

An Automatic End-to-End Chemical Synthesis Development

Platform Powered by Large Language Models

由大型语言模型驱动的自动化端到端化学合成开发平台

Yixiang Ruan, Chenyin Lu, Ning Xu, Yuchen He, Yixin Chen, Jian Zhang, Jun Xuan, Jianzhang Pan, Qun Fang, Hanyu Gao, Xiaodong Shen, Ning Ye, Qiang Zhang, Yiming Mo

Abstract 摘要

The rapid emergence of large language model (LLM) technology presents promising opportunities to facilitate the development of synthetic reactions. In this work, we leveraged the power of GPT-4 to build an LLM-based reaction development framework (LLM-RDF) to handle fundamental tasks involved throughout the chemical synthesis development. LLM-RDF comprises six specialized LLM-based agents, including Literature Scouter, Experiment Designer, Hardware Executor, Spectrum Analyzer, Separation Instructor, and Result Interpreter, which are pre-prompted to accomplish the designated tasks. A web application with LLM-RDF as the backend was built to allow chemist users to interact with automated experimental platforms and analyze results via natural language, thus, eliminating the need for coding skills and ensuring accessibility for all chemists. We demonstrated the capabilities of LLM-RDF in guiding the end-to-end synthesis development process for the copper/TEMPO catalyzed aerobic alcohol oxidation to aldehyde reaction, including literature search and information extraction, substrate scope and condition screening, reaction kinetics study, reaction condition optimization, reaction scale-up and product purification. Furthermore, LLM-RDF's broader applicability and versatility was validated on various synthesis tasks of three distinct reactions (S_NAr reaction, photoredox C-C cross-coupling reaction, and heterogeneous photoelectrochemical reaction).

大语言模型（LLM）技术的迅速发展为合成反应的开发提供了广阔的机遇。在本研究中，我们利用 GPT-4 构建了一个基于 LLM 的反应开发框架（LLM-RDF），以处理化学合成开发过程中涉及的基础任务。LLM-RDF 由六个专门的 LLM 代理组成，包括文献搜索代理、实验设计代理、硬件执行代理、光谱分析代理、分离指导代理和结果解释代理，这些代理通过预设提示词来完成指定任务。我们开发了一个基于 LLM-RDF 的 Web 应用，使化学家能够通过自然语言与自动化实验平台交互并分析结果，从而无需编写代码，确保所有化学家都能轻松使用。我们展示了 LLM-RDF 在铜/TEMPO 催化的有氧醇氧化为醛反应的端到端合成开发过程中的应用，包括文献搜索和信息提取、底物范围和条件筛选、反应动力学研究、反应条件优化、反应放大以及产物纯化。此外，我们进一步验证了 LLM-RDF 在三种不同类型的合成任务中的广泛适用性和灵活性，包括 S_NAr 反应、光氧化还原 C-C 交叉偶联反应以及非均相光电化学反应。

Introduction

Designing proper synthesis reactions and routes towards target compounds is one of core tasks during drug discovery and process development, requiring significant time and cost. Due to the enormous design space and necessity of experimental validation, this process mainly relies on expert chemists and chemical engineers to go through iterative design-make-test-analyze cycles to identify an efficient synthesis route. The multifaceted and complex requirements for synthesis reaction design, such as efficiency, cost, sustainability, safety, scalability, and impurity control, make it hard to formulate this task into a well-defined problem that can be tackled algorithmically and autonomously without customized inputs and decisions from experts.

设计合适的合成反应和路线以获得目标化合物是药物发现和工艺开发中的核心任务之一，需要大量时间和成本。由于设计空间巨大且需要实验验证，该过程主要依赖于化学专家和化学工程师通过迭代的“设计-合成-测试-分析”循环来确定高效的合成路线。合成反应设计涉及多个复杂因素，如效率、成本、可持续性、安全性、可扩展性以及杂质控制，使得该任务难以被明确地定义为一个可以通过算法自动解决的问题，而不需要专家的定制输入和决策。

The recent advancement of machine learning (ML) technologies has shown great potential in expediting various subtasks during the synthesis design. Notable examples include deep learning based quantitative structure-activity relationship (QSAR) model facilitating drug molecule design and catalyst design, rapid identification of promising synthetic routes using machine-learning-aided synthesis planning, guiding automated high-throughput experimental platforms to search for optimal reaction conditions, and direct translation of multistep synthesis procedures from literature to experimental execution via natural language processing (NLP) models. Despite this rapid involvement of machine learning methods in synthesis related tasks, the monolithic input-to-output nature of existing machine learning methods makes them to only function as powerful single-purpose tools for experts, while the goal of fully autonomous end-to-end synthesis reaction design and development still remains to be realized.

近年来，机器学习（ML）技术的进步在加速合成设计的各个子任务方面展现出了巨大潜力。典型案例包括基于深度学习的定量构效关系（QSAR）模型辅助药物分子和催化剂设计，利用机器学习辅助合成规划快速识别潜在合成路线，引导自动化高通量实验平台搜索最优反应条件，以及通过自然语言处理（NLP）模型直接将多步合成方案从文献转换为实验执行。尽管机器学习方法已广泛应用于合成相关任务，但现有方法主要是单一用途的输入-输出工具，仅能作为专家的辅助工具，而实现完全自主的端到端合成反应设计和开发仍然是一个未被解决的挑战。

In November 2022, OpenAI released the large language model (LLM) based ChatGPT tool, marking a significant leap towards the artificial general intelligence (AGI). The enormous knowledge and information packed in the LLM enables it to make decisions flexibly according to the complex and non-standardized inputs (prompts). As of now, various advanced LLMs, such as proprietary Anthropic's Claude, Google's Gemini as well as open-source Llama3.1, Mistral and Qwen2, have emerged and shown continuing performance improvement. LLM-based agents, characterized by their strong generalization abilities and broad applicability, have demonstrated significant advancements in language proficiency and interaction with humans. Motivated by the outstanding performance of these agents, scholars have explored and exploited their capability in the various tasks of chemical and material research, such as literature mining, molecule and material discovery, reaction condition recommendation and optimization, and lab apparatus automation.

2022 年 11 月，OpenAI 发布了基于大语言模型（LLM）的 ChatGPT 工具，标志着向通用人工智能（AGI）迈出了重要的一步。LLM 中蕴含的海量知识和信息使其能够根据复杂和非标准化的输入（提示词）灵活做出决策。截至目前，已有多个先进的 LLM 版本问世，如专有的 Anthropic Claude、Google Gemini，以及开源的 Llama3.1、Mistral 和 Qwen2，并且其性能持续提升。LLM 代理以其强大的泛化能力和广泛的适用性，在语言理解和人与机器的交互方面取得了显著进展。受到这些代理卓越性能的启发，研究人员开始探索并利用它们在化学和材料研究中的潜力，例如文献挖掘、分子与材料发现、反应条件推荐与优化以及实验室设备自动化。

The existing reports of LLM-based agents showed scattered coverage of the stages in chemical synthesis development, but have not presented a path to fully exploit the potential of LLM-based agents in the entire development process. Herein, we proposed a unified LLM-based reaction development framework (LLM-RDF) to demonstrate the versatility and performance of LLM-based agents in the entire of chemical synthesis reaction development process. We selected aerobic alcohol oxidation to the aldehyde, an emerging sustainable aldehyde synthesis protocol as a model transformation to showcase the end-to-end synthesis development facilitated by LLM agents. In addition to this case study, we further demonstrated the applicability of LLM-RDF on three distinct scenarios relevant to chemical synthesis development. The findings of this work serve to map out the viable path to the autonomous end-to-end chemical synthesis development using the emerging LLM technology.

现有的基于 LLM 的代理研究在化学合成开发的各个阶段呈现出零散的应用，但尚未提出一个能够充分发挥 LLM 代理潜力的完整开发路径。因此，我们提出了一个统一的 LLM 反应开发框架（LLM-RDF），以展示 LLM 代理在整个化学合成反应开发过程中的多功能性和性能。我们选择了有氧醇氧化为醛这一新兴的可持续醛合成策略作为模型反应，以展示 LLM 代理在端到端合成开发中的应用。除了这一案例研究，我们还进一步验证了 LLM-RDF 在三个不同合成开发场景中的适用性。本研究的发现为利用新兴的 LLM 技术实现自主端到端化学合成开发绘制了一条可行的路径。

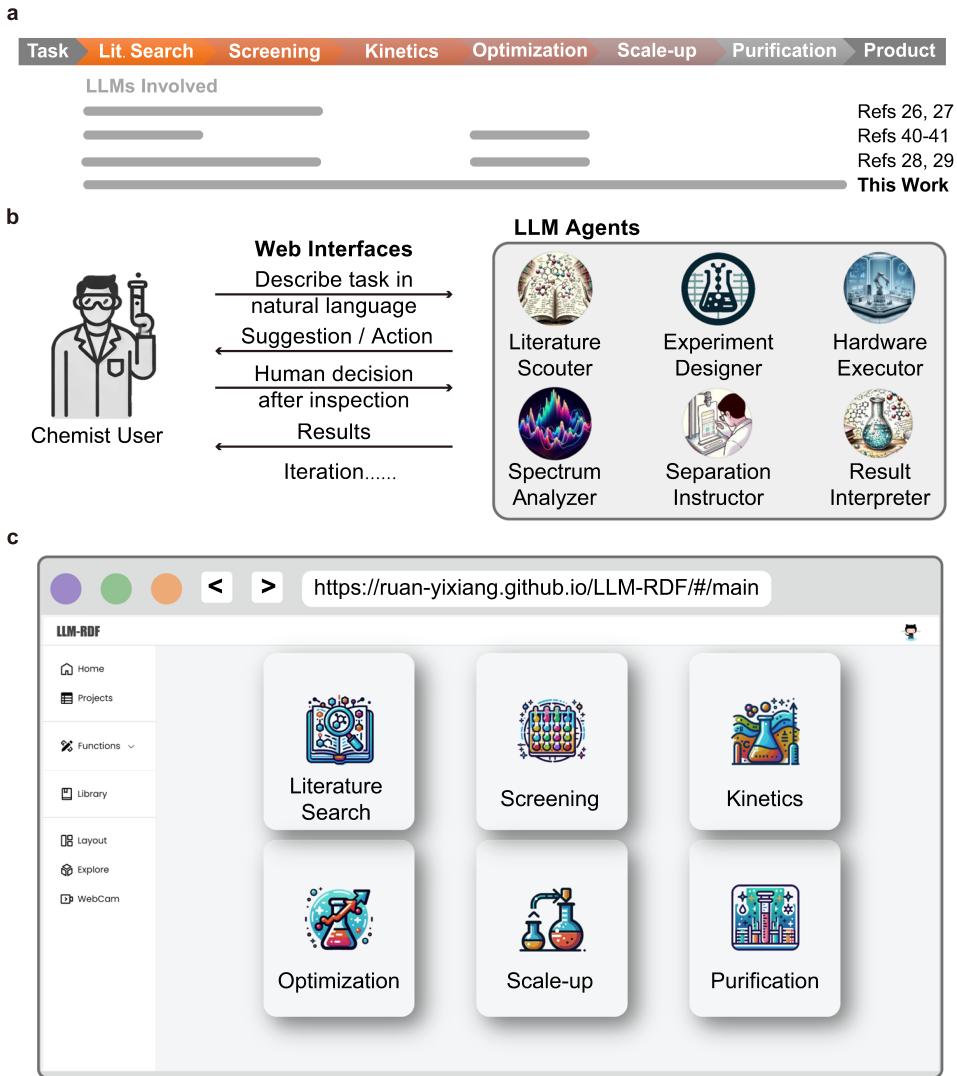


Figure 1: Overview of LLM-based multi-agent system for reaction development.

Results

LLM-based agents for end-to-end chemical synthesis reaction development 基于 LLM 的端到端化学合成反应开发代理

A typical chemical synthesis reaction development workflow consists of five steps: (1) literature search and information extraction, (2) substrate scope and condition screening, (3) reaction kinetics study, (4) reaction condition optimization, and (5) reaction scale-up and product purification. To exploit the capabilities of LLM facilitating this development process, we developed a set of LLM-based intelligent agents in LLM-RDF to handle the fundamental tasks necessary to complete the development steps above. These agents include Literature Scouter, Experiment Designer, Hardware Executor, Spectrum Analyzer, Separation Instructor, and Result Interpreter. We chose to build these agents based on GPT-4 model to maximize their capabilities in context understanding and chemical knowledge reasoning. They were pre-promoted using customized instructions and documents to achieve consistent behavior and performance for a specific task. Detailed LLM agent construction procedures can be found in Methods section and Supplementary Information Section 1.

典型的化学合成反应开发流程包括五个步骤：(1) 文献搜索与信息提取，(2) 底物范围与反应条件筛选，(3) 反应动力学研究，(4) 反应条件优化，(5) 反应放大与产物纯化。为了利用 LLM 促进这一开发过程，我们在 LLM-RDF 框架中开发了一组基于 LLM 的智能代理，以处理完成上述开发步骤所需的基本任务。这些代理包括文献搜索代理、实验设计代理、硬件执行代理、光谱分析代理、分离指导代理和结果解释

代理。我们选择基于 GPT-4 模型构建这些代理，以最大化其在上下文理解和化学知识推理方面的能力。它们通过定制的指令和文档进行预设，以确保在特定任务中的行为一致性和性能稳定性。详细的 LLM 代理构建过程可参见“方法”部分和补充信息第 1 节。

With the set of LLM-based agents developed above, we created a web application to allow users accessing them using natural language in a centralized manner, such that no coding was required during the synthesis reaction development. After agents receive prompts and related reference documents from the users describing the chemical task, they will analyze the requests and infer the appropriate responses or solutions through in-context learning and retrieval-augmented generation (RAG). If necessary, they would employ external tools to enhance their capability to respond information out of the scope of the LLM knowledge itself, including Python interpreter, academic database search, and self-driven reaction optimization algorithms. In addition, there is a chain-of-thought mechanism to allow agents to interact with these tools step-by-step, thus maximizing their reasoning capability. Despite the advanced intelligence of GPT-4 model used for these agents, human chemists are still essential in the decision-making loop, responsible for evaluating the correctness and completeness of agents' responses, interconnecting agents, and deciding whether to directly implement their suggestions or further communicate with them to tweak the responses.

借助上述基于 LLM 的代理，我们开发了一个 Web 应用，使用户能够通过自然语言集中访问这些代理，从而在合成反应开发过程中无需编写代码。当代理接收到用户描述化学任务的提示词和相关参考文献后，它们将通过上下文学习和检索增强生成（RAG）技术分析请求，并推导出适当的响应或解决方案。如有必要，它们还可以使用外部工具来增强自身能力，以应对超出 LLM 知识范围的信息查询，包括 Python 解释器、学术数据库搜索以及自驱动反应优化算法。此外，代理采用链式思维（chain-of-thought）机制，使其能够逐步与这些工具交互，从而最大化其推理能力。尽管这些代理基于 GPT-4 模型，具备较高的智能水平，但人类化学家仍在决策环节中发挥核心作用，负责评估代理响应的正确性和完整性，协调不同代理之间的交互，并决定是直接执行其建议，还是进一步调整其响应。

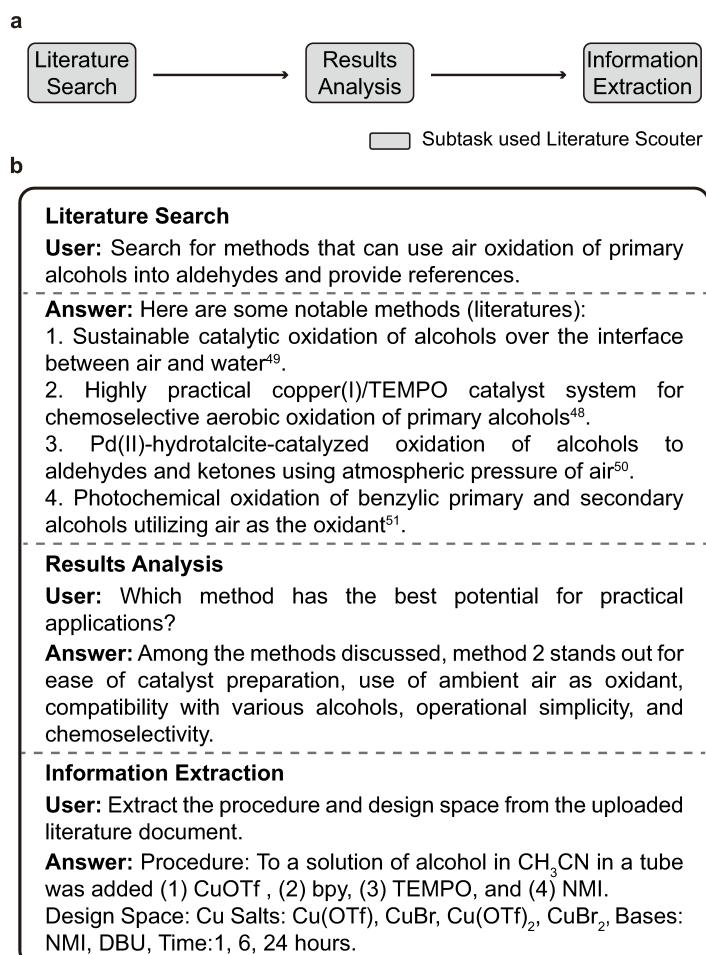


Figure 2: LLM-based agents facilitated literature search and information extraction.

Literature search and information extraction 文献搜索与信息提取

To initiate the synthesis development of the aerobic alcohol oxidation to the corresponding aldehyde, instead of manually finding relevant reports in conventional academic search engines (e.g., SciFinder and Web of Science), we directly input the request to Literature Scouter agent with "Searching for synthetic methods that can use air to oxidize alcohols into aldehydes" prompt. Leveraging vector search technologies, Literature Scouter automatically sifted through the Semantic Scholar database containing over 20 million academic literatures. The use of the Semantic Scholar database instead of relying on the LLM's knowledge (i.e., training data used by OpenAI to train GPT-4) ensured the accuracy of the chemistry details with proper references.

为了启动有氧醇氧化为相应醛的合成开发，我们没有使用传统的学术搜索引擎（如 SciFinder 和 Web of Science）手动查找相关文献，而是直接向文献搜索代理输入提示词：“搜索可利用空气将醇氧化为醛的合成方法”。借助向量搜索技术，文献搜索代理自动筛选了包含 2000 多万篇学术文献的 Semantic Scholar 数据库。选择 Semantic Scholar 数据库而非依赖于 LLM 自身的知识（即 OpenAI 训练 GPT-4 所用的数据）确保了化学细节的准确性，并提供了适当的参考文献。

Among the various methods given by Literature Scouter, we continued to query which method had the greatest potential for practical applications. Literature Scouter recommended the recently developed Cu/TEMPO dual catalytic system developed by Stahl group as this method outpaced others in the aspects of the environmental sustainability, simplicity, safety, chemoselectivity, and substrate compatibility. After manually evaluating other recommended methods, this Cu/TEMPO catalytic chemistry indeed avoids the use of heterogeneous catalysts, high-cost palladium catalysts, or light irradiation used in other approaches, proving to have claimed potentials in practical applications as suggested by the Literature Scouter.

在文献搜索代理提供的多种方法中，我们进一步查询了最具实际应用潜力的方法。文献搜索代理推荐了 Stahl 研究组最近开发的 Cu/TEMPO 双催化体系，该方法在环境可持续性、简便性、安全性、化学选择性和底物兼容性方面优于其他方法。在手动评估其他推荐方法后，我们确认 Cu/TEMPO 催化体系确实避免了使用非均相催化剂、高成本的钯催化剂或光照激发等方法，证明其在实际应用中的潜力符合文献搜索代理的建议。

Having identified the target transformation, we next turned to extract the detailed reaction conditions for this catalytic system. The literature document was provided to Literature Scouter to summarize the detailed experimental procedures and options for various reagents and catalysts. This information served as the basis for the subsequent experimental exploration of this chemistry.

确定了目标反应后，我们进一步提取该催化体系的详细反应条件。文献搜索代理被提供相关文献，并总结了详细的实验步骤以及各种试剂和催化剂的选择。这些信息成为后续实验探索该化学体系的基础。

As demonstrated in the task of method search and information extraction from literature, Literature Scouter demonstrated its capability to assist researchers to identify the possible methodologies necessary to achieve the target transformation under desired conditions, and extracting the required experimental details for executing the reaction. Compared to conventional workflow for identifying the proper chemistry from literature database, Literature Scouter alleviated the labor-intensive tasks of literature searching and reviewing. Especially, when Literature Scouter was connected to an up-to-date academic journal database, it could propose the new chemistries that were not included in the LLM base model training process.

正如在文献中的方法搜索和信息提取任务中所展示的那样，文献搜索器展示了其帮助研究人员确定在所需条件下实现目标转化所需的可能方法以及提取执行反应所需的实验细节的能力。与从文献数据库中识别合适化学反应的传统工作流程相比，文献搜索器减轻了文献搜索和审查的劳动密集型任务。特别是，当文献搜索器连接到最新的学术期刊数据库时，它可以提出 LLM 基础模型训练过程中未包含的新化学反应。

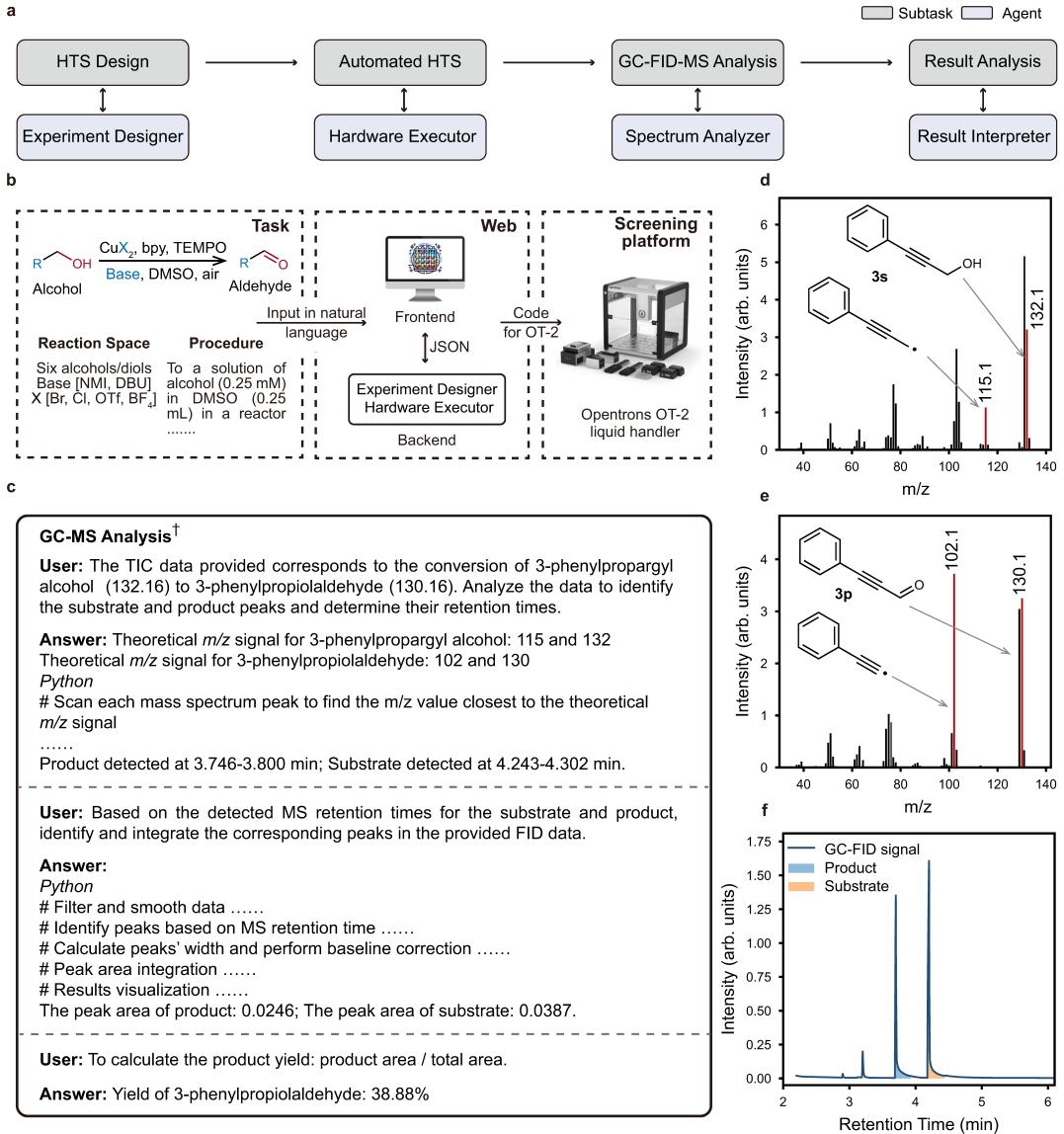


Figure 3: LLM-based agents facilitated substrate scope and condition screening.

Substrate scope and condition screening 底物范围与反应条件筛选

With the literature reported aerobic alcohol oxidation protocol in hand, understanding the substrate scope under various reaction conditions for a methodology is essential for selecting the suitable reaction conditions based on the target compound structure in practical synthesis. It is typically challenging to predict the reaction yield based on first-principle theories, while recently emerging machine learning based methods need a decent amount of experimental data to train the neural model for accurate predictions. 掌握了文献报道的有氧醇氧化方法后，了解该方法在不同反应条件下的底物适用范围对于实际合成中基于目标化合物结构选择合适的反应条件至关重要。通常，基于第一性原理理论预测反应收率具有较大挑战，而近年来兴起的基于机器学习的方法则需要大量实验数据来训练神经网络模型以获得准确预测。

Apart from the high costs of the required HTS hardware, the time-consuming programming for executing the automation platforms and manual analysis of large amount of HTS results create barriers for chemists with minimal coding experience to use HTS technology in their routine workflows.

除了 HTS 设备的高成本外，执行自动化平台所需的编程时间以及对大量 HTS 结果的人工分析也构成了障碍，使得缺乏编程经验的化学家难以在日常工作流程中使用 HTS 技术。

To tackle the above-mentioned challenges, we implemented Experiment Designer, Hardware Executor, Spectrum Analyzer, and Result Interpreter agents to automate HTS investigation of the substrate scope, such that the barrier for routine usage of HTS technology could be significantly lowered. The HTS substrate scope study consists of a series of subtasks, including HTS experiment design, automated HTS

experiments, gas chromatography (GC) analysis, and results analysis.

为了解决上述挑战，我们开发了实验设计代理、硬件执行代理、光谱分析代理和结果解释代理，以自动化 HTS 的底物范围研究，从而大幅降低 HTS 技术在日常应用中的门槛。HTS 底物范围研究包括一系列任务，如 HTS 试验设计、自动化 HTS 试验、气相色谱 (GC) 分析以及结果分析。

The automated HTS of this aerobic oxidation reaction requires the reaction to run in the open-cap vial and continuous operation for an extended period. Consequently, strictly following the procedure and design space extracted by Literature Scouter from the literature leads to two challenges: the high volatility of acetonitrile (MeCN) solvent and the instability of the Cu(I) salts stock solution (Cu(OTf) and CuBr). These issues significantly affect the reproducibility of the experimental results. To address these issues, Experiment Designer suggested switching to a higher boiling point solvent and using the stable Cu(II) salts. Following its recommendation, we replaced acetonitrile with dimethyl sulfoxide (DMSO) as the solvent and used CuCl₂ and Cu(BF₄)₂ as Cu catalysts.

该有氧氧化反应的自动化 HTS 试验要求反应在开口小瓶中运行，并持续较长时间。因此，严格遵循文献搜索代理从文献中提取的实验程序和设计空间会导致两个挑战：乙腈 (MeCN) 溶剂的高挥发性以及 Cu(I) 盐储备溶液 (Cu(OTf) 和 CuBr) 的不稳定性。这些问题显著影响实验结果的可重复性。为了解决这些问题，实验设计代理建议更换沸点较高的溶剂，并使用更稳定的 Cu(II) 盐。按照其建议，我们用二甲基亚砜 (DMSO) 代替乙腈作为溶剂，并采用 CuCl₂ 和 Cu(BF₄)₂ 作为铜催化剂。

In HTS experiment design, Experiment Designer agent parsed the HTS experiment task described in natural language into the standardized JavaScript Object Notation (JSON) experimental procedure and design space that could be displayed on the web application. To execute the HTS task, we chose Opentrons liquid handler (OT-2) as the automated reaction screening platform since the Cu/TEMPO catalyzed aerobic alcohol oxidation reaction only involved soluble reagents. In addition, OT-2 liquid handler has a well-written Python API documentation, based on which Hardware Executor agent could compose liquid handler running code. Thus, Hardware Executor converted the HTS experiment task described in natural language to OT-2 execution codes to load the necessary labware and pipettes, plan the storage locations for stock solutions, prepare the reaction mixtures as dictated by the experimental procedures, and shake the vial plate to perform the aerobic alcohol oxidation.

在 HTS 试验设计过程中，实验设计代理将自然语言描述的 HTS 试验任务解析为标准化的 JavaScript 对象表示 (JSON) 实验程序和设计空间，并可在 Web 应用程序上展示。为了执行 HTS 任务，我们选择 Opentrons 液体处理器 (OT-2) 作为自动化反应筛选平台，因为 Cu/TEMPO 催化的有氧醇氧化反应仅涉及可溶性试剂。此外，OT-2 液体处理器具有完善的 Python API 文档，硬件执行代理可以基于此编写液体处理器的运行代码。因此，硬件执行代理将自然语言描述的 HTS 试验任务转换为 OT-2 执行代码，以加载所需的实验器具和移液器、规划储备溶液的存储位置、按照实验程序配制反应混合物，并震荡小瓶板以进行有氧醇氧化反应。

With this workflow from HTS task described in natural language to automated reaction execution, two rounds of HTS experiments were conducted, and each round contained a full factorial screening of six alcohol substrates (six monohydric alcohols for the first round and six diols for the second round), four copper catalysts [CuCl₂, CuBr₂, Cu(OTf)₂, and Cu(BF₄)₂], and two bases [N-methylimidazole (NMI) and 1,8-diazabicyclo-[5.4.0]undec-7ene (DBU)].

通过这种从自然语言描述 HTS 任务到自动化反应执行的工作流程，我们进行了两轮 HTS 试验。每轮试验均包括对六种醇类底物（第一轮测试六种单醇，第二轮测试六种二醇）、四种铜催化剂 [CuCl₂、CuBr₂、Cu(OTf)₂ 和 Cu(BF₄)₂] 以及两种碱 [N-甲基咪唑 (NMI) 和 1,8-二氮杂双环 [5.4.0] 十一烷-7-烯 (DBU)] 的全因子筛选。

After the HTS experiment, the products were characterized with gas chromatography with parallel flame ionization detector and mass spectrometer (GC-FID-MS). The use of parallel FID and MS detectors enabled the simultaneous identification and quantification of the components in the reaction crudes. Instead of labor-intensive manual identification of peaks for reactants and yield calculation, Spectrum Analyzer agent was used to automate this process. Specifically, GC-FID-MS analysis instructions and the raw chromatogram data, including FID intensity chromatogram and total ion chromatogram (TIC) from MS detector, were provided to Spectrum Analyzer. It could identify the corresponding reactant and product peaks in TIC by looking for their characteristic fragmentation patterns, and calculated the reaction yield based on FID intensity chromatogram.

HTS 试验结束后，产物通过气相色谱-火焰离子化检测器-质谱 (GC-FID-MS) 进行表征。采用并行 FID 和 MS 检测器，可同时对反应粗产物中的组分进行鉴定和定量。为了避免劳动密集型的手动峰识别和收率计算，我们使用光谱分析代理自动执行此过程。具体而言，GC-FID-MS 分析指令和原始色谱数据（包括 FID 强度色谱图和 MS 检测器的总离子色谱图，TIC）被提供给光谱分析代理。该代理通过识别 TIC 中的特征碎片模式来匹配相应的反应物和产物峰，并基于 FID 强度色谱图计算反应收率。

Using 3-phenylpropargyl alcohol (3s) converting to the corresponding product 3-phenylpropiolaldehyde (3p) as an example, Spectrum Analyzer thought that 3s should have a 132 mass to charge (m/z) ratio signal for the molecule itself and 115 m/z signal for the fragment resulting from the loss of a hydroxyl group, and 3p should have 130 m/z signal for the molecule itself and 102 m/z signal for the fragment resulting from the loss of the carbonyl group. Subsequently, Spectrum Analyzer wrote a Python code to search TIC data for mass spectrometry peaks containing the characteristic m/z signals and determine the retention times of the substrate and product. Next, Spectrum Analyzer integrated the FID peak areas at the substrate and product retention times to determine the reaction yield (assuming that the response factors of the products and substrates are the same in FID). The yields obtained by Spectrum Analyzer of all the monohydric alcohols experiments were nearly consistent with those derived from manual analysis using commercial chromatography software.

以 3-苯基丙炔醇 (3s) 氧化为相应产物 3-苯基丙炔醛 (3p) 为例，光谱分析代理认为 3s 应该在质谱中表现出 132 质量电荷比 (m/z) 信号，对应于分子本身，同时由于羟基的丢失还应出现 115 m/z 信号。而 3p 应该表现出 130 m/z 信号，对应于分子本身，并且由于羰基的丢失应出现 102 m/z 信号。随后，光谱分析代理编写了一段 Python 代码，在 TIC 数据中搜索包含这些特征 m/z 信号的质谱峰，并确定底物和产物的保留时间。接下来，光谱分析代理整合了 FID 在底物和产物保留时间处的峰面积，以确定反应收率（假设产物和底物在 FID 中的响应因子相同）。光谱分析代理计算出的所有单醇实验的收率与使用商业色谱软件手动分析得到的结果几乎一致。

Finally, we utilized Result Interpreter agent to summarize HTS results and explain observed patterns based on fundamental chemistry knowledge. Result Interpreter recognized that DBU base significantly outperformed NMI, and the reactivity of copper salt followed the order of $\text{CuCl}_2 = \text{CuBr}_2 = \text{Cu}(\text{OTf})_2 > \text{Cu}(\text{BF}_4)_2$. In addition, it concluded that electron-withdrawing functional groups near the hydroxyl group (e.g., aromatic rings or unsaturated carbon bonds) could increase the oxidation reactivity, which was consistent with chemistry principles. However, Result Interpreter's ability to conduct further in-depth analysis was still limited with existing GPT-4 model as the backend. For example, in explaining why diol 9s and 10s exhibited no reaction in any condition tested, it could only suggest superficially that the arrangement of functional groups or the spatial configuration of the molecules might play a role. The literature-proposed mechanism involves the chelation of copper catalyst by the vicinal diol substrates (9-10s) to form an unreactive Cu-phenolate species, thus deactivating the copper catalyst.

最后，我们使用结果解释代理来总结 HTS 结果，并基于基本化学知识解释观察到的反应趋势。结果解释代理识别出 DBU 碱的催化效果明显优于 NMI，并且铜盐的反应活性顺序为 $\text{CuCl}_2 = \text{CuBr}_2 = \text{Cu}(\text{OTf})_2 > \text{Cu}(\text{BF}_4)_2$ 。此外，它得出结论，羟基附近的电子吸引基团（如芳香环或不饱和碳键）可以增强氧化活性，这一趋势与化学原理一致。然而，结果解释代理在进行更深入的分析时仍然受限于当前以 GPT-4 为后端的模型。例如，在解释为什么二醇 9s 和 10s 在所有测试条件下均无反应时，它只能给出表面的解释，即可能是官能团的排列或分子空间构型影响了反应。而文献中提出的机制表明，邻二醇底物 (9-10s) 可能通过络合铜催化剂形成不活性的 Cu-酚盐物种，从而使铜催化剂失去活性。

Reaction kinetics study 反应动力学研究

As mentioned earlier, this copper/TEMPO catalytic system prefers to oxidize primary hydroxyl group compared to secondary hydroxyl group. We observed that dimethyl sulfoxide (DMSO) solvent (used in the HTS experiment) gave superior primary alcohol (12s) oxidation chemoselectivity compared to acetonitrile (MeCN) solvent (used in the literature). To investigate the observed solvent effects, Experiment Designer agent suggested that we could conduct oxidation kinetics study for different solvent. Recently, automated kinetic profiling has become an efficient tool to help researchers establish reaction kinetic models. However, similar to the HTS technology discussed above, it is still not a routine tool used in process development due to the high entry barrier for mastering automated hardware and intricate programming involved in fitting kinetics models. Experiment Designer, Hardware Executor, Spectrum Analyzer, and Result Interpreter agents orchestrated to complete the kinetic study task, consisting of subtasks including kinetics experiment design, automated sampling experiments, proton nuclear magnetic resonance (^1H NMR) analysis, and kinetic model fitting and analysis.

如前所述，该铜/TEMPO 催化体系更倾向于氧化伯羟基，而非仲羟基。我们观察到，在 HTS 实验中使用的二甲基亚砜 (DMSO) 溶剂相比于文献中使用的乙腈 (MeCN) 溶剂，表现出更优越的伯醇 (12s) 氧化化学选择性。为了研究溶剂效应，实验设计代理建议我们针对不同溶剂进行氧化动力学研究。近年来，自动化动力学分析已成为帮助研究人员建立反应动力学模型的有效工具。然而，与前文讨论的 HTS 技术类似，由于自动化设备的高门槛及动力学模型拟合所需的复杂编程，该技术尚未成为工艺开发中的常规工具。实验设计代理、硬件执行代理、光谱分析代理和结果解释代理协同完成动力学研究任务，包括动力学实验设计、自动取样实验、质子核磁共振 (^1H NMR) 分析以及动力学模型拟合和分析。

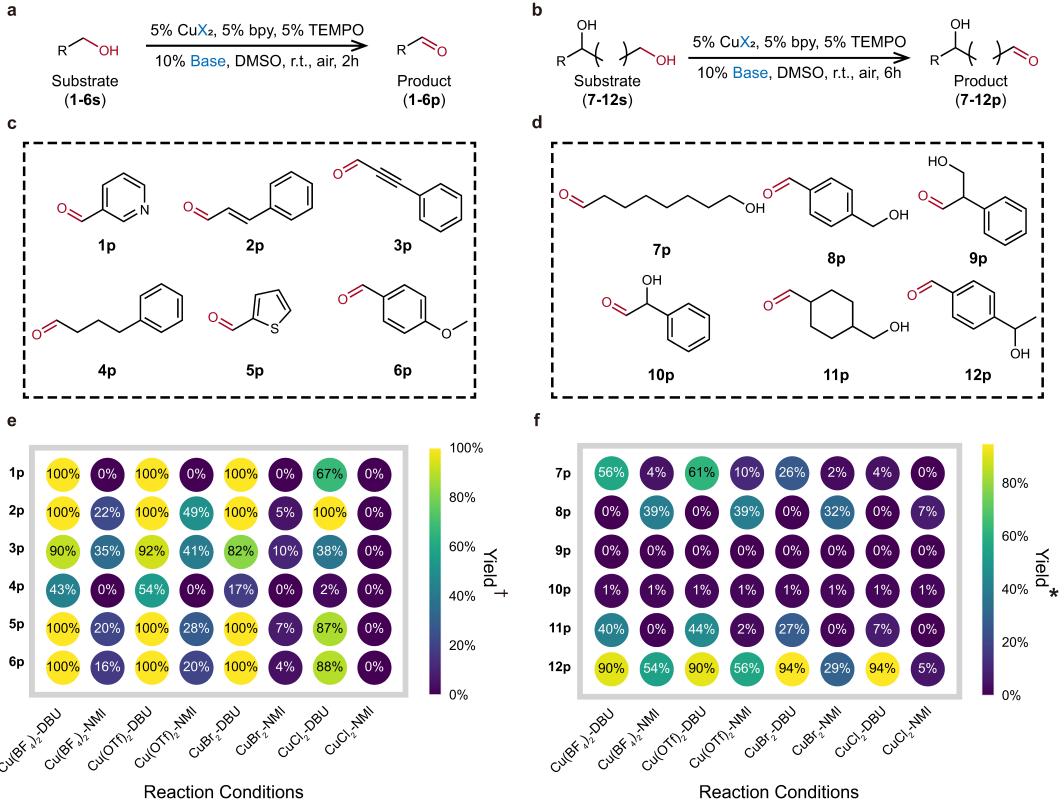


Figure 4: The substrate scope and condition screening results.

In kinetics experiment design, Experiment Designer planned a sampling schedule for time-course data collection. To provide approximate reaction rate information for experimental design, we firstly monitored the reaction via thin-layer chromatography (TLC) and found that substrate 12s was rapidly consumed within the initial first hour reaction time, and the reaction slowed down afterward.

在动力学实验设计中，实验设计代理规划了一个取样时间表以收集时间序列数据。为了提供实验设计所需的反应速率信息，我们首先通过薄层色谱（TLC）监测反应，发现底物 12s 在最初 1 小时内迅速消耗，随后反应速率明显降低。

Based on this observation, Experiment Designer proposed a sampling schedule spanning a 10-hour reaction period. Samples were to be collected at 10, 20, 30, 40, 50, 60, 120, 240, 360, 480, and 600 min, such that denser data points could be obtained during the early stage of the reaction when the reaction rate was large. Subsequently, Hardware Executor agent generated the OT-2 running code based on the experimental design proposed by the Experiment Designer. The coded OT-2 liquid handler procedure contained a series of operations for sampling, such as stopping the reaction's shaking, pipetting to sample, quenching the reaction in the sample, and resuming shaking.

基于这一观察结果，实验设计代理提出了一个涵盖 10 小时反应时间的取样计划。在 10, 20, 30, 40, 50, 60, 120, 240, 360, 480 和 600 分钟时收集样品，以在反应速率较大的早期阶段获得更密集的数据点。随后，硬件执行代理根据实验设计代理的方案生成 OT-2 运行代码。该 OT-2 液体处理程序包括一系列取样操作，例如停止反应摇动、移液取样、终止样品反应并恢复摇动。

The compositions of the sampled reaction crude were analyzed by ^1H NMR. Instead of manual analysis of the NMR data, we provided Spectrum Analyzer with ^1H NMR spectra and approximate chemical shifts for characteristic hydrogen atoms in the substrate, product, and byproducts (overoxidation products). Spectrum Analyzer wrote a Python program according to the API documentation for the TopSpin NMR processing software to automate the analysis of NMR data, the procedure of which included identifying target peaks, performing peak integration, and calculating the compositions of the samples.

取样后的反应粗产物通过 ^1H NMR 进行成分分析。为避免人工分析 NMR 数据，我们将 ^1H NMR 光谱以及底物、产物和副产物（过度氧化产物）中特征氢原子的近似化学位移提供给光谱分析代理。光谱分析代理依据 TopSpin NMR 处理软件的 API 文档编写了 Python 程序，实现 NMR 数据的自动化分析，包括识别目标峰、进行峰积分以及计算样品组成。

Next, providing the obtained kinetics experiment results to Result Interpreter, it fitted the time-course data to the kinetic model equations. The reaction rate for substrate to product followed saturation kinetic dependence on the substrate alcohol (Eq. 1), and in addition, the product overoxidation was assumed to be a first-order reaction (Eq. 2). Result Interpreter calculated the corresponding reaction rate constants (k_1, k_2, k_3), and the proposed kinetic models fitted well with the experimental data.

随后，我们将获得的动力学实验结果提供给结果解释代理，该代理将时间序列数据拟合到动力学模型方程中。底物转化为产物的反应速率遵循对底物醇的饱和动力学依赖关系（方程 1），此外，产物的过度氧化假定为一级反应（方程 2）。结果解释代理计算了相应的反应速率常数 (k_1, k_2, k_3)，并且拟合模型与实验数据吻合良好。

$$\begin{aligned} r_{\text{Product}} &= \frac{k_1 k_2 C_{\text{Substrate}}}{1 + k_1 C_{\text{Substrate}}} \\ r_{\text{Byproducts}} &= k_3 C_{\text{Product}} \end{aligned} \quad (1)$$

Result Interpreter further concluded that the rate constant for the product overoxidation (k_3) was larger in MeCN than that in DMSO, indicating that the product overoxidation rate had strong dependence on the reaction solvent choice. This analysis highlighted that Result Interpreter had the ability to understand the underlying kinetics behind the observed reaction selectivity.

结果解释代理进一步得出结论，产物过度氧化的速率常数 (k_3) 在 MeCN 溶剂中大于 DMSO 溶剂中，表明产物的过度氧化速率对反应溶剂的选择具有显著依赖性。这一分析结果表明，结果解释代理能够理解反应选择性背后的动力学机制。

Reaction condition optimization 反应条件优化

When a specific target compound is determined for process development towards manufacturing, reaction condition optimization is necessary to improve the synthesis efficiency along with other considerations (e.g., costs and impurity generation). Instead of traditional manual one-factor-at-time (OFAT) optimization, the recent development of optimization algorithms, such as Bayesian optimization (BO) and the mixed-integer nonlinear program (MINLP) algorithm, have enabled the automated experimental platforms to perform closed-loop reaction optimization in an autonomous manner. However, akin to the HTS technology mentioned previously, the steep learning curve associated with mastering automated hardware and optimization algorithms prevents the widespread adoption of the self-driven reaction optimization workflow as a routine tool in process development.

当确定特定目标化合物用于工艺开发和生产时，优化反应条件对于提高合成效率及其他关键因素（如成本和杂质生成）至关重要。与传统的单因素优化 (OFAT) 方法不同，近年来发展出的优化算法（如贝叶斯优化 (BO) 和混合整数非线性规划 (MINLP) 算法）使自动化实验平台能够以自主方式执行闭环反应优化。然而，与前文提到的 HTS 技术类似，掌握自动化硬件和优化算法的陡峭学习曲线阻碍了自驱动反应优化工作流程作为工艺开发常规工具的广泛应用。

To address this challenge, we employed Experiment Designer and Hardware Executor as the backend of a reaction optimization module within our developed web application, such that users could interface with the reaction optimization hardware system via natural language. This hardware system is a robotic platform capable of performing end-to-end reaction and analysis, and the closed-loop reaction optimization was driven by a Bayesian optimization algorithm. Specifically, an automated synthesis equipment (Unchained Big Kahuna) conducts the chemical reactions, which are then analyzed by a high-performance liquid chromatography (HPLC) to provide result feedbacks to the BO for suggesting the next-round reaction candidates.

为了解决这一挑战，我们在开发的 Web 应用中将实验设计代理和硬件执行代理作为反应优化模块的后端，使用户可以通过自然语言与反应优化硬件系统交互。该硬件系统是一个能够执行端到端反应和分析的机器人平台，采用贝叶斯优化算法驱动闭环反应优化。具体而言，自动化合成设备 (Unchained Big Kahuna) 执行化学反应，然后通过高效液相色谱 (HPLC) 分析反应产物，并将分析结果反馈给 BO，以推荐下一轮优化的反应条件。

Although the LLMs have been used as an optimizer in recent publications and shown superior performance for optimizing reactions when provided kinetic information or reaction knowledge, they still fell behind statistical optimization algorithms (e.g., BO) for complex reaction systems. Thus, we chose to use BO as the optimizer in this work.

尽管 LLM 近年来被用于优化任务，并在提供动力学信息或反应知识的情况下表现出色，但在复杂反应系统中，它们仍然落后于统计优化算法（如 BO）。因此，在本研究中，我们选择 BO 作为优化器。

To demonstrate LLM-based agents copiloted reaction optimization workflow, we conducted the condition optimization for the selective oxidation of diol (12s) to the corresponding monooxidized aldehyde product (12p). The reaction design space included two continuous variables (i.e., equivalents of base and reaction time) and two categorical variables (i.e., types of bases and copper catalysts). The optimization objective is to maximize the reaction yield of 12p.

为了展示基于 LLM 的代理如何辅助反应优化工作流程，我们对二醇（12s）选择性氧化生成单氧化醛产物（12p）进行了条件优化。反应设计空间包括两个连续变量（即碱当量和反应时间）和两个分类变量（即碱种类和铜催化剂种类）。优化目标是最大化 12p 的反应收率。

First, Experiment Designer translated synthesis procedure description [To a solution of substrate (0.05 mmol) in DMSO (0.25 mL) in a reactor was added sequentially a solution of (1) CuX₂/bpy (0.25 mL, 0.01 M), (2) TEMPO (0.25 mL, 0.01 M), and (3) Base (0.25 mL, 0.02 M).] and workup procedure description [Add 0.75 mL HEDP.] into standardized JSON procedure steps for display on the web application. Hardware Executor generated code templates based on these JSON procedure steps to define the automated synthesis platform operation workflows.

首先，实验设计代理将合成实验步骤描述 [向 DMSO (0.25 mL) 中的底物 (0.05 mmol) 溶液依次加入 (1) CuX₂/bpy (0.25 mL, 0.01 M), (2) TEMPO (0.25 mL, 0.01 M), (3) 碱 (0.25 mL, 0.02 M).] 以及后处理步骤描述 [加入 0.75 mL HEDP。] 转换为标准化的 JSON 实验步骤，并在 Web 应用上展示。随后，硬件执行代理基于这些 JSON 实验步骤生成代码模板，以定义自动化合成平台的操作流程。

Next, Experiment Designer converted the optimization parameter space described in natural language [I want to optimize four variables: 1. Reaction time: 45-90 min; 2. Base volume: 0.125-0.25 mL; 3. Cu catalyst: CuCl₂, CuBr₂, Cu(OTf)₂, Cu(BF₄)₂; 4. Base type: NMI, DBU.] into JSON format that was used as inputs for the Bayesian optimizer. At last, users reviewed the entire experimental plan before running the reaction optimization on the automation hardware.

接下来，实验设计代理将自然语言描述的优化参数空间 [优化 4 个变量：1. 反应时间：45-90 min；2. 碱体积：0.125-0.25 mL；3. 铜催化剂：CuCl₂、CuBr₂、Cu(OTf)₂、Cu(BF₄)₂；4. 碱种类：NMI、DBU。] 转换为 JSON 格式，以供贝叶斯优化器使用。最后，用户在自动化硬件上运行反应优化之前，审查整个实验方案。

The self-driven optimization system iteratively conducted reactions and proposed candidate experiments based on existing reaction results, thus gradually improving the reaction yield of 12p. Multiple high-yield reaction conditions were identified within the design space.

该自驱动优化系统迭代地执行反应，并基于已有的反应结果提出新的实验方案，从而逐步提高 12p 的反应收率。在设计空间内识别出了多个高收率的反应条件。

To automatically stop the reaction optimization task when the expectation of further yield improvement was diminished, we compared the statistical stopping criterion and stopping decision given by the LLM-based agent Result Interpreter. The probability of improvement (PI) metric, a typical statistical stopping criterion, was first examined by stopping the optimization when the cumulative number of proposed reaction conditions with PI values below 0.01 reached two. This PI stopping criterion was met after completing 36 experiments, based on which the optimal conditions should be confidently identified. 为了在反应收率的进一步提升预期降低时自动停止优化任务，我们比较了统计停止准则和基于 LLM 的代理——结果解释代理——给出的停止决策。首先，我们测试了改进概率 (PI) 这一典型的统计停止准则，即当 PI 值低于 0.01 的提议反应条件累计达到两次时，停止优化。该 PI 停止准则在完成 36 次实验后达到，该实验数量足以确定最优反应条件。

Reaction scale-up and product purification 反应放大与产物纯化

In the process development, the scale-up investigation serves as a critical phase to determine whether a small-scale chemistry is suitable for further large-scale synthesis with similar reaction efficiency. Here, we used the high-yield reaction conditions found in the previous reaction optimization task for targeting 1 gram scale synthesis of the compound 12p to demonstrate the utility of LLM-based agents in facilitating the reaction process development.

在工艺开发过程中，反应放大研究是一个关键阶段，用于确定小规模化学反应是否适用于更大规模合成，并能保持相似的反应效率。在本研究中，我们采用先前反应优化任务中找到的高收率反应条件，以 1 克规模合成化合物 12p，展示基于 LLM 的代理在促进反应工艺开发中的应用价值。

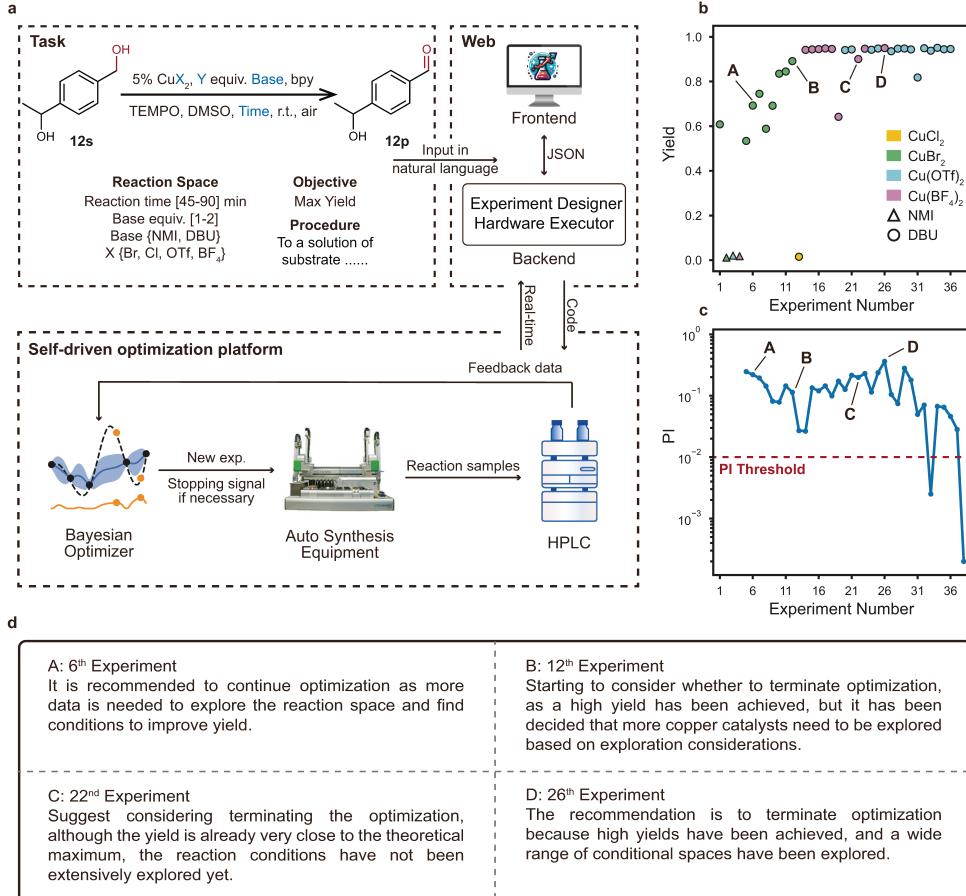


Figure 5: LLM-based agents facilitated self-driven reaction condition optimization.

Among various high-yield ($\geq 94.5\%$) conditions during the condition optimization of diol oxidation, Experiment Designer selected the condition used in 35th experiment for scaling up. The choice of reaction conditions was made based on the preference to the high product yield, short reaction time, and low catalyst and reagent costs. The 35th experiment used a 45-min reaction time, Cu(OTf)₂ catalyst, and 1.34 equivalent DBU base, achieving a high yield of 94.5%.

在二醇氧化的反应条件优化过程中，我们获得了多种高收率（ $\geq 94.5\%$ ）的反应条件。实验设计代理选择了第 35 次实验所采用的条件进行放大。该条件的选择基于对高产率、较短反应时间以及低催化剂和试剂成本的偏好。第 35 次实验采用 45 分钟的反应时间，以 Cu(OTf)₂ 作为催化剂，并使用 1.34 当量的 DBU 碱，最终获得 94.5% 的高收率。

To showcase LLM's ability to facilitate reaction scale-up, we first engaged with Experiment Designer to develop a scale-up strategy for this gas-liquid biphasic reaction. Experiment Designer proposed a two-stage scale-up strategy: first to 1 g to validate the reaction's reproducibility and stability, and then to 100 g to assess feasibility for industrial production. The scale-up process included key considerations such as maintaining efficient gas-liquid contact, ensuring proper oxygen supply, and selecting appropriate reactors for different scales.

为了展示 LLM 在促进反应放大中的能力，我们首先与实验设计代理合作，为该气-液两相反应制定放大策略。实验设计代理提出了两阶段放大策略：首先放大至 1 克规模，以验证反应的可重复性和稳定性；然后进一步放大至 100 克，以评估其工业化生产的可行性。放大过程中包含关键考虑因素，如保持高效的气-液接触、确保适当的氧气供应以及选择适合不同规模的反应器。

For illustrative purpose, we targeted the 1-gram scale in this work, Experiment Designer accurately calculated the stoichiometries of the reagents based on the selected reaction condition for the 1 g scale-up. We then conducted the scale-up experiment according to the parameters proposed by Experiment Designer.

为说明 LLM 代理的作用，我们在本研究中选择 1 克规模进行放大。实验设计代理基于选定的反应条件，精确计算了 1 克放大实验所需的试剂化学计量比。随后，我们按照实验设计代理提出的参数进行放大实验。

Prior to the product purification using flash column chromatography, the optimal eluent composition is typically determined with manual TLC. TLC fine-tunes the eluent polarity to ensure that the retention factor value (R_f value) of the target compound falls within 0.2 – 0.3, and, at the same time, impurities are separated from the target compound.

在使用快速柱色谱进行产物纯化之前，通常需要通过手动薄层色谱（TLC）确定最佳洗脱剂组成。TLC 通过微调洗脱剂的极性，确保目标化合物的保留因子 (R_f 值) 落在 0.2 – 0.3 范围内，同时实现目标化合物与杂质的有效分离。

A recent publication has applied machine learning model to predict the R_f value of a given compound structure in different eluent compositions. However, due to the inevitable prediction inaccuracy, this data-driven prediction model can only serve to provide good initial eluent composition guesses to try, and chemists still need to determine the eluent suitable for practical separation processes by conducting iterative trial-and-error experiments based on their own experience and the polarity-controlled separation principles in TLC.

最近的研究已应用机器学习模型来预测不同洗脱剂组成下特定化合物结构的 R_f 值。然而，由于预测误差的不可避免性，该数据驱动的预测模型只能提供初始洗脱剂组成的合理猜测，而化学家仍需根据自身经验和 TLC 中的极性控制分离原理，通过反复试验来确定适用于实际分离过程的洗脱剂。

To enable automated identification of optimal eluent composition, we implemented Separation Instructor agent to replace chemists for making eluent composition decisions during the iterative TLC experiment. Here, TLC experiments were performed manually, but the automated TLC device is commercially available to achieve closed-loop optimal eluent composition identification in an autonomous manner.

为了实现洗脱剂组成的自动优化，我们开发了分离指导代理，以在 TLC 试验过程中替代化学家进行洗脱剂选择决策。在本研究中，TLC 试验仍由人工执行，但目前已有商用自动化 TLC 设备，可实现自主闭环的洗脱剂组成优化。

Upon inputting the initial TLC outcome of 12p separation at hexane : ethyl acetate = 1:1 ratio into Separation Instructor, it advised to reduce the polarity of the eluent to decrease the R_f value of 12p. 在将 12p 于正己烷:乙酸乙酯 = 1:1 条件下的初始 TLC 结果输入分离指导代理后，该代理建议降低洗脱剂极性，以降低 12p 的 R_f 值。

Following two iterative decision-and-experiment rounds, Separation Instructor finalized the eluent composition (hexane : ethyl acetate = 3:1), under which the product's R_f value was 0.28 with 0.49 R_f value for the impurity, providing a sufficiently large difference for effective separation.

经过两轮决策和实验迭代，分离指导代理最终确定了最佳洗脱剂组成（正己烷:乙酸乙酯 = 3:1）。在该条件下，产物 R_f 值为 0.28，而杂质的 R_f 值为 0.49，提供了足够大的分离度以实现有效纯化。

Subsequently, this optimal eluent composition was used in the automated preparative column chromatography system to successfully separate the product, yielding 915 mg of the product (12p) with the isolated yield of 86% and a purity > 98%.

随后，我们将该优化的洗脱剂组成应用于自动化制备型柱色谱系统，成功分离出目标产物，最终获得 915 mg 的 12p，分离收率为 86%，纯度超过 98%。

Applications 应用

After validating the LLM-RDF copiloted workflow for the end-to-end synthesis development on the case study of the aerobic alcohol oxidation, we sought to explore its utility in real-world chemical synthesis development tasks, including (1) reaction kinetics study of a nucleophilic aromatic substitution (S_NAr) reaction, (2) reaction condition optimization of a photoredox C-C cross-coupling reaction, and (3) scale-up design of a heterogeneous photoelectrochemical reactor.

在验证了 LLM-RDF 协同工作流在有氧醇氧化案例研究中的端到端合成开发能力后，我们进一步探讨了其在实际化学合成开发任务中的应用，包括：(1) 亲核芳香取代 (S_NAr) 反应的反应动力学研究，(2) 光氧化还原 C-C 交叉偶联反应的反应条件优化，(3) 非均相光电化学反应器的放大设计。

S_NAr reaction ranks as the top-3 frequently used reaction types in drug discovery and development for its capability of forming C-X bonds, and understanding its kinetics information is critical for reaction mechanism elaboration, reactor engineering, and impurity control. We employed Experiment Designer, Hardware Executor, and Result Interpreter to accomplish the kinetic study of S_NAr reaction between an electron-deficient aryl fluoride (13) and an amine (14) to form aniline (15).

S_NAr 反应因其能够形成 C-X 键，在药物发现和开发中位列三大常用反应类型之一，理解其动力学信息对于阐明反应机理、反应器工程设计以及杂质控制至关重要。我们使用实验设计代理、硬件执行代理和结果解释代理，完成了电子缺陷芳基氟化物 (13) 与胺 (14) 反应生成苯胺 (15) 的 S_NAr 反应动力学研究。

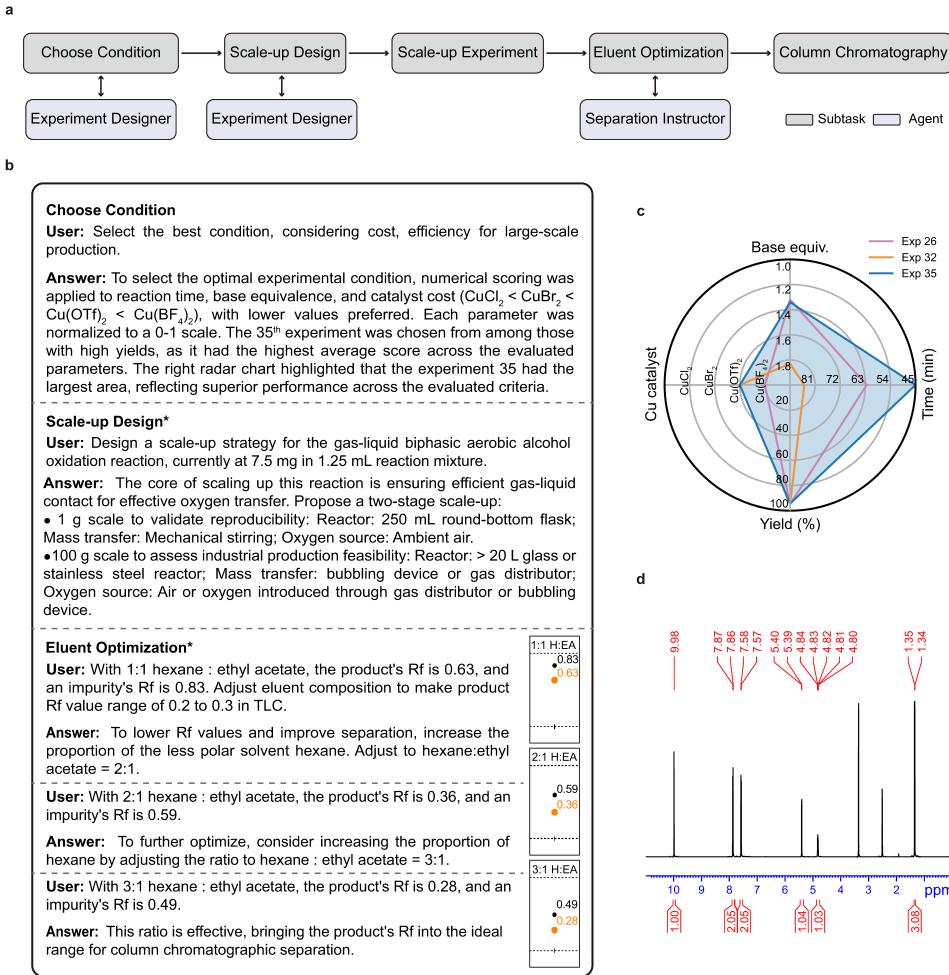


Figure 6: LLM-based agents facilitated reaction scale-up and product purification.

Experiment Designer proposed a sampling schedule over a 2-h reaction period, based on which Hardware Executor generated the OT-2 running code to automate the reaction sampling process. Subsequently, the obtained kinetic data were supplied to Result Interpreter, which identified $r_{25} = 0.2C_{22}C_{23}^2$ was the best-fit kinetic model ($R^2 = 0.995$) among various possible kinetic models.

实验设计代理提出了一个涵盖 2 小时反应时间的取样计划，硬件执行代理据此生成 OT-2 运行代码，以实现自动化反应取样。随后，我们将获得的动力学数据提供给结果解释代理，该代理分析发现，在多个可能的动力学模型中， $r_{25} = 0.2C_{22}C_{23}^2$ 是最佳拟合模型 ($R^2 = 0.995$)。

In terms of mechanistic explanation, Result Interpreter inferred that the second-order dependence on the concentration of N-methylpiperazine (C_{24}) indicated the bifunctional roles of 14 in S_NAr reaction besides being a nucleophile. However, similar to the previous discussion on the diol inhibition mechanism on Cu/TEMPO catalytic system, Result Interpreter based on GPT-4 base model lacks the in-depth chemistry knowledge to propose the specific roles of 14 acting both as a nucleophile and a base catalyst accelerating the reaction.

在反应机理解释方面，结果解释代理推断出，N-甲基哌嗪 (C_{24}) 浓度的二级依赖性表明，14 在 S_NAr 反应中除了作为亲核试剂外，还发挥了双功能作用。然而，与之前讨论的 Cu/TEMPO 催化体系中二醇抑制机理类似，基于 GPT-4 基础模型的结果解释代理缺乏深入的化学知识，无法具体阐明 14 既作为亲核试剂，又作为碱催化剂加速反应的具体作用。

The recently discovered amino radical transfer (ART) strategy enabled $\text{C}(\text{sp}^3)\text{-C}(\text{sp}^3)$ cross-coupling reactions between alkyl boronic esters and aryl halides under mild visible-light irradiation, representing an important advancement in the cross-coupling chemistry.

最近发现的氨基自由基转移 (ART) 策略使得 $\text{C}(\text{sp}^3)\text{-C}(\text{sp}^3)$ 交叉偶联反应能够在温和的可见光照射条件下，在烷基硼酸酯和芳基卤化物之间进行，这代表了交叉偶联化学的重要进展。

Implementing such newly-developed chemistry in practice requires extensive efforts in condition optimization due to the lack of historical collection of experimental data on various substrate structures unlike well-established chemistries.

由于该新开发的化学反应不同于已成熟的化学反应，其底物结构的实验数据积累较少，因此在实际应用中需要投入大量精力进行反应条件优化。

Thus, we chose to employ our LLM-agent copiloted reaction optimization workflow for the cross-coupling of 2-bromo-5-chloropyridine (18) and benzylboronic acid pinacol ester (19).

因此，我们选择使用 LLM 代理协作的反应优化工作流，优化 2-溴-5-氯吡啶 (18) 与苯硼酸频哪醇酯 (19) 的交叉偶联反应条件。

Since the mono-coupled product (20) could further react with remaining 19 to form the bis-coupled byproduct (21), it is desired to find the optimal condition to maximize the yield of 20.

由于单偶联产物 (20) 可能继续与剩余的 19 反应生成双偶联副产物 (21)，因此需要找到最优反应条件，以最大化 20 的产率。

Hardware Executor generated the OT-2 running code based on the optimization task description for automating the execution of the experiments. The experimental outcome was fed into BO to suggest next-round trial.

硬件执行代理根据优化任务描述生成 OT-2 运行代码，以自动化执行实验。实验结果随后被输入贝叶斯优化 (BO) 算法，以建议下一轮实验方案。

After three rounds of iteration, each consisting of five experiments, Result Interpreter concluded that the diminishing gains in yield improvement made it reasonable to terminate the optimization process. 经过三轮迭代（每轮包括五次实验），结果解释代理得出结论，产率的提升幅度逐渐减小，因此合理终止优化过程。

Under the optimal reaction condition of 1.38 equivalents of morpholine, 3 equivalents of 19, and 10 mol% NiCl₂-glyme over 7.1 h, substrate 18 was fully consumed, yielding 87% of product 20 with almost no formation of byproduct 21.

在最优反应条件下，即 1.38 当量吗啉、3 当量 19 和 10 mol% NiCl₂-glyme，反应进行 7.1 小时，底物 18 完全消耗，产物 20 的收率达到 87%，几乎未检测到副产物 21。

Limitations and outlook 局限性与展望

With the extensive evaluation above of the LLM agents copiloted end-to-end synthesis development, we identified several limitations and areas for improvement in the future development of this technology. 通过上述对 LLM 代理协作端到端合成开发的广泛评估，我们识别出了该技术在未来发展中需要改进的若干局限性和优化方向。

Reliability of LLM-based agents' response: The LLM-based agents may provide incorrect responses, which, if without proper inspection, could lead to experimental failure and data inaccuracies. For example, Hardware Executor was only used for generating running codes for automated experimental equipment, and the codes needed to go through manual verification and simulated execution preview before execution to avoid potential equipment damage or even personal injuries. A recent study has demonstrated that introducing another LLM to automatically inspect and modify the responses from LLMs could partially mitigate unreliable response issues.

基于 LLM 代理的响应可靠性：LLM 代理可能会提供错误的响应，如果未经过适当检查，可能导致实验失败和数据不准确。例如，硬件执行代理仅用于为自动化实验设备生成运行代码，这些代码在执行前需要经过人工验证和模拟执行预览，以避免可能的设备损坏甚至人身伤害。最近的一项研究表明，引入另一 LLM 自动检查和修改 LLM 生成的响应，可以在一定程度上缓解不可靠响应的问题。

Lack of domain knowledge: Result Interpreter failed in this work to analyze the underlying mechanisms behind the reaction selectivity and kinetics, indicating the lack of advanced chemistry knowledge for GPT-4-based agents. Recent studies have shown that incorporating domain-specific chemical knowledge into LLMs, typically through finetuning methods, significantly enhances their performance on chemistry-related tasks. RAG can also be employed to help LLM-based agents bridge gaps in specialized knowledge. For example, when Spectrum Analyzer was provided with the documentation of TopSpin Python Interface, it could successfully automate the analysis of NMR raw data.

领域知识的缺乏：在本研究中，结果解释代理未能分析反应选择性和动力学背后的机制，这表明基于 GPT-4 的代理缺乏高级化学知识。最近的研究表明，通过微调方法将特定领域的化学知识整合到 LLM 中，可以显著提升其在化学相关任务中的表现。此外，RAG 方法也可用于帮助基于 LLM 的代理弥补专业知识的不足。例如，当光谱分析代理被提供 TopSpin Python 接口的文档后，它成功地实现了 NMR 原始数据的自动化分析。

Mathematical operations: One of the recognized limitations of LLMs is their inherent difficulty in performing precise mathematical operations and handling numerical data. To address this limitation, we equipped the agents with integrated tools such as Python interpreter and Bayesian optimization algorithms for handling numerical computations, reasoning, and processing. In addition, fine-tuning the LLMs with datasets specifically curated for mathematical operations could improve the model's inherent ability to handle mathematical calculations.

数学运算: LLM 的一个已知局限性是其在执行精确数学运算和处理数值数据方面的固有困难。为了解决这一问题，我们为代理集成了 Python 解释器和贝叶斯优化等工具，以处理数值计算、推理和数据处理。此外，通过专门针对数学运算构建的数据集对 LLM 进行微调，可以提高模型在数学计算方面的固有能力。

Reproducibility and transparency: Closed-source proprietary LLMs such as GPT-4 pose several challenges, including poor long-term reproducibility, lack of transparency, and concerns over data privacy. Building agents based on open-source LLMs would mitigate these issues. In this work, we compared agents constructed using open-source LLMs (Qwen2-72B and Llama3.1-70B) with those based on GPT-4 in the task of reaction kinetics study. The GPT-4-based agents outperformed the two tested open-source models in completing all testing subtasks including kinetic experiment design, automated hardware execution, NMR analysis, and kinetic model fitting. However, the open-source LLM-based agents also demonstrated acceptable performance, despite some minor errors in code generation and document information retrieval. These discrepancies were attributed to the performance differences between the LLMs and the effectiveness of the RAG method using OpenAI's proprietary implementation compared to open-source alternatives. However, with continuing development of open-source LLMs, their capability to function as the base model is expected to improve progressively over time.

可重复性与透明性: 诸如 GPT-4 等闭源 LLM 存在多个挑战，包括长期可重复性较差、缺乏透明度以及数据隐私问题。构建基于开源 LLM 的代理可以缓解这些问题。在本研究中，我们对比了基于开源 LLM (Qwen2-72B 和 Llama3.1-70B) 与基于 GPT-4 的代理在反应动力学研究任务中的表现。GPT-4 代理在完成所有测试子任务（包括动力学实验设计、自动化硬件执行、NMR 分析和动力学模型拟合）方面优于两个测试的开源模型。然而，基于开源 LLM 的代理也表现出可接受的性能，尽管在代码生成和文档信息检索方面存在一些小错误。这些差异归因于 LLM 之间的性能差异，以及 OpenAI 专有 RAG 方法相较于开源替代方案的有效性。然而，随着开源 LLM 的持续发展，其作为基础模型的能力预计将在未来逐步提高。

Communication among LLM-based agents: In this work, all developed agents were connected via human for message passing, since we would like to involve human inspections on the agent-generated experimental plans and results. This approach would avoid any potential errors in agents' response that might lead to hardware malfunction. Moving forward with improved reliability of LLM base models, it would be desired to develop a multi-agent system similar to AutoGen framework that allows direct communication between agents. In this proposed system, human intervention would be only required for critical decisions, such as automated equipment operations or complex experimental designs.

基于 LLM 代理之间的通信: 在本研究中，所有已开发的代理均通过人为干预进行信息传递，因为我们希望对代理生成的实验方案和结果进行人工检查。这种方法可以避免代理响应中的潜在错误，防止导致硬件故障。随着 LLM 基础模型可靠性的提高，未来应开发类似于 AutoGen 框架的多代理系统，以实现代理之间的直接通信。在该系统中，人工干预仅用于关键决策，例如自动化设备操作或复杂实验设计。

Discussion 讨论

In this work, we developed and demonstrated LLM-RDF for the end-to-end development workflow of the sustainable aerobic alcohol oxidation, from methodological search to product purification. Then, its utility was further demonstrated in three real-world chemical synthesis development tasks.

在本研究中，我们开发并展示了 LLM-RDF 在可持续有氧醇氧化端到端开发工作流程中的应用，从方法学搜索到产物纯化。随后，我们进一步在三个实际化学合成开发任务中验证了其实用性。

The specialized LLM-based agents showcased their versatility in autonomous chemical research, undertaking tasks such as synthesis method search, code composing for automated equipment, spectrum signal processing and analysis, reaction stoichiometric calculation, optimization of separation eluent composition, reactor design, and deriving chemically informed conclusions.

专门构建的基于 LLM 的代理在自主化学研究中展现了多功能性，承担了合成方法搜索、自动化设备代码编写、光谱信号处理与分析、反应化学计量计算、分离洗脱剂组成优化、反应器设计以及化学信息推导等任务。

LLM-RDF demonstrates a transformative approach to chemical synthesis that integrates chemist users, LLM-based agents, and automated experimental platforms, significantly streamlining the traditional expert-driven and labor-intensive workflow of reaction development.

LLM-RDF 体现了一种变革性的化学合成方法，将化学家用户、基于 LLM 的代理和自动化实验平台整合在一起，大幅简化了传统依赖专家驱动且劳动密集型的反应开发工作流程。

Methods

Construction of LLM-based agents 基于 LLM 的代理构建

LLM-based agents developed in this work were based on OpenAI's GPT-4 model and two open-source LLMs (Qwen2-72B and Llama3.1-70B). These intelligent agents include:

本研究中开发的基于 LLM 的代理基于 OpenAI 的 GPT-4 模型以及两个开源 LLM (Qwen2-72B 和 Llama3.1-70B)。这些智能代理包括：

(1) Literature Scouter: This agent was developed using Consensus available from OpenAI's GPT store, which can access Semantic Scholar database for academic literatures.

文献搜索代理: 该代理使用 OpenAI GPT 商店中的 Consensus 进行开发，可访问 Semantic Scholar 数据库以检索学术文献。

(2) Experiment Designer: This agent designs chemical experiments and transforms reaction procedures and parameters described in natural language into standardized reaction execution protocols to interface with experimental platforms.

实验设计代理: 该代理负责设计化学实验，并将自然语言描述的反应步骤和参数转换为标准化的反应执行协议，以便与实验平台对接。

(3) Hardware Executor: Specific hardware running code examples or Opentrons Python API manual were provided in the prompt, such that Hardware Executor could generate running codes for the automation platforms according to the standardized execution protocols.

硬件执行代理: 在提示中提供了特定硬件的运行代码示例或 Opentrons Python API 手册，使得硬件执行代理能够根据标准化执行协议为自动化平台生成运行代码。

(4) Spectrum Analyzer: This agent processes raw spectral data obtained from analytical apparatus (e.g., gas chromatograph and NMR), identifies the target compound peaks, and calculates the reaction outcomes.

光谱分析代理: 该代理处理从分析设备（如气相色谱和 NMR）获得的原始光谱数据，识别目标化合物峰，并计算反应结果。

(5) Separation Instructor: This agent instructs on identifying the appropriate TLC eluent composition to be used for subsequent flash column chromatography separation.

分离指导代理: 该代理指导确定适用于后续快速柱色谱分离的 TLC 最佳洗脱剂组成。

(6) Result Interpreter: This agent interprets and concludes experiment results based on fundamental chemical knowledge.

结果解释代理: 该代理基于基本化学知识分析并解读实验结果。

We provided detailed descriptions and instructions as preprompts to teach them to perform chemical synthesis development tasks. For more details, refer to the Supplementary Information Section 1.

我们为这些代理提供了详细的描述和指令作为预设提示，以指导其执行化学合成开发任务。更多细节参见补充信息第 1 节。

Web application

The web application functioned as the interface through which users could interact with agents and experimental platform. The frontend graphical interface was developed using the Vue.js and Node.js frameworks, creating a user-friendly and interactive environment.

该 Web 应用程序作为用户与代理及实验平台交互的接口。前端图形界面采用 Vue.js 和 Node.js 框架开发，提供了一个用户友好且交互性强的环境。

For the backend, the Python FastAPI framework was employed to manage the logics of multi-agent system and experimental platform, including interfacing with the LLM-based agents through the GPT-4 APIs hosted on Microsoft Azure and handling the operations of the experimental platforms.

在后端，采用 Python FastAPI 框架管理多代理系统和实验平台的逻辑，包括通过托管在 Microsoft Azure 上的 GPT-4 API 与基于 LLM 的代理进行交互，并处理实验平台的操作。

In addition, the web application was segmented into individual modules corresponding to each task of the chemical synthesis reaction development workflow.

此外，该 Web 应用程序被划分为多个独立模块，以对应化学合成反应开发工作流程中的各个任务。

OT-2 liquid handler platform 液体处理平台

The experimentation for substrate scope screening, reaction kinetics study, and condition optimization of photocatalytic reaction was conducted using the Opentrons OT-2 liquid handling workstation.

用于底物范围筛选、反应动力学研究和光催化反应条件优化的实验是在 Opentrons OT-2 液体处理工作站上进行的。

In the OT-2, modules including the pipette module (P300 GEN2, 20-300 μL) for liquid transferring, heater-shaker module (200-3000 RPM, 37 – 95°C) for enhancing mixing of reaction mixture, and storage module for storing reaction stock solutions.

OT-2 设备包括多个模块，如移液模块 (P300 GEN2, 20-300 μL) 用于液体转移，加热-震荡模块 (200-3000 RPM, 37 – 95°C) 用于增强反应混合物的混合效果，以及储存模块用于存放反应储备溶液。

Operation codes, generated by the Hardware Executor, were uploaded to the OT-2 via its desktop application or a Jupyter notebook to initiate automated reaction execution.

由硬件执行代理生成的操作代码通过 OT-2 桌面应用程序或 Jupyter notebook 上传至 OT-2，以启动自动化反应执行。

Automated reaction optimization platform 自动化反应优化平台

The reaction condition optimization of the aerobic alcohol oxidation was conducted using this automated hardware. The self-driven reaction condition optimization platform consists of three modules, including an automated synthesis equipment (Unchained Labs, Big Kahuna), a HPLC (Thermo Fisher Scientific Vanquish), and a six-axis robotic arm (AUBO-i5) with a linear track.

有氧醇氧化反应的条件优化是在该自动化硬件平台上进行的。该自驱动反应条件优化平台由三个模块组成，包括自动化合成设备 (Unchained Labs, Big Kahuna)、高效液相色谱仪 (HPLC, Thermo Fisher Scientific Vanquish) 以及带有直线轨道的六轴机械臂 (AUBO-i5)。

Big Kahuna automated experimental procedures, incorporating several components, including an extended tip liquid dispenser (20-3000 μL) for liquid transferring, the vortexing stations (60-3750 RPM) for mixing the reaction mixture, and a vial/plate gripper for transferring reaction vials and plates.

Big Kahuna 设备自动化实验流程，包括多个组件，如大体积移液装置 (20-3000 μL) 用于液体转移，涡旋混合站 (60-3750 RPM) 用于混合反应物，以及小瓶/板夹具用于传输反应小瓶和微孔板。

HPLC analyzed reaction mixtures using a C18 reverse-phase column, with water and MeCN as the mobile phases. The robotic arm was responsible for transferring samples between Big Kahuna and HPLC. HPLC 采用 C18 反相色谱柱，以水和乙腈 (MeCN) 作为流动相，对反应混合物进行分析。机械臂负责在 Big Kahuna 和 HPLC 之间转移样品。

This hardware platform was controlled via a customized LabVIEW software, and experimental procedures and parameters were defined by the JSON method files.

该硬件平台由定制的 LabVIEW 软件控制，实验流程和参数通过 JSON 方法文件进行定义。

Reaction optimization algorithm 反应优化算法

The Bayesian optimization algorithm and the PI stopping criterion was developed and discussed in previous work. In brief, it is composed of Gaussian process (GP) model and acquisition functions (AF): 贝叶斯优化算法和 PI 停止准则的开发及相关讨论已在之前的研究中进行。简而言之，该算法由高斯过程 (GP) 模型和采集函数 (AF) 组成：

$$\begin{aligned} \left\{x_{\text{inter}}^{(k)}\right\}_{k=1}^q &= \operatorname{argmax}_{\mathbf{x}} qEI\left(\left\{x^{(k)}\right\}_{k=1}^q\right) \\ &= \operatorname{argmax} \mathbb{E}_n\left(\operatorname{ReLU}\left(\max _{i=1, \ldots, q} f\left(x_i\right)-f_n\left(x^{+}\right)\right)\right) \end{aligned} \quad (2)$$

where $\left(x_{\text{inter}}^{(k)}\right)_{k=1}^q$ is the q newly proposed reaction conditions, x^{+} is the current optimal condition, and \mathbb{E}_n indicates that the expectation is taken under the posterior distribution at time n .

其中 $\left(x_{\text{inter}}^{(k)}\right)_{k=1}^q$ 为 q 个新提出的反应条件, x^{+} 为当前的最优条件, \mathbb{E}_n 表示期望值在时刻 n 的后验分布下取值。

The probability of improvement (PI) value is a measure of the possibility that the newly proposed reaction candidate could have an improvement over the current optimal value (Eq. 4).

改进概率 (PI) 值衡量了新提出的反应候选者能够比当前最优值有所改进的可能性 (等式 4)。

$$PI(x)=\mathbb{P}\left(f(x) \geq f\left(x^{+}\right)+\xi\right)=\Phi\left(\frac{\mu(x)-f\left(x^{+}\right)-\xi}{\sigma(x)}\right) \quad (3)$$

where $\mu(\cdot)$ is GP's mean, $\sigma(\cdot)$ is GP's standard deviation, $\Phi(\cdot)$ is the normal cumulative distribution function, and ξ is the trade-off parameter of exploitation and exploration.

其中 $\mu(\cdot)$ 是 GP 的均值, $\sigma(\cdot)$ 是 GP 的标准差, $\Phi(\cdot)$ 是正态累积分布函数, ξ 是开发和探索的权衡参数。

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All the relevant data generated in this study have been deposited in the GitHub repository under <https://github.com/Ruan-Yixiang/LLM-RDF>. Source data are provided in this paper. Source data are provided with this paper.

Code availability

All the relevant code are publicly available in the GitHub repository (<https://github.com/Ruan-Yixiang/LLM-RDF>). An online web application demo is available at <https://ruan-yixiang.github.io/LLM-RDF/#/main> (Note: this web application only deploys the frontend for illustrative purpose. For full functionality, both frontend and backend need to be deployed by following the guidelines available in the GitHub repository).

Acknowledgements

We acknowledge National Natural Science Foundation of China (22478335, 22227812, and 22108242) (Y.M.), National Key R&D Program of China (2021YFA1502700) (Y.M.), and Fundamental Research Funds for the Zhejiang Provincial Universities (226-2024-00113) (Y.M.) for providing support for this work for providing support for this work.

Author contributions

Y.R. and Y.M. conceived the project. Y.R. developed and implemented the LLM-based agents. Y.R. and C.L. developed the web application. Y.R., N.X., and J.X. designed and performed the chemical experiments. Y.H. and Y.C. contributed to the scale-up design strategy of the heterogeneous photoelectrochemical reactor. J.Z., H.G., and Q.Z. participated in discussions on the development of LLM-based agents. Y.R., Y.M., C.L., J.P., and Q.F. built the automated reaction optimization platform. Y.R. and Y.M. wrote the manuscript. X.S., N.Y., and Q.Z. reviewed the manuscript. Y.M. supervised the project and secured funding.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-024-54457-x>.

Correspondence and requests for materials should be addressed to Yiming Mo.

Peer review information

Nature Communications thanks Mayk Ramos and the other anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access

This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

本文根据 Creative Commons 署名-非商业-禁止演绎 4.0 国际许可协议授权，允许以任何媒介或格式进行任何非商业性使用、分享、分发和复制，只要您给予原作者和来源适当的认可、提供 Creative Commons 许可链接并表明您是否修改了许可材料。根据此许可，您无权分享根据本文或本文部分内容改编的材料。本文中的图像或其他第三方材料包含在本文的 Creative Commons 许可协议中，除非在材料的署名行中另有说明。如果材料未包含在本文的 Creative Commons 许可协议中，并且您的预期用途不被法律法规允许或超出了允许的用途，则您需要直接从版权所有者处获得许可。要查看此许可协议的副本，请访问 <http://creativecommons.org/licenses/by-nc-nd/4.0/>。