



数据驱动研究范式下材料数据库的构建与应用

封俊杰, 周琨, 李木琛, 王新江*, 张立军*

吉林大学材料科学与工程学院, 集成光电子学国家重点实验室, 汽车材料教育部重点实验室, 物质模拟方法与软件教育部重点实验室, 长春 130012

* 联系人, E-mail: lijun_zhang@jlu.edu.cn; xinjiang_wang@jlu.edu.cn

2024-09-04 收稿, 2024-10-23 修回, 2024-11-04 接受, 2024-11-06 网络版发表

国家自然科学基金(62125402, 62321166653)资助

摘要 近年来, 随着大数据和人工智能技术的发展, 数据驱动的材料研发范式在材料构效关系挖掘和新材料设计与筛选等方面展现出显著优势. 数据作为此研究范式的基础, 在拓展材料设计空间和提升人工智能模型性能方面具有重要意义. 因此, 构建高质量材料数据库是数据驱动材料研发的重要一环. 自“材料基因组计划”启动以来, 理论数据库和实验数据库的数量和规模持续扩大, 数据基础设施和大数据技术也得到了充分发展. 本文围绕数据驱动研究范式下的材料数据库构建展开讨论, 首先从数据生成、数据预处理、数据存储和数据访问4个方面出发重点叙述了数据库构建的主要步骤, 然后基于数据体量、数据种类、访问方式和数据库特色等方面对国内外具有代表性的材料数据库进行总结, 最后从材料数据本身和数据基础设施建设两个视角出发, 对材料数据库未来的发展进行展望.

关键词 数据库构建, 新材料设计, 材料基因工程, 机器学习

材料科学是科技发展的先导, 是解决许多工业问题的关键, 新材料的发现对于推动人类社会具有重要作用. 传统的材料研发高度依赖于科学直觉和实验试错, 受限于理论基础模型和实验表征方法, 研究人员通常难以只通过实验手段准确描述材料复杂的构效关系, 挖掘深层次作用机理. 随着计算机算力的提高和计算算法的优化, 计算成本大幅降低, 密度泛函理论和分子动力学模拟逐渐在物理学、化学和材料科学中得到广泛应用, 推动材料研发范式由实验、理论驱动向计算驱动的转变^[1]. 实验和理论计算积累的大量数据使研究人员看到数据驱动材料研发的可能, “材料基因组计划”(Materials Genome Initiative, MGI)的提出, 更是推动大数据和人工智能技术在材料科学领域的全面应用, 加速材料研发第四范式的发展^[2].

数据驱动的第四研究范式通过建立目标属性与高维数据之间的关联有效弥补传统研究方法在研究效率和参数空间覆盖率上的不足, 已经为生物^[3,4]、医疗^[5]、工业^[6]等领域带来巨大突破. DeepMind推出的AlphaFold2^[3]成功预测98.5%的人类蛋白质结构, 并且将预测精度提升到原子级别; 华为研发的盘古气象大模型^[6], 其预测精度首次超过传统数值预报方法, 并将预测速度提高1万倍以上. 在材料领域, 材料基因工程将人工智能技术与材料科学融合, 推动自动化材料设计与性能预测. 哈佛大学研究团队基于海量实验数据训练的机器学习模型凭借其超过化学家的预测能力成功指导新材料合成, 代表数据驱动方法对传统材料研究模式的革新^[7]. 自此, 数据驱动方法在材料科学领域展现出越来越大的影响力. 微软研发的MatterGen^[8]能够通过

引用格式: 封俊杰, 周琨, 李木琛, 等. 数据驱动研究范式下材料数据库的构建与应用. 科学通报, 2025, 70: 4044–4065

Feng J, Zhou K, Li M, et al. Construction and application of materials database under data-driven research paradigm (in Chinese). Chin Sci Bull, 2025, 70: 4044–4065, doi: [10.1360/TB-2024-0946](https://doi.org/10.1360/TB-2024-0946)

物理化学约束直接生成稳定新材料,其生成结构稳定性与CDVAE相比提高了2.9倍;DeepMind开发的GNoME^[9]更是将材料研发效率提高到80%以上。同时,材料基因工程通过高通量筛选高效建立材料组分、工艺、结构、性能之间的关系,识别潜在候选材料。中国科学院物理研究所团队通过高通量实验研发出具有高强度和良好热塑成型性能的金属玻璃,颠覆了传统金属玻璃研究的实验试错研究模式^[10];在此研究方法基础上,该团队进一步采集了5700种金属玻璃实验数据,发现了同一体系中X射线衍射(X-ray diffraction, XRD)谱第一峰宽与玻璃形成能力的相关性,大幅提高了金属玻璃研发效率^[11]。目前,基于此研究范式,研究人员已经在光电^[12-14]、热电^[15,16]、催化^[17,18]、储能^[19,20]等领域设计并发现一大批新型材料,部分材料经实验验证有望应用于多个工业领域。

材料数据是实现材料研发第四范式的前提,主要表现为两方面:首先,Matrgen、GNoME等高性能模型均基于数以万计的高质量材料数据进行训练,体现了数据量和数据质量在数据驱动方法中的重要地位。其次,高通量计算也需要大量可靠结构作为初始输入,通过计算结果分析材料演化规律。因此,数据库建设则成为数据驱动材料研发过程中重要的一环。在MGI及相关计划的引领下,数据库作为其支柱之一实现了快速发展^[21]。近10年以来,实验数据库和理论数据库齐头并进,以ICSD^[22]、CSD^[23]、COD^[24]为代表的实验数据库数据规模持续增长,同时迅速涌现出大量理论数据库,例如美国的Materials Project^[25]、OQMD^[26-28]、AFLOW^[29]、欧洲的NOMAD^[30]等。依托材料基因工程,国内材料数据库及大数据技术也得到了充分发展,建立起包括Atomly^[31]在内的诸多理论/实验材料数据库,研发基于区块链的无模式数据存储技术^[32],发展出ALKEMIE^[33]、MatCloud^[34]、OCPMDM^[35]等智能数据平台。

本文以数据库构建为主线,综述了材料数据库构建的主要流程,简要总结了国内外常用实验数据库、理论数据库、实验-理论数据库及其使用情况,其中重点叙述了数据库构建中的数据生成、数据预处理、数据存储和数据访问4个部分,然后从数据体量、材料类型、数据种类、访问方式和数据库特色等方面介绍典型材料数据库,最后从材料数据和数据基础设施两个角度出发进行展望。我们希望通过材料数据库的系统性总结,为数据库构建以及数据库在数据驱动研究

范式中的应用提供启发和指导。

1 材料数据库构建

数据基础设施的发展大致历经3个阶段,从最初仅提供数据托管和查询服务的数据库,到提供基本材料数据和数据分析服务的数据中心,再到如今集成自动数据收集、数据存储、数据挖掘的一整套工作流程的材料设计平台,其发展历程的演变反映了材料研发信息化,信息积累自动化,多维度信息相互耦合的趋势^[36,37]。其中,MGI的启动加速了数据基础设施规模的扩大和功能的拓展,是数据驱动材料研发的决定性时刻。如今的材料设计平台相较于早期独立材料数据库和数据中心,除了满足科学数据的FAIR原则(即可发现、可访问、可交互、可重用)外,其数据生成更趋于自动化,数据访问更趋于多元化,数据表示更趋于标准化,同时力求满足大数据背景下数据库高并发、高可用、高扩展的要求。

无论材料设计平台还是数据中心,数据库都是两者的核心部分。本节将从4个方面阐述数据库构建的一般流程(图1)。(1)数据生成:通过高通量计算、高通量实验等方法批量获取理论数据和实验数据。在计算参数设置或者实验合成方法合理的情况下,能快速生成高质量数据。(2)数据预处理:对原始非结构化数据进行标准化,提升数据的可交互性和可重用性;删除异常数据,提高数据质量。(3)数据存储:合理构建数据模型,选择关系型数据库或者非关系型数据库存储数据,提高数据库的查询效率、灵活性和可扩展性。(4)数据检索:设计合理查询策略,精准定位到每一条数据;设置多种访问方式,适应新材料研发范式的要求,提升数据的可发现性和可访问性。

1.1 数据生成

1.1.1 高通量计算

高通量计算代表了一种强大且高效的数据生成技术,它基于量子力学,无需实验合成,在计算资源充足的前提下能保证在短时间内自动生成大量精确数据,这些数据涵盖了材料的热力学稳定性、电子结构等关键性质,为构建精准、全面的材料数据库提供了坚实基础^[38,39]。它包含3个基本要素:材料性质的可算性、计算任务并行化、任务管理自动化^[40]。材料性质可算性要求目标性质可以直接通过理论计算获得或者由可计算物理量直接决定,前者包括晶格参数、能带结构、

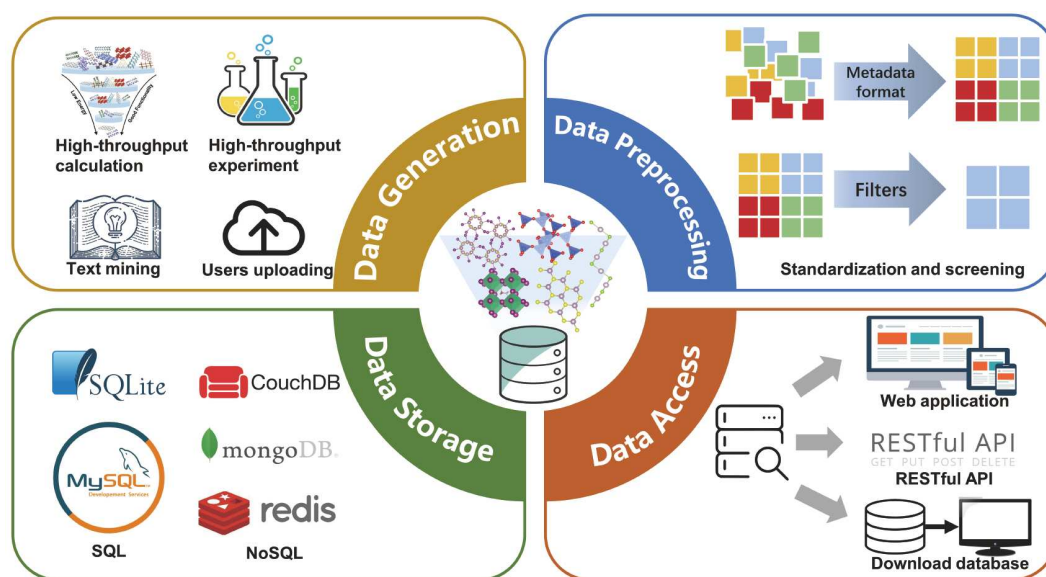


图1 (网络版彩色)数据库构建主要步骤: 数据生成、数据预处理、数据存储、数据访问

Figure 1 (Color online) Main steps of database construction: Data generation, data preprocessing, data storage, and data access

体系能量等可直接由密度泛函理论计算得到的物理量,后者如半导体载流子迁移率,可通过将计算获得的弹性常数、有效质量、形变势常数代入计算公式中获得.计算任务并行化和任务管理自动化分别通过减少单个计算任务计算时间和自动推进计算流程、追踪并恢复计算错误来提高计算效率,快速生成理论数据.对于高通量计算,泛函和参数的选择很大程度上决定了计算的精度和成本^[41],例如,在高通量计算前,应当选择合适的K点网格密度、截断能以平衡精度与成本;在电子性质计算中,局域密度近似和广义梯度近似表现优异但是存在低估带隙的情况^[42,43],此时可以通过引入诸如Heyd-Scuseria-Ernzerhof杂化泛函^[44]等更昂贵的泛函修正误差;在强关联体系计算中,可以引入DFT+U框架修正能量偏差.此外,高通量计算常以其生成的大型理论数据库而闻名,例如AFLOW、Materials Project、OQMD.高通量计算通常具有4个主要步骤:(1)初始结构获取;(2)自动化任务管理;(3)数据筛选;(4)数据存储.本小节将介绍前两部分(图2),数据筛选和数据存储部分将在后文详细叙述.

高通量计算的初始结构可以通过材料数据库获得^[45].诸如Materials Project、OQMD、ICSD之类的材料数据库会提供RESTful API,用户可通过RESTful API以编程的方式访问数据库获得大量结构数据,在过滤掉不合要求和重复结构后,其余部分可以作为高通量

计算的初始输入.通过结构工厂策略同样可以获得大量初始结构^[46].结构工厂指通过结构操作或元素替换构建复杂结构,此策略可以通过诸如插入分子、增删原子、晶格缩放等结构操作构建缺陷、异质结、超晶格等结构,也可以在满足价电子守恒的前提下对确定晶体结构进行单元素或双元素替换扩大搜索空间,目前已被广泛应用于材料高通量筛选.Li等人^[47]基于内含四面体结构的无机半导体和链状有机分子设计有机-无机杂化半导体异质结.Zhao等人^[48]通过将两个二价 Pb^{2+} 转化为一个一价 M^+ 离子和一个三价 M^{3+} 离子的策略设计无机无铅卤化物钙钛矿.材料信息学软件包JAMIP的StructureFactory模块能够满足结构工厂的需求^[46],其整合了两种结构操作算法,基本结构算法能够完成增删原子、晶格缩放、元素替换、平移旋转、构建超胞、构建真空层、晶格重定义等操作,复杂结构算法能够基于上述基本操作的灵活组合生成复杂结构.借助此模块,用户可以批量构建缺陷、异质结、超晶格、合金、界面、表面、孪晶界等结构.除数据库调用和结构工厂外,生成式模型通过学习大量输入结构,对其在潜在空间中的概率分布进行建模,能够生成目标结构作为高通量计算的初始输入^[49].在材料领域,变分自编码器、生成式对抗网络等生成式模型已被广泛用于材料逆向设计与晶体结构生成.Kim等人^[50]利用生成式对抗网络实现多孔材料逆向设计;Lyngby等

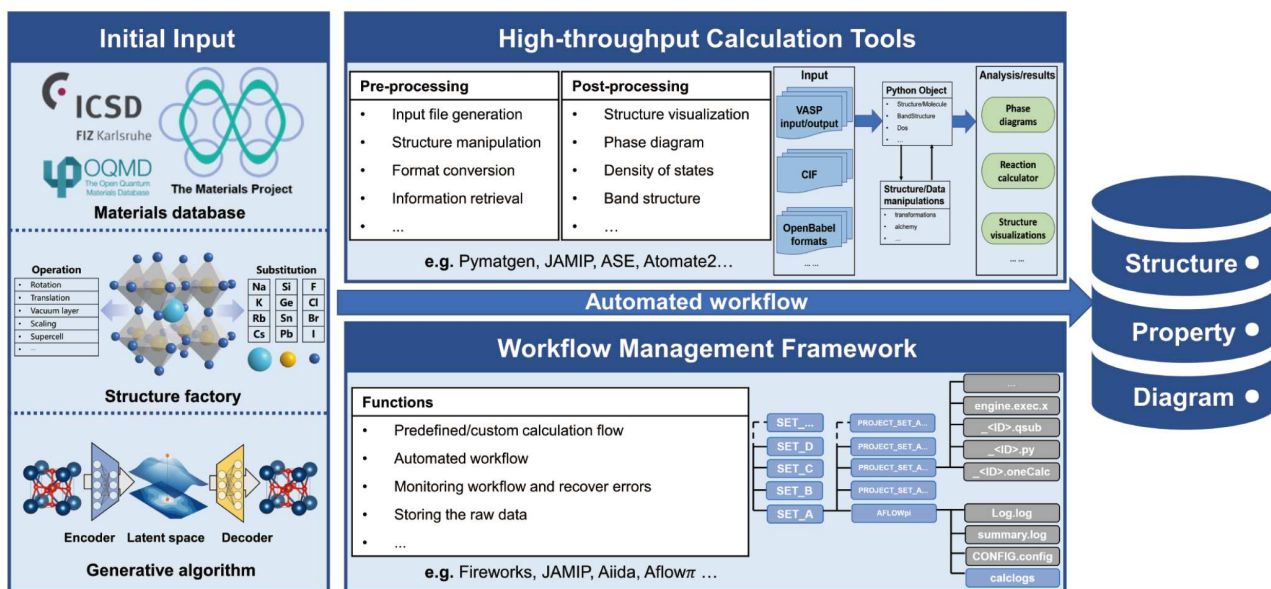


图2 (网络版彩色)高通量计算利用材料数据库、结构工厂、生成式模型提供的结构作为初始输入, 通过高通量框架自动管理任务, 获取理论数据以构建材料数据库

Figure 2 (Color online) High-throughput calculation utilizes structures provided from the materials database, structure factory, and generative models as initial inputs and automatically manages tasks through the high-throughput framework to obtain theoretical data for constructing materials databases

人^[51]通过训练CDVAE生成新型二维结构扩充二维材料数据库; Jacobsen等人^[52]通过生成式模型扩充一维材料数据库。

数据库调用、结构工厂和生成式模型为研究人员提供了大量初始结构的同时, 自动化数据处理和计算任务管理又带来了新的挑战^[53]。高通量计算工具能够批量处理高通量计算的输入输出文件, 同时 workflow 管理框架可以按照预设方案自动推进任务。MGI 启动以来涌现出了大量诸如 Pymatgen^[54]、FireWorks^[55]、JAMIP^[46]、Atomate2^[56]、Aflowr^[57] 等高通量软件包, 为数据处理和计算任务管理提供了极大便利。

Pymatgen 是一款强大的开源 Python 材料分析库, 能够处理包括 VASP、Abinit、Gaussian 在内的多种主流计算程序的输入或输出文件, 可以将晶体结构、分子结构等转换为 Python 对象用于进一步操作和分析, 还能实现序列化对象和 JSON 文件之间的转换, 便于数据的存储和调用。同时 Pymatgen 拥有强大的后处理功能, 融合了相图分析^[58]、电子结构分析及可视化、晶体结构可视化以及对称性分析等功能。此外, Pymatgen 集成了 Materials Project 的 RESTful API^[59], 允许用户通过超文本传输协议 (HyperText Transfer Protocol, HTTP) 直接访问数据库并返回 JSON 格式的数据文件, 便于用户批

量获取数据用于高通量计算和机器学习。

FireWorks 是一款开源 Python 高通量计算 workflow 管理软件, 由用于数据存储的 LaunchPad 和推进任务的 FireWorkers 两部分组成, 主要负责自动运行高通量计算流程, 允许在较长时间跨度内管理计算任务, 监控计算进程, 追踪并恢复计算进程中出现的错误防止任务中断。FireWorks 的动态工作流是由各 FireWork 和控制依赖组成的有向无环图, 其储存在 JSON 或 YAML 文件中, 支持用户基于通用任务设计计算流程。每一个 FireWork 由 FireTask 组成。FireTask 本质上是单个 Python 函数, 可以调用 shell 脚本进行文件的传输、读写、删除操作和调用其他 Python 函数, 每一个 FireTask 返回的 FWAction 对象可以存储数据、传递数据或根据过去输出动态地修改下一层任务。

1.1.2 高通量实验

高通量实验是一种允许同时执行大量实验的技术, 与传统实验方法相比, 每个实验所需工作量更少, 它起源于 20 世纪 50 年代生物学领域, 并快速传播到其他领域^[60]。在材料领域, 高通量实验是材料基因工程理念下促进新材料高效开发的重要一环, 它可以快速积累实验数据, 直接促进材料筛选和应用, 同时也可以作为理论计算的补充, 为机器学习提供更丰富的数据集进行

材料优化设计, 指导实验合成. 此方法采用组合材料合成技术, 进行批量合成和表征以快速建立成分、结构和性质之间的关系, 其整个过程主要由实验设计、样品平行制备、样品快速表征组成^[52](图3).

实验设计是指通过多次实验建立参数与结果之间的经验关系, 它通常结合数据分析方法, 旨在通过最少的实验次数获得最优参数^[61]. 在高通量实验初期, 通常需要基于物理化学直觉或先验知识建立参数与结果之间的大致趋势, 缩小参数空间, 基于这个缩小的参数空间进行后续参数搜索和实验设计. 参数搜索是一个高维优化问题, 可以通过诸如遗传算法之类的全局优化算法寻找近似最优值, 也可以通过机器学习构建参数与结果之间的可解析函数^[62]. 遗传算法旨在通过迭代改进给定问题的解决方案. 在此过程中, 算法通过适应度函数进行个体评估, 选中其中最优个体进行选择、交叉、变异操作以产生后代. 这些后代随后会被评估, 并重复上述流程. Corma等人^[63]成功应用遗传算法结合组合材料学方法对不同的氧化物组成进行筛选, 最终优化了两种催化体系: $\text{SO}_4^{2-}/\text{ZrO}_2$ 和 WO_x/ZrO_2 . 机器学习模型通过学习数据集中高维变量与目标属性之间的线性或者非线性关系, 拟合出一个可解析函数进而指导实验. Liu等人^[64]将贝叶斯优化用于优化钙钛矿太阳能电池合成工艺, 用此策略筛选出来的合成工艺制造的器件实现了最佳光电转换效率. Chen等人^[65]将机器学习算法用于分析实验数据并评估显著影响2D场效应晶体管电子性质的各种关键工艺参数, 优化制造器件的性能.

高通量实验要求实验人员在精细控制合成参数的前提下, 平行制备包含成分梯度的样品. 平行制备追求用尽可能少的工作量生成尽可能多的样品, 且材料样品必须是空间分辨的^[66]. 为了发展高通量实验, 各领域开发了不同平行合成技术. 目前应用最广泛的样品制备方法之一是薄膜气相沉积法, 它可用于制备金属或无机非金属成分梯度库. Löbel等人^[67]通过类似于共溅射的方法在Si晶片上交替沉积楔形Ti、Ni和Cu层制备Ti-Ni-Cu三元薄膜, 发现复合材料区域内存在相变, 揭示了不同的相变路径. Mao等人^[68]通过气相沉积合成 16×16 的Zn-Cd-S半导体材料阵列, 发现向Zn-Cd-S中加入In和Sn可以减小带隙, 提高载流子寿命. 此外, Koinuma等人^[69]开发的组合激光分子束外延方法可用于在样品中产生温度和成分梯度, 也可以在三维空间中控制沉积, 能够合成高掺杂薄膜和热平衡条件下无法获得的人造超晶格等独特材料.

高通量表征是高通量实验的基础, 要求高灵敏、空间分辨、定量、快速^[66]. 近年来随着科学技术的不断进步, 高通量表征手段也得到了长足的发展. 高通量X射线衍射是表征晶体结构的重要手段, 与传统XRD相比具有高亮度、表征速度快的特点. Mao^[68]通过毛细管聚焦机制将X射线束聚焦到直径约为70 μm 的单个材料上, 以二维自动扫描模式批量获取材料的成分和结构. 纳米压痕技术可以在纳米尺度上测得薄膜样品的弹性模量、硬度等力学性能, 被广泛应用于高通量表征. Zarnetta等人^[70]采用高通量纳米压痕技术对Ti-Ni-Cu薄膜进行力学性能表征, 结果表明当Ti含量偏离约

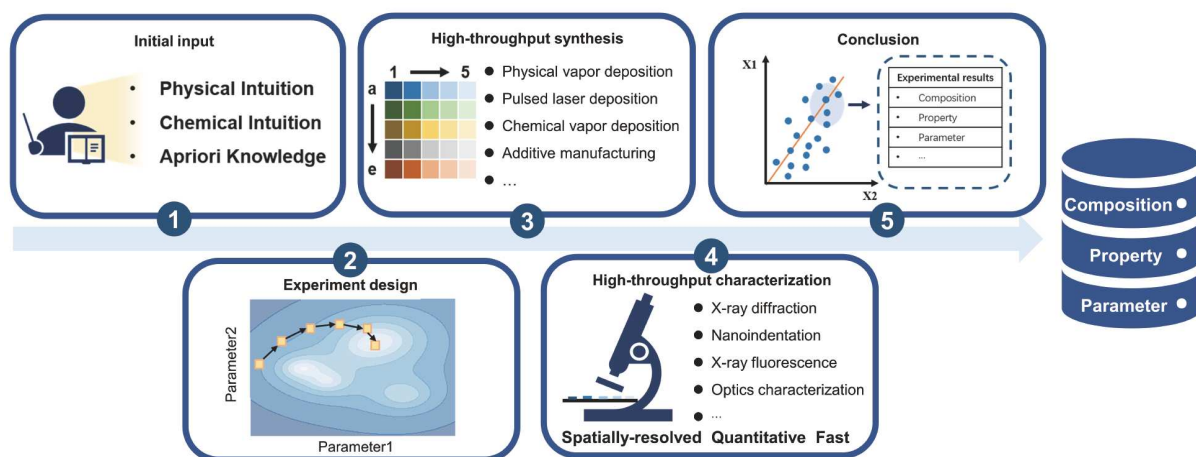


图3 (网络版彩色)高通量实验主要由实验设计、样品平行制备、样品快速表征组成, 可以同时执行大量实验, 快速积累实验数据

Figure 3 (Color online) High-throughput experiment primarily consists of experimental design, parallel sample synthesis, and rapid sample characterization. It enables the execution of a large number of experiments simultaneously, allowing for the rapid accumulation of experimental data

50 at.%时, 薄膜硬度和杨氏模量均有增加. 光学表征方法能够对样品进行无损平行表征, 且具有快速和空间分辨的特征. Sáfrán等人^[71]基于“单样品概念”, 提出了一种基于透射电子显微镜的组合表征方法, 该方法能够制备和表征梯度样品, 实现了二元薄膜相图的构建与分析.

1.1.3 其他方法

除高通量计算、高通量实验以外, 文本挖掘和用户上传也是获取数据的重要手段. 互联网的飞速发展促进全球信息共享与交流, 材料科学领域的科学文献数量也呈指数级增长. 作为科研成果的重要载体, 数量巨大的科学文献中蕴含了丰富的专业知识和数据, 但是这些数据大多以文本、表格之类的非结构化形式存在, 研究人员无法直接调用^[72,73]. 基于自然语言处理的文本挖掘提供了一套自动化信息提取手段, 通过监督、半监督或无监督方式训练的语言模型能够自动提取文献中的目标实体或数据, 生成材料数据库. 文本挖掘工作主要遵循图4所示流程: 文本获取、文本预处理、文本向量化、信息提取.

获取原始文本是文本挖掘的第一步. 大多数科学文献出版商以HTML、XML或JSON格式提供文本, 研

究人员可以基于出版商提供的API编写python脚本自动爬取, 也可以通过软件包实现这一功能. 例如, Cole课题组开发的化学感知自然语言处理工具包ChemDataExtractor^[74], 其中内置的网页爬虫模块支持从Elsevier、Springer、RSC等科学文献出版商网站自动获取文本. 在文本预处理阶段, 段落分割和标记化是关键步骤, 这一过程包括识别并分割句子中的词或短语以便将它们转化为token. 在此阶段, 需要文本规范化、词性标注加以辅助. 以上操作均可以在NLTK、SpaCy等通用自然语言处理软件包的辅助之下完成. 值得注意的是, 材料领域科学文献中包含的大量化学式和专业术语, 容易导致标记错误, 进而提升实体命名识别的难度. 利用化学领域专用的自然语言处理软件包, 例如ChemDataExtractor^[74]、ChemicalTagger^[75]、OSCAR4^[76]基于指定规则和内置词典进行标记化可以一定程度上减少标记错误, 但是解决问题的关键还是在于制定一个普适的化学术语标记规则. 在信息提取之前, 需要将文本映射到向量空间. 词袋模型和tf-idf模型是较为简单的文本向量化方式, 它们可以识别由关键词定义的文本主题, 但无法捕捉上下文语意, 常应用于文本分类和主题建模. 随着自然语言处理技术的发展,

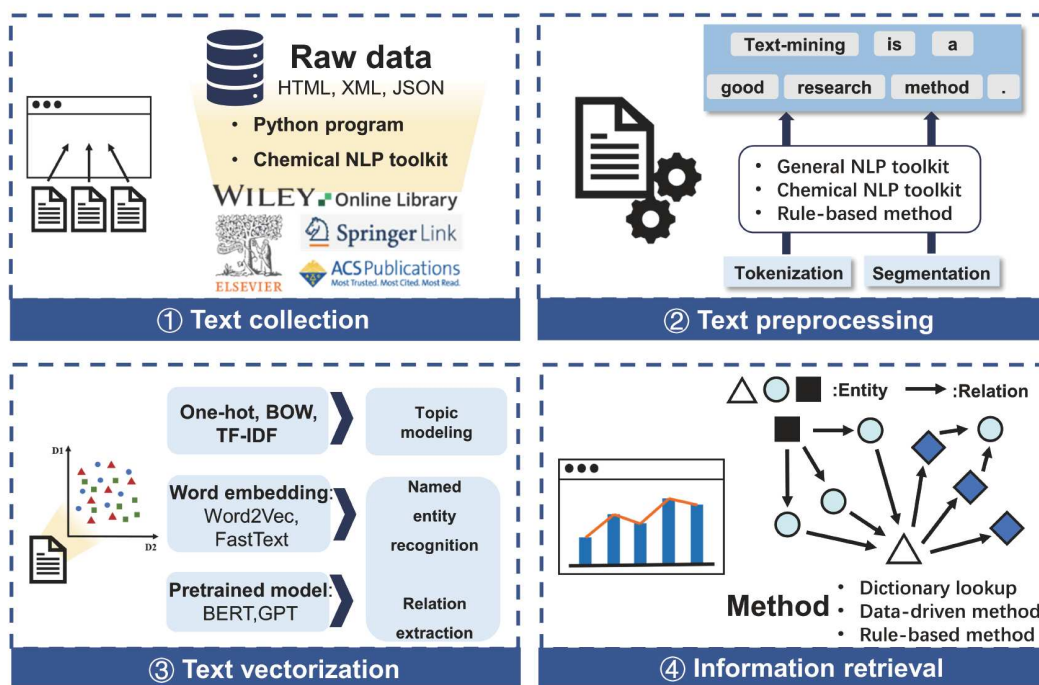


图4 (网络版彩色) 文本挖掘主要流程由文本获取、文本预处理、文本向量化、信息提取组成, 支持从科学文献中自动提取结构化数据
Figure 4 (Color online) The process of text mining consists of text collection, text preprocessing, text vectorization, and information retrieval, supporting the automatic extraction of structured data from scientific literature

词嵌入模型逐渐成为主流向量化方法。以Word2Vec、Fasttext为代表的词嵌入模型^[77,78]以无监督的方式进行训练,能够基于上下文语意捕捉单词之间的相似性和关系,输出的词向量被广泛用作命名实体识别等下游任务的输入。如今,Transformer模型^[79]通过引入自注意力机制成功捕捉文本序列的长程依赖关系,基于Transformer模型发展出的BERT模型^[80]和GPT模型^[81~83]已经引领自然语言处理进入预训练大模型时代,这些模型在处理各类下游任务时展现出前所未有的优势。信息提取是文本挖掘的核心任务,包括命名实体识别和关系提取等子任务,可以从科学文献中提取大量数据用于数据库构建。目前已经有大量工作将文本挖掘应用到材料领域构建材料数据库。剑桥大学Cole课题组利用ChemDataExtractor从文献中提取数据,构建了多个材料数据库,例如电池数据库^[84]、热电材料数据库^[85]、居里温度和尼尔温度数据库^[86]等。Kononova等人^[87]使用双向长短时神经网络结合条件随机场完成材料实体识别,使用Word2Vec词向量结合依赖性分析提取合成操作,从固态合成段落中提取出大量合成路线构建固态合成数据集。Huo等人^[88]利用半监督学习的方式从文献中提取大量无机材料合成信息构建数据库。

同时,部分数据库鼓励用户参与数据库建设,将上传的数据作为重要数据来源。上传数据时,用户需采用数据库平台指定元数据格式表示晶体结构,同时应提供技术细节以保证数据质量和可重用性。COD、CSD等实验数据库支持研究人员上传晶体结构,其中COD开发了一个自动化流程,允许用户上传已发表或待发表的原子坐标文件或者结构因子文件,每一个结构文件都会根据IUCr数据验证标准自动或手动检查语法和有效性。CSD也具有相似的上传流程,上传者必须在上传过程中修复所有可能出现的问题,而数据库平台也会提供关键元数据以供研究人员进一步检查、完善数据。包括NOMAD在内的一些理论数据库同样支持用户将合格的计算数据上传到网站,促进材料数据的共享与交换。

1.2 数据预处理

1.2.1 数据标准化

基于上述方法得到的原始数据通常是非结构化的,而非结构化数据既不能直接用于机器学习也无法进行精准检索,因此进行数据存储之前需要进行数据标准化。数据标准化指的是将原始数据转化为统一格式的

结构化数据,以实现更有效的数据管理和查询,其中统一格式,即元数据框架,就是完全表征、精确定位并最终再现已识别数据全部属性所必须的描述性信息,是数据标准化的关键^[37]。在开放共享过程中元数据框架必须符合数据FAIR原则。2016年,FORCE11对FAIR原则进一步细化,从中可以总结出通过元数据实现FAIR原则的具体要求^[89]。可检索性可以通过为数据分配唯一且永久的标识符、用丰富的元数据描述数据、在可检索资源中注册或索引元数据实现。标识符例如DOI、mp-ID、ICSD-ID等可以消除检索中产生的歧义,精确定位每一条数据;采用丰富的元数据描述数据的意义在于丰富检索手段,在缺少标识符的情况下用户仍可根据元数据进行数据筛选和检索;将元数据在可检索资源中注册保证数据资源或元数据是可发现的。可访问性要求元数据可以通过标准化通信协议规定的标识符来检索,此原则可以由支持查询或管理整个数据库的API实现。可交互性旨在建立数据与数据以及数据与文献之间的联系,此原则可以通过使用正式的、可获取的、共享的且广泛适用的元数据框架表示知识来实现,这些元数据格式需要适应储存数据的类型,并为每个结构提供丰富的描述手段以便精准定位,例如晶体学数据库常采用晶体学信息框架(Crystallographic Information Framework, CIF)格式^[90,91]、理论数据库NOMAD采用NOMAD Meta Info格式^[92,93]。可重用性是以上三原则的最终目的,为实现这一目标,开发人员应使用充分的属性描述数据和元数据,例如添加实验方案、计算方法、技术细节、关键参数等描述数据生成方法的元数据,同时还需包含详细的出处信息和数据使用协议,例如引用文献、数据源链接。

一般来说,开发人员会根据数据库关注的材料领域和性质,选择或者开发合适的元数据框架进行数据标准化(图5)。例如,ICSD、COD等晶体学数据库采用CIF作为标准元数据框架。CIF是化学和材料领域开发的第一个元数据框架,由数据标签和模块化的相关数据组成,其中每个数据标签与所储存数据一一对应。文件有严格定义的语法和指定的元数据项,主要包含如下几个方面的信息:(1) 文件来源、编辑和使用信息;(2) 结构编号及引用信息;(3) 结构信息,包括晶胞参数、空间群、对称位置等;(4) 原子参数信息,包括原子坐标、占位度、Wyckoff标记等;(5) 原子类型及价态。CIF文件支持以XML格式储存,可读性较高,研究人员和计算机都可以对文件进行编码,目前也被广泛应用

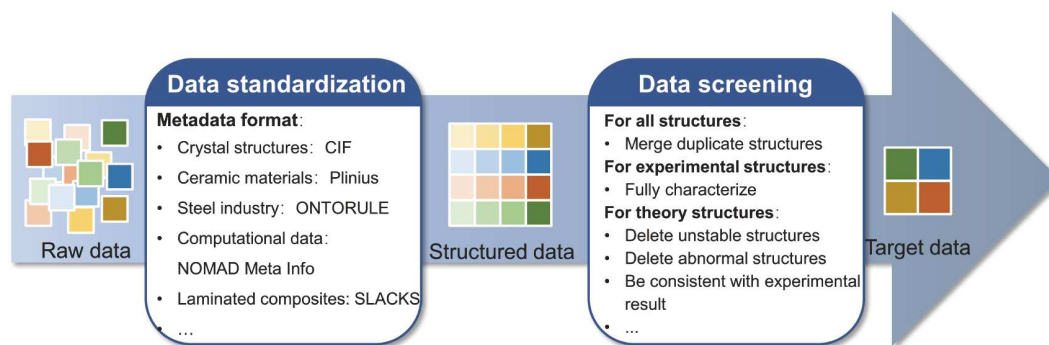


图5 (网络版彩色)数据预处理包括数据标准化和数据筛选两部分, 数据标准化包括选择恰当的元数据框架将非结构化数据转化为标准格式, 数据筛选包括通过校对结构、去除重复结构等操作提高数据质量, 两者均旨在提升数据质量和科学数据的FAIR属性

Figure 5 (Color online) Data preprocessing consists of two main components: Data standardization and data filtering. Data standardization involves selecting an appropriate metadata framework to transform unstructured data into a standard format, while data filtering includes operations such as checking structures and removing duplicates to enhance data quality. Both components aim to improve the quality of the data and enhance the FAIR attributes of scientific data

于其他领域如结构生物学和光谱学。计算材料数据库 NOMAD 采用 NOMAD Meta Info 作为标准元数据框架。NOMAD Meta Info 的数据存储模式高度层次化和模块化, 它将元数据组织成多个模块, 包括计算方法模块、模拟算法信息模块、结构信息模块、输出数据模块、 workflow 模块、计算参数模块等, 每个模块采用“键-值”的数据存储模式且为每个实例分配独特的统一资源定位器(Uniform Resource Locator, URL)。这种数据存储模式为元数据框架带来极大可扩展性, 尤其是在储存 workflow 时, 用户可以任意嵌套子 workflow 同时通过 URL 精准定位目标数据。物理信息文件(Physical Information File, PIF)^[94]是由 Citrination 平台开发的一种灵活的数据存储架构, 该架构通过信息分层全面描述材料、设备、物理系统的结构、制造方式及性能。区别于刚性存储模式, PIF 的顶层系统可以拥有任意数量的子系统, 通过嵌套这些子系统可以存储任意尺度、任意层次的信息, 灵活地调整数据结构。

1.2.2 数据筛选

为保证数据的准确性和可靠性, 在存入数据之前需要先进行数据筛选(图5)。在数据标准化后, 对于实验数据库和理论数据都应根据化学式、空间群和局域结构相似性等判据合并重复结构, 减少数据冗余。特别地, 对于实验数据库, 收录结构需要经过充分表征, 有确定的化学成分和原子坐标。对于理论数据库, 尽量保证收录与实验结果更接近的计算数据; 对于来自元素替换或者机器学习生成的理论结构, 需要在筛选过程中设置动力学稳定性判据和热力学稳定性判据, 删除

不稳定结构, 同时检查原子间距, 删除不合理结构。

1.3 数据存储

1.3.1 基于关系型数据库的数据存储

关系型数据库(图6)是指建立在关系模型上的数据库。关系模型可以理解为由行(记录)和列(属性)构成的二维表格, 表格之间通过主键和外键来保持数据的一致性和完整性, 这些表格及其相互之间的关系构成了整个关系型数据库。关系型数据库的设计需要遵循三大范式, 后一范式都是对前一范式的进一步规范。第一范式是指表格中每一字段的值均为不可拆分的最小数据单元, 要求存储的数据不可再分; 第二范式在第一范式的基础上进一步规定, 表格里的每一条记录都是唯一可表示的, 且所有非主键字段必须完全依赖于主键, 旨在保持数据的唯一性。第三范式要求所有非主键字段不能依赖于其他非主键字段, 保持非主键字段之间的独立性, 消除依赖传递。以上三大范式意在避免数据冗余, 保证数据一致性。当然, 在数据库构建中, 需要根据实际情况调整关系模型, 出于数据库的运行效率和使用友好度考虑, 允许降低范式标准, 保留冗余数据。

关系型数据库自1969年诞生伊始, 凭借其处理结构化数据的强大能力在各领域都有广泛应用。几乎所有的关系型数据库都可以用结构化查询语言(Structured Query Language, SQL)进行增、改、删、查的基本操作, 支持多个表格之间的复杂查询^[95], 关系型数据库采用二维表格进行数据存储, 其数据结构清晰, 整体架构易于理解; 通过外键约束大大减少了数据冗余和

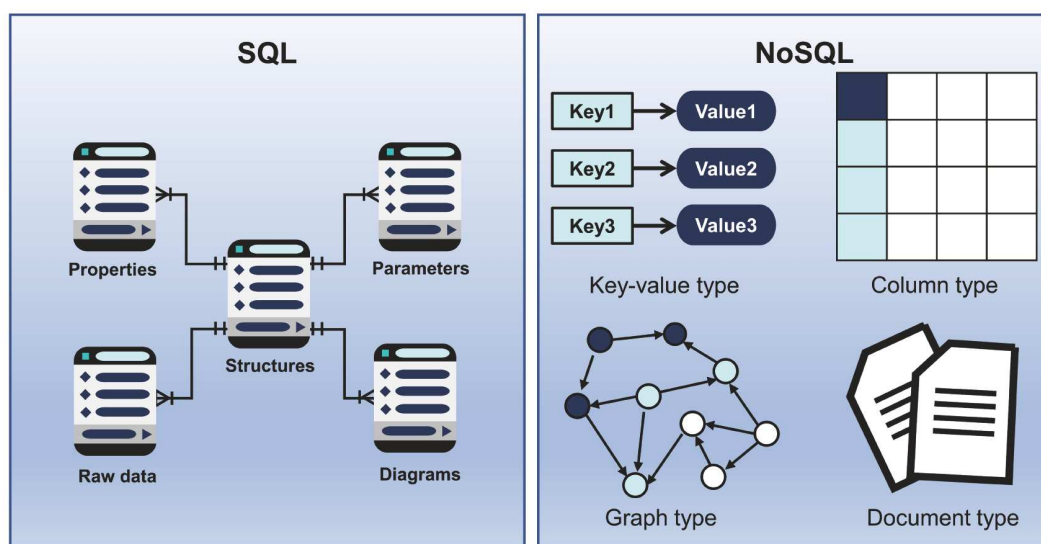


图 6 (网络版彩色) 关系型数据库和非关系型数据库, 其中关系型数据库由二维数据表以及表间关系组成, 非关系型数据库按数据模型分为键值型数据库、列存储数据库、文档型数据库、图存储数据库

Figure 6 (Color online) Relational databases and NoSQL, where relational databases consist of two-dimensional data tables and relationships between tables, non-relational databases are divided into key-value databases, column-family databases, document databases, and graph databases according to the data model

数据不一致, 易于进行数据库维护和更新; 能够保证系统中事务的正确执行, 具备处理事务回滚、恢复、并发控制以及死锁问题的能力。但是, 关系型数据库只能垂直扩展, 即通过添加磁盘、内存、CPU来升级数据库, 可扩展性较差; 关系型数据库由于其部署方式和查询机制等问题, 在高并发请求、大数据量的情况下性能较差; 由于其数据模型定义严格, 数据高度结构化, 导致存储数据灵活性较差。目前, 常用关系型数据库主要有MySQL、Oracle、SQLite、PostgreSQL等。ICSD、AFLOW、OQMD等材料数据库都使用关系型数据库存储数据。

1.3.2 基于非关系型数据库的数据存储

在互联网飞速发展, 数据量激增的背景下, 由于传统关系型数据库存在可扩展性差、灵活性差、读写效率低的缺点, 非关系型数据库逐渐受到研究人员青睐。非关系型数据库是一类不使用传统关系模型的数据库, 在处理非结构化数据方面有得天独厚的优势, 主要包含4种类型, 包括键值存储数据库、文档型数据库、列存储数据库、图形数据库^[9](图6)。

键值存储数据库以哈希表的方式实现数据存储, 允许用户将数据存储为键-值模式, 键作为数据库索引, 值用于存储数据。如此简单的存储模式为数据库带来了高并发性、高查询效率。目前常用键值存储数据库

有Redis、Memcached等。文档型数据库以文档集合的方式存储数据, 可以完美处理结构化、半结构化、非结构化数据, 是最常用的非关系型数据库。文档采用XML、PDF、JSON等标准格式储存, 可以包含多个键-值对、键-数组对, 甚至还可以包含嵌套文档, 储存方式极为灵活。每个文档有一个独特的键用于文档查询, 这个键可以是URL或者存储路径的字符串。目前, 常用的文档型数据库有MongoDB、CouchDB等。列存储数据库按列存储数据, 而不是像传统数据库按行存储数据, 这种存储模式在执行查询时只需要访问特定列而不是整张数据表, 有效减少系统I/O活动次数, 提高查询效率; 此外, 列存储数据库支持分布式储存架构, 可以实现数据库系统的并行查询, 且具有高可扩展性。目前, 常用的列存储数据库有Cassandra、HBase等。图形数据库以图的形式存储数据, 图由节点和边组成, 节点表示数据对象, 边表示数据之间的关系。该存储模型支持在相对较短时间内进行复杂查询, 同时保持传统关系型数据库的ACID属性(即原子性、一致性、隔离性、持久性)。目前常用的图形数据库有Neo4j、AllegroGraph等。

与传统关系型数据库相比, 非关系型数据库在处理海量数据时拥有巨大优势。非关系型数据库支持分布式存储, 可以实现查询请求的并行处理, 能够处理高

并发或者大数据量的任务;数据模型简单,虽然牺牲了数据一致性,但是存储灵活性和数据读写效率得到巨大提升,同时能避免全表扫描,利用索引查询提高查询效率;能够通过添加更多服务器实现水平扩展,提升了数据库的可扩展性。但是非关系型数据库缺少通用查询语言,增加了用户的学习成本;由于牺牲传统数据库的ACID属性以提升数据库性能,开发人员需要额外编程施加ACID约束。目前,包括Materials Project在内的材料数据库和ALKEMIE^[33]等理论计算平台使用非关系型数据库储存数据。

1.4 数据访问

1.4.1 在线访问

大多数材料数据库提供多种在线访问的方式,其中网页访问最为直观(图7)。通过访问网页图形界面,用户可以设置元素组成、化学式进行结构搜索,通过空间群、温度、压力、带隙、形成能等条件筛选搜索结果。不同材料数据库提供了不同的查询策略和筛选条件。Materials Project包含Materials Explorer、Battery Explorer、Molecules Explorer等网页应用程序和Phase Diagram等分析工具。每种应用程序针对存储数据的特点定制了不同搜索策略,例如,Materials Explorer允许

用户根据元素组成进行材料检索,根据热稳定性、对称性等过滤条件进行材料筛选并探索其性质;Battery Explorer在根据元素组成基础上增加了电压、电容等条件,有针对性地搜索锂电池电极材料;Molecules Explorer则提供化学式、元素数量、净电荷、自旋多重度等条件,便于用户进行分子检索。每种分析工具允许用户以交互的方式访问并分析数据,例如Phase Diagram允许根据元素组成为任何封闭系统和开放系统构建低温相图。ICSD允许基于5个方面进行搜索:化学成分、晶体数据、约化胞、对称性、参考文献。在化学成分搜索中,数据库设计了“exclusive OR”策略。在第一次搜索材料失败时,用户通常可以确定目标材料合成过程中涉及的成分,基于该搜索策略,用户可以根据多个起始材料自动搜索所有可能的元素组合。晶体数据和约化胞搜索允许根据作者实验确定的晶胞信息和约化晶胞信息进行搜索。对称性搜索的筛选条件中包含了所有对称性性质,包括空间群、Wyckoff位置、Pearson符号。参考文献搜索支持搜索实验条件、出版物标题、作者姓名等。基于上述搜索条件,用户还可以使用AND、OR、NOT进行布尔式搜索,产生多个结果集。此外,ICSD^[97]还引入关键词和自由文本搜索,其中关键词描述了材料特性、所用分析方法、技术应用领

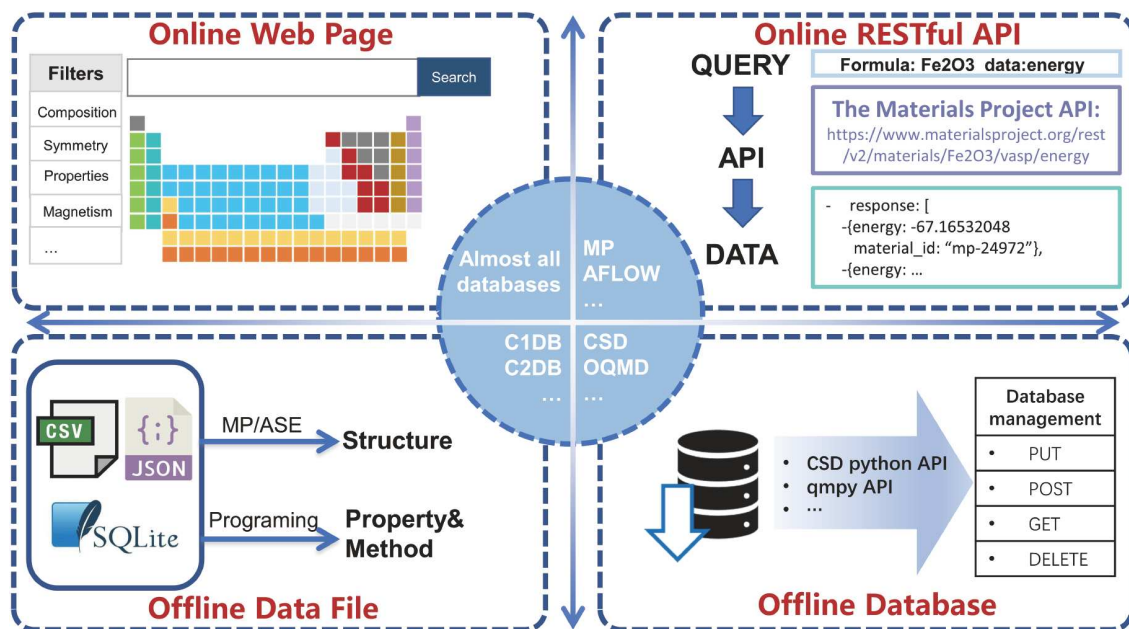


图7 (网络版彩色)数据访问方式可以分为在线访问和离线访问,其中在线访问又可分为网页访问和在线API访问,离线访问又可分为离线数据文件和离线数据库

Figure 7 (Color online) Data access methods can be divided into online access and offline access. Online access includes web page access and online API access, while offline access consists of offline data files and offline databases

域。HTEM^[98]支持通过“全部”或“任意”选项搜索包含目标元素的材料，“全部”选项表示搜索结果必须包含所有目标元素，“任意”选项则表示只要存在任何目标元素就能返回搜索结果。用户可以根据包括合成条件、数据质量/完整性在内的多个标准筛选搜索结果，找到数据量和数据质量之间的平衡。此外，用户还可以通过交互的方式可视化和分析搜索结果。

在数据驱动的材料研发范式下，网页访问已经难以满足高通量计算和机器学习的海量数据需求，材料数据库通常会向用户提供API以实现在线自动数据抓取^[99] (图7)。API基于REST协议实现，允许用户无需依赖网页图形界面，直接以编程的方式与服务器交互，进而批量获取信息。例如，Materials Project向用户提供RESTful API直接访问数据库，每一个材料都可以用唯一的URL结合HTTP方法进行访问^[59]。Materials API的URL设计简单直观，通常包含4个主要部分：序言、资源关键字、标识符、请求数据类型。Materials API通常返回JSON格式的结构化数据文件，也支持XML、YAML格式文件。同样，OQMD、AFLOW、COD等材料数据库也支持用户通过RESTful API在线访问数据库。

各材料数据库均配备了独特的API用于处理数据访问和管理底层数据，然而由于API的语法设计不同，导致它们之间存在或大或小的差异，同时各材料数据库在材料分类、术语表示上的不同进一步加剧了API之间的差异，为跨数据库访问带来巨大障碍。OPTIMADE API通过定义规范URL设计了一个访问不同材料数据库的API，以实现材料数据库间的无缝访问和互操作^[100]。其URL由基础URL、endpoint、OPTIMADE filter三部分组成，基础URL由数据库定义，endpoint用于指定目标数据类型，最后OPTIMADE filter部分定义了搜索条件的标准语法。OPTIMADE API可以将通用请求附加到基础URL之后发送给不同数据库，由支持OPTIMADE API的数据库将其转化为自己相应的查询语句，最后返回响应数据。目前，AFLOW、COD、TCOD^[101]、Materials Cloud^[102]、Materials Project、NOMAD、OQMD等数据库均支持通过OPTIMADE API进行访问。

1.4.2 离线访问

除了在线访问，部分数据库也支持离线访问(图7)。以OQMD、CSD为代表的部分数据库支持将整个数据库离线到本地，通过官方提供的API管理数据库，例如OQMD离线到本地后，用户可以通过qmpy的Django

API查询数据。一些小型数据库则是以CSV、JSON文件的形式分发数据，用户可以通过ASE^[103]等成熟的软件包提取结构信息，也可以根据数据结构编写程序提取材料数据。对于一些未提供Python访问模式的商业闭源数据库，需要手动下载结构文件和提取性质。

2 国内外常用材料数据库

当前，全球范围内已经建立了众多材料数据库。这些数据库收录了广泛的材料类型，涵盖了从无机晶体、有机分子再到工程材料等多样化的材料体系；包含丰富的材料数据，横跨微观性质，例如晶体结构、电子结构、输运性能等，到宏观性能，例如密度、蠕变、弹性模量等。这些大规模、多尺度、多样化的材料数据为科研人员提供了宝贵的资源，推动了材料的高通量筛选和数据驱动的新材料设计。本文简要总结了国内外材料数据库(表1)，并将典型材料数据库分为实验数据库、理论数据库、实验-理论数据库，从数据体量、数据种类、结构来源、查询方式、访问方式等方面对其进行介绍。

2.1 实验数据库

CSD^[23]：CSD是由英国剑桥晶体数据中心开发并维护的世界上最大、最全面的有机小分子和金属有机晶体结构数据库，可以弥补ICSD缺少有机物的不足。目前，数据库已经收录了超过129万个已发表的结构以及它们的单晶或者粉末的X射线或中子衍射分析结果，数据种类包括晶格参数、原子坐标、精修结构、参考文献信息等。它支持网页访问和在线API访问，其中网页访问可以根据参考文献信息、结构信息、晶胞信息和化学式进行检索，而基于Python的API访问允许用户使用自定义脚本访问数据库。CSD不仅是一个材料数据库，更是一个数据共享和标准化平台，它鼓励用户上传数据，同时对上传数据严格审查。在审查过程中，系统通过算法自动检查CIF语法并提取关键数据供研究人员进一步审核。数据会被自动分配一个CCDC编号，并在发表后自动链接到论文。此外，CSD还提供查找、分析、可视化数据的专业应用程序。例如，Mercury能够可视化三维晶体结构和分子结构，探索分子网络与晶胞堆积。Mogul可以快速提取结构的键长、键角、环构象等信息来指导几何结构预测和验证实验结果。

HTEM^[98]：HTEM是由美国国家可再生能源实验室(National Renewable Energy Laboratory, NERL)构建

表1 国内外常用实验、理论、实验-理论数据库的总结

Table 1 Summary of commonly used experimental, theoretical, experimental-theoretical databases

数据库	类型	材料	结构数量	网址
The Cambridge Structure Database (CSD) ^[23]	实验	有机分子, 金属有机化合物	~1290000	https://www.ccdc.cam.ac.uk
MatWeb	实验	聚合物, 金属, 陶瓷, 半导体, 纤维	~180000	https://matweb.com
Springer Materials	实验	无机固体, 金属, 合金, 聚合物	~90000	https://materials.springer.com/
High throughput experimental materials (HTEM) ^[98]	实验	无机固体	~65000	https://hitem.nrel.gov/
Database of Zeolite Structures	实验	沸石	-	https://www.iza-structure.org/databases/
ChemSpider ^[104]	实验	分子	~129000000	https://chemspider.com
PubChem ^[105]	实验	分子	~118000000	https://pubchem.ncbi.nlm.nih.gov/
Materials Project ^[25]	计算	无机固体	~150000	https://next-gen.materialsproject.org/
Open Quantum Materials Database (OQMD) ^[26-28]	计算	无机固体	~1220000	https://www.oqmd.org/
Automatic-FLOW (AFLOW) ^[29]	计算	无机固体, 合金	~3530000	https://aflowlib.org/
The Computational 2D Materials Database (C2DB) ^[106]	计算	二维材料	~16000	https://c2db.fysik.dtu.dk/
The Computational 1D Materials Database (C1DB) ^[107]	计算	一维材料	3144	https://cmrdb.fysik.dtu.dk/c1db
NRELMatDB ^[108]	计算	无机固体	~20000	https://materials.nrel.gov/
Materials cloud ^[102]	计算	无机固体, 金属有机化合物	~11000000	https://www.materialsccloud.org/home
Alloy Database ^[109]	计算	合金	-	https://alloytest.phys.cmu.edu
Clean Energy Project Database(CEPDB) ^[110]	计算	有机光伏材料	~2300000	https://www.molecularspace.org/clean-energy_subdomain/index.html
Electronic Structure Project (ESP) ^[111]	计算	无机固体	>100000	http://gurka.fysik.uu.se/esp
Novel Materials Discovery (NOMAD) ^[30]	计算	无机固体	~3000000	https://nomad.fhi.mpg.de/
Joint Automated Repository For Various Integrated Simulations (JARVIS-DFT) ^[112]	计算	无机固体	~77000	https://www.ctcms.nist.gov/~knc6/JVASP.html
2DMatPedia ^[113]	计算	二维材料	6351	http://www.2dmatpedia.org/
Atomly ^[31]	计算	无机固体	~350000	https://www.atomly.net/
Theoretical Crystallography Open Database (TCOD) ^[101]	计算	无机固体, 金属有机化合物, 矿物	~7000	https://www.crystallography.net/tcod/index.php
The Cambridge Energy Landscape Database	计算	团簇	-	https://www-wales.ch.cam.ac.uk/CCD.html
Inorganic Crystal Structure Database (ICSD) ^[22]	计算, 实验	无机固体	~300000	https://icsd.fiz-karlsruhe.de/
Crystallography Open Database (COD) ^[24]	计算, 实验	有机分子, 无机固体, 金属有机化合物, 矿物	~510000	http://www.crystallography.net
Computational Chemistry Comparison and Benchmark Database (CCCBDB)	计算, 实验	气相原子, 分子	2186	https://cccbdb.nist.gov/
MatNavi (NIMS)	计算, 实验	聚合物, 无机固体, 金属	-	https://mits.nims.go.jp/
Pauling File ^[114]	计算, 实验	无机固体	~350000	https://www.paulingfile.com/

的高通量实验材料数据库。目前, 数据库共包含近8万余种薄膜样品, 收录数据种类包括合成条件、化学成分、晶体结构、光学性质、电子性质。目标样品通常来自计算预测、已发表文献或者特定应用(例如太阳能电池、压电材料等)。每个样品均采用组合物理气相

沉积法合成, 并通过空间分辨的表征技术测量。用户可以通过数据库网页界面按元素进行搜索, 然后基于合成条件、数据质量等条件进行筛选, 从而在数据质量和数量之间找到平衡。最后, 可以对目标数据进行可视化和交互式分析。数据库也支持用户通过API批量获取

数据,进行数据挖掘,HTEM会以CSV、MDCS、MIF等多种格式返回数据.近年,致力于结合高通量实验、数据库、数据挖掘,开发数据基础设施^[115].通过分析高通量实验和数据挖掘的需求,工作人员构建了由两条工作流组成的数据基础设施,实验工作流用于收集、分类、储存原始实验数据,数据工作流通过定制元数据框架将原始实验数据标准化,使其易于存储和检索,可以直接用于数据挖掘.

MatWeb: MatWeb是由美国公司Automation Creation开发的在线材料数据库.数据库专注于工程材料,收录的材料和性质十分全面.它包含超过18万种材料,收录材料种类包括热塑性聚合物、热固性聚合物、金属、陶瓷、半导体、纤维等工程材料,涵盖材料的物理性质、力学性质、热学性质、电学性质、光学性质、加工参数等,同时提供制造商和供应商的联系方式. MatWeb提供强大的搜索工具,在基本检索模式下,用户可以基于最多三项性质的数值搜索材料,可以基于材料类别、商标名称和生产商名称进行检索,还可以通过关键词、短语等文本内容进行检索;在高级检索模式下,用户可以对上述检索方式进行组合实现综合搜索.

ChemSpider^[104]: ChemSpider是由英国皇家化学会运营的在线开放化学数据库.从某种意义上来说,此数据库可以说是一个TB级化学数据搜索引擎,它整合了来自近400个高质量数据源的数据,收录了超过1亿个化学结构,包含丰富的实验数据、理论数据和光谱.除此之外,数据库在已收集数据的基础上添加了相关信息、制造商链接、数据源链接.此数据库还具有化学数据共享功能,它鼓励注册用户上传、注释自己的化学数据,而平台会通过自动结构检查、人工检查两种方式保证数据质量.用户可以基于网页界面根据化学名称、SMILE字符串或化学结构进行基本检索,或者综合结构、元素、性质、数据源等信息进行高级检索.此外,数据库还提供更简便的RESTful API以供用户批量获取数据.

PubChem^[105]: PubChem是美国国家生物技术中心开发和维护的化学数据库.作为世界上最大的免费化学信息库,它整合了来自1000余个独立数据源例如政府机构、化学品生产商、文献出版社等的化学信息,目前共收录1亿余种化合物,其中大部分化合物都是小分子,除此之外还储存有siRNA、miRNA、脂质、碳水化合物等物质.数据库收录了这些化合物详细的化

学信息,包括结构信息、分子式、同分异构体、物理化学性质、毒性等. PubChem提供了强大的搜索和分析工具,用户可以通过化学元素、分子结构、分子式、分子ID等条件进行检索,也可以基于分类进行浏览型检索.

SpringerMaterials: SpringerMaterials是世界上最大的商业材料数据库之一.此数据库由Springer Nature开发和维护,收录了29万余种材料,例如金属、高分子材料、半导体、陶瓷、玻璃,包含多种性质,例如,电性质、热性质、磁性质、机械性质等,其内容涵盖物理、化学、工程以及材料科学.数据库提供交互式晶体结构、相图和数据表,为用户深入分析材料结构和性质提供便利.此外,它还提供强大的搜索工具,可以根据化学元素、化学结构、文本进行搜索,然后按照数据源或性质优化搜索结果.

2.2 理论数据库

Materials Project^[25]: Materials Project是由美国劳伦斯伯克利实验室以及麻省理工学院联合发起的理论材料数据库,旨在通过创建针对计算材料的开放、协作平台来加速材料设计.此数据库绝大多数结构来自ICSD,目前数据库中约包含15万种材料、17万余种分子、4千余种嵌入式电极,计算了包括结构能量、电子结构、热稳定性、相图、弹性张量、压电张量在内的多种性质.此数据库采用非关系型数据库MongoDB储存数据,具有出色的可扩展性和灵活的储存模式,在操作大量数据时相较于关系型数据库具有更高的运行效率.用户可以通过网页访问数据库,通过设置元素组成、化学式进行搜索,通过设置结构信息、对称性、电子结构等条件加以筛选.此外,数据库还提供在线RESTful API和OPTIMADE API访问服务,在此访问方式下,每一个数据请求或操作都可以表示为URL,收到URL后数据库会返回结构化数据.目前,Materials Project凭借其数据量和数据质量上的优势在数据驱动的材料研究中发挥重要作用. Gopakumar等人^[116]以Materials Project中所包含的1864个介电张量作为数据集训练神经网络模型,并结合全局优化算法和第一性原理计算从OQMD的11102种稳定非金属材料中筛选高介电常数材料,最终在仅对17种材料进行DFPT计算的情况下筛选出3种具有高介电常数和带隙的材料. Jang等人^[117]在PU learning的框架下,基于Materials Project训练图卷积神经网络分类模型分析材料的可合成性,在测试集上实现了

87.4%的准确率,并在预测2015至2019年新合成材料的可合成性任务中实现了86.2%的准确率。

OQMD^[26-28]: OQMD是美国西北大学研究团队开发的理论材料数据库。数据库有约10%结构来自ICSD, 剩余90%结构通过基于已有结构原型进行元素替换等其他操作获得。目前, 数据库共包含122万余种材料, 其中大多数化合物是三元化合物, 其次是二元化合物和四元化合物; 涵盖晶体结构、能量信息、电子结构、热力学相图等性质, 其中清晰的热力学相图是数据库一大特色。数据库提供建立二元、三元、四元基态热力学相图的quickhull算法和用于分析金属氢化物、锂离子电池阴极材料基态热力学性质的GCLP算法。OQMD支持多种访问方式, 通过网页访问, 用户可以设置化学组成、结构原型、DFT计算性质等条件进行检索; 除网页访问外, 用户还可以通过RESTful API或OPTIMADE API在线访问批量获取数据; 用户也可以将整个数据库离线到本地, 通过研究团队自主开发的高通量工具qmpy中Django API接口管理数据库。该数据库收录的大量结构为研究人员提供了巨大的数据集和广阔的材料搜索空间。Al-Fahdi等人^[15]结合化学成分描述符和轨道场矩阵设计的卷积神经网络在高功率因数热电材料筛选中表现出优异性能, 并通过分析随机森林模型的特征重要性发现组成元素具有高平均融化温度 and 低电负性平均偏差的化合物具有高功率因数, 为快速筛选潜在热电材料提供了新途径。Kim等人^[118]基于OQMD数据库的DFT数据训练机器学习模型, 根据晶体结构和化学成分预测材料稳定性, 并应用此模型快速预测了55种新型四元Heusler化合物。

AFLOW^[29]: AFLOW是由美国杜克大学研究团队主导开发和维护的理论材料数据库。目前, 数据库共收录约353万种材料, 其中最多的四元体系约有15万种, 其次是三元体系和二元体系。它储存了超过7亿条材料计算性质, 包括热力学稳定性、电子结构、巴德电荷、磁学性质、弹性性质、热学性质等。其所有计算性质使用从头算软件包例如VASP、Quantum Espresso在自主开发的高通量计算框架下获得^[119], 计算数据经处理后储存在关系型数据库MySQL中。AFLOW集成了多个强大的网页应用, 可供用户进行在线数据分析和数据挖掘。其中, 结构原型百科全书包含约1100种晶体结构原型, 提供每个结构的完整描述; AFLOW-ML模块仅需用户输入结构信息就可利用网站提供的3个模型进行在线机器学习; AFLOW-CHULL支持用户在线构建

相图, 分析相稳定性。研究团队开发了数据库专用API——AFLUX用于数据管理或检索, 用户可以借助AFLUX通过编程的方式访问数据库。同时, 数据库也提供更直观的网页搜索, 用户可通过元素组成进行查询, 通过计算性质过滤结果。Roberts等人^[120]基于AFLOW数据库中的结构训练CHfTa、CHfZr、CMoW、CTaTi的矩张量势, 随后用这个势函数对约6300个随机对称性结构进行结构优化并通过主动学习进行改进, 最终得到用于结构优化的势函数和用于优化低能量系统的精确势。这些势函数的预测结果与AFLOW数据库保持较高一致性, 且在CMoW系统中发现了新稳定结构。

NOMAD^[30]: NOMAD是一个以FAIR为导向建立的用户驱动的数据共享平台, 用于储存、共享和分析材料理论数据, 推动新材料发现。迄今为止, 数据库已经储存超过300万种材料和1300万条数据, 上传文件多达104 TB。注册用户可以将原始数据以压缩文件的形式上传到NOMAD, 解析器会自动处理上传文件, 以NOMAD Meta Info将它们转换为规范格式后存入数据库供所有用户浏览。NOMAD解析器支持处理多种从头算计算软件的输出文件, 例如VASP、Gaussian、LAMMPS、CP2K等。NOMAD Meta Info包含丰富信息, 包括模拟的成分/系统、计算方法、参数细节、材料分类、计算数据等, 用户可以基于这些元数据进行在线检索。此外, 数据库也可以通过OPTIMADE API以编程的方式访问。

2DMatPedia^[113]: 2DMatPedia是由新加坡国立大学研究团队开发的二维材料理论数据库。开发人员基于Materials Project数据库中的结构, 利用自上而下的方法和自下而上的方法生成二维材料。自上而下的方法指的是通过基于几何的方法识别具有层状结构的体相材料, 然后剥离成二维单层, 而自下而上的方法则是对自上而下方法获得的一元和二元二维材料进行元素替换。目前, 数据库中共包含6351种二维材料, 其中2940种由自上而下方法获得, 3409种由自下而上方法获得, 2种取自文献, 涵盖了包括晶体结构、电子结构、剥离能、磁矩在内的性质。2DMatPedia支持基于化学组成的网页访问, 也可以将数据库离线到本地。离线数据以JSON格式储存, 需要自行编程解析, 其中结构信息以字典形式提供, 可通过Pymatgen转换为其他格式。Wan等人^[121]基于2DMatPedia数据库中的二维结构构建了12个描述符, 训练了集成树、支持向量机、多元

线性回归和回归树4个模型以预测剥离能,其中集成树和回归树模型拥有最佳预测性能且展现出良好的外推预测精度.同时,通过敏感性分析发现特征的物理意义和物理直觉有很好的一致性.

C2DB^[106]: C2DB是丹麦科技大学Computational Materials Repository项目下的二维材料理论数据库.开发人员通过已知二维晶体结构原型进行元素替换生成候选结构,对候选结构进行结构优化后再进行维度检查、重复性检查、稳定性检查,最后对合格二维结构进行性质计算并将计算结果存入数据库.材料性质通过密度泛函理论和多体微扰理论进行计算,整个流程采用半自动化方式推进,最大程度保证数据一致性.目前,数据库包含1.6万余种二维材料,储存性质包括结构信息、热力学稳定性、弹性性质、电子结构、磁学性质、光学性质和输运性质.数据库支持在线访问,可以通过化学式、磁性、稳定性、带隙等信息进行搜索.用户也可以直接将数据库离线到本地,通过ASE查询或管理数据库.最近, Lyngby等人^[51]通过生成式模型发现的新二维材料已经加入C2DB,该数据库规模还在持续扩大.

JARVIS-DFT^[112]: 作为JARVIS项目的重要组成部分, JARVIS-DFT是美国国家标准技术研究院开发的晶体材料理论数据库.目前,数据库收录了约7万种体相材料和约1000种低维材料,除包含多种常规性质外,例如电子结构、形成能、磁矩、弹性常数、压电常数和介电常数,还包含剥离能、频率依赖介电函数等性质,其规模还在持续扩大.该数据库针对不同目标性质采用不同计算方法以保证结果的准确性,例如采用vdW-DF-OptB88准确预测晶格参数和形成能,采用TBmBJ势修正带隙,采用HSE06、G0W0等方法进行激发态计算. JARVIS-DFT提供网页访问方式,用户可以根据化学组成或者JARVIS-ID检索;同时,此数据库还为用户提供JARVIS-API下载或上传材料数据;不熟悉Python编程的用户可以从Figshare、NIST-Materials、MDF下载JARVIS-DFT的所有原始输入和输出文件. Choudhary等人^[16]通过DFT计算结合半经典输运方法基于JARVIS-DFT进行高通量计算筛选热电候选材料,并训练机器学习模型预测功率因数和塞贝克系数以加速筛选,从36000种3D结构和900中2D结构中筛选出2932种3D和148种2D热电候选材料.

Atomly^[31]: Atomly是由中国科学院物理研究所和松山湖材料实验室开发的无机材料理论数据库.数据

库收录近35万种无机化合物,涵盖包括电子结构、力学性质、介电性质、压电性质、输运性质、相图在内的多种数据,所储存材料数量已达到Materials Project的两倍,电子结构数量已达到Materials Project的4倍.数据库的Run4U功能允许用户提交结构,系统会自动计算并将结果加入数据库.用户可通过网页直接访问数据库,基于化学式、化学元素、结构ID进行材料检索.目前, Atomly已经被广泛应用于无机化合物稳定性判断、反应路径推断,作为数据集助力基于机器学习的物性预测与构效关系挖掘.

CEPDB^[110]: CEPDB是由哈佛大学清洁能源项目所创建的有机光伏材料理论数据库,其目的是通过高通量筛选识别下一代碳基太阳能电池候选材料.该项目研究人员开发了一个组合分子生成器,该生成器能够基于26种基础分子单元并遵循特定的连接规则,有效地构建由4~5个单元链接而成的潜在候选分子. CEPDB利用关系型数据库MySQL存储数据,性质计算主要依赖于商业软件Q-Chem.据报道,该数据库收录了约230万种有机光伏候选材料,包含了分子结构、填充因子、光学带隙、HOMO/LUMO、功率转换效率、开路电压等性质.用户可以通过CEPDB官方网站输入SMILE字符串或者太阳能电池性能参数进行搜索.

Alloy Database^[109]: Alloy Database是Mihalkovic和Widom开发的基于VASP计算的合金理论数据库.数据库包含多种二元、三元和四元合金,向用户提供合金的结构文件、结构原型、赝势类型、截断能、稳定性、形成焓等信息.用户可以选择任意列出的二元、三元或四元合金体系进行搜索.同时,数据库还提供广义搜索,即用户可以选择最多4种元素、结构原型、Pearson符号进行组合搜索.

2.3 实验-理论数据库

ICSD^[22]: ICSD是由德国FIZ Karlsruhe开发并维护的全球最大的无机晶体结构数据库,包含自1913年以来几乎所有已报道的无机晶体.目前,数据库已经收录了约30万种结构,包括22万余种实验无机晶体结构,4万余种金属-有机结构,近3万种理论结构.它每年更新约12000种新结构,数据规模和数据质量都在持续提高.收录进数据库的实验结构都经过充分表征,有确定的原子坐标和化学组成,数据库收录的数据类型涵盖化学成分、晶体结构信息、各原子参数信息、对称性信息、参考文献信息.除实验结构外, ICSD也收录了部分

理论结构及其关键计算参数^[97]。数据库为规范理论结构制定了严格的标准: (1) 理论结构需在同行评审的期刊上发表; (2) 选择总能量更低的理论结构; (3) 如果同一理论结构使用多种方法计算, 则收录和实验结果最接近的数据。它可以通过不同途径访问, 其中网页访问最为直观, 可设置唯一标识符、参考文献、化学式、晶格参数、对称性等条件进行检索, 也可根据关键词进行检索。该数据库收录大量高质量实验材料数据, 已成为机器学习加速材料筛选的基石。Jin等人^[122]基于从文献检索获得的98种光伏材料和98种非光伏材料进行特征工程并训练分类模型, 利用此模型对ICSD中的187093种已合成材料进行筛选获得了3011种光伏候选材料, 然后通过维度识别和DFT验证最终得到了26种二维光伏候选材料。Cai等人^[123]基于ICSD等数据库训练多步机器学习模型, 成功识别了1000余种潜在负热膨胀化合物并预测了它们的负热膨胀系数和温度范围。

COD^[24]: COD是主要由立陶宛维尔纽斯大学研究人员负责开发和维护的开放材料数据库。目前, 数据库已经收录超过50万条数据, 涵盖矿物、无机晶体、有机晶体和金属有机化合物等多种类型化合物, 包含所收录结构的详细结构信息和引用文献, 其规模还在不断扩大。所储存结构主要来自同行评审的出版物且均经过充分表征, 其结构数据通常以CIF格式储存, 因此数据库开发人员能够开发自动化流程检查并修正文件的语法错误、检查重复结构。COD推出了一个数据存储网站, 鼓励其他科研人员共享数据, 所有晶体结构文件均以发表、预发表或者个人通信的形式接受, 系统会检查数据文件并联系用户进行修改。用户可以通过网页界面在线访问数据库, 根据晶胞参数、化学成分、引用信息、SMILE字符串进行检索, 也可以通过API直接连接到数据库, 通过SQL读取和搜索整个数据库。COD还可以被离线到本地以供用户查询。

MatNavi: MatNavi是由日本国立材料科学研究所建立的综合材料数据库。该数据库主要由PolyInfo(收录高分子材料的化学结构、聚合方法、样品处理条件、核磁共振谱、物理性质等)、AtomWork(涵盖从已发表文献中提取的无机材料和金属材料的晶体结构、X射线衍射谱、性质、相图等数据)、CPDDB(储存700余个一元、二元和高阶系统的计算相图)、CompES-X(涵盖基于AtomWork中收录结构计算的电子结构)、Kakusan(收录来自文献的金属材料、无机材料的基本扩散数据)、Thermophysical Property Database(包含金

属、合金、陶瓷材料的密度、表面张力和黏度系数数据)六大基本性质数据库和KinzoKu(包含金属材料的密度、弹性模量、泊松比、拉伸、蠕变等性能)、CCTD(储存焊接连续冷却转变图)两大工程数据库组成, 此外还有5个在线数据表格和3个数据库应用, 其涵盖材料种类、数据种类之广泛, 是世界上最具有影响力的专业材料数据库之一。数据库支持注册用户通过化学式、关键词、材料ID、性质等条件进行跨数据库检索, 或根据每个子数据库提供的检索方式获取数据, 不支持网页爬虫。Long等人^[124]从AtomWork数据库中收集铁磁和反铁磁金属间化合物, 为每个结构构建了Magpie化学描述符和SOAP结构描述符, 利用随机森林模型进行训练以识别磁基态和预测铁磁体的居里温度, 在分类任务和回归任务中均表现出良好性能。

Pauling File^[114]: Pauling File是由瑞士公司MPDS主导开发和维护的无机固体材料数据库, 主要包含相图、晶体结构、物理性质、衍射模式四类数据。所有晶体结构数据都提取自原始文献且经过标准化, 其中的实验结构均经过充分表征。目前, 该数据库共收录一元和二元体系7万余种, 三元体系15万余种, 四元及以上体系近20万种, 收录数据种类包括晶胞参数、空间群、原子位置、位移参数。数据库相图部分主要包含二元相图、三元相图和部分多元体系(四元及以上)的相图, 用户可以在网站给出的二元或三元体系中选择自己感兴趣的体系浏览相图和引用文献。物理性质部分收录了实验和计算获得的无机固体化合物的物理性质, 包括电子性质、热力学稳定性、磁学性质、机械性质、光学性质、热学性质、相变相关参数, 条件允许的情况下还会收录样品制备、实验条件或计算方法相关信息。

3 总结与展望

材料数据库作为数据驱动材料研发范式的基础, 对加速材料构效关系挖掘和新材料优化设计起到至关重要的作用。本文从数据生成、数据预处理、数据存储、数据检索四方面综述了数据库构建的一般流程, 总结了国内外典型实验材料数据库和理论材料数据库。在MGI及其相关计划实施后, 实验材料数据库和理论材料数据库在多个维度上都实现了长足的进步, 具体表现为数据生成方式自动化, 除了手动从出版物中收集已充分表征结构的传统方式外, 高通量计算、高通量实验、文本挖掘等自动化方法拓宽数据来源, 大大提高了数据生成的效率; 数据表示方式标准化, 各数据库通过数据标

准化提高数据质量,同时致力于开发统一数据库API接口促进数据共享;数据存储方式多元化,开发人员可以根据材料数据库的自身定位选择关系型数据库或者非关系型数据库存储数据,找到可扩展性、存储灵活性和数据一致性、完整性之间的平衡点;数据访问方式多样化,除了简便直观的网页访问,为了适应数据驱动材料研发的巨大数据量需求,开发人员还提供离线访问和基于RESTful API的在线访问。然而,为了进一步扩大数据库在材料研发中的影响力,需要在材料数据和数据基础设施建设等方面取得更显著的突破。

目前,材料数据的数据体量和数据生成速度方面仍有提升空间。以欧洲生物信息研究所为例,其存储的数据量高达约90 PB^[125],相较之下,Materials Project最终向用户呈现的数据量约为1 PB, NOMAD数据库所收录的数据量仅约为104 TB。当下数据驱动材料研发范式对数据体量和数据生成速度提出了更高要求,研究人员仍需要继续扩大数据生产规模,借助高性能计算集群和高通量工具自动化生成理论数据,借助高通量实验平台快速生成实验数据,将人工智能模型与高通量方法结合加速数据生成。中国科学院物理研究所开发的理论数据库Atomly.net已经取得显著进展,在数据量上超越了Materials Project。

发展高质量材料数据库将极大助力材料领域大语言模型的开发,从而加速新材料发现。以ChatGPT为代表的大语言模型(large language model, LLM)在文本分类、信息提取、文案生成、常识问答等领域展现出的卓越能力,让研究人员看到LLM在材料领域应用的广阔前景。将材料数据库与LLM紧密结合,开发出材料领域大模型将极大增强材料信息自动提取、查询、分析的能力,提高智能实验设计和材料性能预测的水平。基于基础模型进行微调是训练领域模型的有效方法,例如基于开源模型LLaMA^[126]或OpenAI的商业闭源模型GPT利用大量材料数据和领域文本进行微调,能够得

到性能优异的领域模型。当前,国外的DARWIN^[127]和国内的ChemLLM^[128]、ChemDFM^[129]等模型已经在部分物理、化学、材料科学的任务中表现出与GPT-4相当甚至超过GPT-4的性能,为材料研发提供了便利。

材料数据的标准化也是未来的一个重要趋势,标准化的元数据框架能够促进不同来源的数据的无缝访问和聚合,提升数据的可重用性和互操作性。目前材料领域还未开发出标准的元数据框架,无机晶体领域常用CIF表示数据,计算材料学中则倾向于使用NOMAD Meta Info等格式存储计算相关数据,陶瓷材料领域常用Plinius^[130],钢铁领域常用ONTORULE^[131]。开发标准元数据框架关键在于构建通用材料本体。自然语言处理可用于从科学文献中挖掘知识,帮助自动结构分类、统一专业术语和注释数据,进而助力开发通用本体。此外,通过观察文献中使用的命名法和概念关系也有助于构建本体。目前各国的科研机构正在为统一材料科学的命名标准积极合作,并开发出上文提到的多数据平台通用访问接口OPTIMADE。

面对数据驱动材料研发的趋势,构建高度系统化的数据基础设施能够促进数据生产规模化、规范化,保证材料数据FAIR属性。数据基础设施以数据为中心,实现自动数据生成、数据存储、数据挖掘三大功能。它通过集成高性能计算硬件、高通量计算工具和计算模拟软件,批量生成多尺度计算数据,或基于大科学装置或者高通量实验平台进行批量合成和表征,产生实验数据,这些材料数据具有明确的生成参数和流程,具有高度一致性和完整性。数据基础设施的数据库应采用统一元数据框架储存数据,构建合理数据模型便于数据高效存储和检索;平台应提供多种访问方式,采用统一数据传输标准,促进数据共享与数据库间数据融合。数据基础设施中应集成数据挖掘算法,基于数据库进行模型训练,进而指导实验合成或直接生成新结构以扩充数据库。

参考文献

- 1 Lin L. Materials databases infrastructure constructed by first principles calculations: A review. *Mater Performance Charact*, 2015, 4: 148–169
- 2 Agrawal A, Choudhary A. Perspective: Materials informatics and big data: Realization of the “fourth paradigm” of science in materials science. *APL Mater*, 2016, 4: 053208
- 3 Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 2021, 596: 583–589
- 4 Hao M, Gong J, Zeng X, et al. Large-scale foundation model on single-cell transcriptomics. *Nat Methods*, 2024, 21: 1481–1491
- 5 Tu T, Azizi S, Driess D, et al. Towards generalist biomedical AI. 2023, arXiv: 2307.14334 arXiv: 2307.14334
- 6 Bi K, Xie L, Zhang H, et al. Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 2023, 619: 533–538

- 7 Raccuglia P, Elbert K C, Adler P D F, et al. Machine-learning-assisted materials discovery using failed experiments. *Nature*, 2016, 533: 73–76
- 8 Zeni C, Pinsler R, Zügner D, et al. MatterGen: A generative model for inorganic materials design. 2024, arXiv: 2312.03687 arXiv: 2312.03687
- 9 Merchant A, Batzner S, Schoenholz S S, et al. Scaling deep learning for materials discovery. *Nature*, 2023, 624: 80–85
- 10 Li M X, Zhao S F, Lu Z, et al. High-temperature bulk metallic glasses developed by combinatorial methods. *Nature*, 2019, 569: 99–103
- 11 Li M X, Sun Y T, Wang C, et al. Data-driven discovery of a universal indicator for metallic glass forming ability. *Nat Mater*, 2022, 21: 165–172
- 12 Lu S, Zhou Q, Ouyang Y, et al. Accelerated discovery of stable lead-free hybrid organic-inorganic perovskites via machine learning. *Nat Commun*, 2018, 9: 3405
- 13 Li Y, Zhu R, Wang Y, et al. Center-environment deep transfer machine learning across crystal structures: From spinel oxides to perovskite oxides. *npj Comput Mater*, 2023, 9: 109
- 14 Bartel C J, Sutton C, Goldsmith B R, et al. New tolerance factor to predict the stability of perovskite oxides and halides. *Sci Adv*, 2019, 5: eaav0693
- 15 Al-Fahdi M, Yuan K, Yao Y, et al. High-throughput thermoelectric materials screening by deep convolutional neural network with fused orbital field matrix and composition descriptors. *Appl Phys Rev*, 2024, 11: 021402
- 16 Choudhary K, Garrity K F, Tavazza F. Data-driven discovery of 3D and 2D thermoelectric materials. *J Phys-Condens Matter*, 2020, 32: 475501
- 17 Guo Y, He X, Su Y, et al. Machine-learning-guided discovery and optimization of additives in preparing Cu catalysts for CO₂ reduction. *J Am Chem Soc*, 2021, 143: 5755–5762
- 18 Chen D, Chen L, Zhao Q C, et al. Square-pyramidal subsurface oxygen [Ag₄OAg] drives selective ethene epoxidation on silver. *Nat Catal*, 2024, 7: 536–545
- 19 Sendek A D, Cubuk E D, Antoniuk E R, et al. Machine learning-assisted discovery of solid Li-ion conducting materials. *Chem Mater*, 2019, 31: 342–352
- 20 Liu B, Yang J, Yang H, et al. Rationalizing the interphase stability of Li[doped-Li₇La₃Zr₂O₁₂] via automated reaction screening and machine learning. *J Mater Chem A*, 2019, 7: 19961–19969
- 21 Su Y J, Fu H D, Bai Y, et al. Progress in materials genome engineering in China (in Chinese). *Acta Metall Sin*, 2020, 56: 1313–1323 [宿彦京, 付华栋, 白洋, 等. 中国材料基因工程研究进展. *金属学报*, 2020, 56: 1313–1323]
- 22 Belsky A, Hellenbrandt M, Karen V L, et al. New developments in the Inorganic Crystal Structure Database (ICSD): Accessibility in support of materials research and design. *Acta Crystlogr B Struct Sci*, 2002, 58: 364–369
- 23 Groom C R, Bruno I J, Lightfoot M P, et al. The Cambridge Structural Database. *Acta Crystlogr B Struct Sci Cryst Eng Mater*, 2016, 72: 171–179
- 24 Gražulis S, Chateigner D, Downs R T, et al. Crystallography Open Database—An open-access collection of crystal structures. *J Appl Crystlogr*, 2009, 42: 726–729
- 25 Jain A, Ong S P, Hautier G, et al. The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater*, 2013, 1: 011002
- 26 Saal J E, Kirklin S, Aykol M, et al. Materials design and discovery with high-throughput density functional theory: The Open Quantum Materials Database (OQMD). *JOM*, 2013, 65: 1501–1509
- 27 Kirklin S, Saal J E, Meredig B, et al. The Open Quantum Materials Database (OQMD): Assessing the accuracy of DFT formation energies. *npj Comput Mater*, 2015, 1: 15010
- 28 Shen J, Griesemer S D, Gopakumar A, et al. Reflections on one million compounds in the Open Quantum Materials Database (OQMD). *J Phys Mater*, 2022, 5: 031001
- 29 Esters M, Oses C, Divilov S, et al. aflow.org: A web ecosystem of databases, software and tools. *Comput Mater Sci*, 2023, 216: 111808
- 30 Scheidgen M, Himanen L, Ladines A N, et al. NOMAD: A distributed web-based platform for managing materials science research data. *JOSS*, 2023, 8: 5388
- 31 Liu M, Meng S. Atomly.net materials database and its application in inorganic chemistry (in Chinese). *Sci Sin-Chim*, 2023, 53: 19–25 [刘森, 孟胜. Atomly.net数据平台及其在无机化学中的应用. *中国科学: 化学*, 2023, 53: 19–25]
- 32 Liu S, Su Y, Yin H, et al. An infrastructure with user-centered presentation data model for integrated management of materials data and services. *npj Comput Mater*, 2021, 7: 88
- 33 Wang G, Peng L, Li K, et al. ALKEMIE: An intelligent computational platform for accelerating materials discovery and design. *Comput Mater Sci*, 2021, 186: 110064
- 34 Yang X, Wang Z, Zhao X, et al. MatCloud: A high-throughput computational infrastructure for integrated management of materials simulation, data and resources. *Comput Mater Sci*, 2018, 146: 319–333
- 35 Zhang Q, Chang D, Zhai X, et al. OCPMDM: Online computation platform for materials data mining. *Chemometrics Intelligent Lab Syst*, 2018, 177: 26–34
- 36 Xu Y. Accomplishment and challenge of materials database toward big data. *Chin Phys B*, 2018, 27: 118901
- 37 Himanen L, Geurts A, Foster A S, et al. Data-driven materials science: Status, challenges, and perspectives. *Adv Sci*, 2019, 6: 1900808

- 38 Jain A, Hautier G, Moore C J, et al. A high-throughput infrastructure for density functional theory calculations. *Comput Mater Sci*, 2011, 50: 2295–2310
- 39 Curtarolo S, Hart G L W, Nardelli M B, et al. The high-throughput highway to computational materials design. *Nat Mater*, 2013, 12: 191–201
- 40 Liu X J, He Y D. Design and implementation of a high throughput computing and data management platform for two dimensional semiconductor materials (in Chinese). *J Southwest Univ Natl*, 2023, 49: 671–678 [刘小久, 何远德. 二维半导体材料的高通量计算及数据管理平台的设计与实现. *西南民族大学学报*, 2023, 49: 671–678]
- 41 Li M, Wang X, Xie J, et al. Theoretical design of optoelectronic semiconductors (in Chinese). *Chin Sci Bull*, 2023, 68: 2221–2238 [李木琛, 王新江, 颜家豪, 等. 光电半导体材料的理论设计. *科学通报*, 2023, 68: 2221–2238]
- 42 Perdew J P, Burke K, Ernzerhof M. Generalized gradient approximation made simple [Phys. Rev. Lett. 77, 3865 (1996)]. *Phys Rev Lett*, 1997, 78: 1396
- 43 Becke A D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys Rev A*, 1988, 38: 3098–3100
- 44 Krukau A V, Vydrov O A, Izmaylov A F, et al. Influence of the exchange screening parameter on the performance of screened hybrid functionals. *J Chem Phys*, 2006, 125: 224106
- 45 Luo S, Li T, Wang X, et al. High-throughput computational materials screening and discovery of optoelectronic semiconductors. *WIREs Comput Mol Sci*, 2021, 11: e1489
- 46 Zhao X G, Zhou K, Xing B, et al. JAMIP: An artificial-intelligence aided data-driven infrastructure for computational materials informatics. *Sci Bull*, 2021, 66: 1973–1985
- 47 Li Y, Yang J, Zhao R, et al. Design of organic–inorganic hybrid heterostructured semiconductors via high-throughput materials screening for optoelectronic applications. *J Am Chem Soc*, 2022, 144: 16656–16666
- 48 Zhao X G, Yang J H, Fu Y, et al. Design of lead-free inorganic halide perovskites for solar cells via cation-transmutation. *J Am Chem Soc*, 2017, 139: 2630–2638
- 49 Sanchez-Lengeling B, Aspuru-Guzik A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science*, 2018, 361: 360–365
- 50 Kim B, Lee S, Kim J. Inverse design of porous materials using artificial neural networks. *Sci Adv*, 2020, 6: eaax9324
- 51 Lyngby P, Thygesen K S. Data-driven discovery of 2D materials by deep generative models. *npj Comput Mater*, 2022, 8: 232
- 52 Moustafa H, Lyngby P M, Mortensen J J, et al. Hundreds of new, stable, one-dimensional materials from a generative machine learning model. *Phys Rev Mater*, 2023, 7: 014007
- 53 Wang Z, Sun Z, Yin H, et al. Data-driven materials innovation and applications. *Adv Mater*, 2022, 34: 2104113
- 54 Ong S P, Richards W D, Jain A, et al. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput Mater Sci*, 2013, 68: 314–319
- 55 Jain A, Ong S P, Chen W, et al. FireWorks: A dynamic workflow system designed for high-throughput applications. *Concurr Comput*, 2015, 27: 5037–5059
- 56 Mathew K, Montoya J H, Faghaninia A, et al. Atomate: A high-level interface to generate, execute, and analyze computational materials science workflows. *Comput Mater Sci*, 2017, 139: 140–152
- 57 Supka A R, Lyons T E, Liyanage L, et al. AFLOW π : A minimalist approach to high-throughput *ab initio* calculations including the generation of tight-binding hamiltonians. *Comput Mater Sci*, 2017, 136: 76–84
- 58 Ong S P, Wang L, Kang B, et al. Li–Fe–P–O₂ phase diagram from first principles calculations. *Chem Mater*, 2008, 20: 1798–1807
- 59 Ong S P, Cholia S, Jain A, et al. The Materials Application Programming Interface (API): A simple, flexible and efficient API for materials data based on REpresentational State Transfer (REST) principles. *Comput Mater Sci*, 2015, 97: 209–215
- 60 Shevlin M. Practical high-throughput experimentation for chemists. *ACS Med Chem Lett*, 2017, 8: 601–607
- 61 Liu X, Liu B, Ding J, et al. Building a library for catalysts research using high-throughput approaches. *Adv Funct Mater*, 2022, 32: 2107862
- 62 Maier W, Stöwe K, Sieg S. Combinatorial and high-throughput materials science. *Angew Chem Int Ed*, 2007, 46: 6016–6067
- 63 Corma A, Serra J M, Chica A. Discovery of new paraffin isomerization catalysts based on SO₄²⁻/ZrO₂ and WO_x/ZrO₂ applying combinatorial techniques. *Catal Today*, 2003, 81: 495–506
- 64 Liu Z, Rolston N, Flick A C, et al. Machine learning with knowledge constraints for process optimization of open-air perovskite solar cell manufacturing. *Joule*, 2022, 6: 834–849
- 65 Chen X, Xie Y, Sheng Y, et al. Wafer-scale functional circuits based on two dimensional semiconductors with fabrication optimized by machine learning. *Nat Commun*, 2021, 12: 5953
- 66 Gao C, Bao J, Luo Z L, et al. Recent progresses in the combinatorial materials science (in Chinese). *Acta Physico-Chim Sin*, 2006, 22: 899–912 [高琛, 鲍俊, 罗震林, 等. 组合材料科学研究进展. *物理化学学报*, 2006, 22: 899–912]
- 67 Löbel R, Thienhaus S, Savan A, et al. Combinatorial fabrication and high-throughput characterization of a Ti–Ni–Cu shape memory thin film composition spread. *Mater Sci Eng-A*, 2008, 481–482: 151–155

- 68 Mao S S. High throughput combinatorial screening of semiconductor materials. *Appl Phys A*, 2011, 105: 283–288
- 69 Koinuma H, Kawasaki M, Itoh T, et al. Concept and development of combinatorial laser MBE for oxide electronics. *Physica C-Supercond*, 2000, 335: 245–250
- 70 Zarnetta R, Kneip S, Somsen C, et al. High-throughput characterization of mechanical properties of Ti–Ni–Cu shape memory thin films at elevated temperature. *Mater Sci Eng-A*, 2011, 528: 6552–6557
- 71 Sáfrán G. “One-sample concept” micro-combinatory for high throughput TEM of binary films. *Ultramicroscopy*, 2018, 187: 50–55
- 72 Kononova O, He T, Huo H, et al. Opportunities and challenges of text mining in materials research. *iScience*, 2021, 24: 102155
- 73 Olivetti E A, Cole J M, Kim E, et al. Data-driven materials research enabled by natural language processing and information extraction. *Appl Phys Rev*, 2020, 7: 041317
- 74 Swain M C, Cole J M. ChemDataExtractor: A toolkit for automated extraction of chemical information from the scientific literature. *J Chem Inf Model*, 2016, 56: 1894–1904
- 75 Hawizy L, Jessop D M, Adams N, et al. ChemicalTagger: A tool for semantic text-mining in chemistry. *J Cheminform*, 2011, 3: 17
- 76 Jessop D M, Adams S E, Willighagen E L, et al. OSCAR4: A flexible architecture for chemical text-mining. *J Cheminform*, 2011, 3: 41
- 77 Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space. 2013, arXiv: 1301.3781 arXiv: 1301.3781
- 78 Joulin A, Grave E, Bojanowski P, et al. Bag of tricks for efficient text classification. 2016, arXiv: 1607.01759 arXiv: 1607.01759
- 79 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. 2023, arXiv: 1706.03762 arXiv: 1706.03762
- 80 Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding. 2019, arXiv: 1810.04805 arXiv: 1810.04805
- 81 Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners. OpenAI blog, 2019, 1: 9
- 82 Radford A, Narasimhan K, Salimans T, et al. Improving Language Understanding by Generative Pre-Training. 2018, <https://hayate-lab.com/wp-content/uploads/2023/05/43372bfa750340059ad87ac8e538c53b.pdf>
- 83 Brown T B, Mann B, Ryder N, et al. Language models are few-shot learners. 2020, arXiv: 2005.14165 arXiv: 2005.14165
- 84 Huang S, Cole J M. A database of battery materials auto-generated using ChemDataExtractor. *Sci Data*, 2020, 7: 260
- 85 Sierpeklis O, Cole J M. A thermoelectric materials database auto-generated from the scientific literature using ChemDataExtractor. *Sci Data*, 2022, 9: 648
- 86 Court C J, Cole J M. Auto-generated materials database of Curie and Néel temperatures via semi-supervised relationship extraction. *Sci Data*, 2018, 5: 180111
- 87 Kononova O, Huo H, He T, et al. Text-mined dataset of inorganic materials synthesis recipes. *Sci Data*, 2019, 6: 203
- 88 Huo H, Rong Z, Kononova O, et al. Semi-supervised machine-learning classification of materials synthesis procedures. *npj Comput Mater*, 2019, 5: 62
- 89 Wilkinson M D, Dumontier M, Aalbersberg I J, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*, 2016, 3: 160018
- 90 Hall S R, Allen F H, Brown I D. The crystallographic information file (CIF): A new standard archive file for crystallography. *Acta Crystlogr Found Crystlogr*, 1991, 47: 655–685
- 91 Bernstein H J, Bollinger J C, Brown I D, et al. Specification of the Crystallographic Information File format, version 2.0 *J Appl Crystlogr*, 2016, 49: 277–284
- 92 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. *npj Comput Mater*, 2017, 3: 46
- 93 Ghiringhelli L M, Baldauf C, Bereau T, et al. Shared metadata for data-centric materials science. *Sci Data*, 2023, 10: 626
- 94 Michel K, Meredig B. Beyond bulk single crystals: A data format for all materials structure–property–processing relationships. *MRS Bull*, 2016, 41: 617–623
- 95 Roman C, Michal K. Comparison of query performance in relational a non-relation databases. In: 13th International Scientific Conference on Sustainable, Modern and Safe Transport (TRANSCOM 2019). Transportation Research Procedia. Amsterdam: Elsevier, 2019, 40: 170–177
- 96 Nayak A. Type of NOSQL Databases and its comparison with relational databases. *Int J Appl Inf Syst*, 2013, 5: 4
- 97 Zagorac D, Müller H, Ruehl S, et al. Recent developments in the Inorganic Crystal Structure Database: Theoretical crystal structure data and related features. *J Appl Crystlogr*, 2019, 52: 918–925
- 98 Zakutayev A, Wunder N, Schwarting M, et al. An open experimental database for exploring inorganic materials. *Sci Data*, 2018, 5: 180053
- 99 Jain A, Persson K A, Ceder G. Research Update: The materials genome initiative: Data sharing and the impact of collaborative *ab initio* databases. *APL Mater*, 2016, 4: 053102
- 100 Andersen C W, Armiento R, Blokhin E, et al. OPTIMADE, an API for exchanging materials data. *Sci Data*, 2021, 8: 217
- 101 Merkys A, Mounet N, Cepellotti A, et al. A posteriori metadata from automated provenance tracking: Integration of AiiDA and TCO. *J Cheminform*, 2017, 9: 56

- 102 Talirz L, Kumbhar S, Passaro E, et al. Materials Cloud, a platform for open computational science. *Sci Data*, 2020, 7: 299
- 103 Hjorth Larsen A, Jørgen Mortensen J, Blomqvist J, et al. The atomic simulation environment—A Python library for working with atoms. *J Phys-Condens Matter*, 2017, 29: 273002
- 104 Pence H E, Williams A. ChemSpider: An online chemical information resource. *J Chem Educ*, 2010, 87: 1123–1124
- 105 Kim S, Chen J, Cheng T, et al. PubChem 2023 update. *Nucleic Acids Res*, 2023, 51: D1373–D1380
- 106 Hastrup S, Strange M, Pandey M, et al. The Computational 2D Materials Database: High-throughput modeling and discovery of atomically thin crystals. *2D Mater*, 2018, 5: 042002
- 107 Moustafa H, Larsen P M, Gjerding M N, et al. Computational exfoliation of atomically thin one-dimensional materials with application to Majorana bound states. *Phys Rev Mater*, 2022, 6: 064202
- 108 Stevanović V, Lany S, Zhang X, et al. Correcting density functional theory for accurate predictions of compound enthalpies of formation: Fitted elemental-phase reference energies. *Phys Rev B*, 2012, 85: 115104
- 109 Widom M, Mihalkovic M. Stability of Fe-based alloys with structure type C_6Cr_{23} . *J Mater Res*, 2005, 20: 237–242
- 110 Hachmann J, Olivares-Amaya R, Atahan-Evrenk S, et al. The Harvard Clean Energy Project: Large-scale computational screening and design of organic photovoltaics on the world community grid. *J Phys Chem Lett*, 2011, 2: 2241–2251
- 111 Klintonberg M, Derenzo S E, Weber M J. Potential scintillators identified by electronic structure calculations. *Nucl Instrum Methods Phys Res Sect A-Accel Spectrom Deect Assoc Equip*, 2002, 486: 298–302
- 112 Choudhary K, Garrity K F, Reid A C E, et al. The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design. *npj Comput Mater*, 2020, 6: 173
- 113 Zhou J, Shen L, Costa M D, et al. 2DMatPedia, an open computational database of two-dimensional materials from top-down and bottom-up approaches. *Sci Data*, 2019, 6: 86
- 114 Villars P, Berndt M, Brandenburg K, et al. The pauling file, binaries edition. *J Alloys Compd*, 2004, 367: 293–297
- 115 Talley K R, White R, Wunder N, et al. Research data infrastructure for high-throughput experimental materials science. *Patterns*, 2021, 2: 100373
- 116 Gopakumar A, Pal K, Wolverton C. Identification of high-dielectric constant compounds from statistical design. *npj Comput Mater*, 2022, 8: 146
- 117 Jang J, Gu G H, Noh J, et al. Structure-based synthesizability prediction of crystals using partially supervised learning. *J Am Chem Soc*, 2020, 142: 18836–18843
- 118 Kim K, Ward L, He J, et al. Machine-learning-accelerated high-throughput materials screening: Discovery of novel quaternary Heusler compounds. *Phys Rev Mater*, 2018, 2: 123801
- 119 Curtarolo S, Setyawan W, Hart G L W, et al. AFLOW: An automatic framework for high-throughput materials discovery. *Comput Mater Sci*, 2012, 58: 218–226
- 120 Roberts J, Rijal B, Divilov S, et al. Machine learned interatomic potentials for ternary carbides trained on the AFLOW database. *npj Comput Mater*, 2024, 10: 142
- 121 Wan Z, Wang Q D. Machine learning prediction of the exfoliation energies of two-dimension materials via data-driven approach. *J Phys Chem Lett*, 2021, 12: 11470–11475
- 122 Jin H, Zhang H, Li J, et al. Discovery of novel two-dimensional photovoltaic materials accelerated by machine learning. *J Phys Chem Lett*, 2020, 11: 3075–3081
- 123 Cai Y, Wang C, Yuan H, et al. Exploring negative thermal expansion materials with bulk framework structures and their relevant scaling relationships through multi-step machine learning. *Mater Horiz*, 2024, 11: 2914–2925
- 124 Long T, Fortunato N M, Zhang Y, et al. An accelerating approach of designing ferromagnetic materials via machine learning modeling of magnetic ground state and Curie temperature. *Mater Res Lett*, 2021, 9: 169–174
- 125 Thakur M, Buniello A, Brooksbank C, et al. EMBL's European Bioinformatics Institute (EMBL-EBI) in 2023. *Nucleic Acids Res*, 2024, 52: D10–D17
- 126 Touvron H, Lavril T, Izacard G, et al. LLaMA: Open and efficient foundation language models. 2023, arXiv: 2302.13971 arXiv: 2302.13971
- 127 Xie T, Wan Y, Huang W, et al. DARWIN Series: Domain specific large language models for natural science. 2023, arXiv: 2308.13565 arXiv: 2308.13565
- 128 Zhang D, Liu W, Tan Q, et al. ChemLLM: a chemical large language model. 2024, arXiv: 2402.06852 arXiv: 2402.06852
- 129 Zhao Z, Ma D, Chen L, et al. ChemDFM: Dialogue foundation model for chemistry. 2024, arXiv: 2401.14818 arXiv: 2401.14818
- 130 Van der Vet P E, Speel P H, Mars N J. The Plinius ontology of ceramic materials. In: Proceedings of the 11th European Conference on Artificial Intelligence (ECAI'94) Workshop on Comparison of Implemented Ontologies, Amsterdam, 1994. 187–205
- 131 Rudolph S, Gutierrez C. Web Reasoning and Rule Systems: 5th International Conference, RR 2011, Galway, Ireland, August 29–30, 2011, Proceedings. Berlin, Heidelberg: Springer, 2011, 6902: 24–29

Summary for “数据驱动研究范式下材料数据库的构建与应用”

Construction and application of materials database under data-driven research paradigm

Junjie Feng, Kun Zhou, Muchen Li, Xinjiang Wang* & Lijun Zhang*

Key Laboratory of Material Simulation Methods & Software of MOE, Key Laboratory of Automobile Materials of MOE, State Key Laboratory of Integrated Optoelectronics, College of Materials Science and Engineering, Jilin University, Changchun 130012, China

* Corresponding authors, E-mail: lijun_zhang@jlu.edu.cn; xinjiang_wang@jlu.edu.cn

In recent years, with the rapid development of big data and artificial intelligence technologies, the data-driven paradigm for materials research has increasingly demonstrated significant advantages in areas such as mining structure-property relationships and screening and designing novel materials. It relies on vast amounts of data, employing high-throughput screening or artificial intelligence aided methods to shorten the materials research and development cycle while reducing scientific research costs. In this paradigm, data serves not only as the foundation but also plays a crucial role in expanding the search space for high-throughput screening and enhancing the performance of artificial intelligence models. Consequently, establishing high-quality materials databases has become an essential component of data-driven materials research. Since the launch of the Materials Genome Initiative, both theoretical and experimental databases have seen continuous growth in number and scale, characterized by automatic data generation, diversified data access, and standardized data representation. These databases strive to meet the demands for high concurrency and scalability in the context of big data while improving data reliability and usability. This review focuses on materials database construction under data-driven research paradigm, outlining the key of database construction from four aspects: Data generation, data preprocessing, data storage, and data access. In the data generation phase, methods like high-throughput computation and experimentation facilitate batch acquisition of theoretical and experimental data. When calculation or synthesis methods are reasonable, both data accuracy and efficiency of data acquisition can be achieved simultaneously. Text mining, based on natural language processing technology, can extract synthetic information, material entities, or properties from massive scientific literature, which is also an important supplementary means of data generation. Data preprocessing is vital for ensuring compliance with FAIR (findable, accessible, interoperable, and reusable) principles and enhancing data quality which can be achieved through data standardization and data filtering. During the data storage phase, it is necessary to construct a rational data model and choose between relational and non-relational databases for storing data, ensuring efficient and flexible data storage. Data retrieval can be facilitated by designing effective query strategies and offering diverse access methods. Subsequently, a brief summary is offered on representative materials databases from around the world, considering aspects such as the volume and variety of data, access methods, and distinctive features of the databases. We will also integrate selected case studies from database applications to underscore the pivotal role and advantages of databases within the data-driven research paradigm. Finally, we discuss future developments in materials databases from two perspectives: The material data itself and the construction of data infrastructure.

database construction, novel material design, materials genome engineering, machine learning

doi: [10.1360/TB-2024-0946](https://doi.org/10.1360/TB-2024-0946)