科学通报 2026 年 第 71 卷 第 23 期: 5546 ~ 5563
人工智能-能源材料交叉前沿专题 评 述
《中国科学》杂志社 SCIENCE CHINA PRESS

基于本体工程的大语言模型驱动化工研究范式革新

何福鹏[1], 顾紫怡[1], 顾欣颖[2], 齐弼卿[2], 白磊[2], 周天航[1,3*], 李玉强[2*], 徐春明[1]

1. 中国石油大学(北京), 重质油全国重点实验室, 北京 102249
2. 上海人工智能实验室, 上海 200023
3. 中海储能科技(北京)有限公司, 北京 102200
* 联系人, E-mail: zhouth@cup.edu.cn; liyuqiang@pjlab.org.cn

2025-11-16 收稿, 2025-12-24 修回, 2025-12-25 接受, 2025-12-26 网络版发表
新一代人工智能国家科技重大专项(2025ZD0121802)、上海市"通用人工智能大模型"基础研究专项、国家自然科学基金(22308376)、黄河三角洲学者(DYRC20230107)和AI4S攀登者计划资助

摘要 化工研究横跨分子至系统的多尺度层级, 具有高度复杂性与非线性特征. 传统依赖经验与迭代试验的方法, 已难以满足绿色化、智能化和可持续发展的新需求. 本文综述了以大语言模型为核心的"四维一体"智能化SOTA研究范式, 系统阐释其在重构化工研究体系中的实现路径. 该范式以大语言模型为认知中枢, 整合知识引擎、分子推理、过程模拟与智能决策四个核心维度: 在知识层面, 结合检索增强生成与知识图谱技术构建动态知识库, 实现文献、专利与实验数据的语义统一; 在分子层面, 融合生成与推理模型, 加速新分子和新材料的发现与性能预测; 在过程层面, 引入物理信息神经网络与数字孪生技术, 构建机理-数据融合的跨尺度建模体系; 在决策层面, 以本体工程为语义基石, 结合强化学习构建可解释、可追溯的知识驱动闭环. 上述四个维度协同构成了贯通"分子—过程—系统"的智能化认知链条, 为化工研究由经验驱动向智能推理驱动的范式转型提供了系统的理论支撑. 本文进一步提出"本体驱动的多智能体协同系统"概念, 展望了化工研究从单点智能向体系级协同和知识级自治演进的方向, 为实现数据、模型与决策的语义互操作提供基础, 并为绿色化工与智能制造的深度融合奠定方法论与战略支撑.

关键词 大语言模型, 智能化SOTA研究范式, 本体工程, 化工智能

能源化工领域正面临绿色转型、工业规模化创新及多学科协同等前所未有的挑战[1]. 尽管传统化学工程方法在过去数十年中取得了重要进展, 但随着新技术的快速迭代与产业需求的持续升级, 其内在局限性逐渐显现, 主要体现在知识获取滞后、分子与材料设计效率不足以及复杂机理难以与多源数据协同融合等方面[2]. 这些局限性并非仅来自方法不足, 而是来源于化工研发本身所具有的高维度、多尺度、强耦合的科学本质. 从化工研发的实际工作流程来看, 传统范式在多个关键环节存在难以突破的效率约束: (1) 在文献筛选与知识获取阶段, 研究仍依赖人工阅读与经验归纳, 面对指数级增长的文献与专利数据, 难以实现高效、系统化的知识提取与更新[3,4]; (2) 分子与材料设计效率不足, 传统方法主要基于经验驱动, 缺失智能化推理和自动化筛选机制, 导致高昂的试错成本, 并难以发掘潜在的创新材料或反应路径[5,6]; (3) 三是机理与数据的整合难题, 化学反应机理往往复杂且高度耦合, 现有的数值模拟方法与实验技术难以从理论和实验数据中同步提取全面、准确的机理信息, 导致理论与实践脱节, 无法构建高效的反馈与优化机制[7]; (4) 在过程建模与仿真环节, 传统CFD(computational fluid dynamics, CFD)、MD(molecular dynamics, MD)与机理模型虽具有较高

引用格式: 何福鹏, 顾紫怡, 顾欣颖, 等. 基于本体工程的大语言模型驱动化工研究范式革新. 科学通报, 2026, 71: 5546–5563
He F, Gu Z, Gu X, et al. Ontology engineering–based large language model revolution in chemical research paradigms (in Chinese). Chin Sci Bull, 2026, 71: 5546–5563, doi: 10.1360/CSB-2025-5792

© 2025《中国科学》杂志社

物理可信度, 但计算代价高昂, 难以支撑复杂工况下的快速预测与实时优化. 这些核心科学难题构成了传统化工范式难以逾越的瓶颈, 迫切需要引入“可计算知识”和“可泛化推理”的新方法体系予以突破[8]. 在此背景下, 人工智能(artificial intelligence, AI), 尤其是以大语言模型(large language models, LLMs)为代表的模型体系, 为突破上述瓶颈提供了新的方法论基础与技术可能[9]. LLMs在语义理解、跨领域迁移、复杂推理以及多模态数据处理等方面展现出显著优势[10], 能够构建从知识获取、逻辑推断到多模态融合的端到端智能化支撑体系. 这不仅为化工研究注入了创新驱动力, 更拓展了传统研究方法的可行性边界[10]. 由此, 化工研究正经历从“模拟驱动”向“智能驱动”的范式转型, 并在绿色低碳发展目标与工业化大规模应用之间构建起更高效的协同转化机制.

基于上述背景, 本文提出并系统综述了一套贯穿化工研发全流程的“四维一体”智能化SOTA研究范式(图1). 该范式以LLMs为核心认知引擎, 通过多层次语义协同与算法—机理耦合构建化工研究的新型智能框架, 涵盖四个关键维度: (1) 知识引擎: 面向化工知识碎片化, 通过检索增强生成与知识图谱构建, 实现文献、专利与实验数据的语义一致化, 形成可计算、可推理的知识体系; (2) 分子推理: 针对经验驱动设计效率低的问题, 依托LLMs与逻辑推理能力, 在可合成性、性质预测及反应可达性等多重约束下, 高效探索超大化学空间, 压缩研发周期, 推动分子与材料设计向自动化与智能化跃迁; (3) 过程模拟: 基于LLMs与数字孪生技术, 融合机理知识与实验数据, 实现跨尺度过程的高保真仿真、参数辨识与动态优化, 为复杂化工体系建立灵活响应和精准预测的模型基础; (4) 结果分析: 针对工业过程多目标耦合与控制困难, 以本体工程为语义基础, 结合不确定性量化与多目标强化学习, 构建可解释、可追溯的智能决策闭环, 并接触自动化/半自动化实验平台实现快速迭代. 通过“四维协同”机制, 该范式在“分子—过程—系统”多尺度上实现知识、模型与决策的有机耦合, 系统提升研发效率、绿色效益与系统鲁棒性, 为能源化工的低碳转型与智能化发展提供新的方法论基础与技术路径.

1 知识引擎: 从碎片化信息到结构化知识

化工领域长期积累了海量学术文献、专利等, 呈现出多源异构、高度碎片化的复杂特征[11]. 传统基于

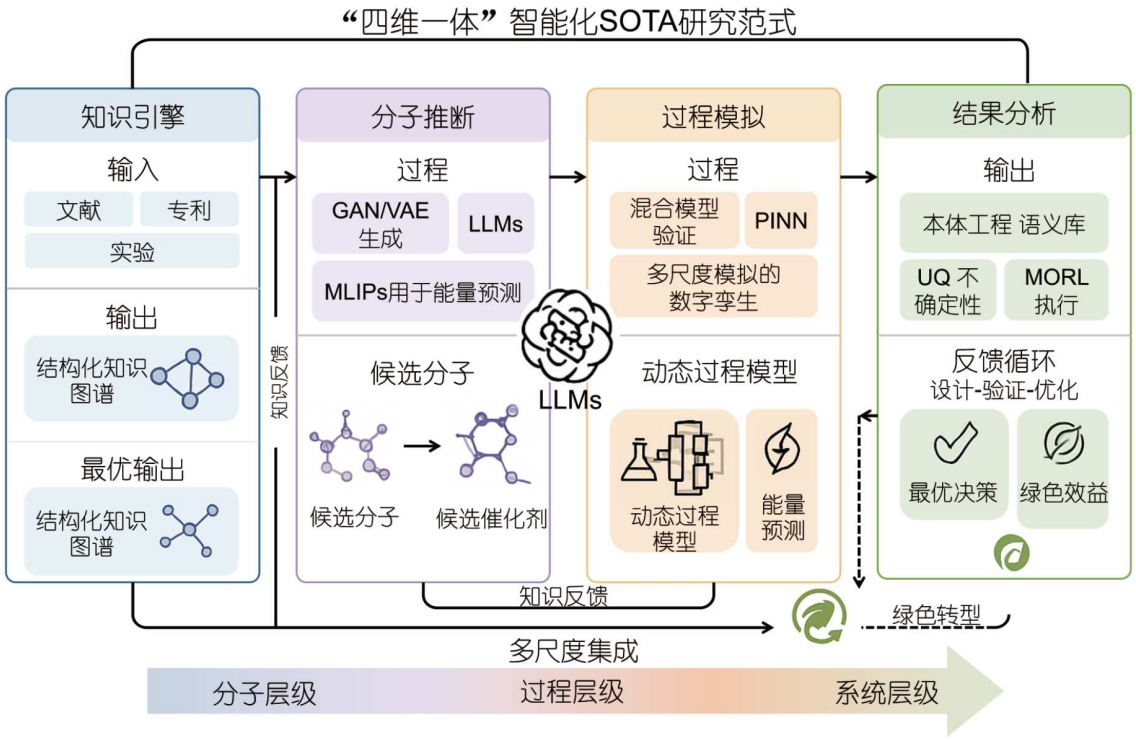


图 1 “四维一体”智能化SOTA研究范式

Figure 1 Four-dimensional integrated intelligent SOTA research paradigm

“关键词—倒排索引”的检索方式难以准确理解文本语义和底层逻辑[12], 导致检索效率低下, 且难以挖掘跨领域的隐性关联知识; 同时, 化工问题通常具有复杂的因果关系和多重约束特征, 传统检索手段难以支撑基于证据的科学决策. 因此, 化工领域亟需从“信息检索”迈向“知识发现”, 即将检索行为从粗放的浅层搜索向基于证据和推理的研究级索引转型.

1.1 知识检索

近年来, 科研检索范式已由“关键词匹配”逐步演进为以大模型为核心的“语义理解—多步推理—证据整合”新模式[13]. 研究者可直接提出问题, 模型能解析问题意图, 自动执行语义扩展与因果关联挖掘, 进而系统性覆盖反应机理、催化剂活性位等维度, 最终以结构化证据链支撑结论生成. 当前, 该范式的核心技术路径主要依赖向量化表征与检索增强生成(retrieval-augmented generation, RAG)技术[14], 相比传统关键词检索, 该方法显著提升了跨文献关联能力和逻辑推理深度. 与之相补充的是知识图谱(knowledge graph, KG)技术, 相关解释如图2所示, 系统化地揭示分子、反应、工艺等多维复杂关联, 实现从碎片化信息到结构化知识的跃迁[15]. 上述能力的提升, 本质上源于大语言模型在上下文建模、深层语义解析与因果推理方面的突破. 通过对化工领域专有知识的学习, LLMs能够执行语义扩展、同义替换及因果关系挖掘等认知操作, 支持跨领域、多维度的综合信息分析. OpenAI推出的Deep Research功能, 具有自主浏览网络、解析多模态数据并生成结构化研究报告的能力. 与传统综述写作相比, 该范式将传统的“文献搜集”升级为“可审计证据网络构建”, 更契合化工领域跨尺度强关联的研究特性. 在此基础上, 新一代智能辅助工具生态迅速发展. 以秘塔AI、Coze等为代表的平台, 集成对话式检索、任务化流程与可解释输出等功能, 可显著降低信息获取成本, 有效提升了学术检索的整体效率. 例如, 进行催化剂筛选时, 能提取诸如反应条件范围、载体效应及副反应路径等多源信息, 并能结合安全数据表及具体的工艺危害进行交叉分析, 为后续的试验方案设计与工艺优化路径提供坚实的科学决策依据.

总体而言, 大模型驱动的知识发现正促使化工研究者由“信息搜集者”转变为“证据组织者”与“决策引导者”. 秘塔AI、Coze等平台为快速理解领域前沿提供高效入口. 同时, Deep Research给出了“如何将检索转化为研究”的可操作框架. 通过语义检索与代理式研究, 研究活动由“找得到”跃迁为“看得懂、用得上、可验证”, 为后续的自动化假设生成与知识图谱推理奠定数据与方法学基础.

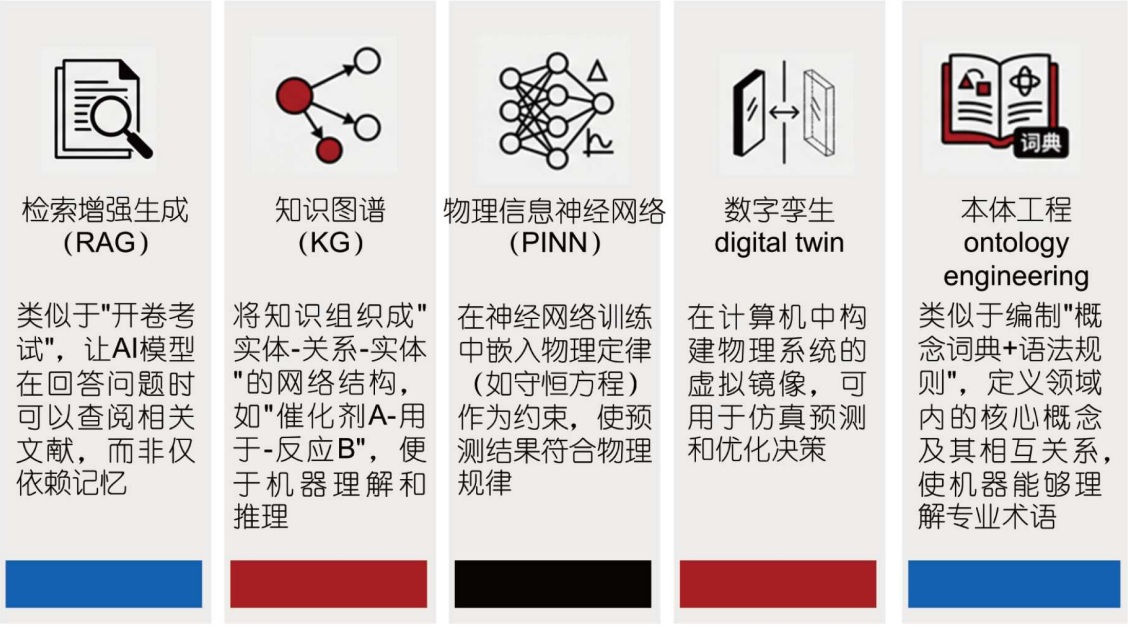


图 2 相关名词解释示意图
Figure 2 Diagram of relevant Noun explanations

1.2 知识管理与知识库构建

知识管理作为化工研究的基础环节，核心任务在于高效整合海量的异构信息资源[16]. 据统计, PubChem数据库涵盖化合物已超过1.1亿, 且Web of Science每年新增化工相关文献约50万篇[17]. 信息规模的快速膨胀使传统依赖人工筛选与归纳的知识管理方式难以满足高效研发需求. 在此背景下，LLMs的出现为知识管理注入了新动力. 凭借其语义理解、多源数据融合与自适应学习能力，LLMs从碎片化信息向结构化知识的自转化，形成了“语义感知—动态推理”的新型管理范式[18].

当前, Zotero、EndNote、ima等主流的知识管理平台正加速融合大语言模型能力，实现由“文献管理工具”向“智能知识中枢”的升级. 一方面，平台普遍具备语义检索、自动摘要、多文献对比与证据回溯能力, 降低信息管理成本；另一方面，部分系统已支持基于RAG的问答和本地知识库构建，为知识提炼与团队协作提供支撑. 例如，Zotero作为开源的参考文献管理工具, 凭借其丰富的插件和数据抓取功能, 在集成大模型后, 能构建基于RAG的问答及证据表, 提升多论文比对与写作引用的准确性；Mendeley的学术社交及多设备同步功能突出，通过与Elsevier数据库的协同，可实现结构化摘要统一；EndNote具有高精度格式化与稿件整合等功能, 结合大模型技术后, 能自动完成证据约束写作和图表说明初稿的生成，有效缩减论文产出周期和格式校正时间；Obsidian作为本地知识库构建工具，通过插件支持笔记同步与高亮回流，接入大模型后实现本地RAG检索与跨笔记语义联想，满足深度知识提炼；Notion则以团队模板管理和项目层级看板著称，引入大模型后可实现周期性AI汇总及证据网络日志自动化, 为小团队规划与研究复现提供保障. 在底层工具层面, 面向文档解析与知识抽取的一类技术工具亦在快速发展，包括多格式文档解析、结构化信息抽取及公式与图像识别等能力. 如表1所示, MarkItDown专注于Word文档的快速批量转换，操作轻量便捷，但功能较单一, 难以处理PDF及复杂视觉布局；Marker表现出较高的综合解析精度, 尤其对代码块和公式处理优异, 但其对GPU资源依赖较大, 计算成本较高. 这类工具为动态知识库构建提供了数据入口，但在复杂版式解析、计算资源消耗及部署成本等方面仍存在一定工程化约束. 总体来看，知识管理正由单一的文献管理工具演化为集文献挖掘、知识抽取、知识图谱构建与智能科研助手于一体的多维智能体系. 与此同时, 该体系仍面临数据异构融合不足、动态更新滞后、模型可解释性有限及深度推理能力受限等突出问题，亟需在关键技术层面实现进一步突破.

综上所述，知识引擎正由传统的人工检索模式向以大语言模型为核心的结构化、智能化体系加速演进. 该新范式不仅显著提升了信息整合与利用效率，也为分子设计、过程优化与智能决策提供了稳定的数据与知识支撑. 知识引擎正由“信息服务工具”转变为驱动“设计—验证—优化”全流程智能化的核心枢纽，为化工领域的智能研发提供关键基础设施支撑.

2 分子推理: 加速新分子与新材料发现

化工研究的核心目标之一，是通过分子层级的理

表 1 工具对比及关键应用场景

Table 1 Tool comparison and key application scenarios

工具	优势	权衡点/劣势	关键应用场景
Unstructured.io	通用性强, 生态集成度高, 支持文件类型广泛	速度相对较慢, 处理特定复杂格式的精度有待提升	多格式数据源接入: 作为ETL流程的起点
MarkItDown	Word处理速度极快, 轻量易用, 针对性强	功能单一, 无法处理PDF、扫描件或复杂视觉布局	标准化Word文档的快速、批量转换
Marker	综合精度高, Markdown输出质量优秀, 对代码块、公式处理良好	依赖GPU资源以发挥最佳性能, 计算成本较高	图文混排复杂的技术手册、产品白皮书
MinerU	数学/化学公式识别突出, 对中文文档优化良好	高精度解析复杂PDF需依赖GPU加速和复杂配置, 牺牲了轻量化部署能力与处理速度	科技、教育、专利类PDF文档
DoclingAI	表格提取精度极高, 对文档层次结构理解深入	更专注于表格场景	包含大量复杂表格的文档
DeepDoc	端到端整合, 对中文场景深度优化, 功能全面	非独立库, 需部署RAGFlow服务并通过API调用	构建高质量中文RAG系统, 如一体化的深度方案

解与设计, 发现具备特定功能的新分子和新材料. 面向绿色化工与智能制造的战略需求, 分子推理正逐步从经验驱动走向数据与模型协同驱动. 以LLMs为核心的新范式, 通过多模态理解、结构生成、能量刻画与知识约束的协同集成, 构建起从“分子构想”到“工程可行”的闭环逻辑链条: LLMs负责高层语义建模与跨模态推理, 分子生成模型与图神经网络建立结构—性质映射关系, 机器学习原子势(MLIPs)实现高保真能量与动力学刻画, KG与RAG则提供证据约束与可追溯性支撑, “可想象的分子”得以转化为“可计算、可合成、可放大”的现实对象. 基于上述框架, 本节将围绕分子生成与多目标优化、反应预测与合成规划、势能面与原子间势、材料与催化性能预测四个方面展开论述, 如图3所示.

2.1 分子生成与反应预测

在指数级扩张的分子设计空间中, 实现“可控、可证、可合成”的分子生成, 关键在于对物化约束与合成可行性的协同建模. 以BatGPT-Chem为代表的化工专用基础模型, 依托百亿级参数与亿级化学实例, 实现“化学序列—自然语言序列”的统一建模, 在逆合成、产物/产率预测等核心任务上展现出零样本与小样本泛化能力[19]. 在此基础上, 研究重点逐步转向“如何引导模型生成满足特定需求的分子”. MOLLM等框架通过上下文学习与提示工程, 将溶解度、稳定性等多目标纳入统一优化过程[20]; ChatChemTS、GeLLM3O与DrugAssist则分别以对话式交互、多属性指令对齐与对抗生成等路径出发, 提升了从“初始构想”到“多属性可行解”的生成效率与多样性. M^{2}LLM与Chem3D LLM工作将空间结构等先验知识耦合进LLMs, 显著增强了语义描述与真实结构的映射一致性[21,22]. 在可控生成方面, Bhattacharya等人通过特定的提示工程, 设计了从“基本生成”到“引导生成”再到“相似度控制”的递进式提示策略, 验证了仅通过自然语言指令, 即可对生成分子的结构特征实现精准、可量化的调控[23]. 该范式正快速向生物分子领域扩展. Chen等人提出的InstructBioMol不仅能理解复杂的生物分子指令, 还能在药物发现、酶设计等任务中生成具有更优结合亲和力或催化活性的分子结构[24]. 大模型推理速度慢已成为限制实际落地的关键瓶颈之一, 其根源在于自回归(autoregressive model, AR)模型依赖逐字生成的串行解码. 为此, 上海AI实验室[25]提出了混合扩散语言模型SDAR(synergistic diffusion-autoRegression), 通过“训练-推理解耦”将AR模型的高性能与扩散模型的并行推理优势相结合, 可在较低成本下将任意AR模型升级为并行解码模型. 在此基础上构建的SDAR-30B-A3B-Sci

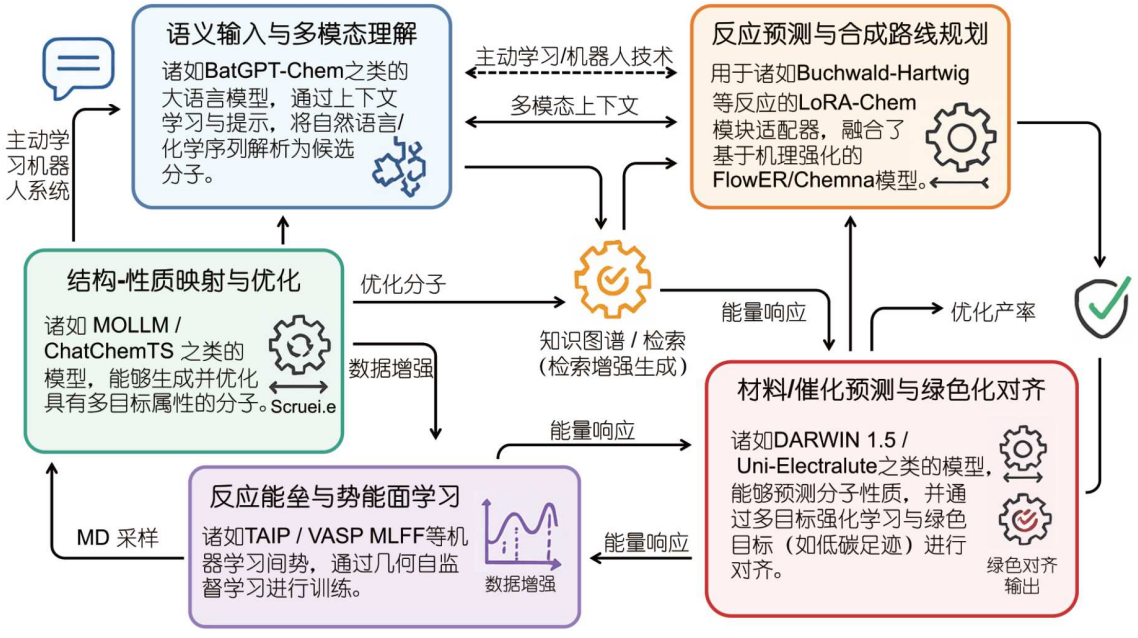


图 3 大模型助力分子推理框架图

Figure 3 Large models empowering the molecular reasoning framework diagram

在化学推理任务中表现尤为优异：在ChemBench上取得超过10%的性能提升，并在GPQA-diamond等高难度基准上同样实现显著增益，验证了其在处理化学反应、分子结构与科学符号推理方面的独特优势，有望共同推动LLMs从“知识驱动”迈向“知识+推理双驱动”的新阶段.

分子生成之后的核心问题在于“如何反应、在何种条件下最优”. 在反应预测层面，如图4(a，b)所示，LoRA-Chem首次实现了单基础模型跨多类有机反应的通用预测[26]. 该方法在保留大模型通用推理能力的同时，为Buchwald–Hartwig、Suzuki–Miyaura、C(sp^3)–C(sp^3)、NiH等反应加入特定参数，使其在专业反应预测精度上可比肩甚至超过专用模型，同时具备显著的迁移能力与更低的训练成本. 针对传统AI在反应预测中易违背物理守恒的问题，Joung等人提出了Flow ER模型[27]，将原子守恒/电子守恒显式嵌入生成过程，使复杂有机反应的预测可靠性获得数量级提升. 面向有机合成的Chemma模型进一步展示了AI驱动实验闭环的潜力[28]：在N-杂环Suzuki反应中，仅用15次实验即将产率优化至67%，体现了“策略生成—机器人执行”的高效协同能力. 结合ChemCrow等工具型代理[29]，从反应设计到实验执行的自动化门槛被进一步降低.

尽管上述模型在分子生成、性质优化与跨模态语义对齐方面展现出显著的技术进展，但必须指出，LLM本质上仍是概率生成模型，在化学空间中不可避免地面临幻觉风险. 生成的SMILES并不等同于分子在物理

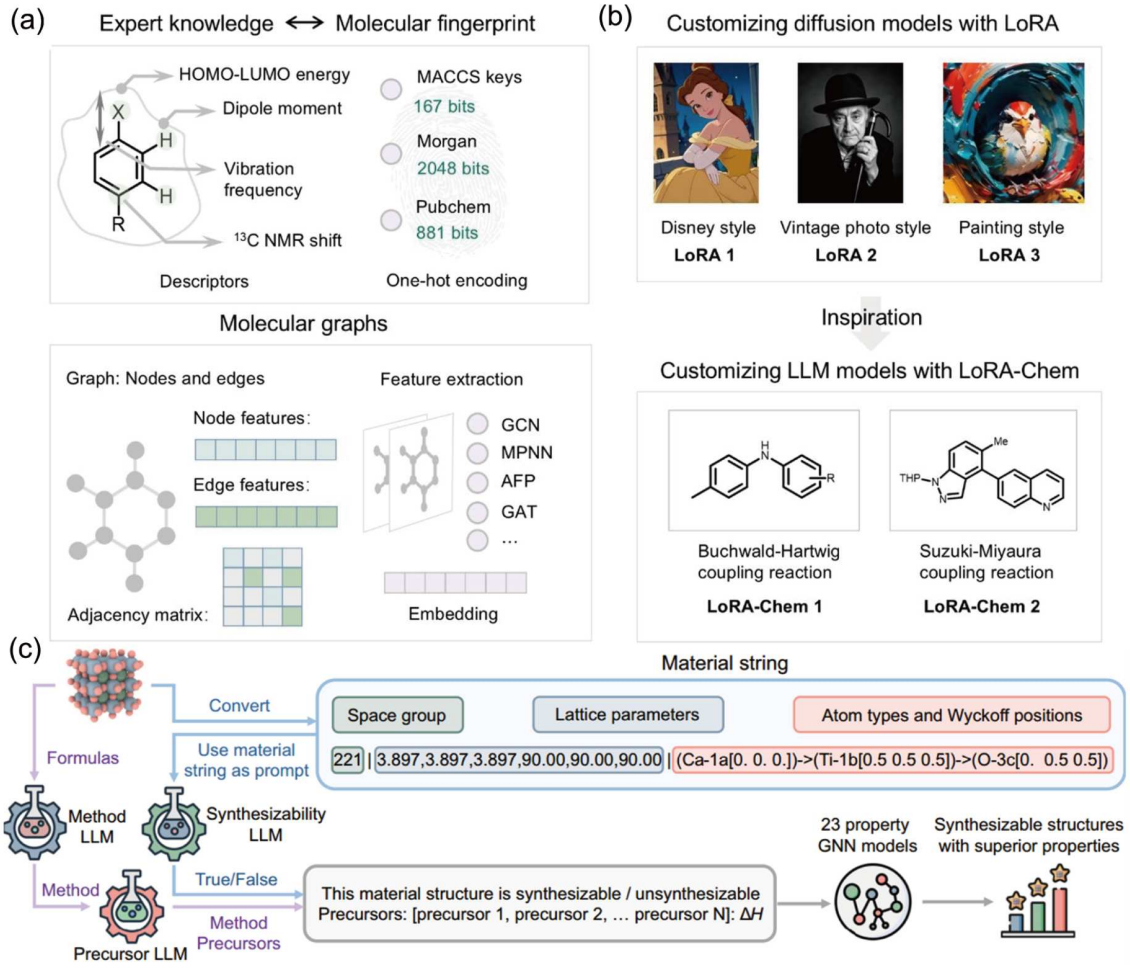


图 4 传统研究范式与基于大语言模型的反应预测流程示意图. (a) 传统有机化学反应预测范式[26]; (b) LoRA模块化的理念[26]; (c) 使用LLMs预测可合成性的总工作流程图[33]

Figure 4 Schematic illustration of the conventional research paradigm and the large language model-based workflow for reaction prediction. (a) Traditional paradigm for organic chemical reaction prediction[26]; (b) concept of LoRA modularization[26]; (c) overall workflow diagram for synthesizability prediction using LLMs[33]

层面稳定、在化学层面可合成, 也不意味着其满足目标性质要求. 提示工程虽然能够改善结构多样性, 但无法从根源上避免违反价键规则、忽略电子效应或生成不可合成结构等问题. 因此, 近年来研究逐步将结构约束、本体知识与物理模型引入到模型中, 以提升生成化学空间的可信度. 一方面, 通过嵌入真实价键信息等约束信息, 使生成候选结构满足最基本的化学可行性; 另一方面, 利用ChEBI、PubChem Ontology等化学本体以及知识图谱对生成分子进行语义一致性验证, 确保其所属类别、基团兼容性与可合成性不违背常识. 此外, 部分工作通过耦合DFT/半经验方法、机器学习原子势(machine-learned interatomic potentials, MLIPs)等快速能量评估工具, 对生成结构进行物理稳定性筛选, 并借助可合成性预测模型进一步过滤不可操作的方案. 总体来看, AI驱动的分子设计正从“结构合法性”向“物理一致性、机理合理性与可合成性”演化, 只有在本体语义与物理化学规则的多层约束下, 才能真正缓解幻觉并为后续反应规划与自动化实验闭环奠定坚实基础.

2.2 分子能量与势函数学习

反应与合成规划要实现可靠落地, 必须建立在高精度的势能面与能量曲线之上. 然而, 经验力场对多体相互作用适用有限, 而从头算方法又难以覆盖真实工况所需的时空尺度. 在此背景下, MLIPs成为连接LLMs语义生成与分子动力学验证的关键“能量引擎”. VASP的MLFF模块已支持on-the-fly主动学习[30], 使力场在动态模拟中随数据分布自适应更新, 兼顾长程相互作用精度与GPU加速的可扩展性. 针对MLIPs在跨分布预测中易出现精度退化甚至模拟崩溃的问题, Zhang等人提出了在线测试时适应原子间势函数(TAIP)策略, 旨在提升模型在测试数据上的泛化能力[31], 并设计了双重自监督学习策略, 利用编码器同时提取全局和局部结构信息, 在推断阶段针对测试样本进行参数更新, 使模型能够自适应提取更匹配的结构特征. 在MD17、ISO17、水体系及电解质溶液等数据集上, TAIP在不引入额外数据的条件下将预测误差平均降低约30%, 并显著提升了分子动力学模拟在极端条件下的稳定性, 有效缓解了训练分布与测试分布之间的偏移问题. 在兼顾精度与效率方面, E^2GNN模型实现了MLIPs在“高精度—高效率”上的平衡[32], 该模型以标量—向量双通道编码替代传统高阶张量表示, 在保持旋转、平移对称性的同时大幅降低了参数规模与计算复杂度. 在多类分子与晶体数据集上, E^2GNN在能量、力预测及真实分子动力学模拟稳定性方面均优于多种主流模型, 验证了其在大尺度高精度模拟中的应用潜力.

在材料与催化场景中, 可合成性是连接理论设计与工程落地的关键瓶颈. 传统依赖声子谱等能量参数的筛选方法精度有限. 如图4(c)所示, Wang等人以“第四范式”材料设计为理念, 构建了贯通结构生成、可合成性判定与工艺路线推荐的智能闭环框架[33]. 该研究以“材料字符串”作为高密度语义表示, 将三维晶体结构信息压缩为模型可直接理解的输入, 并构建由可合成性判定、合成方法推荐与前驱体筛选三类专用大模型组成的CSLLM体系. 模型在测试集上实现98.6%的可合成性预测准确率, 在方法分类与前驱体识别上亦超过90%, 并能在复杂大胞结构上保持约98%的泛化能力. 该团队结合等变Transformer图神经网络CGTNet, 对45632个被CSLLM选出的潜在可合成理论晶体的23

表 2 基于LLMs不同分子推理模型对比

Table 2 Comparison of different molecular reasoning models based on LLMs

模型名称	核心任务	技术特点	优势	局限性
BatGPT-Chem	逆合成、分子生成、产率预测	百亿参数、亿级化学实例	零样本/小样本泛化能力强	计算资源需求高
MOLLM	多目标分子优化	上下文学习、提示工程	多指标同时优化	依赖提示设计质量
ChatChemTS	对话式分子生成	对话交互、蒙特卡罗搜索	用户友好、可交互	生成多样性有限
GeLLM3O	多属性分子生成	多属性指令对齐	属性控制精确	训练数据依赖强
DrugAssist	药物分子设计	对抗自编码	生成多样性高	可合成性验证不足
M^2LLM	多视角分子表征	多模态融合	结构-性质映射准确	3D信息利用有限
Chem3DLLM	3D分子理解	3D结构编码	空间信息保留完整	计算复杂度高
InstructBioMol	蛋白质-小分子设计	跨模态对齐	生物分子适用	领域迁移性待验证
LoRA-Chem	多反应类型预测	模块化低秩适配	迁移灵活、成本低	依赖基座模型质量

项关键物性进行批量预测，形成从“结构生成—可合成性评估—工艺建议—性质预测”的一体化工作流，使人工智能由“高算力计算工具”迈向“具备材料语义理解与决策能力的虚拟化学家”，为加速新材料发现提供了可扩展、可落地的技术范式.

综上所述，以LLMs为核心的AI技术正在系统性重构分子发现路径，构建了一个从“语言构想→结构生成→反应规划→能量验证→可合成性评估”的完整智能推理链条. 通过整合专用模型、高保真模拟引擎、自动化工作流与实验验证平台，该范式不仅显著提升了研发效率，也推动了科研模式由经验驱动向“数据—机理—智能”融合驱动转型. 未来研究将聚焦于内嵌物理化学定律的多模态基础模型构建，并致力于其与自动化实验平台形成紧密的“预测—实验—反馈”闭环.

3 过程模拟: 跨越机理与数据的桥梁

过程模拟是化工研究的核心工具，其本质在于融合物理机理、实验数据与数学模型，对化学过程进行定量描述与预测，为流程优化与装置放大提供理论支撑. 传统的CFD和MD构成了过程模拟的基础，但其计算成本高、求解周期长，限制了其在实时优化等应用场景中的可行性. 在此背景下，融合了AI的过程模拟技术应运而生，形成“数据驱动+机理导向”的混合范式，在保证物理一致性的同时显著提升计算效率. 如图5所

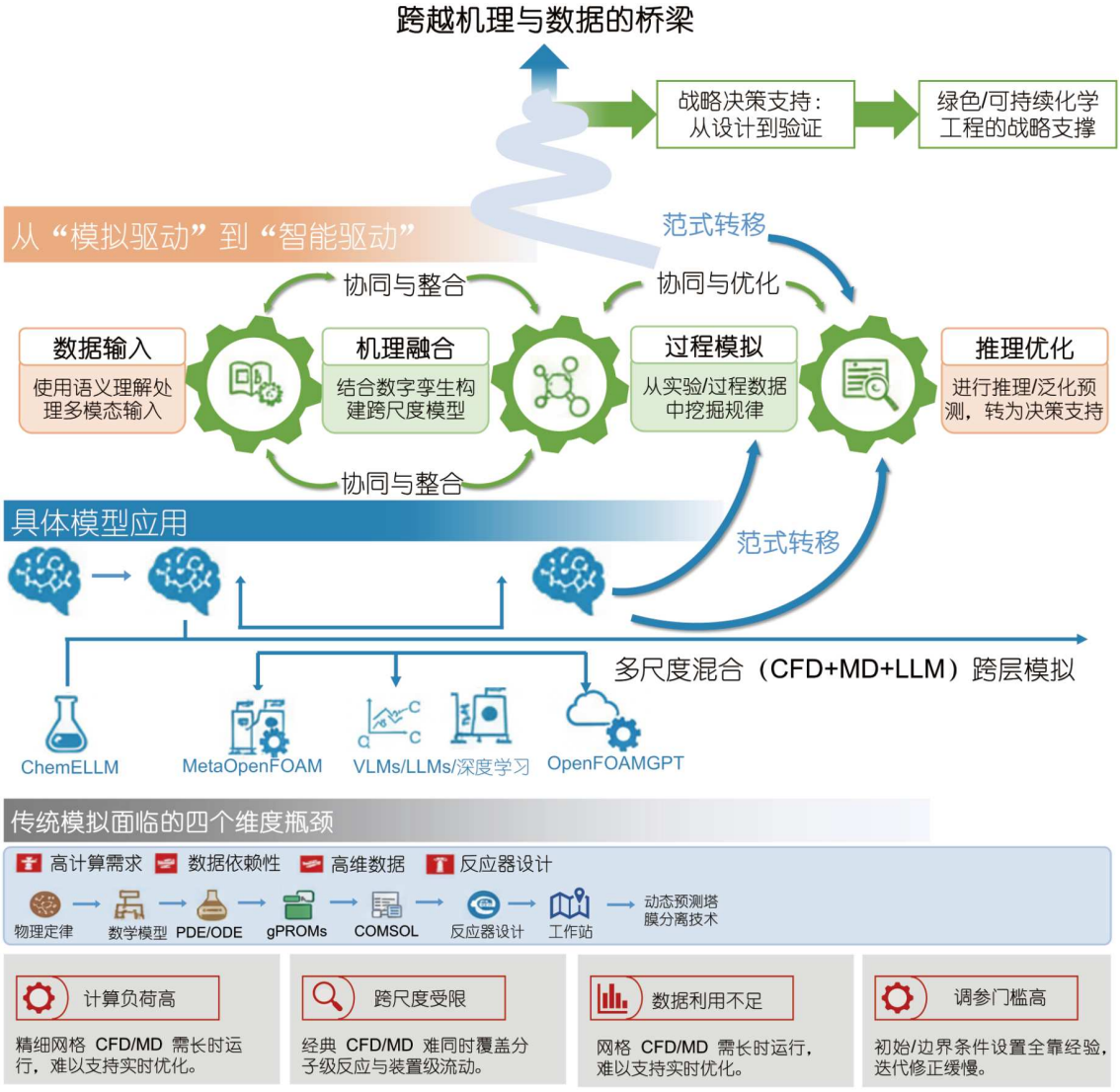


图 5 过程模拟: 跨越机理与数据的桥梁

Figure 5 Process simulation: a bridge connecting mechanisms and data

示, LLMs正推动过程模拟由单一计算工具向智能化决策系统演进, 促进化工研究范式由“模拟驱动”向“智能驱动”转型.

3.1 多物理场层: 从高成本CFD到毫秒级代理模型

针对传统多相流与CFD计算成本高、难以支撑快速参数扫描与工程优化的问题, 研究者提出了以深度学习为核心的代理建模策略. 以三相鼓泡塔反应器为代表的研究中, 构建了CFD-DNN混合建模框架[34], 利用CFD生成的大规模流场数据训练神经网络模型, 对速度场、压力场及气液固滞留率等关键变量进行快速预测. 结果表明, 该代理模型在不同操作工况下均保持较高精度(R^2≈0.97~0.99), 计算速度较传统CFD提升数个数量级, 显著降低了结构设计与操作优化的计算成本, 并具备良好的工程可用性. 在此基础上, Mosavi等人引入ANFIS与CFD的集成建模框架, 用于提升多相鼓泡塔流动预测的精度与泛化能力[35]. 该类ANFIS-CFD集成模型在保持较低计算成本的同时, 提高了对复杂非线性流动行为的描述能力, 体现了推理与数据驱动方法在多相流建模中的互补优势. 此外, 在面向高超声速正常激波后非平衡化学流动等更高维、多物理强耦合场景中, 神经算子方法被引入以突破传统CFD在多尺度、多物理耦合求解中的算力瓶颈. 基于DeepONet构建的算子学习框架, 对包含速度、温度及多组分浓度在内的七个耦合物理场分别建立映射关系, 实现了对复杂多物理场的快速预测[36], 可对速度、温度及多组分浓度等多个物理场进行统一映射, 实现对复杂多物理场的快速预测. 在充分训练后, 该方法相较原始CFD求解器可获得超过五个数量级的加速, 并在未见工况下仍保持良好精度.

3.2 自动化层: 从专家操作到模型驱动仿真

在多物理场代理模型显著降低计算成本之后, 过程模拟面临的另一核心瓶颈来自高保真仿真建模流程本身的复杂性与高度专业化特征. 几何建模、网格划分、边界条件设定与求解器参数配置等环节存在技术门槛, 使CFD与FEM(finite element method, FEM)等工具在应用中具有高成本. 随着LLMs的快速增强, 过程模拟正由“人工主导的建模流程”向“模型驱动的自动化仿真”转变. ChemELLM是面向化工过程工程构建的专用大语言模型, 涵盖3个级别、15个维度和101项任务, 通过ChemEData数据集支持模型开发, 并公开可用, 在过程建模、单元操作与过程控制等任务上展现出稳定的领域推理能力[37], 对未来桥接实验室创新与工业实施、加速化学工程技术进步提供新范式. 在此基础上, Chen等人提出了多代理自动化框架MetaOpenFOAM[38], 该框架利用LLMs实现仅通过自然语言输入完成CFD模拟任务, 包括网格预处理、模拟和后处理等仿真流程, 在涵盖多维流动问题任务上达到了85%的通过率. Pandey等人开发了OpenFOAMGPT[39], 过引入RAG与链式思考机制, 实现了算例搭建、边界条件修正及模型选择的迭代闭环优化, 在单相/多相流、热传输和工程场景中表现出色. 在更广泛的工程设计与有限元分析领域, Elrefaie等人构建了融合视觉语言模型(vision language models, VLMs)、LLM与几何深度学习的多代理设计框架[40], 实现了从概念结构生成到CFD分析的自动衔接, 在空气动力学设计任务中显著缩短了设计周期. 类似地, 在有限元模拟领域, Zhang等人设计了基于LLMs的多代理框架MooseAgent[41], 通过多代理协同实现MOOSE输入文件的自动生成与迭代验证, 在热传导、力学与多物理耦合问题中表现出良好稳定性, 有效降低了复杂FEM模拟对人工配置的依赖程度.

3.3 过程系统层: 从数学规划到语义代理决策

在过程建模与仿真逐步实现自动化之后, 化工研究进一步面临的核心问题在于, 如何在复杂约束、多目标与强耦合条件下, 实现流程系统的高效优化与自主化决策. 传统过程系统工程主要依赖数学规划与机理模型进行求解, 但在可行域未知、目标函数高度非线性或组合复杂度极高的情形下, 往往面临建模困难与计算效率受限的问题. 强化学习与LLMs的引入注入了新动力.

Zeng等人构建了一种面向稳态化工流程的多代理LLM优化框架[42], 实现了传统优化方法难以处理的“无先验可行域”问题. 该框架通过Context Agent对流程描述进行语义解析, 自主生成设备操作上下限; 结合Simulation Agent以稳态模型迭代验证可行性; 并由Suggestion Agent在目标函数驱动下导航参数空间. 模型能够识别加热炉温度、闪蒸条件与能耗之间的因果链关系. 在此基础上, Pajak等人提出了面向流程工业可持续运行的多代理LLMs决策框架[43], 构建了由优化、经济、环境、决策与操作智能体组成的层级结构, 通过语义推理实现生产规划、工厂级设点与装置运行约束

的跨层级协调. 在Gas-Oil Separation Plant案例中, 该方法实现了秒级的决策—验证闭环, 展现了其在可持续运行管理与自主化决策支持中的应用潜力. 除直接参与流程优化之外, LLMs也被用于复杂系统的可解释性分析. Pang等人构建了基于LLMs辅助的多目标建模框架[44], 通过提示工程引导GPT-4自动构建BP(backpropagation neural network)、LSTM(long short-term memory)与CNN(convolutional neural network)等模型结构, 在真空瓦斯油加氢处理过程的应用中, 该方法实现了对产品密度、蒸馏范围与杂质含量等关键属性的准确预测, 并结合Grad-CAM对模型决策依据进行可视化解释.

在具体过程层面, 强化学习被广泛引入换热网络(heat exchanger network, HEN)与复杂传热结构的优化设计中, 用于突破传统数学规划在高度非线性问题上的局限. 针对HEN优化问题, Tan等人提出了融合Q-learning与数学规划的混合强化学习策略[45], 通过构建刻画网络拓扑特征的状态表示与定制化奖励函数, 实现了对网络结构的高效搜索, 在多规模算例中均优于传统优化方法. Lee等人提出了一种将DRL(deep reinforcement learning, DRL)融入换热器拓扑优化的新型设计方法[46], 该研究以微型核反应堆项目中的印刷电路换热器(PCHE)为对象, 形成“智能生成—高精优化—3D打印验证”的智能设计链条, 相较传统拓扑优化, 换热效率提升达14.8%. 此外, Queiroz等人提出了基于Q-learning的壳管式换热器优化方法[47], 通过Q-learning训练的模型能够在较短的时间内进行多次迭代, 显著提高了换热器的设计效率, 为换热器的进一步优化与实际工业应用提供了一个有效的智能优化工具. 除结构设计外, 强化学习也被用于换热系统的动态运行控制. Sun等人提出了基于机理模型与强化学习融合的换热器动态控制方法[48], 通过实时采集换热器两侧流量与温度数据, 对整体传热系数K进行在线识别. 在多工况条件下, 该模型预测温度与实测值的平均相对误差低于5%, 且传热系数变化趋势与实际过程一致. 该模型在不同操作条件下的验证, 证明了其良好的可扩展性与适用性, 为热交换系统的精确控制优化提供了重要的动态参数支持.

综上所述, 过程模拟正围绕“算力—自动化—系统决策”三个层级发生系统性变革: 在多物理场层面, 代理模型与神经算子突破了传统CFD的算力瓶颈; 在自动化层面, 自然语言驱动的仿真框架显著削弱了对人工建模的依赖; 在过程系统层面, LLM多代理与强化学习协同推动流程优化、结构设计与动态控制由数学规划迈向语义代理决策. 三者共同构成了智能过程模拟的完整技术闭环, 推动化工研究范式由“模拟驱动”向“智能驱动”转型. 未来仍需在多尺度统一建模、人机协同机制与小样本可靠学习等方向持续突破, 以支撑智能过程模拟在复杂工程场景中的稳定落地.

4 结果分析: 从数据到智能决策

在知识构建、分子推理与过程模拟的基础上, 最终目标是实现从数据到决策的转化. 然而, 传统决策模式受限于数据壁垒与知识碎片化, 往往优化周期长、适应性不足. 要构建贯通“分子—过程—系统”多尺度的智能决策架构, 核心挑战在于打通不同尺度、不同系统之间的语义鸿沟. 基于此, 本节提出以本体工程为语义基石、大语言模型为认知核心的智能决策框架. 本体工程通过系统化定义化工领域的核心概念及其逻辑关系, 构建机器可理解的统一知识表示体系, 使实验室数据、控制系统数据与制造执行系统数据等多源异构信息, 得以转化为可计算、可推理的结构化知识图谱. 在此基础上, LLMs才能进行有据可依的推理与策略生成, 实现从数据感知到智能决策的端到端转化.

4.1 基于本体工程的智能决策框架

实现从数据到决策的有效转化, 首先需要在语义层面建立统一的理解框架. 本体工程通过系统定义化物质、设备、工艺单元与控制回路等核心概念及其相互关系, 构建起跨系统、跨部门的语义共识基础. 中国石油大学(北京)徐春明院士团队构建了名为Cyber Academia-Chemical Engineering(CA-ChemE)的多智能体协同平台[49], 可视作一个具备自主演化能力的数字学术共同体, 如图6(a)所示. 系统内集成分子设计、中试放大、工程验证、实验研发、理论机制、工艺安全与质量控制等七类领域专家智能体, 具备面向复杂化工问题的发现—讨论—建模—验证全流程协同能力. 为有效消弭跨学科语义障碍, 团队引入具备本体工程知识库与协同策略推理能力的协作智能体(CA), 以本体及其语义关系为核心构件, 对术语体系、知识结构与决策逻辑进行统一建模, 显著增强专家智能体之间的语义对齐与协作效率. 该系统的核心贡献不在于单一智能体的知识扩展, 而在于以本体工程为语义桥梁, 将原本孤立的专业模块嵌入统一本体框架之中, 实现持续互动、自适应调度与策略优化, 由此催生跨领域的创

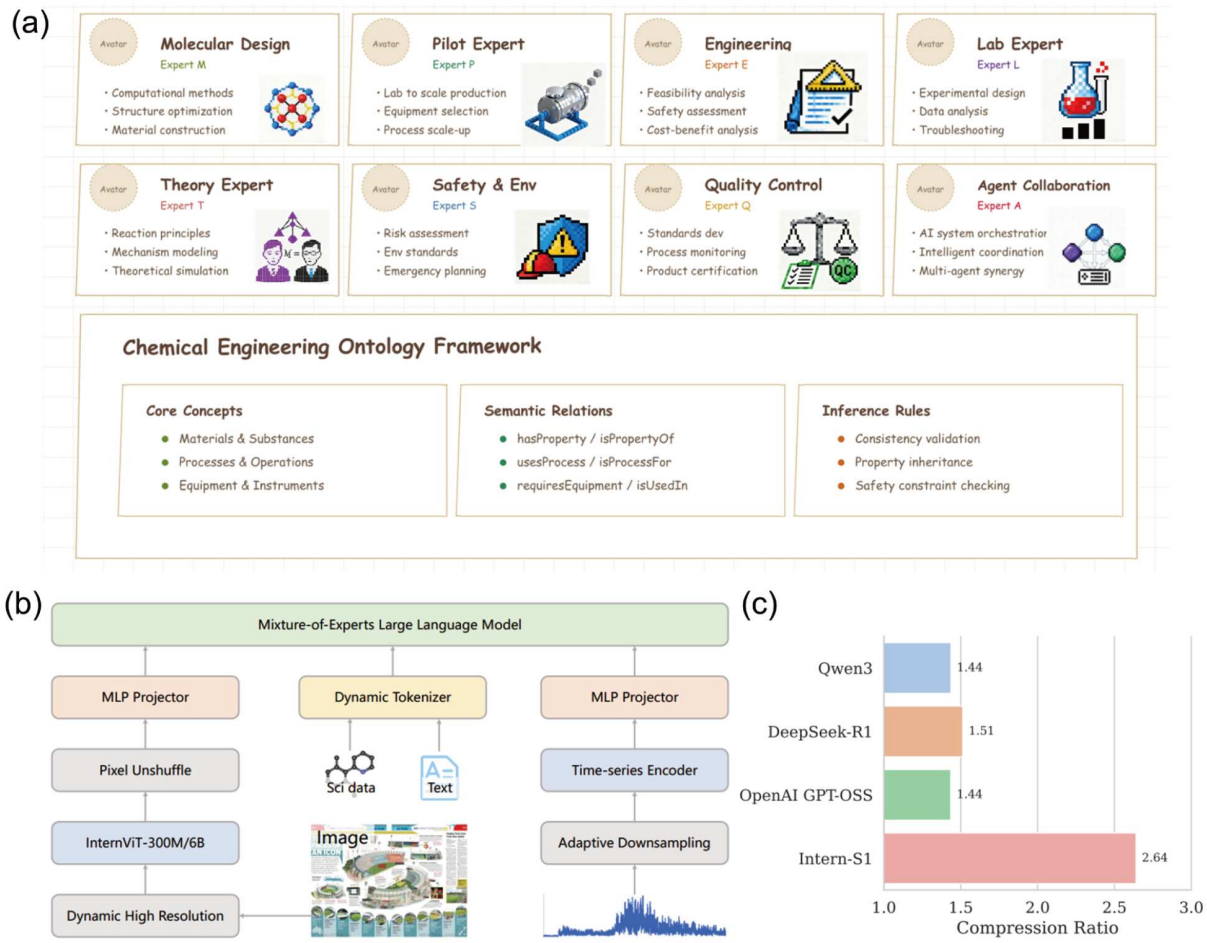


图 6 基于本体工程的智能决策框架. (a) 化学工程本体框架示意图[49]; (b) Intern-S1的架构框架[50]; (c) Intern-S1与其他模型的压缩率对比图[50]

Figure 6 Intelligent decision-making framework based on ontology engineering. (a) Schematic diagram of chemical engineering ontology framework[49]; (b) architectural framework of Intern-S1[50]; (c) comparison diagram of compression ratios among Intern-S1 and other models[50]

新洞见与决策路径. 实践表明, CA-ChemE正依托本体驱动的语义协调机制, 推动化工场景多智能体系统由“任务级执行”迈向“自主化协同”演化, 展示了本体工程作为顶层认知框架的强大能力. 然而, 任何强大的认知框架都依赖于对物理世界高质量、多维度的感知输入. 传统的本体构建往往局限于处理文本数据, 难以消化海量的、多模态的科学数据, 这构成了知识获取的瓶颈. 上海人工智能实验室发布的书生科学多模态大模型Intern-S1提供了关键的底层支撑[50], 如图6(b)所示. 该模型首创“跨模态科学解析引擎”, 可精准解读化学分子式、蛋白质结构等多种复杂科学模态数据, 并具备多项前沿科研能力, 如预测化合物合成路径、判断化学反应可行性等, 真正让AI从“对话助手”进化为“科研搭档”, 扮演了本体工程的“智能感知前端”. 体系化的技术创新为Intern-S1的能力突破提供了有效支撑. 自书生大模型首次发布以来, 上海AI实验室已逐步构建起丰富的书生大模型家族, 包括大语言模型书生·浦语InternLM、多模态模型书生·万象InternVL、强推理模型书生·思客InternThinker等. 如图6(c)所示, 用数据证明了这一设计的优越性. 在处理SMILES数据时, Intern-S1的分词器压缩比达到了2.64, 而其他主流模型(如Qwen3、DeepSeek-R1)仅在1.5左右. 这意味着Intern-S1可以用远少于其他模型的token来表示相同的化学信息, 效率提升超过70%. 这在处理数万亿token的预训练中, 计算成本显著节省.

4.2 智能决策框架在典型化工过程中的落地应用

基于本体工程的智能决策框架已在语义统一与知识推理层面构建了跨系统协同的基础, 但其真正价值需在工艺、装置与控制层面实际落地. 为使结构化知

识转化为可执行策略，体系必须进一步与化工过程的动态行为、操作约束和安全逻辑深度融合，从而实现可量化的性能提升. 本体驱动智能决策并非替代传统过程系统工程，而是作为其“语义+推理增强层”提升优化的自动化、可解释性与适应性. 以下将通过DCS集成(distributed control system, DCS)、报警管理与批次优化三个典型场景展示其在实际化工过程中的落地方式.

(1) DCS集成：语义驱动的操作决策与自动执行. 在DCS层面，本体工程为模型的控制推理提供语义基础，使其不仅能识别装置运行状态，还能在工程约束下生成可执行的控制策略. 本体通过对设备、控制回路、操作指令和工艺约束等概念的形式化建模，LLMs的分析对象由原始信号跃迁为“具有因果关系的语义事件”，从而构建起“系统状态→语义解释→策略生成→指令输出”的闭环控制链路. Pribylskii等人利用本体化的电力系统描述[51]，使LLMs能够理解“负荷上升”“备用电源切换”等语义事件，并据此生成如投入发电机、切除负荷等符号化控制动作，其行为与人类调度员的逻辑结构一致，体现了本体约束下推理链条的可解释性与可控性. Xia等则将本体作为“控制语义层”[52]，统一数字孪生中设备、接口与功能描述，使LLMs能够基于任务语义在DCS/MES之间规划并组合操作步骤，实现跨模块技能编排与自适应控制. 本体在此充当了从高层任务目标到底层服务接口的映射规则集，使LLMs具备“知道能做什么”“应该如何做”的能力. 在本研究提出的架构中，本体工程因此承担三重功能：(a) 将DCS数据统合为具有逻辑关系的语义事件，使LLMs具备可推理的输入；(b) 通过形式化定义操作边界等，本体确保LLMs的策略生成可追溯、可验证；(c) 自然语言推理能够被转译为工业可执行指令. 本体驱动的DCS集成不再是“LLMs对控制系统的附加能力”，而是一种语义增强型控制模式，推动化工控制系统向自主化与自解释方向发展.

(2) 报警管理：基于本体与知识图谱的关联分析与根因诊断. 在复杂化工装置中，报警事件呈现多源多因的特征，传统系统难以对多源报警进行有效归因. 本体工程通过形式化定义语义关系，使报警信号转化为嵌入知识结构的语义事件. 与知识图谱结合，可对信号进行自动聚类等，系统能够从多个报警链中识别实体及其因果关系. 这种机制在典型化工场景中展现出显著优势. 例如，当精馏塔出现泛塔趋势时，常伴随塔压上升、塔底液位波动、再沸器负荷异常等多条报警同时出现，传统系统难以判断这些报警的先后顺序与主因，而基于本体的语义推理可识别“气液负荷失衡→塔板液泛→压降升高→温度偏移”的因果链，从而将泛塔识别为根本原因；对于反应釜升温异常，知识图谱可将催化剂投加误差、搅拌效率下降或放热反应失控等与实时报警关联，实现可解释的诱因判定. Wu等人利用知识图谱-LLM结构识别多源设备报警的根因，诊断精度达97.5%[53]；Wang等人在图结构增强的光网络报警研究中，将拓扑与传播路径编码为知识模块，使根因定位准确率超过86.8%[54]；在催化裂化工业场景下，Sun等利用知识图谱与LLMs的因果推理能力，实现了催化裂化装置异常根因的可靠识别[55]. 总体而言，本体驱动的报警管理体系使LLMs能够在复杂、多源、多因报警环境下实现语义理解与因果整合，将传统“信号响应式”报警系统升级为“基于知识的主动诊断系统”，为工业安全提供关键支撑.

(3) 批次优化：语义增强的质量预测与过程优化. 在间歇化工过程中，工艺变量随时间呈现强非线性、耦合性与阶段性变化，静态模型难以描述其动态演化. 本体工程通过形式化描述批次过程的单元操作、物料属性、设备约束与质量指标之间的语义关系，使批次运行轨迹从原始时间序列转化为具有逻辑结构的“语义化过程”. Shen等人针对间歇及半连续批次过程运行轨迹长度不等、动态行为强非线性的问题，提出了一种融合多变量轨迹相似度与长短期记忆网络(long short-term memory network, LSTM)的批次软测量建模方法[56]，研究表明，以“完整运行轨迹”为建模对象、结合动态深度模型进行软测量，是解决复杂批次过程质量预测问题的有效途径，为轨迹级建模与批次优化提供了重要技术基础. Wang等人针对复杂多阶段批次过程变量非线性等问题，构建了一种融合分阶段LSTM与堆叠自编码器(stacked autoencoder, SAE)的深度学习质量预测框架[57]. 该方法首先按照实际工艺流程将批次过程划分为多个阶段，并在各阶段内分别采用LSTM网络对多变量时间序列进行动态建模，以提取阶段末端的隐含状态特征；随后，通过SAE对多阶段隐含特征进行进一步降维与特征融合，最终建立质量预测模型，在非等长、多阶段批次数据条件下表现出优良的预测精度与鲁棒性. 此外，稀疏特征聚类-多任务学习模型进一步实现任务结构自动聚类，有效应对多品种批次过程的数据不足与变量冗余问题，能在样本数量受限条件

下提升多牌号产品的质量预测精度与模型泛化能力[58]. 批次优化的知识语义层与动态学习层的结合, 使系统不仅能够从运行轨迹中识别偏差源并预测终点质量, 还能生成具有可解释性的优化策略, 为波动性强、风险高、机理复杂的间歇过程提供可自适应、可解释、可验证的智能化优化路径.

4.3 本体驱动的化工决策优势

基于本体工程的智能决策体系正成为化工行业迈向智能化与绿色化的重要驱动力. 本体工程通过形式化定义领域概念、属性及其逻辑关系, 建立统一的语义知识体系, 使实验、装置、控制与管理系统中分散的数据得以语义化整合和结构化表达, 从而为跨部门、跨层级、跨系统的决策提供可计算、可推理的知识基础, 如图7所示. 这种知识驱动的推理方式不仅减少了重复试验与人工判断, 还促使经验知识转化为可迁移、可扩展的决策资源. 在与数字孪生系统结合后, 本体模型进一步拓展了决策的实时性与动态适应能力, 通过语义层映射装置运行状态, 实现基于知识驱动的自适应优化. 总体来看, 本体工程通过语义统一、知识融合、逻辑推理与动态优化的协同作用, 构建了贯通数据、模型与决策的智能化体系, 使化工决策从经验驱动迈向知识驱动与本体驱动的科学化新时代. 值得注意的是, 实现“知识驱动”的路径并非唯一. 除了构建符号化知识体系, 另一条重要的路径是训练领域专用的知识模型. 本研究团队通过广泛收集和整理液流电池相关文献资料, 并结合数据驱动与机理驱动的方法, 在液流电池领域构建了FlowBD大语言模型[59]. 通过在海量专业文献上进行深度训练, FlowBD核心优势在于将非结构化的文本知识内化于模型参数之中, 使其能够针对液流电池储能这一垂直领域的特定问题, 生成比通用模型更精确、幻觉更少的答案. FlowBD可作为高效的自动化知识提取工具, 从海量文献中识别出电解液类型、电极材料改性、流动设计等关键信息, 并将其系统化地整合, 为本体知识库的快速构建与动态更新提供高质量的“知识原料”.

总体而言, 基于本体工程的智能决策框架将分散的知识资源转化为可计算的结构化资产, 通过语义理解、证据推理和闭环优化, 实现了从数据到决策的智能化跃升. 这一变革不仅提升了企业的运营效率, 更为化工行业向绿色化、智能化转型提供了坚实的技术支撑.

5 展望与未来框架: 全流程智能化研究模式

化工研究以复杂性和多尺度耦合为特征, 贯穿从分子设计、过程模拟到系统决策的全链条环节. 本文提出“四维一体”的SOTA智能化研究范式, 系统阐释了大语言模型在化工研究中的协同作用, 涵盖知识引擎、分子推理、过程模拟与智能决策四个维度: 在知识层面, 知识引擎实现了从碎片化文献到结构化知识的语义整合, 赋予科学知识以可计算与可推理的特性; 在分子层面, 分子推理融合生成模型与推理模型, 显著提升了新分子与新材料的筛选与发现效率; 在过程层面, 过程模拟借助物理信息神经网络与数字孪生技术, 促成机理建模与数据驱动的双向融合, 推动工艺仿真向实时化与智能化方向发展; 在决策层面, 智能决策以本体工程为语义基础, 结合强化学习与不确定性量化方法, 构建了兼具可解释性与可追溯性的知识驱动决策闭环. 上述四个维度协同作用, 共同形成了贯通“分子—过程—系统”的完整认知链条, 为化工研究从经验驱动向智能推理驱动的范式转型提供了系统的理论框架与方法论支持.

随着智能化研究不断扩展, 化工领域的研究范式正从局部任务优化迈向跨层级、跨主体的系统级协同. 伴随大模型能力的提升与知识体系的语义化重构, 化工研究正在迈向以本体工程为核心约束的多智能体协同阶段(ontology-driven multi-agent system, OD-MAS). 在该框架下, 统一的本体知识图谱作为语义中枢, 将分子设计、过程模拟、决策优化以及安全与绿色评估等多类智能体连接到同一认知空间中, 使其能够围绕同一研究目标开展分工协作、信息互补与策略生成. 智能体在本体约束下具备语义对齐、自适应学习与策略协调能力, 可在动态情境中实现任务分解与协同推理, 逐渐形成可演化的知识驱动科研生态. 由此, “本体+多智能体”模式成为智能化研究范式的重要发展方向, 推动化工研究从数据驱动的发现向知识驱动的协同创造转变, 并在本质安全、绿色低碳与全流程优化方面展现出新的潜力.

尽管OD-MAS为未来化工智能化提供了清晰的发展路径, 但其真正落地仍面临一系列现实挑战. 在技术层面, 多智能体协同推理依赖高质量、跨尺度的一致化本体, 而当前的知识构建仍存在语义不统一的问题; 在复杂的化工场景中, 智能体之间的推理结果是否一致、生成的策略是否可解释、是否始终处于安全边界之内, 仍需更加可靠的验证机制加以保障; 在数据与安

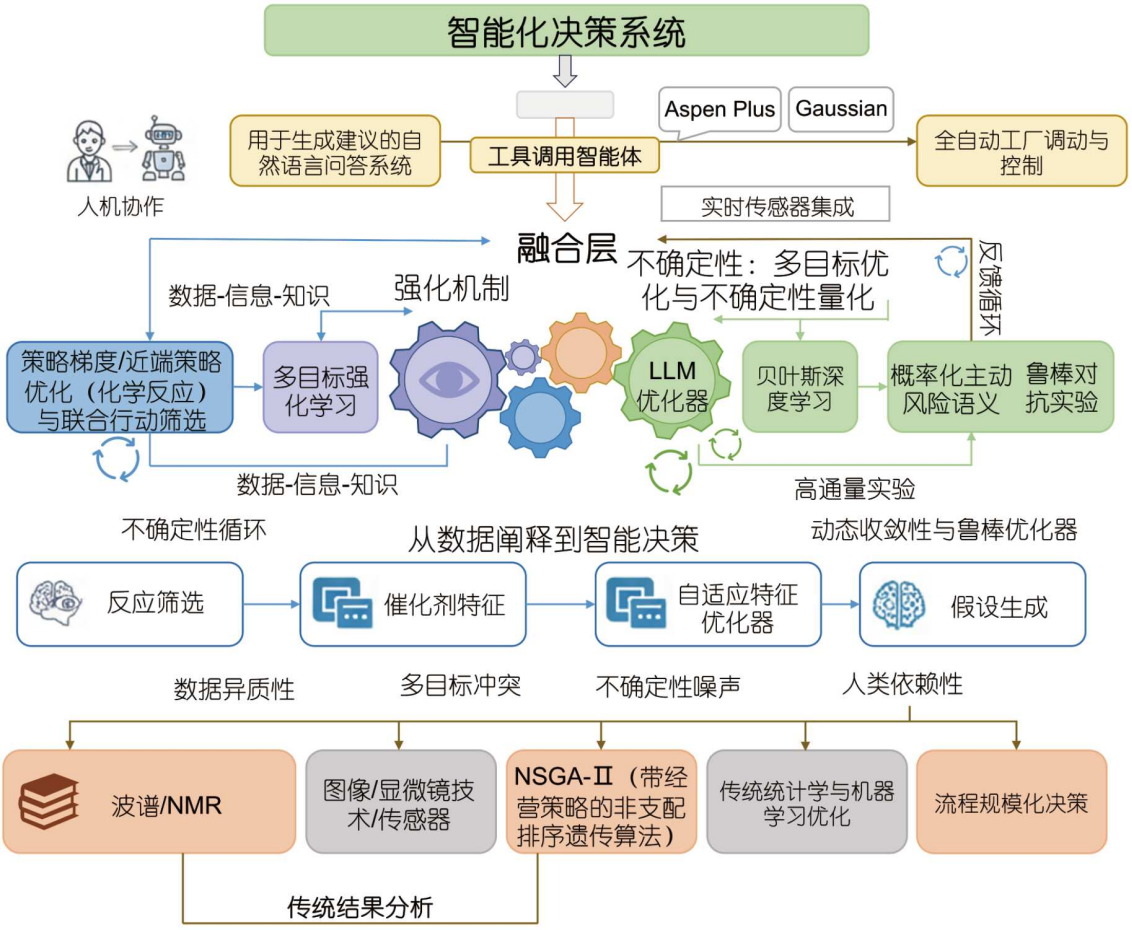


图 7 智能化决策系统框架图

Figure 7 Schematic diagram of the intelligent decision-making system framework

全层面, 化工过程涉及商业机密与关键装置参数, 大规模数据共享可能带来隐私泄露风险, 若数据本身如果存在缺失或记录不全, 也可能影响模型判断的准确性; 在伦理与责任层面, 当多智能体能够在一定程度上自主生成优化方案和控制策略时, 如何明确“谁对结果负责”、如何保证决策过程可追溯、可解释, 都是必须面对的现实问题; 此外, 大模型训练和推理带来的能耗问题, 以及自动化水平提升对传统岗位结构的潜在影响, 也需要在工程应用中统筹考虑. 因此, 迈向OD-MAS的过程, 不仅是算法和算力层面的升级, 更是研究方法、治理体系与产业生态的整体重构. 未来可以从三个方向持续推进: 其一, 构建结构清晰、可持续扩展的跨尺度化工本体体系, 为多智能体协同提供稳定统一的语义基础; 其二, 建立多层级的安全保障机制, 如推理过程校验、异常识别与风险提示, 确保系统在复杂工况下稳定可靠运行; 其三, 逐步完善数据管理、模型审查与伦理规范, 使OD-MAS在“安全、可信、可控”的前提下实现跨主体的深度协同创新. 在技术、治理与伦理三方面的共同支撑下, 本体驱动的多智能体系统有望成为下一代智能化化工体系的重要基础.

参考文献

1 Liu Z M. Current status and future trends of artificial intelligence technology application in energy and chemical industry (in Chinese). Frontiers, 2025, (2): 20–25 [刘中民. 人工智能技术在能源化工行业的应用现状及未来趋势. 人民论坛·学术前沿, 2025, (2): 20–25]

2 Liu D Y, He F P, Qiu W, et al. Low-carbon intelligent transformation: future development and challenges in the energy and chemical sectors (in

Chinese). Chem Ind Eng Prog, 2025, 44: 3305–3323 [刘东阳, 何福鹏, 邱伟, 等. 低碳智慧化转型: 能源化工领域的未来发展与挑战. 化工进展, 2025, 44: 3305–3323]
3 Ciatto G, Sabbatini F, Agiollo A, et al. Symbolic knowledge extraction and injection with sub-symbolic predictors: a systematic literature review. ACM Comput Surv, 2024, 56: 1–35
4 Atagong S D, Tonnang H, Senagi K, et al. A review on knowledge and information extraction from PDF documents and storage approaches. Front Artif Intell, 2025, 8: 1466092
5 Kang J, Mi X X, Wang H L, et al. Research progress of artificial neural networks in material science (in Chinese). Mater Rep, 2020, 34: 21172–21179 [康靓, 米晓希, 王海莲, 等. 人工神经网络在材料科学中的研究进展. 材料导报, 2020, 34: 21172–21179]
6 Xie J X, Su Y J, Xue D Z, et al. Machine learning for materials research and development (in Chinese). Acta Metall Sin, 2021, 57: 1343–1361 [谢建新, 宿彦京, 薛德祯, 等. 机器学习在材料研发中的应用. 金属学报, 2021, 57: 1343–1361]
7 Gui W, Zeng Z, Chen X, et al. Knowledge-driven process industry smart manufacturing (in Chinese). Sci Sin-Inf, 2020, 50: 1345–1360 [桂卫华, 曾朝晖, 陈晓方, 等. 知识驱动的流程工业智能制造. 中国科学: 信息科学, 2020, 50: 1345–1360]
8 Wang W Y, Luo Y P, Yu J H et al. Analysis and forecasting of Chinese methanol price based on the intelligent chemical engineering large language model (in Chinese). Chem Ind Eng Prog, 2025, 44: 5673–5688 [王文洋, 罗玉平, 余佳洹, 等. 基于智能化工大模型的中国甲醇价格分析与预测. 化工进展, 2025, 44: 5673–5688]
9 Wang Y B. Artificial intelligence has gradually become a trend in applications across vertical fields such as energy and chemical engineering (in Chinese). Sinopec Monthly, 2025, (8): 61–62 [王一冰. 人工智能在能源化工等垂直领域应用渐成趋势. 中国石化, 2025, (8): 61–62]
10 Bi C, Xia M, Li J F, et al. Exploration and prospect of the application of DeepSeek in energy intelligent development (in Chinese). China Energy Environ Protect, 2025, 47: 142–150 [毕超, 夏敏, 李继峰, 等. DeepSeek在能源智能化发展中的应用探索与展望. 能源与环保, 2025, 47: 142–150]
11 Li Y F, Liu C Y. Chemical information retrieval method based on comprehensive semantic similarity (in Chinese). Adhesion, 2025, 52: 121–124 [李玉芬, 刘传银. 基于综合语义相似度的化工信息检索方法. 粘接, 2025, 52: 121–124]
12 Meng D, Wang W, Li J J. Applying the downscoping search mode in patent search of chemical engineering devices filed (in Chinese). China Invention Patent, 2018, 15: 121–124 [孟东, 王维, 李晶晶. 渐缩式检索模式在化工装置领域专利检索中的应用. 中国发明与专利, 2018, 15: 121–124]
13 Wu Y H, Liu Y H, Li Q J, et al. Open sharing of library data based on large language models: logic, path and strategy (in Chinese). J Library Inf Sci Agric, 2025, doi: 10.13998/j.cnki.issn1002-1248.25-0436 [吴玉浩, 刘艺浩, 李庆军, 等. 基于大语言模型的图书馆数据开放共享: 逻辑、路径与策略. 农业图书情报学报, 2025, doi: 10.13998/j.cnki.issn1002-1248.25-0436]
14 Gupta S, Ranjan R, Singh S N. A comprehensive survey of retrieval-augmented generation (RAG): evolution, current landscape and future directions. 2024, arXiv: 2410.12837
15 Xu S, Bi K X, Qiu T et al. Intelligent process systems engineering (iPSE): when process systems engineering meets new generation artificial intelligence (in Chinese). Cheml Ind Eng Prog, 2025, doi: 10.16085/j.issn.1000-6613.2025-1221 [徐硕, 毕可鑫, 邱彤, 等. 智能过程系统工程(iPSE): 当过程系统工程与新一代人工智能相遇. 化工进展, 2025, doi: 10.16085/j.issn.1000-6613.2025-1221]
16 Jiang Y J. Application prospect of AI technology in chemical industry design (in Chinese). Chem Eng Design Commun, 2025, 51: 103–105 [姜宜君. AI技术在化工行业设计工作中的应用展望. 化工设计通讯, 2025, 51: 103–105]
17 Zhang G Q, Miao W, Liu X B. A brief analysis of knowledge management work in the era of artificial intelligence (in Chinese). Modern Office, 2022, (20): 179–182 [张光全, 苗卫, 刘晓波. 人工智能时代知识管理工作浅析. 办公室业务, 2022, (20): 179–182]
18 Luo D, Wen L, Chen B, et al. Key technologies and applications of large language models for materials science (in Chinese). Sci Sin-Inf, 2025, 55: 1599–1621 [罗丹宇, 温李阳, 陈博, 等. 材料科学大语言模型关键技术及其应用. 中国科学: 信息科学, 2025, 55: 1599–1621]
19 Yang Y F, Shi R H, Li Z C, et al. BatGPT-Chem: a foundation large model for retrosynthesis prediction. 2024, arXiv: 2408.10285
20 Ran N, Wang Y, Zhang X Y, et al. MOLLM: mollm: multi-objective large language model for molecular design optimizing with experts. 2025, arXiv: 2502.12845
21 Ju J X, Zheng Y Z, KOH H Y, et al. LLM: multi-view molecular representation learning with large language models. 2025, arXiv: 2508.08657
22 Jiang L, Sun S Z, Qi B Q, et al. Chem3DLLM: 3D multimodal large language models for chemistry. 2025, arXiv: 2508.10696
23 Bhattacharya D, Cassady H J, Hickner M A, et al. Large language models as molecular design Engines. J Chem Inf Model, 2024, 64: 7086–7096
24 Zhuang X, Ding K, Lyu T, et al. Advancing biomolecular understanding and design following human instructions. Nat Mach Intell, 2025, 7: 1154–1167
25 Cheng S, Bian Y H, Liu D W, et al. SDAR: a synergistic diffusion-autoregression paradigm for scalable sequence generation. 2025, arXiv: 2510.06303
26 Gao B, Li P H, Zhang D, et al. LoRA-Chem: modular machine learning for multitask prediction in organic reactions. 2025, ChemRxiv, doi: 10.26434/chemrxiv-2025-p7sxn-v2

27 Joung J F, Fong M H, Casetti N, et al. Electron flow matching for generative reaction mechanism prediction. Nature, 2025, 645: 115–123
28 Zhang Y, Han Y, Chen S, et al. Large language models to accelerate organic chemistry synthesis. Nat Mach Intell, 2025, 7: 1010–1022
29 Bran A M, Cox S, Schilter O, et al. Augmenting large language models with chemistry tools. Nat Mach Intell, 2024, 6: 525–535
30 Kresse G, Furthmüller J. Efficiency of *ab-initio* total energy calculations for metals and semiconductors using a plane-wave basis set. Comput Mater Sci, 1996, 6: 15–50
31 Cui T, Tang C, Zhou D, et al. Online test-time adaptation for better generalization of interatomic potentials to out-of-distribution data. Nat Commun, 2025, 16: 1891
32 Yang Z, Wang X, Li Y, et al. Efficient equivariant model for machine learning interatomic potentials. npj Comput Mater, 2025, 11: 49
33 Song Z, Lu S, Ju M, et al. Accurate prediction of synthesizability and precursors of 3D crystal structures via large language models. Nat Commun, 2025, 16: 6530
34 Dhakane V, Yadav A. Computational fluid dynamics–deep neural network (CFD-DNN) surrogate model with graphical user interface (GUI) for predicting hydrodynamic parameters in three-phase bubble column reactors. Ind Eng Chem Res, 2024, 63: 11670–11685
35 Mosavi A, Shamshirband S, Salwana E, et al. Prediction of multi-inputs bubble column reactor using a novel hybrid model of computational fluid dynamics and machine learning. Eng Appl Comput Fluid Mech, 2019, 13: 482–492
36 Mao Z, Lu L, Marxen O, et al. DeepM&Mnet for hypersonics: predicting the coupled flow and finite-rate chemistry behind a normal shock using neural-network approximation of operators. J Comput Phys, 2021, 447: 110698
37 Zhou J, Xu F, Chang Z, et al. From lab to fab: a large language model for chemical engineering. Chin J Catal, 2025, 73: 159–173
38 Chen Y X, Zhu X, Zhou H, et al. Metaopenfoam: an LLM-based multi-agent framework for CFD. 2024, arXiv: 2407.21320
39 Pandey S, Xu R, Wang W, et al. OpenFOAMGPT: a retrieval-augmented large language model (LLM) agent for OpenFOAM-based computational fluid dynamics. Phys Fluids, 2025, 37: 035120
40 Elrefaie M , Qian J , Wu R , et al. AI agents in engineering design: a multi-agent framework for aesthetic and aerodynamic car design. 2025, arXiv: 2503.23315
41 Zhang T, Liu Z H, Xin Y, et al. MooseAgent: a LLM based multi-agent framework for automating moose simulation. 2025, arXiv: 2504.08621
42 Zeng T, Badrinarayanan S, Ock J, et al. LLM-guided chemical process optimization with a multi-agent approach. 2025, arXiv: 2506.20921
43 Pajak E, Bahamdan A, Hellgardt K, et al. Multi-agent LLMs for automating sustainable operational decision-making. Syst Control Transact, 2025, 4: 1824–1829
44 Pang Z, Liu S, Lin Y, et al. Large language model-based multi-objective modeling framework for vacuum gas oil hydrotreating. Chin J Chem Eng, 2025, 84: 133–145
45 Tan H, Hong X, Liao Z, et al. Combining reinforcement learning with mathematical programming: an approach for optimal design of heat exchanger networks. Chin J Chem Eng, 2024, 69: 63–71
46 Lee G, Joo Y, Lee S U, et al. Design optimization of heat exchanger using deep reinforcement learning. Int Commun Heat Mass Transfer, 2024, 159: 107991
47 Queiroz L P, Bruaset O R, Ribeiro A M, et al. Optimization of shell and tube heat exchangers using reinforcement learning. In: The 35th European Symposium on Computer Aided Process Engineering. Ghent, Belgium, 2025. 1818–1823
48 Sun H, Jia Z, Zhao M, et al. Dynamic modeling of heat exchangers based on mechanism and reinforcement learning synergy. Buildings, 2024, 14: 833
49 Jiang Z, Xu C, Zhou T. Cyber Academia-Chemical Engineering (CA-ChemE): a living digital town for self-directed research evolution and emergent scientific discovery. 2025, arXiv: 2510.01293
50 Bai L, Cai Z, Cao Y, et al. Intern-S1: a scientific multimodal foundation model. 2025, arXiv: 2508.15763
51 Liu Z Y, Jiang Z K, Qiu W. Application of artificial intelligence in long-duration redox flow batteries storage systems (in Chinese). Energy Storage Sci Technol, 2024, 13: 2871–2883 [刘子玉, 姜泽坤, 邱伟, 等. 人工智能在长时液流电池储能中的应用: 性能优化和大模型. 储能科学与技术, 2024, 13: 2871–2883]
52 Pribylskii I V, Armeev D V. Application of large language models in decision-making systems of power engineering tasks. In: 2024 IEEE 3rd International Conference on Problems of Informatics, Electronics and Radio Engineering (PIERE). Novosibirsk, Russian Federation: IEEE, 2024. 1230–1235
53 Xia Y, Jazdi N, Zhang J, et al. Control industrial automation system with large language model agents. In: 2025 IEEE 30th International Conference on Emerging Technologies and Factory Automation (ETFA), 2025. 1–8
54 Wu P, Mou X, Gong L, et al. An automatic machine fault identification method using the knowledge graph–embedded large language model. Int J Adv Manuf Technol, 2025, 138: 725–739
55 Wang Y, Pang Y, Liu Y, et al. Graph structure-enhanced large language model for optical network fault diagnosis: an explainable alarm root cause localization approach. IEEE Internet Things J, 2025, 12: 31493–31510

56 Sun Q, Li Y, Zhou C, et al. Root cause analysis for industrial process anomalies through the integration of knowledge graph and large language model. In: 2024 43rd Chinese Control Conference (CCC). IEEE, 2024. 6855–6860

57 Shen F, Zheng J, Ye L, et al. LSTM soft sensor development of batch processes with multivariate trajectory-based ensemble just-in-time learning. IEEE Accraph Structure-Enhanced Large Language Model for Optical Network Fault Diagnosis: An Explainable Alarm Root Cause Localization Approachess, 2020, 8: 73855–73864

58 Wang K, Gopaluni R B, Chen J, et al. Deep learning of complex batch process data and its application on quality prediction. IEEE Trans Ind Inf, 2020, 16: 7233–7242

59 Yamaguchi T, Yamashita Y. Quality prediction for multi-grade batch process using sparse flexible clustered multi-task learning. Comput Chem Eng, 2021, 150: 107320

Summary for “基于本体工程的大语言模型驱动化工研究范式革新”

Ontology engineering–based large language model revolution in chemical research paradigms

Fupeng He[1], Ziyi Gu[1], Xinying Gu[2], Biqing Qi[2], Lei Bai[2], Tianhang Zhou[1,3*], Yuqiang Li[2*] & Chunming Xu[1]

[1] *State Key Laboratory of Heavy Oil Processing, China University of Petroleum (Beijing), Beijing 102249, China*
[2] *Shanghai Artificial Intelligence Laboratory, Shanghai 200023, China*
[3] *ZhongHai Energy Storage Technology Co., LTD, Beijing 102200, China*
* Corresponding authors, E-mail: zhouth@cup.edu.cn; liyuqiang@pjlab.org.cn

Chemical research spans multiple scales from molecules to processes and industrial systems, and is inherently characterized by high dimensionality, strong coupling, and pronounced nonlinearity. Traditional research paradigms relying on empirical heuristics, isolated simulations, and iterative experimentation are increasingly inadequate to meet the emerging demands for green transformation, intelligent manufacturing, and sustainable development. In this work, we systematically review a state-of-the-art “four-dimensional integrated” intelligent research paradigm driven by large language models (LLMs), and elucidate its theoretical foundation, technical architecture, and implementation pathways for reconstructing the chemical research framework. Centered on LLMs as the cognitive core, the proposed paradigm integrates four key dimensions: knowledge engine, molecular reasoning, process simulation, and intelligent decision-making. At the knowledge level, retrieval-augmented generation (RAG) and knowledge graph technologies are synergistically combined to construct a dynamic and computable knowledge infrastructure that semantically unifies literature, patents, and experimental data. This enables the transition from conventional information retrieval to evidence-based knowledge discovery and reasoning-oriented scientific research.

At the molecular level, generative foundation models, multi-objective optimization models, and multimodal molecular representation learning are integrated with physical constraints and chemical ontologies, forming a closed intelligent loop from molecular conception and structure generation to reaction planning, energy verification, and synthesizability assessment. This framework significantly accelerates the discovery of functional molecules and advanced materials while mitigating chemical hallucination risks. At the process level, physics-informed neural networks (PINNs), neural operators, and digital twin technologies are introduced to bridge mechanistic models with data-driven surrogates, enabling cross-scale, high-fidelity, and computationally efficient modeling of complex multiphase transport, reaction, and energy conversion processes. In parallel, LLM-driven multi-agent automation frameworks enable end-to-end workflows for CFD/FEM model generation, simulation execution, and intelligent post-analysis, substantially lowering the technical barrier for high-level process simulations. At the decision-making level, ontology engineering serves as the semantic foundation that formalizes chemical entities, equipment, unit operations, and control logic into a unified machine-interpretable structure. On this basis, reinforcement learning and uncertainty-aware optimization construct an interpretable, traceable, and transferable knowledge-driven decision loop, enabling semantic-level integration of process control, alarm diagnosis, batch optimization, and system-wide sustainable operation. The four dimensions work synergistically to form an intelligent cognitive chain that links molecules, processes, and systems, driving the transformation of chemical research from experience-driven to inference-driven paradigms. Furthermore, we propose the emerging concept of an ontology-driven multi-agent collaborative system (OD-MAS), which envisions a transition from isolated task-level intelligence toward system-level collaboration and knowledge-level autonomy. OD-MAS enables multiple specialized intelligent agents to operate within a unified semantic space for coordinated discovery, simulation, optimization, and safety evaluation. This framework provides a fundamental semantic infrastructure for the interoperability of data, models, and decisions, and establishes methodological and strategic foundations for the deep integration of green chemistry, intelligent manufacturing, and autonomous scientific discovery.

large language model (LLM), intelligent SOTA research paradigm, ontology engineering, chemical engineering intelligence

doi: 10.1360/CSB-2025-5792