

数据智能驱动的机器化学家探索

徐锦程, 陈林江, 江俊*

中国科学技术大学精准智能化学全国重点实验室, 合肥 230026

*通讯作者, E-mail: jiangjl@ustc.edu.cn

收稿日期: 2025-03-25; 接受日期: 2025-05-8; 网络版发表日期: 2025-05-26

中国科学院战略性先导科技专项(编号: XDB0450302)和国家自然科学基金(编号: 22025304, 22033007)资助项目

摘要 近年来, 人工智能与自动化技术的深度融合为化学研究范式的革新提供了全新路径。本文系统构建了集理论计算、机器学习、自动化实验与云端基础设施于一体的智能化学实验室系统, 通过“数据生成-模型优化-实验验证”的理实迭代机制实现研究流程的动态闭环优化, 并创新性地引入大语言模型驱动的多智能体协作架构, 完成从自然语言指令解析到跨设备任务调度的全流程自主操作。本文详细阐述了智能实验室的硬件集成方案、多模态数据融合策略及云端资源调度机制, 通过火星陨石催化剂开发、胶体纳米晶形貌调控及高熵催化剂筛选等典型案例验证了系统的全局优化能力, 最后对智能化学研究网络的基础设施建设与协同创新模式进行了前瞻性探讨。

关键词 智能化学, 机器化学家, 理实迭代学习, 大语言模型, 智能化学家基础设施

1 引言

化学研究的核心挑战在于如何建立微观机制与宏观性能之间的可靠关联。传统研究范式长期面临方法论困境: 经验驱动的试错法难以系统探索复杂参数空间, 导致新材料研发周期漫长; 局部搜索策略使优化陷入局部最优陷阱, 无法获得全局最优解; 实验数据的碎片化与非标准化特性进一步阻碍构效关系的深入解析。这些问题共同导致基础研究成果向工程应用的转化效率低下, 凸显研究范式革新的迫切需求。

自 McCarthy 于 1956 年首次提出“人工智能(artificial intelligence, AI)”概念以来, AI 技术已发展成为推动科学研究范式变革的重要驱动力^[1,2]。2024 年诺贝尔物理学奖^[3,4]与诺贝尔化学奖^[5,6]的授予, 凸显了 AI 在复杂物理系统建模与大分子结构预测中的革命性贡献。

AI 的核心优势体现在其处理高维复杂数据的能力, 通过解析多维特征间的非线性关联, 可加速高精度模拟过程以阐明构效关系^[7,8], 并实现重要目标体系的高效预测与筛选^[9,10]。在合成化学领域, AiZynthFinder^[11]、AIDDISON^[12]与 SYNTHIA™^[13] 等 AI 辅助合成系统通过优化实验路径规划, 可智能推荐目标材料的最优合成策略。当前, AI 与化学研究的深度融合正推动研究范式转型, 展现出巨大的发展潜力。

作为 AI 技术栈的关键组成, 大语言模型(large language models, LLMs)通过架构创新推动科研协同范式升级。以 GPT-4o 为代表的多模态架构突破, 通过跨模态数据解析与智能体协作框架构建, 显著提升复杂科研任务的自主决策能力。典型应用如 ChemCrow 系统^[14], 依托 GPT-4 实现有机合成路径的智能规划与动态优化, 有效衔接理论预测与实验验证环节。国内

引用格式: Xu J, Chen L, Jiang J. Data-intelligent-driven exploration of robotic chemist systems. *Sci Sin Chim*, 2025, 55: 1606–1622, doi: 10.1360/SSC-2025-0093

DeepSeek 团队开发的 R1 模型创新性采用基于强化学习的调优策略, 实现模型参数自动微调与高效压缩, 显著降低了 LLMs 开发成本。这些进展标志着 LLMs 正从辅助工具演变为驱动实验设计、数据解析与资源调度的核心智能中枢。

有效的 AI 驱动方法需以大规模高质量结构化数据为基础构建高可靠性预测模型, 而传统实验数据普遍存在非标准化、碎片化与可重复性不足等缺陷^[15,16]。在此背景下, 自动化机器人平台迅速发展, 能以标准化、高通量方式生成高质量实验数据^[17~20]。进一步地, 此类平台与 LLMs 及 AI 算法的深度融合形成协同效应: 不仅实现常规实验任务的自动化执行, 更可完成复杂决策流程构建、合成方法优化及实验路径规划等智能功能^[20~27]。利物浦大学 Cooper 团队^[23]开发的移动化学家平台, 通过贝叶斯优化算法自主完成高通量光催化材料筛选, 其筛选效率显著优于传统的人工方法。Cronin 团队^[25]开发的 Chemputer 平台通过整合文献解析、方案设计、有机合成与表征检测, 展现出化学合成的全流程自动化能力。Google DeepMind 研发的 A-Lab^[21,27]集成计算模拟、文献挖掘、机器学习与主动学习算法, 通过自主规划机器人实验与实时结果解析, 攻克了固态无机粉末处理与表征难题。Boiko 等构建了 Coscientist^[28], 这是一个由 GPT-4 驱动的人工智能系统, 能够通过互联网和文档搜索、代码执行和实验自动化半自主地设计、规划和执行实验。

当前智能化学研究面临三个关键挑战: 理论模拟与实验验证的线性工作流导致数据反馈滞后, 阻碍动态优化机制建立; 单一模型架构难以协调多步骤、多站点、多机器人实验任务, 导致复杂实验难以自主执行; 现有实验平台多孤立运行, 跨实验室数据共享与资源协同机制缺失, 制约整体研究效能。为此, 我们提出三重系统性解决方案应对上述挑战: (1) 构建理实迭代学习方法论, 构建“数据生成-模型优化-实验验证”闭环优化体系, 通过机器学习智能推荐高潜力样本, 经计算与机器人实验双重验证动态校准模型, 最终逼近复杂化学体系的全局最优解; (2) 开发大语言模型驱动的多智能体协作架构, 通过高效人机交互解析研究需求, 实现任务智能分解与多模块协同调度, 最终完成自主化实验执行; (3) 设想智能科研云平台, 通过统一设备与数据标准打破系统隔阂, 利用智能任务分配和资源调度整合跨实验室的设备、数据与算力资源。

2 理实迭代学习的方法论

数据驱动的化学研究对高质量数据源提出了系统性要求。理论计算通过简化假设构建低维解空间, 在揭示反应普适规律与支撑高通量模拟方面具有显著优势, 应用量子化学方法可高效解析分子间相互作用的内在机制。然而, 此类方法因多重近似处理而难以精准描述高维复杂体系(图 1a)。实验研究虽能获取物质的高维本征信息, 但受限于数据通量, 易陷入局部最优解而忽略全局最优方案。研究表明, 融合二者优势的协同策略是突破数据驱动研究瓶颈的关键路径^[29,30], 但传统线性工作流中理论模拟与实验验证的分立操作模式^[31~34], 使得二者难以形成动态互馈机制。

为此, 我们提出基于机器化学家的理实迭代学习的方法论^[35](图 1b), 其核心在于构建“数据生成-模型优化-实验验证”的闭环优化链路: 首先通过预训练模型建立理论基准, 继而利用高通量计算验证模型推荐的高潜力样本, 再经机器人实验获取真实数据校准模型参数, 最终实现全局最优解的渐进逼近。该方法论的创新价值主要体现在三方面: (1) 突破理论高维建模与实验低维验证的维度壁垒; (2) 通过数据驱动发现与化学直觉相悖的新规律; (3) 捕获传统方法难以稳定的亚稳态功能材料^[36~38]。

2.1 机器化学家的理论计算

电子行为的基本规律为化学性质预测奠定了理论基础。基于波函数或电子密度的量子化学方法(如 Hartree-Fock 方法^[39]、密度泛函理论^[40]和量子蒙特卡罗方法^[41])已成功应用于从分子体系到晶体材料的多尺度模拟。其中, 密度泛函理论(DFT)通过 Kohn-Sham 方程将多电子问题转化为单电子近似, 并借助交换-相关泛函平衡计算精度与效率^[40], 现已成为高通量计算的核心工具。该方法的广泛应用催生出 Materials Project^[42]、Open Quantum Materials Database^[43]等主流材料数据库, 其标准化数据为材料设计研究提供了系统化数据支撑。

随着计算能力的提升, 高通量计算展现出三重优势: (1) 高效生成大量数据; (2) 建立理论参考框架以指导实验设计; (3) 构建高质量数据集驱动机器学习模型训练。为支撑大规模计算需求, pymatgen^[44]、原子模拟环境(ASE)^[45]等分析工具与 atomate^[46]自动化工作流

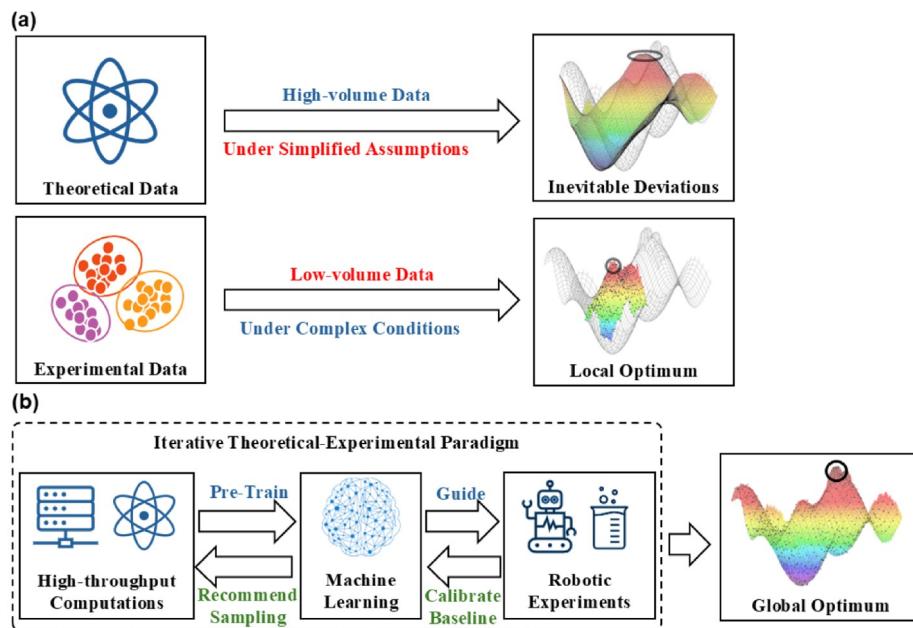


图 1 (网络版彩图)理论或实验方法与理实迭代学习方法论的比较, 突出显示了各自的优点和局限性. (a) 理论化生成大量数据, 但输出为低维度, 受简化假设的限制, 导致不可避免的偏差. 实验化产生高维度、现实世界的结果, 但数据量较小, 受限于局部最优解. (b) 理实迭代学习的方法论示意图, 旨在通过持续反馈进行全局优化^[35]

Figure 1 (Color online) Comparison between theoretical or experimental approaches and iterative theoretical-experimental learning methodology in research highlights their respective strengths and limitations. (a) Theoretical chemistry generates large volumes of data but with low-dimensional outputs, limited by simplified assumptions, leading to inevitable deviations. Experimental chemistry produces high-dimensional, real-world results but yields low data volume, constrained by locally optimal solutions. (b) Schematic of iterative theoretical-experimental learning methodology designed for global optimization through continuous feedback [35].

等工具应运而生, 实现了材料模拟流程的标准化管理. 基于这些基础设施, 高通量计算已在光谱模拟^[47~49]、电池材料开发^[50~54]、热电性能预测^[55,56]等领域取得系列突破.

值得注意的是, 深度学习模型训练通常需要约 10000 个数据点才能保证预测精度^[57]. 针对中小规模数据集, 可通过迁移学习^[58]、多保真度建模^[59]等技术进行数据增强. 通过维护和优化大型数据库、开发专用软件库以及搭建自动化数据处理流程, 生成可直接用于 AI 训练的材料化学数据, 为化学与材料领域的研发效率带来突破性提升.

2.2 机器化学家的智能建模

在化学与材料科学领域, 机器学习(ML)通过构效关系建模机制, 建立物质结构、性质及光谱间的定量关联, 提升理论模型与实验数据的动态耦合效率. 其技术应用主要聚焦于性质预测、方案优化和逆向设计三大方向, 其中化学描述符作为数值化表征分子与材

料特征的核心要素, 可分为以下四类.

(1) 性质描述符: 利用分子量、分配系数(LogP)、氢键供受体数目及拓扑极性表面积等物理化学性质实现物质定量表征. 在药物分子研究中, 此类性质参数因与目标特性的物理关联性常被选作描述符^[60], 在单原子电催化剂领域, 自旋矩描述符通过关联电子自旋态特征, 被证实能有效指导电催化剂设计^[61]. 性质描述符的优势在于其物理相关性赋予模型良好的可解释性, 但需依赖领域专业知识筛选参数, 且对复杂体系中的高阶相互作用捕捉能力有限.

(2) 序列描述符: 简易分子线性输入规范(simplified molecular input line entry system, SMILES)、自引用嵌入式字符串(self-referencing embedded strings, SELFIES)等符号系统通过将分子键合结构编码为字符串, 实现了化学信息与文本处理技术的兼容. 基于 SMILES 描述符的预测模型已成功解析分子能量、电子性质和热力学性质^[62]. 在此基础上, Transformer 和 GPT 架构进一步实现了新药物分子生成^[63~66]. 序列描

述符虽适配自然语言处理(NLP)模型,但由于它们本身缺乏直接的可解释性,理解这些符号化编码所对应的物理化学规律仍存在显著挑战。

(3) 结构描述符: 通过键长、键角、内坐标及分子拓扑等结构特征构建机器学习输入参数。研究表明,包含键长、键角、二面角与拓扑信息的组合结构描述符可有效预测键解离能^[67,68]。结构描述符能精确表征分子几何构型与拓扑特征,但针对大分子或复杂体系的结构参数计算成本显著增加,限制了其实际应用范围。

(4) 光谱描述符: 红外(IR)、质谱(MS)和核磁共振(NMR)等光谱通过实验测量或理论计算的光谱数据构建分子与材料的数值化特征。例如,基于振动光谱(红外、拉曼)的特征峰位移可定量解析表面-吸附物相互作用的吸附能与电荷转移特性^[69]; 二维红外光谱与预训练模型结合实现了蛋白质构型演变的动态追踪^[30,70]; 机器学习驱动的红外光谱分析可实时识别C-C偶联等反应路径^[71,72]; 核磁共振化学位移与理论计算联用则揭示了催化剂电子结构的调控机制^[73],并进一步辅助反应类型的精细分类^[74]。光谱描述符的优势源于其理论计算与实验测量的双重数据来源,使其在模型训练与实验验证中具有广泛适用性; 但光谱特征与物化性质的直接关联常难以建立,需通过复杂分析提取关键特征方能解译。

监督学习依赖标注数据集,在分子性质^[75~77]、反应产率^[78~80]及蛋白质光谱预测^[70,81]中展现显著优势,为理解化学机理及开发新型化合物提供了关键支持。典型模型涵盖线性回归、支持向量机(SVM)、随机森林及神经网络等算法。无监督学习通过K-means聚类等方法,在材料性质关联分析^[82~84]、太阳能电池组分优化^[85]等领域实现未标注数据的规律挖掘。强化学习通过策略迭代优化,显著提升了新材料设计^[86]、色谱分析^[87]等实验效率。

生成对抗网络(GANs)和变分自编码器(VAEs)等生成模型突破传统试错模式,已成功实现半导体材料^[88]、功能多孔材料^[89]等体系的逆向设计。这些技术进展不仅加速了研究进程,更通过降低实验成本推动定制化功能分子与材料的开发,显著推动了化学与材料领域的发展。

2.3 机器化学家的实验探索

为实现从大量机器学习预测中筛选全局最优解,

开发具有高重复性的快速实验系统至关重要。我们构建的机器人化学家平台^[29]由配备6轴机械臂的移动实验操作机器人与二十多个自动化工作站协同组成,支持用户自定义实验方案的执行。工作站按功能分为三类:(1) 自动化合成;(2) 自动化取样;(3) 自动化表征与测试。三类工作站通过机器人调度实现全流程协同运作,完整复现无机化学实验室的核心工作流程。

自动化合成工作站通过可定制的标准化操作模板实现多样化实验需求,具体涵盖液体与固体反应物的精确分配、混合及反应条件调控。液体处理采用泵体驱动的精准加样系统,固体进样则依托自动化称量装置完成。样品混合功能通过超声波、磁力搅拌和摇床三类工作站协同实现,其模块化设计允许研究人员根据溶液或悬浮液的特性灵活选择操作模式。相较于传统固态反应处理系统,该平台通过全流程液体操作显著降低了样品残留风险。

自动化取样工作站集成基底分配与样品制备功能,通过机械臂与轨道系统的空间协同实现电化学测试、IR及X射线衍射(XRD)的标准化样品准备。样品处理流程包含溶剂干燥、基底匹配和参数优化等步骤,其中干燥过程可通过加热或红外辐照方式完成。系统通过动态路径规划算法协调机械臂运动轨迹与工作站操作时序,有效平衡多变量实验需求与机器人操作效率之间的矛盾。

自动化表征与测试系统由光谱分析模块和性能检测模块构成。光谱分析模块支持紫外-可见吸收光谱与气相色谱测试,通过自动进样装置执行连续采样,配合专用软件实现光谱数据的实时采集与基线校正。性能检测模块包含真空封装单元和光催化反应系统,前者完成样品瓶的惰性气氛封装,后者通过程序化光照控制实现多通道光催化测试。所有实验数据与合成参数通过XML文件实现时空关联,形成包含操作日志的标准化数据集。

机器人系统通过全向移动平台与机械臂的协同作业,实现样品在不同工作站间的精准转运。机器人系统基于全向移动平台和机械臂构成,通过激光雷达建图与视觉识别算法实现实验室自主导航与精准操作。研究人员通过可视化界面拖拽配置实验流程,系统自动将方案转化为两类XML指令:机器人端负责样品跨区域转运与任务调度,工作站端执行预设参数的具体实验操作。这种物理操作与流程控制的解耦设计,使机

器人专注样品瓶抓取/运输, 工作站专注加样/测试等专业操作, 通过状态反馈实现工序衔接。同时, 系统引入多机器人多任务调度框架, 基于约束规划算法将复杂实验步骤拆解为可并行操作单元, 自动协调不同机器人之间的任务分配与执行顺序。当有新实验任务加入时, 系统能够实时调整调度方案并优化机器人移动路线, 有效减少多任务并行时的等待与冲突时间^[90]。

传统实验条件优化涉及众多变量的系统筛选, 通常依赖人工试错完成。机器化学家通过自动化并行实验显著提升筛选效率, 其产生的高通量数据为机器学习模型迭代优化提供可靠基础, 构成了理实迭代学习的关键环节(图2a~d)。

2.4 理实迭代的应用方法

我们建立的理实迭代学习的方法论通过机器人实

验与智能算法深度融合, 实现了理论指导与实验验证的动态交互。该方法以自动化实验数据为纽带, 构建“数据生成-模型优化-实验验证”的动态闭环优化机制, 显著提升了新材料开发的效率与可靠性。

胶体纳米晶体(NCs)在电化学、光化学及药物研发等领域具有重要的应用价值, 其性能与微观形貌密切相关, 然而传统合成与表征方法存在资源消耗大、效率低等瓶颈; 针对此问题, 我们通过集成1300余篇文献的数据挖掘、机器人高通量合成与机器学习建模, 开发了面向形貌调控的自动化平台(图3a~d)^[30]: 该工作基于文献数据确定关键合成参数(如浓度范围), 通过原位表征与离线验证构建实验数据库, 并利用机器学习模型解析形貌特征与结构导向剂之间的构效关系, 最终在金纳米晶体与双钙钛矿材料的制备中实现了合成效率优化与形貌可控性的提升。

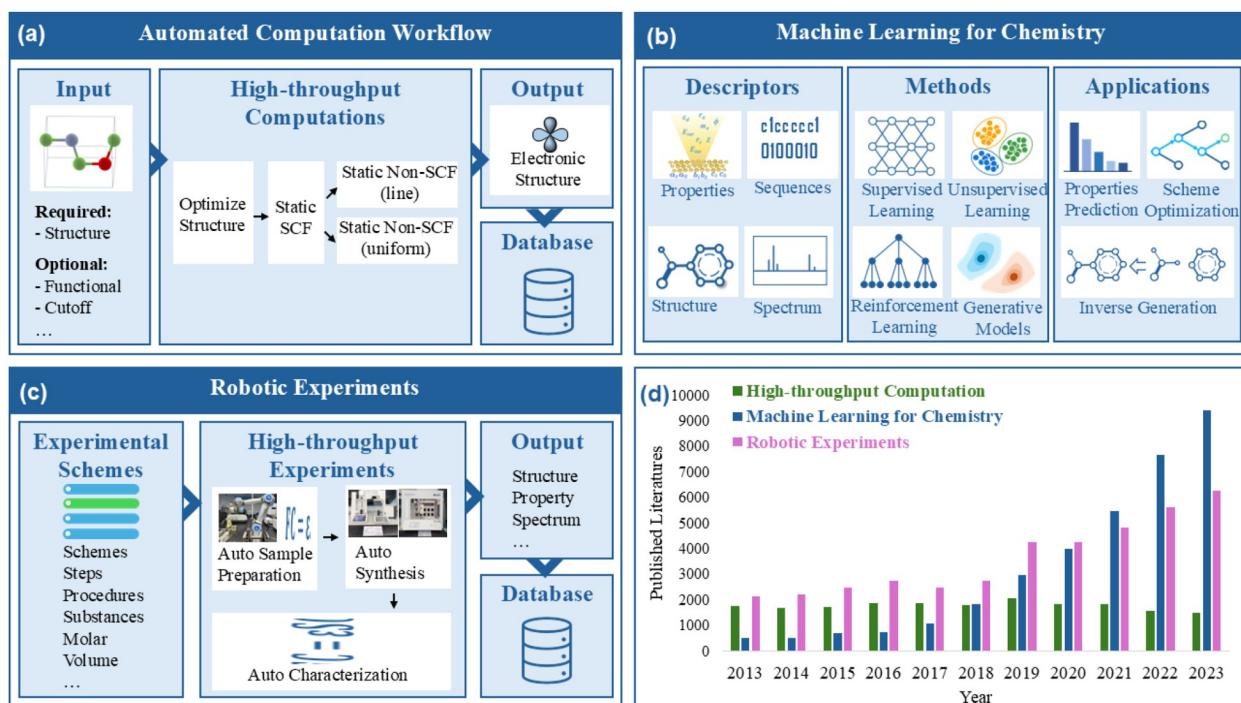


图 2 (网络版彩图)自动化高通量计算、机器学习模型和机器人实验的示意图, 展示了基于过去十年出版趋势, 其应用日益增加。(a) 使用密度泛函理论(DFT)计算电子结构的自动化计算工作流。(b) 化学中机器学习的关键组成部分: 描述符、方法和应用。(c) 通过自动化工作站连接的机器人实验方案和工作流。(d) 高通量计算、化学中的机器学习和机器人实验的趋势, 展示了 Web of Science 中相关关键词('High-throughput Computation/Framework/Screening + Chemistry'、'Machine Learning + Chemistry' 和 'Robotic Experiments')的年度出版趋势(2013~2023年)^[35]

Figure 2 (Color online) Schematic for automated high-throughput computations, ML models, and robotic experiments, illustrating their growing use based on publication trends in the past ten years. (a) The automated computational workflow using DFT to calculate electronic structure. (b) Key components of ML in chemistry: descriptors, methods, and applications. (c) Schemes and the workflow for robotic experiments connected by automated workstations. (d) Trends in high throughput computations, ML in chemistry, and robotic experiments, as shown by annual publications (2013–2023) from Web of Science using relevant keywords (“High-throughput Computation/Framework/Screening + Chemistry”, “Machine Learning + Chemistry”, and “Robotic Experiments”) [35].

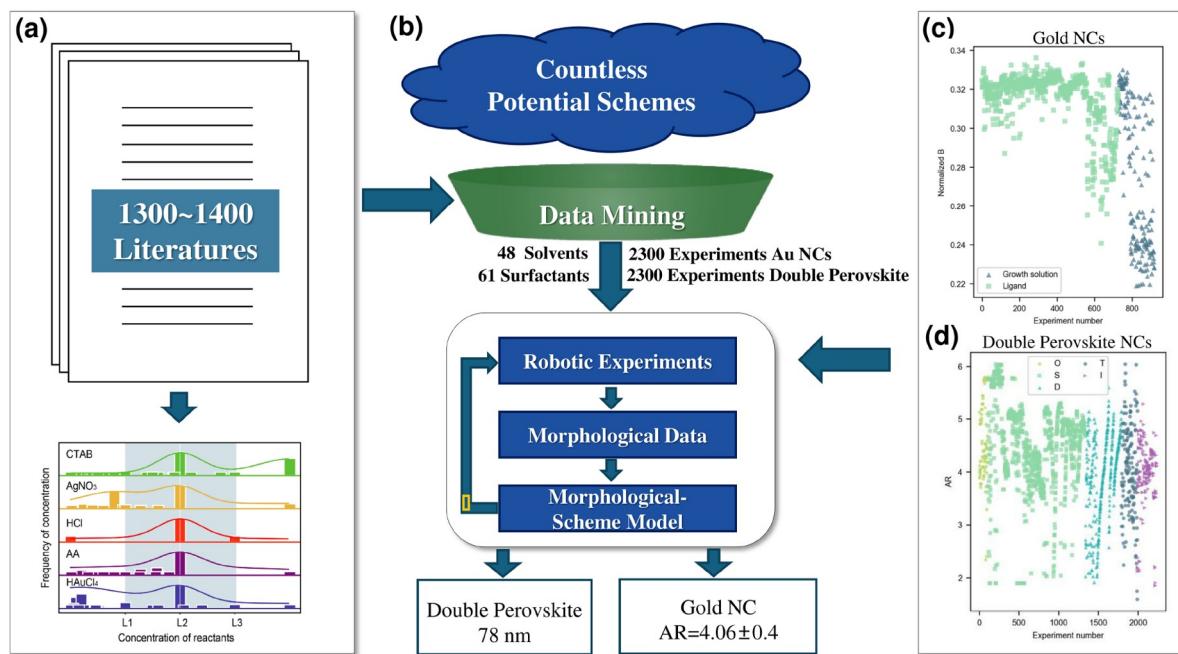


图 3 (网络版彩图)金与双钙钛矿纳米晶形貌控制. (a) 基于1300~1400篇文献的数据挖掘确定关键合成参数(CTAB/AgNO₃/HCl/AA/HAuCl₄), 建立参数边界(L1~L3). (b) 机器人实验与机器学习结合实现优化形态(钙钛矿尺寸78 nm; 金纳米晶纵横比4.06). (c) 双钙钛矿纳米晶的自主机器人合成. (d) 金纳米晶合成的五种实验设计方法对比: 正交法(O)、单/双/三因素法(S/D/T)及逆向设计法(I)^[35].

Figure 3 (Color online) Morphology control of Au and double-perovskite nanocrystals. (a) Literature data mining (1300–1400 papers) identifies key synthesis parameters (CTAB/AgNO₃/HCl/AA/HAuCl₄) and establishes parameter boundaries (L1–L3). (b) Robotic experiments integrated with ML achieve optimized morphologies (78 nm size for perovskite; 4.06 aspect ratio for Au NCs). (c) Autonomous robotic synthesis of double-perovskite NCs. (d) Comparative evaluation of five experimental design approaches for Au NC synthesis: orthogonal (O), single/double/triple-factor (S/D/T), and inverse design (I) [35].

随着人类对地外生存空间与资源需求的增长, 火星低氧环境成为制约长期驻留的关键瓶颈, 基于原位资源制备氧气成为火星移民的首要任务。我们开发的智能化学家系统^[91](图4)通过全自动化流程实现了火星陨石衍生析氧反应(oxygen evolution reaction, OER)催化剂的高效合成: 首先利用激光诱导击穿光谱(LIBS)解析陨石元素组成, 通过机器人平台提取金属氢氧化物前驱体并完成243组比例调控实验测定过电位; 结合高通量分子动力学模拟与第一性原理计算, 获取了29902种金属组分的催化描述符(ΔG_{OH^*} 、 $\Delta G_{\text{O}^*\text{-OH}^*}$ 、 Δq), 构建机器学习模型揭示组分-性能定量关联; 通过迁移学习框架融合理论模型与实验数据预测过电位, 采用贝叶斯优化(B-OPT)算法从300多万种候选配方中筛选最优组分, 所得催化剂在10 mA cm⁻²电流密度下过电位为445.1 mV, 并保持550000 s稳定运行。该工作证实了AI化学家在火星环境中自主合成

功能材料的可行性, 其动态闭环优化机制显著加速了催化剂筛选进程。

3 化学科研智能体

作为AI在化学研究中的前沿探索方向, 大语言模型(large language model, LLM)近年来在自然语言处理、自动化任务执行及跨领域协作等方面取得了突破性进展^[92,93]。通过集成化学与自动化领域的专业工具及平台, LLM代理已成功应用于加速化学发现流程^[14]和优化机器人操作^[94–96]。然而, 面对多步骤、多站点、多机器人协同的复杂实验场景, 具备复杂化学任务执行能力的智能化系统尚未完全实现。为此, 我们开发了基于“多智能体-多大模型”协同架构的ChemAgents系统(图5)^[97]。该方案通过搭建具有明确专业分工的智能体协作框架, 构建基于任务复杂度的分级响应

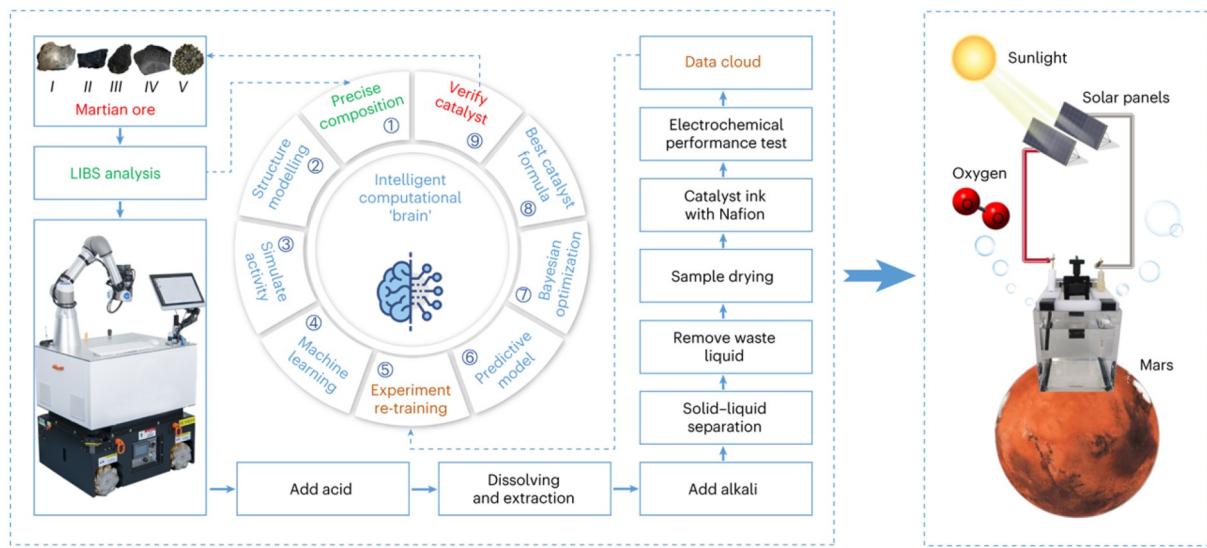


图 4 (网络版彩图) AI 化学家从火星陨石合成 OER 催化剂。通过 LIBS 分析陨石成分, 经酸碱处理与电化学测试(243组实验)自动化提取金属氢氧化物。高通量模拟生成 29902 组理论数据集, 与实验数据联合训练机器学习模型预测过电位。贝叶斯优化(B-OPT)筛选 300 万种配方, 获得最优催化剂(过电位 $445.1 \text{ mV} @ 10 \text{ mA cm}^{-2}$), 并经 550000 s 稳定性测试验证^[35]

Figure 4 (Color online) AI-chemist synthesizes OER catalyst from Martian meteorites. LIBS analysis determines meteorite composition, followed by automated extraction of metal hydroxides via acid/alkali processing and electrochemical testing (243 experiments). High-throughput simulations generate 29902 theoretical datasets, integrated with experimental data to train a machine learning (ML) model predicting overpotentials. Bayesian optimization (B-OPT) screens 3 million formulations, yielding an optimal catalyst (445.1 mV overpotential $@ 10 \text{ mA cm}^{-2}$) validated through 550000-s stability tests [35].



图 5 (网络版彩图) 化学科研智能体 ChemAgents^[97]

Figure 5 (Color online) Chemical research agents: ChemAgents [97].

机制, 实现复杂化学研究任务的全流程自主执行.

3.1 ChemAgents的系统架构

本研究基于开源Llama-3-70B大语言模型构建智能体系统, 并采用华为MindSpore框架完成系统部署。ChemAgents核心架构(图6)围绕任务调度智能体展开: 该智能体通过自然语言接口(由LLM与语义关系处理工具LangGraph共同支撑)与研究人员交互, 精准解析指令语义并生成可调度任务; 同时协调管理文献阅读、实验设计、机器人操作、计算模拟及智算优化五大专业智能体。各智能体通过以下分工实现跨模块任务流的高效协同: 专业文献解析、实验方案设计、机器人指令生成、计算工具调用及智能推演优化, 并通过标准化输出结果支持智能体间数据交互与研究人员理解。而任务调度智能体通过预定义系统提示模板, 运用大语言模型的规划能力, 结合例如基于规则的优先级判定、资源分配策略及任务依赖分析等机制, 实现对多智能体及实验资源的动态调度与优化, 并依托标准化工作流与应用程序编程接口(application programming interface, API)协议构建冲突消解机制, 最终驱动系统完成从需求解析到协同执行的全自动化实验。

各专业型智能体均配备独立LLM实例与专用工具集, 分别对接五大科研基座(图7): 数据库构建化学与材料科学领域的多源知识库, 涵盖学术论文与专利

数据; 实验模板库提供标准化操作流程, 包括分子合成、材料表征及多相反应体系; 自动化实验室配置全流程智能装备集群, 支持合成-表征测试联用; 计算软件库提供多种量子化学计算与分子动力学模拟工具; 智能模型库涵盖物性预测与材料设计算法。科研智能体与科研基座可灵活快速部署于不同实验任务、各类机器人及多种自动化实验设备的实验室环境中, 为能源催化、有机合成等多领域提供了智能化解决方案。

3.2 全流程自动化实验系统的构建与执行

通过对五个专业智能体工作流的细节构建(图7), 我们完成了一套集自动化知识获取, 自动化实验设计、规划、执行, 自动化理论计算、机器学习、迭代优化于一身的全流程自动化实验系统, 下面分别介绍每个智能体的工作流程。

文献阅读智能体集成文献检索(LiteratureSearch)与文献挖掘(LiteratureMine)两大功能模块。前者根据输入关键词在数据库中精准检索文献, 后者采用无监督句法距离分析方法^[98]提取化学物质、物化性质及实验条件等关键参数。其工作流程为: (1) 接收任务指令后, 通过检索工具获取文献标题与摘要; (2) 利用分析工具进行结构化处理并生成统计结果(如高频溶剂使用分布), 为实验设计提供先验知识支撑。

实验设计智能体由协议编写模块与协议评审模块



图 6 (网络版彩图)大模型集群驱动的科研智能体架构

Figure 6 (Color online) Architecture of the multi-agent-driven robotic AI chemist.

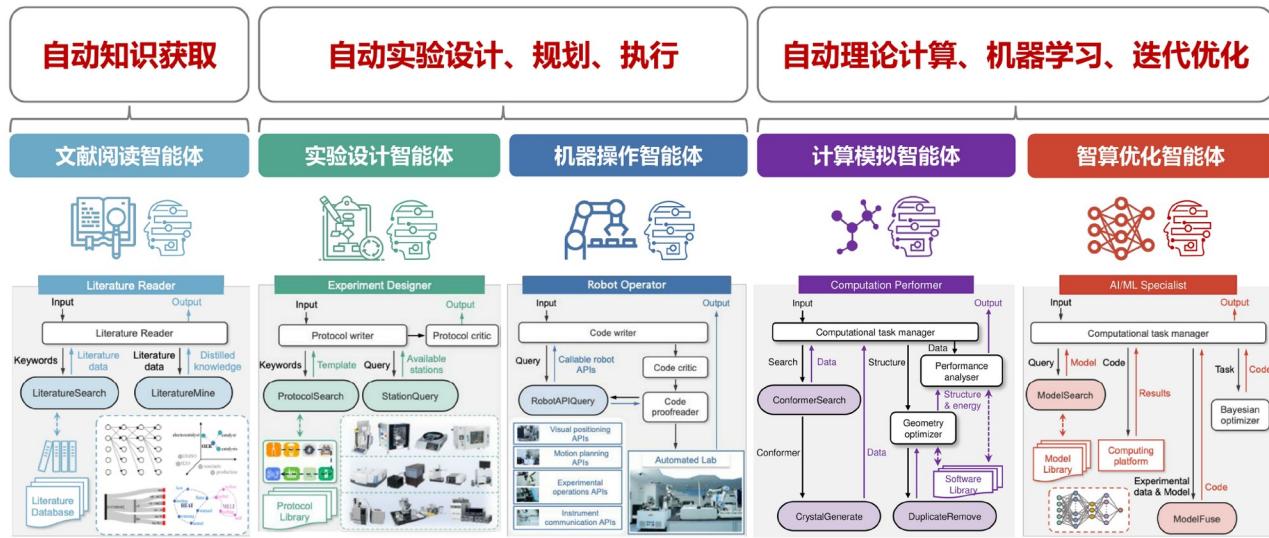


图 7 (网络版彩图)多智能体协同驱动的科研工作流

Figure 7 (Color online) Multi-agent collaborative-driven scientific research workflow.

协同运作。协议编写模块依托LLM技术，通过方案检索(ProtocolSearch)匹配知识库中的实验模板，或借助装置查询(StationQuery)获取实时设备参数；协议评审模块作为独立实例，依据专家规则库对实验流程进行合规性审查。其工作流程为：(1) 接收实验目标描述并匹配最优模板；(2) 若无可用模板，基于装置资源清单自主构建操作序列；(3) 经多维度校验后输出标准化实验程序，最终转化为机器人可执行指令，驱动自动化实验平台的精准执行。

机器操作智能体包含代码编写、代码审查与代码校验三大功能单元。代码编写单元配备设备接口查询工具(RobotAPIQuery)，实时生成初步控制指令；代码审查单元依据安全规则库优化操作逻辑；代码校验单元执行语法校准。其工作流程为：(1) 代码编写单元接收实验方案生成初始代码；(2) 代码审查单元通过多轮迭代优化确保安全性；(3) 代码校验单元完成语法结构与格式校准，输出可直接驱动实验平台执行的标准化代码。

系统针对任务复杂度实施分级处理：对于常规的“合成-表征-测试”序列任务(如测定分子红外光谱)或固定设计空间探索任务(如对某些变量进行全因子实验分析)，系统通过调用实验设计智能体与机器操作智能体即可完成；而对于需数据驱动发现或闭环优化的复杂任务(如广泛搜索空间中的催化剂筛选)，则需要

调用计算模拟智能体与智算优化智能体。

计算模拟智能体由任务管理中枢、构象生成模块与性能分析模块协同构成。任务管理中枢负责接收用户计算任务(如分子结构预测)并协调资源分配；构象生成模块集成构象搜索(ConformerSearch)与几何优化工具，调用标准化算法生成候选结构集；性能分析模块则通过物化性质评估与数据库比对实现结果验证。其工作流程分为：(1) 解析用户需求并启动构象生成链路；(2) 对候选结构进行能量最小化优化，输出稳定构象及其能量参数；(3) 比对验证后输出稳定构型，未达标任务触发迭代优化循环。各模块通过标准化接口衔接，形成“生成-验证-反馈”闭环优化机制。

智算优化智能体包含任务管理中枢、模型检索(ModelSearch)、模型融合(ModelFuse)工具及贝叶斯优化模块，依托深度学习计算平台执行任务。模型检索工具通过关键词匹配定位最优预训练模型；模型融合工具通过扩展网络结构构建融合模型；贝叶斯优化模块可根据人类研究者的任务描述自主生成算法代码。其工作流程为：(1) 接收任务指令后匹配适用模型；(2) 根据需求选择是否需要模型融合或贝叶斯优化策略；(3) 在计算平台执行代码并输出预测结果。

3.3 按需自主化学研究

“多智能体-多大模型”的机器化学家系统可以按

需执行多种实验任务，其执行能力主要受限于自动化工作站的可用性和能力。系统通过任务调度智能体解析研究任务需求，自主生成实验工作流，并通过智能选择与组合功能模块及智能体实现任务目标。为验证系统性能，我们指示机器人AI化学家执行六个具有不同复杂性和难度的实验任务(图8a~f)：前五项任务由实验规划器与机器人操作单元协同完成，最复杂的第六项任务则额外调用文献阅读智能体和智算优化智能体。系统执行流程如下。

“制备与表征”任务。此类任务是机器化学家系统

的核心基础功能。任务调度智能体接收指令后解析实验目标，通过实验设计智能体匹配协议库中的标准化流程，并由机器操作智能体编译为可执行指令驱动实验平台运行。本研究设计三类典型任务验证系统效能：(1) 偶氮苯分子红外光谱分析，系统成功获取N=N键($1600\sim1500\text{ cm}^{-1}$)、苯环C=C键($1500\sim1400\text{ cm}^{-1}$)及C-H键($3100\sim3000\text{ cm}^{-1}$)的特征吸收峰(图8a)；(2) 金属氧化物(ZrO_2 、 ZnO 等)合成与X射线衍射表征，实验结果与标准卡片匹配度超过99% (图8b)；(3) 钙钛矿量子点墨水制备及荧光光谱测试，系统通过精确控制APbX₃

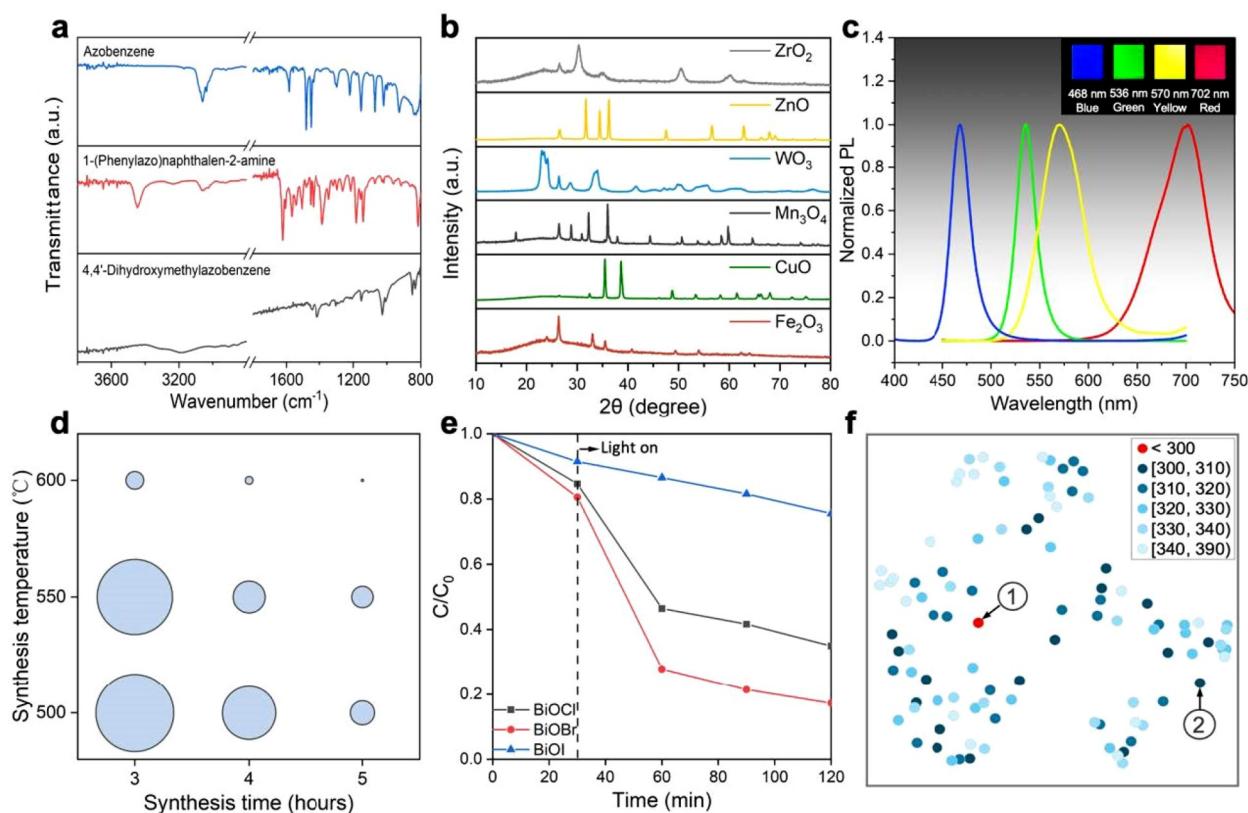


图 8 (网络版彩图)按需自主化学研究。(a) 任务1: 三种偶氮苯衍生物的傅里叶变换红外光谱(FT-IR)。(b) 任务2: 六种金属氧化物的粉末X射线衍射(PXRD)图谱。(c) 任务3: 钙钛矿量子点(PQD)薄膜的归一化光致发光发射光谱(插图为对应颜色样品实物图)。(d) 任务4: 石墨相氮化碳(g-C₃N₄)的析氢反应(HER)活性分布($9.28 \times 10^{-5}\sim2.10 \times 10^{-3}\text{ mmol/g}$), 符号尺寸与HER活性正相关。(e) 任务5: 锰氧卤化物(BiOX, X = Cl/Br/I)对四环素水溶液的光催化降解性能。(f) 任务6: 101种金属有机高熵催化剂(MO-HECs)的二维均匀流形逼近与投影(uniform manifold approximation and projection, UMAP)组分空间投影, 颜色标度表示过电位值(mV); 标签1和2分别标识贝叶斯优化(BO)推荐催化剂与随机筛选最优催化剂^[97]

Figure 8 (Color online) On-demand autonomous chemical research. (a) Task 1: measured Fourier transform infrared spectroscopy (FT-IR) of three azobenzene molecules. (b) Task 2: measured PXRD patterns of six metal oxides. (c) Task 3: normalized fluorescence emission spectra of perovskite quantum dots films displaying different colours, shown as insets. (d) Task 4: measured HER performances (ranging from 9.28×10^{-5} to $2.10 \times 10^{-3}\text{ mmol/g}$) for g-C₃N₄; the symbol sizes are proportional to the corresponding HER performances. (e) Task 5: measured photocatalytic degradation of tetracycline in water by bismuth oxyhalides. (f) Task 6: two-dimensional (2D) uniform manifold approximation and projection (UMAP) embedding of the 101 MO-HECs' compositional space, with symbol colours indicating the value ranges to which their measured overpotentials (in mV) belong; labels 1 and 2 indicate the BO-discovered catalyst and the best catalyst from the random sampling, respectively [97].

组分实现蓝、绿、黄、红四色高纯度发射(图8c). 这些任务均验证了系统在标准化实验流程中的稳定性与可靠性.

“探索和筛选”任务. 该类任务需要更高级的LLM能力, 具备动态调整实验参数与多因素协同分析能力. 以石墨相氮化碳($\text{g-C}_3\text{N}_4$)合成优化为例, 任务调度智能体解析“全因子实验”指令后, 实验设计智能体自主设定温度(500、550和600℃)与时间(3、4和5 h)变量组合, 生成9组合成方案. 实验结果表明, 高温(600℃)导致材料过烧结, 且延长加热时间显著降低析氢性能(图8d). 在铋卤化物光催化降解研究中, 系统通过对比 BiOCl 、 BiOBr 与 BiOI 的带隙特性, 发现 BiOBr 因适中带隙(2.7~3.1 eV)表现最优降解效率, 而 BiOI 因载流子复合速率过快导致活性最低(图8e). 此类任务验证了系统在多变量调控与实验方案设计中的智能化水平.

“发现和优化”任务. 作为系统最高阶功能, 该类任务需文献阅读、实验设计、机器操作与智算优化四类智能体协同工作. 以金属有机高熵催化剂(MO-HECs)开发为例: (1) 文献阅读智能体筛选出Co、Ni、Fe、Mn、Cu为核心金属元素; (2) 实验设计智能体生成100种随机组分合成方案, 机器操作智能体完成制备与过电势测试; (3) 智算优化智能体融合预训练模型与贝叶斯算法, 从十万级组分空间中锁定最优比例(过电势266.1 mV), 较随机筛选组分性能提升显著. 二维UMAP投影显示, 优化组分与随机样本呈显著空间分离, 且相似组分性能差异显著, 印证了传统试错法的局限性(图8f). 该系统通过“实验-计算”闭环迭代, 实现了高性能催化材料的定向发掘, 为复杂材料设计提供了新思路.

4 智能化学家基础设施

目前, 机器化学家领域发展迅猛, 世界各国纷纷部署智能化的化学实验室, 实验室间、不同任务间的信息与资源共享成为关键技术需求. 随着云端系统的加入, 机器化学家实验室正逐步迈入协同作业的新阶段. 通过分布式架构与异步化调度机制, 系统能够对多类型实验流程进行智能拆分与资源适配, 突破单一任务序列与地域限制, 极大扩充了科研探索的可能性. 为应对这一技术演进需求, 我们提出多实验室协同共建的智能科学家基础设施框架, 通过设备协议标准化与

数据云交互规范消除异构系统壁垒, 结合任务分解算法与资源动态调度机制, 构建开放协同的云平台生态体系, 实现跨机构设备集群、多源数据与算力资源的全局优化配置.

我们设想的智能科学家基础设施包含人机交互、机器实验员、化学工作站、化学科学数据库、化学大脑与智能管理决策系统等核心节点(图9), 通过云设施实现协同运作^[199]. 其中, 人机交互模块解析自然语言指令并分发任务, 机器实验员集群执行标准化实验流程, 化学工作站提供模块化反应与检测单元. 这些组件通过云平台标准化API实现跨实验室协议转换, 构建跨地域联动的物理执行网络.

在此物理执行网络的支持下, 化学科学数据库、化学大脑与智能管理决策系统通过云架构实现功能增强: 化学科学数据库整合跨机构实验数据、模拟结果与文献资料, 通过智能清洗模型完成质量评估与格式标准化, 解决多源数据分散存储问题; 基于关联分析技术构建分子结构-性能-合成路径知识图谱, 实现文本、表格、图像等异构数据的语义级互联. 化学大脑依托该知识图谱, 融合符号推理与深度学习架构, 通过化学描述符体系编码专家经验生成可解释假设, 并基于实时实验数据验证预测结果, 形成理实迭代学习的闭环优化机制. 智能管理决策系统构建标准化云平台, 统一设备协议与数据格式, 通过联邦学习框架实现跨实验室模型迁移与隐私保护, 同时集成可视化界面支持远程实验配置与数据分析, 动态调度算法协调多机器人任务序列及工作站资源时序分配, 最终将分散的智能实验室整合为多方共建的云端协同研究网络.

这种新组织形态通过标准化数据协议与分布式资源调度机制, 实现跨区域智能实验室的高效协同运作, 其核心优势体现在三方面: 首先, 云设施支持实验数据实时共享与模型动态优化, 结合多机器人任务编排系统显著提升设备利用率, 加速数据采集、知识发现与技术验证的转化循环; 其次, 自动化实验执行与智能分析工具替代传统科研中的重复性操作, 使研究人员从重复性体力与脑力劳动中解放, 聚焦于创新性科学问题探索与前沿领域攻关; 最后, 通过构建全球化的智能科研网络基础设施, 汇聚全球顶尖实验室资源, 统一知识共享标准与实验能力调度机制, 形成支撑规模化化学创新的跨机构联合创新生态, 增强我国在尖



图 9 (网络版彩图)机器化学家云设施组织形态^[99]

Figure 9 (Color online) Framework of robotic AI-chemist cloud facility [99].

端材料研发等领域的国际竞争力。随着云端协同技术与自主决策系统的持续进化，这种新型科研范式将推动化学研究从离散个体探索向云端聚合的智能网络化协作的跨越发展，为应对全球性科学挑战提供系统性解决方案。

5 总结与展望

本研究融合理论计算、机器学习、自动化实验与云端基础设施等多重技术，构建智能化学研究的全流程支撑体系：基于动态数据反馈的理实迭代范式，通过机器学习推荐高潜力样本并同步校准优化模型，实现复杂化学体系的全局高效搜索；大模型集群驱动的多智能体协作架构，将自然语言指令转化为实验任务分解逻辑，驱动多模块设备自主协同完成操作闭环；跨平台智能云系统通过多实验室共建的标准化协议整合分布式资源，重构科研资源的协同调度模式。三项工作从算法优化、系统集成到生态构建逐层推进，形成“智能决策-自主执行-云端协同”的完整研究范式，为智能化研究提供了从理论方法到工程实践的全链条解决方案。

为更清晰地定位本系统特色，将其与领域内代表

性平台进行了核心功能对比(表1)。

如表1所示，本研究构建的机器化学家平台相较于同类系统，其核心创新在于：通过理实迭代学习实现了理论计算与实验验证的深度协同和闭环优化；并借助大语言模型驱动的多智能体架构，赋予了系统高级别的自主规划与决策智能。这些特性使得本平台能够按需执行从基础表征到复杂优化发现的全流程、多层次科研任务，这与侧重于高通量筛选(如Cooper^[23]系统)、特定测量任务(如Walker^[19]系统)或特定流程自动化(如Lunt^[24]系统)的平台形成了显著差异。这种将深度智能决策与全流程自主执行能力相结合的范式，构成了本系统的核心技术壁垒与创新价值。

智能化学研究体系发展遵循三大演进阶段：第一阶段以独立智能实验室建设为核心，聚焦解决特定领域挑战，但其封闭架构导致数据孤岛与资源利用率低等问题；我们正在快速进入第二阶段，通过云端智能调度系统构建分布式协同网络，突破单点实验室的时空限制，实现跨机构设备资源动态分配与实验数据标准化共享，拓展科学发现的维度；展望第三阶段，我们提出构建全国乃至全球多实验室联合的智能科学家系统网络的愿景，通过深度整合分布式智能化学实验室资源，构建更高效的多机构协作的“数据生成-模型优

表 1 不同智能化学平台对比**Table 1** Comparison of different intelligent chemical platforms

对比维度	中科大机器化学家	Cooper移动化学家	Walker双臂机器人系统	Lunt多机器人系统
技术核心	理实迭代闭环与多智能体的自主决策	移动机器人的高通量实验能力和优化算法	机器人操作与视觉机器学习	多机器人协同
理论与实验协同水平	深度融合: 运用理实迭代闭环, 理论与实验相互指导反馈, 驱动全局优化.	弱: 实验数据驱动优化为主, 理论计算参与不足.	无: 仅关注特定物理性质的实验测量.	初步连接: 实验结果与计算预测匹配, 但未形成迭代.
智能化与决策能力	高: 大语言模型驱动多智能体, 实现复杂任务的自主规划与决策.	中等: 基于贝叶斯优化进行搜索决策, 但高级规划能力有限.	特定任务智能: 卷积神经网络通过视觉数据进行推断, 决策限于特定任务.	偏重执行: 擅长自动化执行预定复杂工作流, 智能决策限于预设流程.
任务范围与复杂度	可按需自主执行制备与表征、探索和筛选、发现和优化三个层次全流程的复杂任务.	聚焦于特定目标(如光催化剂)的高通量筛选与优化.	仅关注非侵入式黏度测量或溶剂识别.	专注于固态化学中的结晶、样品制备和表征流程.

化-实验验证”理实迭代闭环。科学大脑依托跨域知识融合自主生成研究假设, 驱动多模态机器人集群协同执行验证实验, 实时数据通过自适应优化模型动态校准研究方向, 形成自进化研究流程。这种高度互联的协同网络将突破传统实验室的物理边界与资源限制, 实现多实验室任务智能协同、资源全局调度与数据无缝共享, 最终构建跨地域、跨机构、跨学科的多方共建研究生态, 通过规模化协同效应加速复杂科学问题的系统性突破, 为全球性重大挑战提供持续创新的基础设施支撑。

补充材料

本文的补充材料见网络版chemcn.scichina.com。补充材料为作者提供的原始数据, 作者对其学术质量和内容负责。

参考文献

- 1 McCarthy J, Minsky M L, Rochester N, Shannon C E. *AI Magazine*, 2006, 27: 12–20
- 2 Van Noorden R, Perkel JM. *Nature*, 2023, 621: 672–675
- 3 Hopfield JJ. *Proc Natl Acad Sci USA*, 1982, 79: 2554–2558
- 4 Ackley D, Hinton G, Sejnowski T. *Cogn Sci*, 1985, 9: 147–169
- 5 Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. *Nature*, 2021, 596: 583–589
- 6 Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, Ronneberger O, Willmore L, Ballard AJ, Bambrick J, Bodenstein SW, Evans DA, Hung CC, O'Neill M, Reiman D, Tunyasuvunakool K, Wu Z, Žemgulytė A, Arvaniti E, Beattie C, Bertolli O, Bridgland A, Cherepanov A, Congreve M, Cowen-Rivers AI, Cowie A, Figurnov M, Fuchs FB, Gladman H, Jain R, Khan YA, Low CMR, Perlin K, Potapenko A, Savy P, Singh S, Stecula A, Thillaisundaram A, Tong C, Yakneen S, Zhong ED, Zielinski M, Žídek A, Bapst V, Kohli P, Jaderberg M, Hassabis D, Jumper JM. *Nature*, 2024, 630: 493–500

从单机智能化实验到全球协同创新网络的构建, 这种全新的科研范式不仅能显著降低科学的研究的入门门槛, 还将让科研人员从重复的劳动中极大解脱出来, 转而专注于更具创造性和战略性的科学探索。更为重要的是, 这一范式的成熟和广泛应用将通过实验室集群的协同效应加速全球科学发现和技术进步的步伐, 让科研不再受限于地域、资源或时间。我们正站在一个由数据智能驱动科学的新时代入口, 由机器学习与自动化实验开启的创新模式只是序章, 多方共建的云端科研生态支撑的机器化学家实验探索时代已然来临。

- 7 Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, dos Santos Costa A, Fazel-Zarandi M, Sereu T, Candido S, Rives A. *Science*, 2023, 379: 1123–1130
- 8 Qiao Z, Nie W, Vahdat A, Miller Iii TF, Anandkumar A. *Nat Mach Intell*, 2024, 6: 195–208
- 9 Zhang J, Lang M, Zhou Y, Zhang Y. *Trends Genet*, 2024, 40: 94–107
- 10 Yang Z, Zeng X, Zhao Y, Chen R. *Sig Transduct Target Ther*, 2023, 8: 115
- 11 Saigiridharan L, Hassen AK, Lai H, Torren-Peraire P, Engkvist O, Genheden S. *J Cheminform*, 2024, 16: 57
- 12 Rusinko A, Rezaei M, Friedrich L, Buchstaller HP, Kuhn D, Ghogare A. *J Chem Inf Model*, 2023, 64: 3–8
- 13 Klucznik T, Mikulak-Klucznik B, McCormack MP, Lima H, Szymkuć S, Bhowmick M, Molga K, Zhou Y, Rickershauser L, Gajewska EP, Toutchkine A, Dittwald P, Startek MP, Kirkovits GJ, Roszak R, Adamski A, Sieredzińska B, Mrksich M, Trice SLJ, Grzybowski BA. *Chem*, 2018, 4: 522–532
- 14 Bran AM, Cox S, Schilter O, Baldassari C, White AD, Schwaller P. *Nat Mach Intell*, 2024, 6: 525–535
- 15 Hassija V, Chamola V, Mahapatra A, Singal A, Goel D, Huang K, Scardapane S, Spinelli I, Mahmud M, Hussain A. *Cogn Comput*, 2024, 16: 45–74
- 16 Esterhuizen JA, Goldsmith BR, Linic S. *Nat Catal*, 2022, 5: 175–184
- 17 Chatterjee S, Guidi M, Seeberger PH, Gilmore K. *Nature*, 2020, 579: 379–384
- 18 Volk AA, Epps RW, Yonemoto DT, Masters BS, Castellano FN, Reyes KG, Abolhasani M. *Nat Commun*, 2023, 14: 1403
- 19 Walker M, Pizzuto G, Fakhruldeen H, Cooper AI. *Digital Discov*, 2023, 2: 1540–1547
- 20 Yang J, Ahmadi M. *Nat Synth*, 2023, 2: 462–463
- 21 Peplow M. *Nature*, 2023, doi: 10.1038/d41586-023-03745-5
- 22 Abolhasani M, Kumacheva E. *Nat Synth*, 2023, 2: 483–492
- 23 Burger B, Maffettone PM, Gusev VV, Aitchison CM, Bai Y, Wang X, Li X, Alston BM, Li B, Clowes R, Rankin N, Harris B, Sprick RS, Cooper AI. *Nature*, 2020, 583: 237–241
- 24 Lunt AM, Fakhruldeen H, Pizzuto G, Longley L, White A, Rankin N, Clowes R, Alston B, Gigli L, Day GM, Cooper AI, Chong SY. *Chem Sci*, 2024, 15: 2456–2463
- 25 Mehr SHM, Craven M, Leonov AI, Keenan G, Cronin L. *Science*, 2020, 370: 101–108
- 26 Seifrid M, Pollice R, Aguilar-Granda A, Morgan Chan Z, Hotta K, Ser CT, Vestfrid J, Wu TC, Aspuru-Guzik A. *Acc Chem Res*, 2022, 55: 2454–2466
- 27 Szymanski NJ, Rendy B, Fei Y, Kumar RE, He T, Milsted D, McDermott MJ, Gallant M, Cubuk ED, Merchant A, Kim H, Jain A, Bartel CJ, Persson K, Zeng Y, Ceder G. *Nature*, 2023, 624: 86–91
- 28 Boiko DA, MacKnight R, Kline B, Gomes G. *Nature*, 2023, 624: 570–578
- 29 Zhu Q, Zhang F, Huang Y, Xiao H, Zhao LY, Zhang XC, Song T, Tang XS, Li X, He G, Chong BC, Zhou JY, Zhang YH, Zhang B, Cao JQ, Luo M, Wang S, Ye GL, Zhang WJ, Chen X, Cong S, Zhou D, Li H, Li J, Zou G, Shang WW, Jiang J, Luo Y. *Natl Sci Rev*, 2022, 9: nwac190
- 30 Zhao H, Chen W, Huang H, Sun Z, Chen Z, Wu L, Zhang B, Lai F, Wang Z, Adam ML, Pang CH, Chu PK, Lu Y, Wu T, Jiang J, Yin Z, Yu XF. *Nat Synth*, 2023, 2: 505–514
- 31 Greeley J, Jaramillo TF, Bonde J, Chorkendorff I, Nørskov JK. *Nat Mater*, 2006, 5: 909–913
- 32 Broach J R, Thorner J. *Nature*, 1996, 384: 14–16
- 33 Kirklin S, Meredig B, Wolverton C. *Adv Energy Mater*, 2013, 3: 252–262
- 34 Zhu Z, Chu IH, Ong SP. *Chem Mater*, 2017, 29: 2474–2484
- 35 Zhang B, Zhu Z, Li H, Cao J, Jiang J. *CCS Chem*, 2025, 7: 345–360
- 36 Sun W, Dacek ST, Ong SP, Hautier G, Jain A, Richards WD, Gamst AC, Persson KA, Ceder G. *Sci Adv*, 2016, 2: e1600225
- 37 Sun W, Holder A, Orvañanos B, Arca E, Zakutayev A, Lany S, Ceder G. *Chem Mater*, 2017, 29: 6936–6946
- 38 Aykol M, Dwarakanath SS, Sun W, Persson KA. *Sci Adv*, 2018, 4: eaq0148
- 39 Hartree DR. *Math Proc Camb Phil Soc*, 1928, 24: 89–110
- 40 Kohn W, Sham LJ. *Phys Rev*, 1965, 140: A1133–A1138
- 41 Foulkes WMC, Mitas L, Needs RJ, Rajagopal G. *Rev Mod Phys*, 2001, 73: 33–83
- 42 Jain A, Ong SP, Hautier G, Chen W, Richards WD, Dacek S, Cholia S, Gunter D, Skinner D, Ceder G, Persson KA. *APL Mater*, 2013, 1: 011002

- 43 Saal JE, Kirklin S, Aykol M, Meredig B, Wolverton C. *JOM*, 2013, 65: 1501–1509
- 44 Ong SP, Richards WD, Jain A, Hautier G, Kocher M, Cholia S, Gunter D, Chevrier VL, Persson KA, Ceder G. *Comput Mater Sci*, 2013, 68: 314–319
- 45 Larsen AH, Jørgen Mortensen J, Blomqvist J, Castelli IE, Christensen R, Dulak M, Friis J, Groves MN, Hammer B, Hargus C, Hermes ED, Jennings PC, Bjerre Jensen P, Kermode J, Kitchin JR, Leonhard Kolsbjerg E, Kubal J, Kaasbjerg K, Lysgaard S, Bergmann Maronsson J, Maxson T, Olsen T, Pastewka L, Peterson A, Rostgaard C, Schiøtz J, Schütt O, Strange M, Thygesen KS, Vegge T, Vilhelmsen L, Walter M, Zeng Z, Jacobsen KW. *J Phys-Condens Matter*, 2017, 29: 273002
- 46 Mathew K, Montoya JH, Faghaninia A, Dwarakanath S, Aykol M, Tang H, Chu I, Smidt T, Bocklund B, Horton M, Dagdelen J, Wood B, Liu ZK, Neaton J, Ong SP, Persson K, Jain A. *Comput Mater Sci*, 2017, 139: 140–152
- 47 Bagheri M, Komasa HP. *Sci Data*, 2023, 10: 80
- 48 Zheng C, Mathew K, Chen C, Chen Y, Tang H, Dozier A, Kas JJ, Vila FD, Rehr JJ, Piper LFJ, Persson KA, Ong SP. *npj Comput Mater*, 2018, 4: 12
- 49 Chen Y, Chen C, Zheng C, Dwarakanath S, Horton MK, Cabana J, Rehr J, Vinson J, Dozier A, Kas JJ, Persson KA, Ong SP. *Sci Data*, 2021, 8: 153
- 50 Aydinol MK, Kohan AF, Ceder G, Cho K, Joannopoulos J. *Phys Rev B*, 1997, 56: 1354–1365
- 51 Ceder G, Chiang YM, Sadoway DR, Aydinol MK, Jang YI, Huang B. *Nature*, 1998, 392: 694–696
- 52 Van der Ven A, Aydinol MK, Ceder G. *J Electrochem Soc*, 1998, 145: 2149–2155
- 53 Zhou F, Cococcioni M, Kang K, Ceder G. *Electrochem Commun*, 2004, 6: 1144–1148
- 54 He X, Bai Q, Liu Y, Nolan AM, Ling C, Mo Y. *Adv Energy Mater*, 2019, 9: 1902078
- 55 Plata JJ, Nath P, Usanmaz D, Carrete J, Toher C, de Jong M, Asta M, Fornari M, Nardelli MB, Curtarolo S. *npj Comput Mater*, 2017, 3: 45
- 56 Zhu Z, Park J, Sahasrabuddhe H, Ganose AM, Chang R, Lawson JW, Jain A. *npj Comput Mater*, 2024, 10: 258
- 57 Dunn A, Wang Q, Ganose A, Dopp D, Jain A. *npj Comput Mater*, 2020, 6: 138
- 58 Chen C, Ong SP. *npj Comput Mater*, 2021, 7: 173
- 59 Chen C, Zuo Y, Ye W, Li X, Ong SP. *Nat Comput Sci*, 2021, 1: 46–53
- 60 Carracedo- Reboreda P, Liñares-Blanco J, Rodríguez- Fernández N, Cedrón F, Novoa FJ, Carballal A, Maojo V, Pazos A, Fernandez-Lozano C. *Comput Struct Biotechnol J*, 2021, 19: 4538–4558
- 61 Zhong W, Qiu Y, Shen H, Wang X, Yuan J, Jia C, Bi S, Jiang J. *J Am Chem Soc*, 2021, 143: 4405–4413
- 62 Pinheiro GA, Mucelini J, Soares MD, Prati RC, Da Silva JLF, Quiles MG. *J Phys Chem A*, 2020, 124: 9854–9866
- 63 Bagal V, Aggarwal R, Vinod PK, Priyakumar UD. *J Chem Inf Model*, 2021, 62: 2064–2076
- 64 Arús-Pous J, Johansson SV, Prykhodko O, Bjerrum EJ, Tyrchan C, Reymond JL, Chen H, Engkvist O. *J Cheminform*, 2019, 11: 71
- 65 Krenn M, Häse F, Nigam AK, Friederich P, Aspuru-Guzik A. *Mach Learn-Sci Technol*, 2020, 1: 045024
- 66 Kaneko H. *ACS Omega*, 2023, 8: 21781–21786
- 67 St. John PC, Guan Y, Kim Y, Kim S, Paton RS. *Nat Commun*, 2020, 11: 2328
- 68 Huo YY, Jiang J. *Chin J Chem Phys*, 2024, 37: 51–58
- 69 Wang X, Jiang S, Hu W, Ye S, Wang T, Wu F, Yang L, Li X, Zhang G, Chen X, Jiang J, Luo Y. *J Am Chem Soc*, 2022, 144: 16069–16076
- 70 Wu F, Huang Y, Yang G, Ye S, Mukamel S, Jiang J. *Proc Natl Acad Sci USA*, 2024, 121: e2409257121
- 71 He Y, Jiang J, Huang Y, Wang S. *CCS Chem*, 2024, 1–10
- 72 Yang L, Zhao Z, Yang T, Zhou D, Yue X, Li X, Huang Y, Wang X, Zheng R, Heine T, Sun C, Jiang J, Ye S. *Natl Sci Rev*, 2025, 12: nwae389
- 73 Yue X, Song T, Cao J, et al. IR-Bot: an autonomous robotic system for real-time chemical mixture analysis via infrared spectroscopy and machine learning. 2025, ChemRxiv, doi: 10.26434/chemrxiv-2025-8xp04
- 74 Zhang B, Zhang X, Du W, Song Z, Zhang G, Zhang G, Wang Y, Chen X, Jiang J, Luo Y. *Proc Natl Acad Sci USA*, 2022, 119: e2212711119
- 75 Li X, Fourches D. *J Cheminform*, 2020, 12: 27
- 76 Li Z, Jiang M, Wang S, Zhang S. *Drug Discov Today*, 2022, 27: 103373
- 77 Duan Y, Yang X, Zeng X, Wang W, Deng Y, Cao D. *J Med Chem*, 2024, 67: 9575–9586
- 78 Zhu XY, Ran CK, Wen M, Guo G, Liu Y, Liao L, Li Y, Li M, Yu D. *Chin J Chem*, 2021, 39: 3231–3237
- 79 Schwaller P, Vaucher AC, Laino T, Reymond JL. *Mach Learn-Sci Technol*, 2021, 2: 015016

- 80 Żurański AM, Martinez Alvarado JI, Shields BJ, Doyle AG. *Acc Chem Res*, 2021, 54: 1856–1865
- 81 Ye S, Zhong K, Zhang J, Hu W, Hirst JD, Zhang G, Mukamel S, Jiang J. *J Am Chem Soc*, 2020, 142: 19071–19077
- 82 Wen M, Blau SM, Xie X, Dwaraknath S, Persson KA. *Chem Sci*, 2022, 13: 1446–1458
- 83 Zhang L, Wang Z, Wei Z, Li J. *Cell Rep Phys Sci*, 2020, 1: 100269
- 84 Rovira M, Engvall K, Duwig C. *Chem Eng J*, 2022, 438: 135250
- 85 Zhang L, He M. *J Phys-Condens Matter*, 2022, 34: 095902
- 86 Luo S, Xing B, Faizan M, Xie J, Zhou K, Zhao R, Li T, Wang X, Fu Y, He X, Lv J, Zhang L. *J Phys Chem A*, 2022, 126: 4300–4312
- 87 Kensert A, Collaerts G, Efthymiadis K, Desmet G, Cabooter D. *J Chromatogr A*, 2021, 1638: 461900
- 88 Qin C, Liu J, Ma S, Du J, Jiang G, Zhao L. *J Mater Chem A*, 2024, 12: 22689–22702
- 89 Park J, Gill APS, Moosavi SM, Kim J. *J Mater Chem A*, 2024, 12: 6507–6514
- 90 Zhou J, Luo M, Chen L, Zhu Q, Jiang S, Zhang F, Shang W, Jiang J. *Digital Discov*, 2025, 4: 636–652
- 91 Zhu Q, Huang Y, Zhou D, Zhao L, Guo L, Yang R, Sun Z, Luo M, Zhang F, Xiao H, Tang X, Zhang X, Song T, Li X, Chong B, Zhou J, Zhang Y, Zhang B, Cao J, Zhang G, Wang S, Ye G, Zhang W, Zhao H, Cong S, Li H, Ling LL, Zhang Z, Shang W, Jiang J, Luo Y. *Nat Synth*, 2024, 3: 319–328
- 92 Wang L, Ma C, Feng X, Zhang Z, Yang H, Zhang J, Chen Z, Tang J, Chen X, Lin Y, Zhao WX, Wei Z, Wen J. *Front Comput Sci*, 2024, 18: 186345
- 93 Shinn N, Cassano F, Gopinath A, Narasimhan K, Yao S. Reflexion: language agents with verbal reinforcement learning. In: *Advances in Neural Information Processing Systems 36: Proceedings of the 37th Annual Conference on Neural Information Processing Systems (NeurIPS 2023)*. New Orleans, LA, 2023. 8634–8652
- 94 Vemprala S, Bonatti R, Bucker A, Kapoor A. *IEEE Access*, 2023, 11: 21161–21191
- 95 Wu J, Antonova R, Kan A, Lepert M, Zeng A, Song S, Bohg J, Rusinkiewicz S, Funkhouser T. *Auton Robot*, 2023, 47: 1087–1102
- 96 Ahn M, Brohan A, Brown N, Chebotar Y, Cortes O, David B, Finn C, Fu C, Gopalakrishnan K, Hausman K, Herzog A, Ho D, Hsu J, Ibarz J, Ichter B, Irpan A, Jang E, Ruano RJ, Jeffrey K, Jesmonth S, Joshi NJ, Julian R, Kalashnikov D, Kuang Y, Lee KH, Levine S, Lu Y, Luu L, Parada C, Pastor P, Quiambao J, Rao K, Rettinghouse J, Reyes D, Sermanet P, Sievers N, Tan C, Toshev A, Vanhoucke V, Xia F, Xiao T, Xu P, Xu S, Yan M, Zeng A. arXiv: 2204.01691, 2022
- 97 Song T, Luo M, Zhang X, Chen L, Huang Y, Cao J, Zhu Q, Liu D, Zhang B, Zou G, Zhang G, Zhang F, Shang W, Fu Y, Jiang J, Luo Y. *J Am Chem Soc*, 2025, 147: 12534–12545
- 98 Zhang B, Xiao H, Ye G, Song Z, Han T, Sharman E, Luo M, Cheng A, Zhu Q, Zhao H, Zhang G, Wang S, Jiang J. *J Phys Chem Lett*, 2023, 15: 212–219
- 99 Chong Y, Feng S, Wang S, Jiang J. *Bull Chin Acad Sci*, 2024, 39: 41–49 (in Chinese) [崇媛媛, 冯硕, 王嵩, 江俊. 中国科学院院刊, 2024, 39: 41–49]

Data-intelligent-driven exploration of robotic chemist systems

Jincheng Xu, Linjiang Chen, Jun Jiang^{*}

State Key Laboratory of Precision and Intelligent Chemistry, University of Science and Technology of China, Hefei 230026, China

**Corresponding author (email: jiangj1@ustc.edu.cn)*

Abstract: In recent years, the deep integration of artificial intelligence and automation technologies has opened new pathways for revolutionizing chemical research paradigms. This paper systematically constructs an intelligent chemistry laboratory system that synergizes theoretical computation, machine learning, automated experimentation, and cloud-based infrastructure. By implementing a “data generation-model optimization-experimental validation” theory-experiment iteration mechanism, we achieve dynamic closed-loop optimization of research workflows. Innovatively, we introduce a large language model-driven multi-agent collaborative architecture that enables end-to-end autonomous operations from natural language instruction parsing to cross-device task scheduling. The paper elaborates on hardware integration solutions, multimodal data fusion strategies, and cloud resource coordination mechanisms, with case validations including Martian meteorite catalyst development and high-entropy material screening demonstrating the system’s global optimization capabilities. Finally, we provide forward-looking insights into infrastructure development and collaborative innovation models for intelligent chemical research networks.

Keywords: intelligent chemistry, robotic chemist, iterative theoretical-experimental learning, large language models, intelligent chemist infrastructure

doi: [10.1360/SSC-2025-0093](https://doi.org/10.1360/SSC-2025-0093)