。随着产业竞争从技术竞争逐步向生态竞争演进, 一个开放、可信、可持续的数据生态, 不仅能够构建国内产学研用协同的创新闭环, 更将吸引和汇聚全球智力与算力资源, 推动材料学科从传统实验科学升级为深度融合计算、数据与智能的“第四范式”科学, 甚至开启人机协同、自主推理与智能决策的全新科研模式——AI for Research(AI4R)^[8], 为高端制造、新能源、信息技术等战略领域的跨越式发展提供源源不断的创新引擎。

参考文献(References)

- [1] Sanchez-Lengeling B, Aspuru-Guzik A. Inverse molecular design using machine learning: Generative models for matter engineering[J]. *Science*, 2018, 361(6400): 360–365.
- [2] Vasudevan R K, Choudhary K, Mehta A, et al. Materials science in the artificial intelligence age: High-throughput library generation, machine learning, and a pathway from correlations to the underpinning physics[J]. *MRS Communications*, 2019, 9(3): 821–838.
- [3] Kimmig J, Zechel S, Schubert U S. Digital transformation in materials science: A paradigm change in material's development[J]. *Advanced Materials*, 2021, 33(8): 2004940.
- [4] Dinic F, Singh K, Dong T, et al. Applied machine learning for developing next-generation functional materials[J]. *Advanced Functional Materials*, 2021, 31(51): 2104195.
- [5] Duan C R, Nandy A, Kulik H J. Machine learning for the discovery, design, and engineering of materials[J]. *Annual Review of Chemical and Biomolecular Engineering*, 2022, 13: 405–429.
- [6] Fang J H, Xie M, He X Q, et al. Machine learning accelerates the materials discovery[J]. *Materials Today Communications*, 2022, 33: 104900.
- [7] Shang Y, Xiong Z Y, An K, et al. Materials genome engineering accelerates the research and development of organic and perovskite photovoltaics[J]. *Materials Genome Engineering Advances*, 2024, 2(1): e28.
- [8] 李国杰. 智能化科研(AI4R): 第五科研范式[J]. *中国科学院院刊*, 2024, 39(1): 1–9.
- [9] Jensen Z, Kim E, Kwon S, et al. A machine learning approach to zeolite synthesis enabled by automatic literature data extraction[J]. *ACS Central Science*, 2019, 5(5): 892–899.
- [10] Ward L, Wolverton C. Atomistic calculations and materials informatics: A review[J]. *Current Opinion in Solid State and Materials Science*, 2017, 21(3): 167–176.
- [11] Liu Y L, Niu C, Wang Z, et al. Machine learning in materials genome initiative: A review[J]. *Journal of Materials Science & Technology*, 2020, 57: 113–122.
- [12] Song Z L, Chen X W, Meng F B, et al. Machine learning in materials design: Algorithm and application[J]. *Chinese Physics B*, 2020, 29(11): 116103.
- [13] Takahashi K, Takahashi L. Toward the golden age of materials informatics: Perspective and opportunities[J]. *The Journal of Physical Chemistry Letters*, 2023, 14(20): 4726–4733.
- [14] Sivan D, Satheesh Kumar K, Abdullah A, et al. Advances in materials informatics: A review[J]. *Journal of Materials Science*, 2024, 59(7): 2602–2643.
- [15] Wang R Y, Strong D M. Beyond accuracy: What data quality means to data consumers[J]. *Journal of Management Information Systems*, 1996, 12(4): 5–33.
- [16] English L P. Improving data warehouse and business information quality: Methods for reducing costs and increasing profits[M]. New York: Wiley, 1999.
- [17] Jarke M, Jeusfeld M A, Quix C, et al. Architecture and quality in data warehouses: An extended repository approach[J]. *Information Systems*, 1999, 24(3): 229–253.
- [18] Chen H, Hailey D, Wang N, et al. A review of data quality assessment methods for public health information systems[J]. *International Journal of Environmental Research and Public Health*, 2014, 11(5): 5170–5207.
- [19] Ekegren C L, Gabbe B J, Finch C F. Sports injury surveillance systems: A review of methods and data quality[J]. *Sports Medicine*, 2016, 46(1): 49–65.
- [20] 刘悦, 马舒畅, 杨正伟, 等. 面向材料领域机器学习的数据质量治理[J]. *硅酸盐学报*, 2023, 51(2): 427–437.
- [21] Lu Y C, Wang H, Zhang L T, et al. Unleashing the power of AI in science—key considerations for materials data preparation[J]. *Scientific Data*, 2024, 11: 1039.
- [22] Liu Y, Yang Z W, Zou X X, et al. A general framework to govern machine learning oriented materials data quality[J]. *Materials Science and Engineering: R: Reports*, 2025, 166: 101050.
- [23] Liu Y, Ma S C, Yang Z W, et al. Domain knowledge-assisted materials data anomaly detection towards constructing high-performance machine learning models[J]. *Journal of Materiomics*, 2025, 11(6): 101066.
- [24] He B, Wang K X, Jiang Z, et al. Towards a sustainable high-quality materials data ecosystem: Frameworks and strategies[J]. *National Science Review*, 2026, 13(7): nwag108.
- [25] 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): 053208.
- [26] Himanen L, Geurts A, Foster A S, et al. Data-driven materials science: Status, challenges, and perspectives[J]. *Advanced Science*, 2020, 7(2): 1903667.
- [27] 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: 011002.
- [28] Oses C, Esters M, Hicks D, et al. Aflow++: A C++framework for autonomous materials design[J]. *Computational Materials Science*, 2023, 217: 111889.
- [29] Esters M, Oses C, Divilov S, et al. Aflow.org: A web ecosystem of databases, software and tools[J]. *Computational Materials Science*, 2023, 216: 111808.
- [30] Zakutayev A, Wunder N, Schwarting M, et al. An open experimental database for exploring inorganic materials[J]. *Scientific Data*, 2018, 5: 180053.
- [31] Kumar M, Yar Khan S, Liu S L, et al. Scientific mapping and data analysis of the research landscape in perovskite solar cell technology[J]. *Solar Energy*, 2024, 273: 112509.
- [32] Raccuglia P, Elbert K C, Adler P D F, et al. Machine-learning-assisted materials discovery using failed experiments[J]. *Nature*, 2016, 533(7601): 73–76.
- [33] Moosavi S M, Chidambaram A, Talirz L, et al. Capturing chemical intuition in synthesis of metal-organic frameworks[J]. *Nature Communications*, 2019, 10: 539.
- [34] Taniike T, Takahashi K. The value of negative results in data-driven catalysis research[J]. *Nature Catalysis*, 2023, 6(2): 108–111.
- [35] Ghiringhelli L M, Carbogno C, Levchenko S, et al. Towards efficient data exchange and sharing for big-data driven materials science: Metadata and data formats[J]. *npj Computational Materials*, 2017, 3: 46.
- [36] Voigt S P, Kalidindi S R. Materials graph ontology[J]. *Materials Letters*, 2021, 295: 129836.
- [37] Bayerlein B, Schilling M, Birkholz H, et al. PMD Core Ontology: Achieving semantic interoperability in materials science[J]. *Materials & Design*, 2024, 237: 112603.
- [38] Mrdjénovich D, Horton M K, Montoya J H, et al. Propnet: A knowledge graph for materials science[J]. *Matter*, 2020, 2(2): 464–480.
- [39] Statt M J, Rohr B A, Guevarra D, et al. The materials experiment knowledge graph[J]. *Digital Discovery*, 2023, 2(4): 909–914.
- [40] Venugopal V, Olivetti E. MatKG: An autonomously generated knowledge graph in Material Science[J]. *Scientific Data*, 2024, 11: 217.
- [41] Gupta T, Zaki M, Anoop Krishnan N M, et al. MatSciBERT: A materials domain language model for text mining and information extraction[J]. *npj Computational Materials*, 2022, 8: 102.
- [42] Wilkinson M D, Dumontier M, Aalbersberg I J J, et al. The FAIR Guiding Principles for scientific data management and stewardship[J]. *Scientific Data*, 2016, 3: 160018.
- [43] 刘悦, 刘大晖, 葛献远, 等. 高质量的材料科学文本挖掘数据集构建方法[J]. *物理学报*, 2023, 72(7): 070701.
- [44] Roy A, Grisan E, Buckeridge J, et al. ComProScanner: A multi-agent based framework for composition-property structured data extraction from scientific literature[J]. *Digital Discovery*, 2026, 5(4): 1794–1808.
- [45] Andersen C W, Armiento R, Blokhin E, et al. OPTIMADE, an API for exchanging materials data[J]. *Scientific Data*, 2021, 8: 217.
- [46] Evans M L, Bergsma J, Merkys A, et al. Developments and applications of the OPTIMADE API for materials discovery, design, and data exchange[EB/OL]. (2024-02-01). <https://arxiv.org/abs/2402.00572>.
- [47] He B, Wang Q B, Jiang Z, et al. A lifecycle-centered semantic toolkit for provenance-aware materials data and workflows management[J]. *Journal of Energy Chemistry*, 2026, 121: 584–599.

Reshaping the data foundation for AI for Materials: Full lifecycle data quality control framework and trusted ecosystem construction

SHI Siqi^{1,2}

1. State Key Laboratory of Materials for Advanced Nuclear Energy, School of Materials Science and Engineering, Shanghai University, Shanghai 200444, China
2. Materials Genome Institute, Shanghai University, Shanghai 200444, China

Abstract High-quality data acts as a pivotal driver for the intelligent transformation of research paradigms in materials science. However, the structural deficiencies in the current materials data production system not only lead to quality issues such as data fragmentation, siloing, format heterogeneity and incomplete records, but also result in an excessive transfer of quality control responsibilities to downstream users due to weak producer accountability and the absence of a systematic governance framework. Although users attempt to enhance data governance effectiveness by integrating domain knowledge into various corresponding methods, their limited insight into data production mechanisms hinders the generalizability of their governance strategies and the reliability of results. To this end, this paper focuses on the sustainable generation of high-quality materials data as the central theme, systematically examining the causes of data quality degradation across the data full lifecycle stages, and outlining a comprehensive materials data quality control framework. It aims to guide parties involved in data production and use to reevaluate the entire process from data acquisition and management to application, while providing referential insights for implementation pathways that promote efficient delivery of high-quality materials data, thereby laying foundational support for cultivating a materials data trusted ecosystem.

Keywords materials science; data quality control framework; trusted ecosystem; full lifecycle ●



(责任编辑 魏来)