nature communications

Article

https://doi.org/10.1038/s41467-024-54457-x

An automatic end-to-end chemical synthesis development platform powered by large language models

Received: 3 May 2024

Accepted: 7 November 2024

Published online: 23 November 2024

Check for updates

Yixiang Ruan[1,2], Chenyin Lu[2], Ning Xu[1,2], Yuchen He[1,2], Yixin Chen[1,2], Jian Zhang[2], Jun Xuan[2], Jianzhang Pan[2,3], Qun Fang[2,3], Hanyu Gao[4], Xiaodong Shen[5], Ning Ye[6], Qiang Zhang[2,7] & Yiming Mo[1,2]

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 versability was validated on various synthesis tasks of three distinct reactions (S_NAr reaction, photoredox C-C cross-coupling reaction, and heterogeneous photoelectrochemical reaction).

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[1]. 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[2,3]. 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[4].

[1]College of Chemical and Biological Engineering, Zhejiang University, Hangzhou 310027, China. [2]Zhejiang-Hong Kong Joint Laboratory for Intelligent Molecule and Material Design and Synthesis, ZJU-Hangzhou Global Scientific and Technological Innovation Center, Hangzhou 311215, China. [3]Institute of Microanalytical Systems, Department of Chemistry, Zhejiang University, Hangzhou 310058, China. [4]Department of Chemical and Biological Engineering, The Hong Kong University of Science and Technology, Hong Kong 999077, China. [5]Chemical & Analytical Development, Suzhou Novartis Technical Development Co. Ltd., Changshu 215537, China. [6]Rezubio Pharmaceuticals Co. Ltd., Zhuhai 519070, China. [7]College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China. e-mail: yimingmo@zju.edu.cn

The recent advancement of machine learning (ML) technologies has shown great potential in expediting various subtasks during the synthesis design[5,6]. Notable examples include deep learning based quantitative structure–activity relationship (QSAR) model facilitating drug molecule design[7,8] and catalyst design[9], rapid identification of promising synthetic routes using machine-learning-aided synthesis planning[10,11], guiding automated high-throughput experimental platforms to search for optimal reaction conditions[12–14], and direct translation of multistep synthesis procedures from literature to experimental execution via natural language processing (NLP) models[15]. 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.

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[16] and Google's Gemini[17] as well as open-source Llama3.1[18], Mistral[19] and Qwen2[20], 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[21]. 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[22–29], molecule and material discovery[27,29–38], reaction condition recommendation and optimization[26–29,39–41], and lab apparatus automation[27–29,40–43].

The existing reports of LLM-based agents showed scattered coverage of the stages in chemical synthesis development (Fig. 1a), 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 (Fig. 1a). We selected aerobic alcohol oxidation to the aldehyde, an emerging sustainable aldehyde synthesis protocol[44] as a model

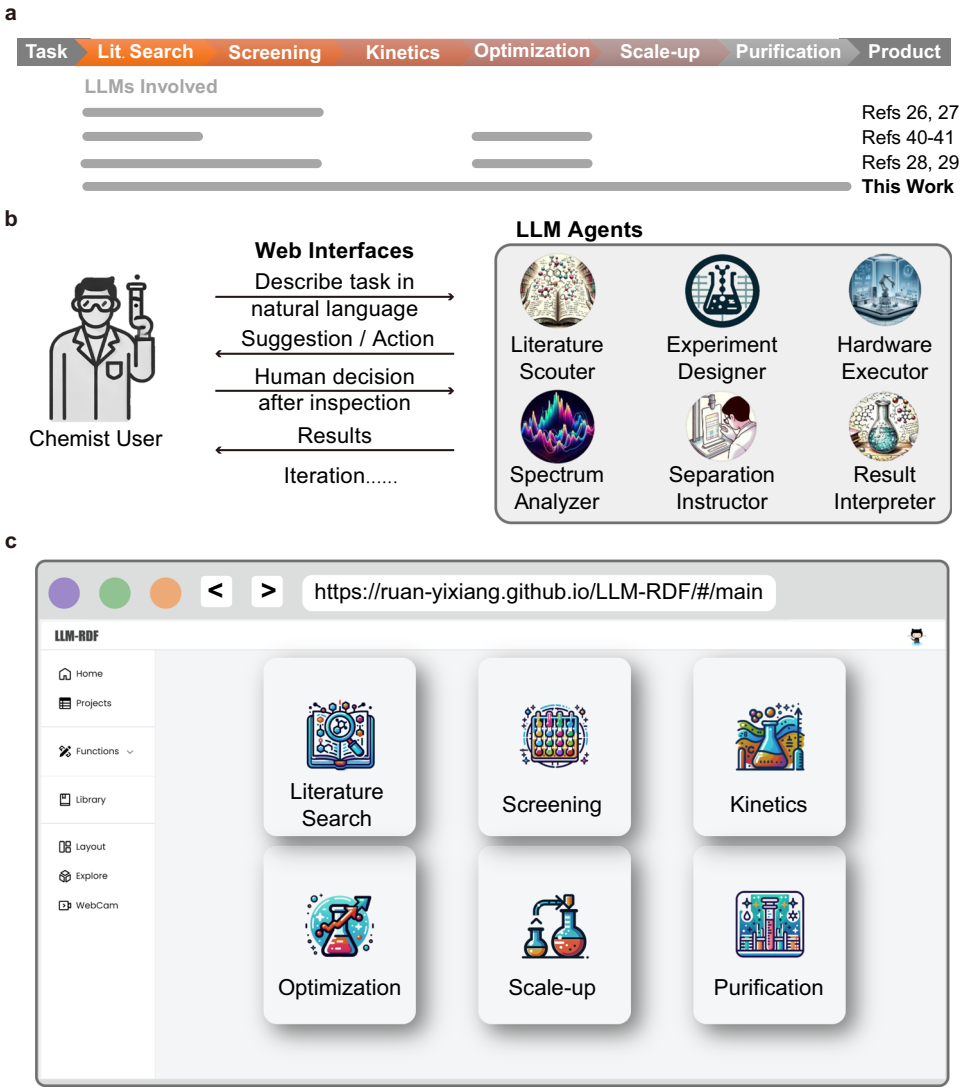

Fig. 1 | Overview of LLM-based multi-agent system for reaction development. **a** Workflow for chemical synthesis development facilitated by large language model (LLM) technology, and comparison with representative published works. (The gray lines denote the involvement of LLMs). **b** Diagram illustrating the interactions between human chemists and the LLM-based agents for performing tasks in the chemical synthesis development. **c** The web application with LLM-based agents as backend for end-to-end reaction development.

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.

Results

LLM-based agents for end-to-end chemical synthesis reaction development

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 (Fig. 1b). 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[45] 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.

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 (Fig. 1c and Supplementary Movie 1). 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[46] and retrieval-augmented generation (RAG)[47]. 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.

Literature search and information extraction

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 (Fig. 2b).

Among the various methods given by Literature Scouter (Fig. 2b and Supplementary Table 1), 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[48] as this method outpaced others in the aspects of the environmental sustainability, simplicity, safety, chemoselectivity, and substrate compatibility. After manually evaluating other

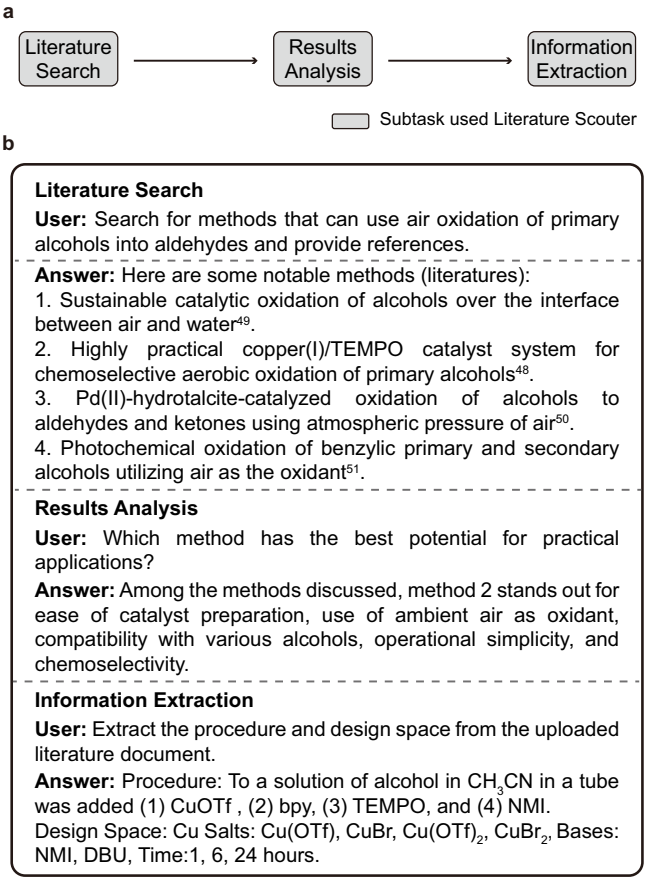

Fig. 2 | LLM-based agents facilitated literature search and information extraction. a Workflow for literature search and information extraction copiloted by Literature Scouter agent. **b** The interaction between human chemists with Literature Scouter. The dialog presented in the figure is simplified for the illustrative purpose, and see details in Supplementary Table 1.

recommended methods, this Cu/TEMPO catalytic chemistry indeed avoids the use of heterogeneous catalysts[49], high-cost palladium catalysts[50], or light irradiation[51] used in other approaches, proving to have claimed potentials in practical applications as suggested by the Literature Scouter. In addition, the chemoselective oxidation of the target hydroxyl group in diols or polyols is attractive in practice as function group protection and deprotection would not be required. The Literature Scouter recognized the capability of Cu/TEMPO catalytic system was able to selectively oxidize primary alcohols in presence of the secondary alcohols on the same molecule (Supplementary Table 1).

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 (Fig. 2b and Supplementary Table 1). This information served as the basis for the subsequent experimental exploration of this chemistry.

As demonstrated in the task of method search and information extraction from literature (Fig. 2a), 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 (Supplementary Table 3).

Substrate scope and condition screening

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[52,53]. The recent development of automated high-throughput screening (HTS) technology has been proven as a powerful tool to accelerate the experimental data acquisition for these substrate scope studies[54,55]. However, HTS technology is still not a routine tool that synthesis practitioners would use on their daily reaction development workflows. 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.

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 (Fig. 3a).

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 (Fig. 2b) 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 (Supplementary Table 4). Following its recommendation, we replaced acetonitrile with dimethyl sulfoxide (DMSO) as the solvent and used $CuCl_2$ and $Cu(BF_4)_2$ as Cu catalysts.

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 (Fig. 3b, and see details in Supplementary Tables 5–6 and Supplementary Fig. 9-11). To execute the HTS task, we chosen 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 (Supplementary Table 7). With this workflow from HTS task described in natural language to automated reaction execution, two rounds of HTS experiments were conducted (Fig. 4a-d), 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 [$CuCl_2$, $CuBr_2$, $Cu(OTf)_2$ and $Cu(BF_4)_2$], and two bases [N-methylimidazole (NMI) and 1,8-diazabicyclo-[5.4.0]undec-7-ene (DBU)].

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 automated this process (Fig. 3c). 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. 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 (Fig. 3d-e). 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) (Fig. 3f). 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 (Supplementary Fig. 30).

Finally, we utilized Result Interpreter agent to summarize HTS results (Fig. 4e-f) and explain observed patterns based on fundamental chemistry knowledge (Supplementary Table 18). Result Interpreter recognized that DBU base significantly outperformed NMI, and the reactivity of copper salt followed the order of $CuCl_2 < CuBr_2 < Cu(OTf)_2 \sim Cu(BF_4)_2$. 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[56]. 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[48].

Reaction kinetics study

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[48]) (Fig. 5b). To investigate the observed solvent effects, Experiment Designer agent suggested that we could conduct oxidation kinetics study for different solvent (Supplementary Table 19). Recently, automated kinetic profiling has become an efficient tool to help researchers establish reaction kinetic models[57]. 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

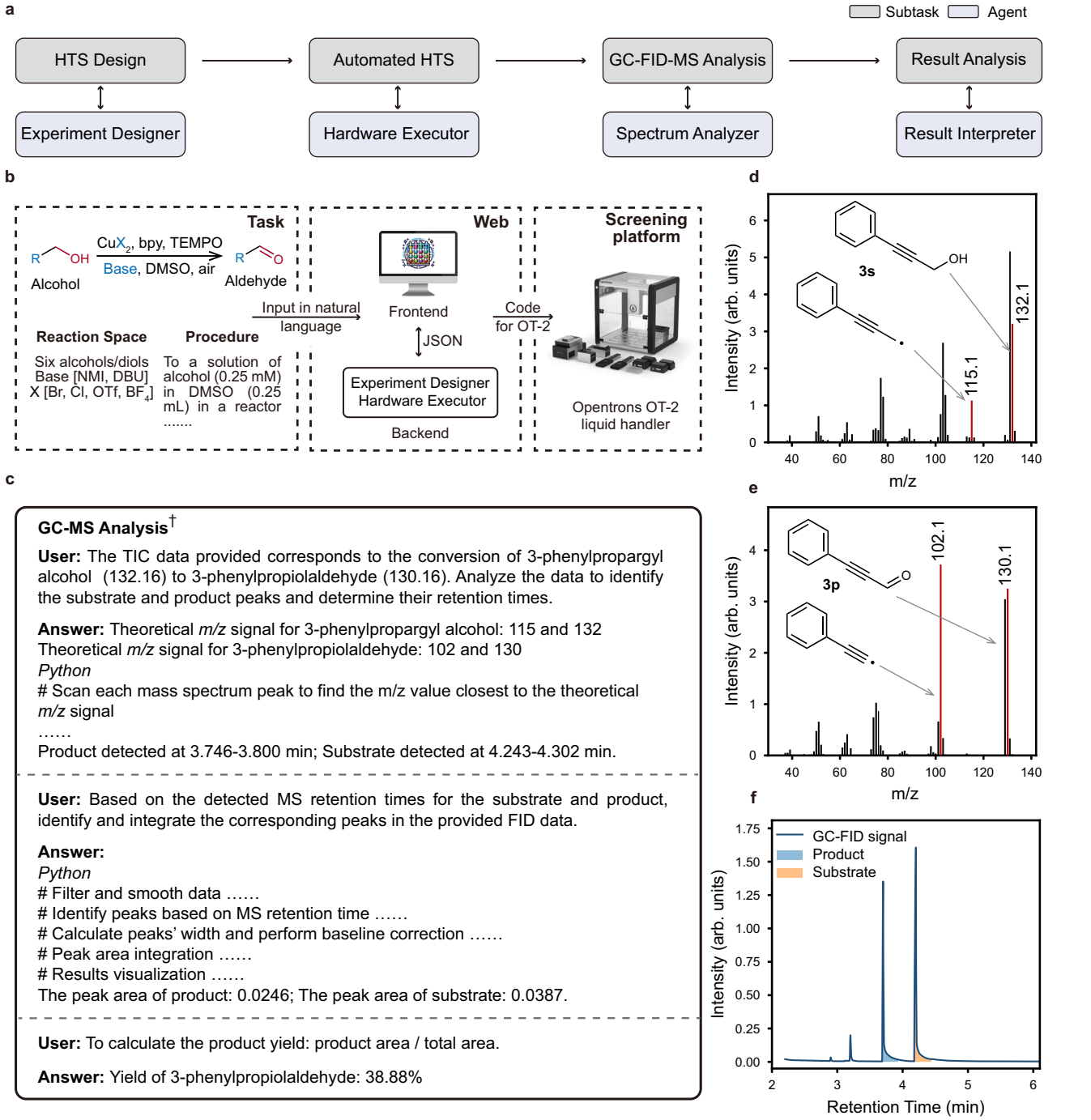

Fig. 3 | LLM-based agents facilitated substrate scope and condition screening. **a** Workflow for substrate scope and condition screening copiloted by Experiment Designer, Hardware Executor, Spectrum Analyzer, and Result Interpreter agents. **b** The aerobic alcohol oxidation reaction screening task described in natural language for subsequent large language model (LLM)-based agent understanding and OT-2 liquid handler reaction execution. The exact transcript of the natural language description of the task is provided in Supplementary Table 5-7. The image of Opentrons OT-2 liquid handler was obtained from the Opentrons website (www.opentrons.com.cn). **c** The interaction between human chemists with Spectrum Analyzer for gas chromatography with parallel flame ionization detector and mass spectrometer (GC-FID-MS) result analysis for **3 s** to **3p** in $CuCl_2$ and DBU condition (see details in Supplementary Table 11). The dagger symbol indicates that the numerical results were generated by the agents' code interpreter. **d** The mass spectra for the retention time within 4.243-4.302 min, matched with substrate **3s** by Spectrum Analyzer. **e** The mass spectra for retention time within 3.746-3.800 min, matched with product **3p** by Spectrum Analyzer. The red lines represent the characteristic charge ratio signals detected by the Spectrum Analyzer in mass spectrometry, which match the molecule fragments. **f** Visualization of GC-FID spectra peak integration for substrate **3s** and product **3p** given by Spectrum Analyzer.

resonance (^{1}H NMR) analysis, and kinetic model fitting and analysis (Fig. 5a).

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.

https://doi.org/10.1038/s41467-024-54457-x

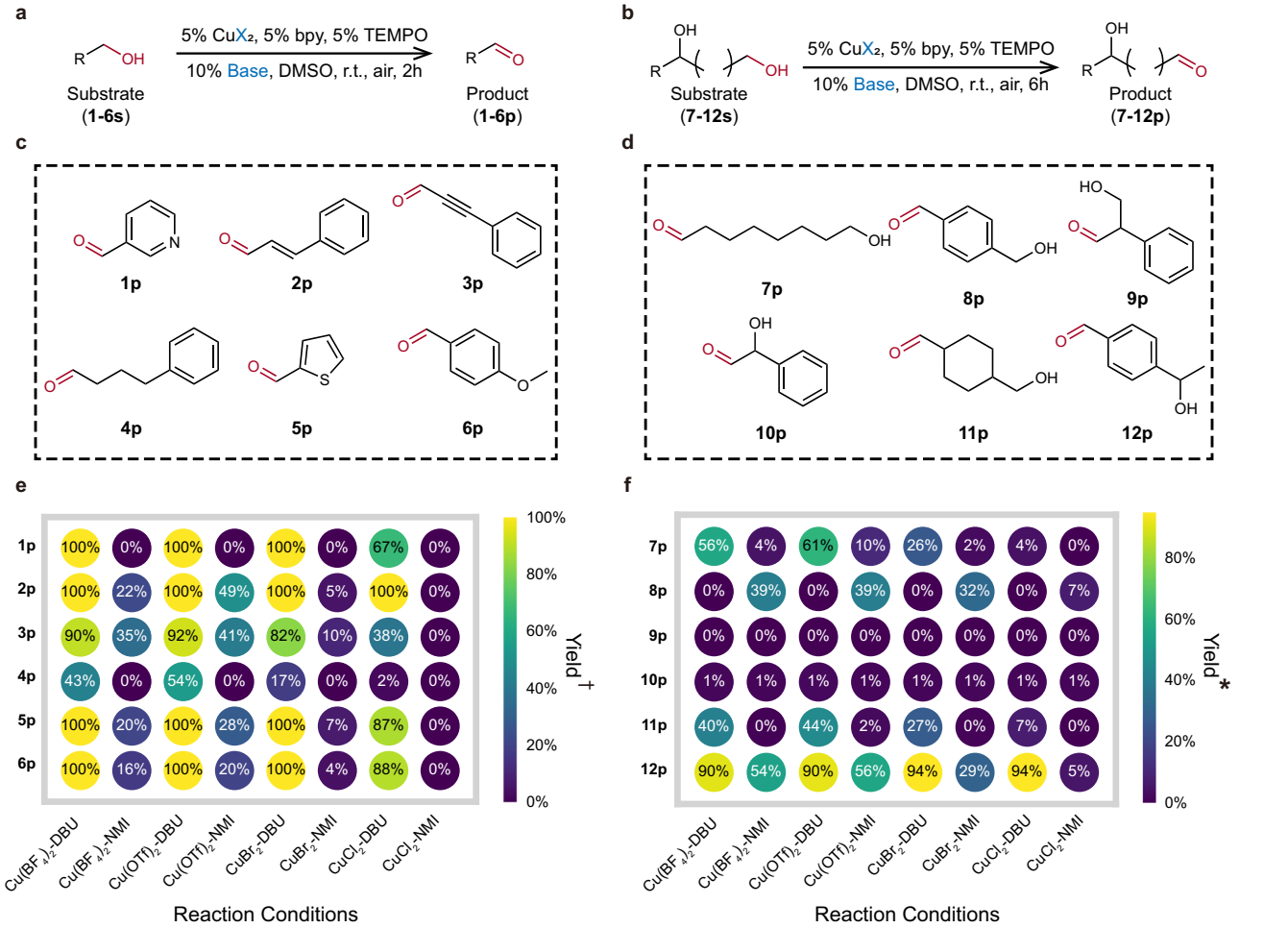

Fig. 4 | The substrate scope and condition screening results. The copper/TEMPO-catalyzed aerobic oxidation reaction of (**a**) monohydric alcohols and (**b**) diols to the corresponding aldehydes in the screening task. Reaction condition: substrate (0.25 mmol), 5 mol% Cu catalyst, 5 mol% bpy, 5 mol% TEMPO, and 10 mol % base were dissolved in DMSO solvent (1.25 mL), and reaction was performed under room temperature and open to air for 2 (monohydric alcohols) or 6 (diols) hours. The aldehyde products derived from the oxidation of the corresponding (**c**) monohydric alcohol and (**d**) diols. **1-12 s** represent the 12 alcohol substrates, and **1-12p** correspond to the respective aldehyde products. Yield heat maps for the oxidation of (**e**) monohydric alcohols and (**f**) diols under various combinations of copper catalyst and base. The dagger symbol indicates that yields were analyzed by Spectrum Analyzer (Supplementary Information Section 3.2.2). The asterisk indicates that yields were calculated by manual processing due to peak overlapping issues for diol oxidations (Supplementary Information Section 3.2.3).

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 (Fig. 5c and Supplementary Table S20). 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 (Supplementary Table 21). The compositions of the sampled reaction crude were analyzed by ^{1}H NMR. Instead of manual analysis of the NMR data, we provided Spectrum Analyzer with ^{1}H 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 (Fig. 5d and Supplementary Table 22).

Next, providing the obtained kinetics experiment results to Result Interpreter, it fitted the time-course data to the kinetic model equations (Supplementary Table 23). The reaction rate for substrate to product followed saturation kinetic dependence on the substrate alcohol (Eq. 1)[58], 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 (Fig. 5e and Supplementary Fig. 39).

$$r_{Product} = \frac{k_1 k_2 C_{Substrate}}{1 + k_1 C_{Substrate}} \quad (1)$$

$$r_{Byproducts} = k_3 C_{Product} \quad (2)$$

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 (Supplementary Table 24). This analysis highlighted that Result Interpreter had the ability to understand the underlying kinetics behind the observed the reaction selectivity.

https://doi.org/10.1038/s41467-024-54457-x

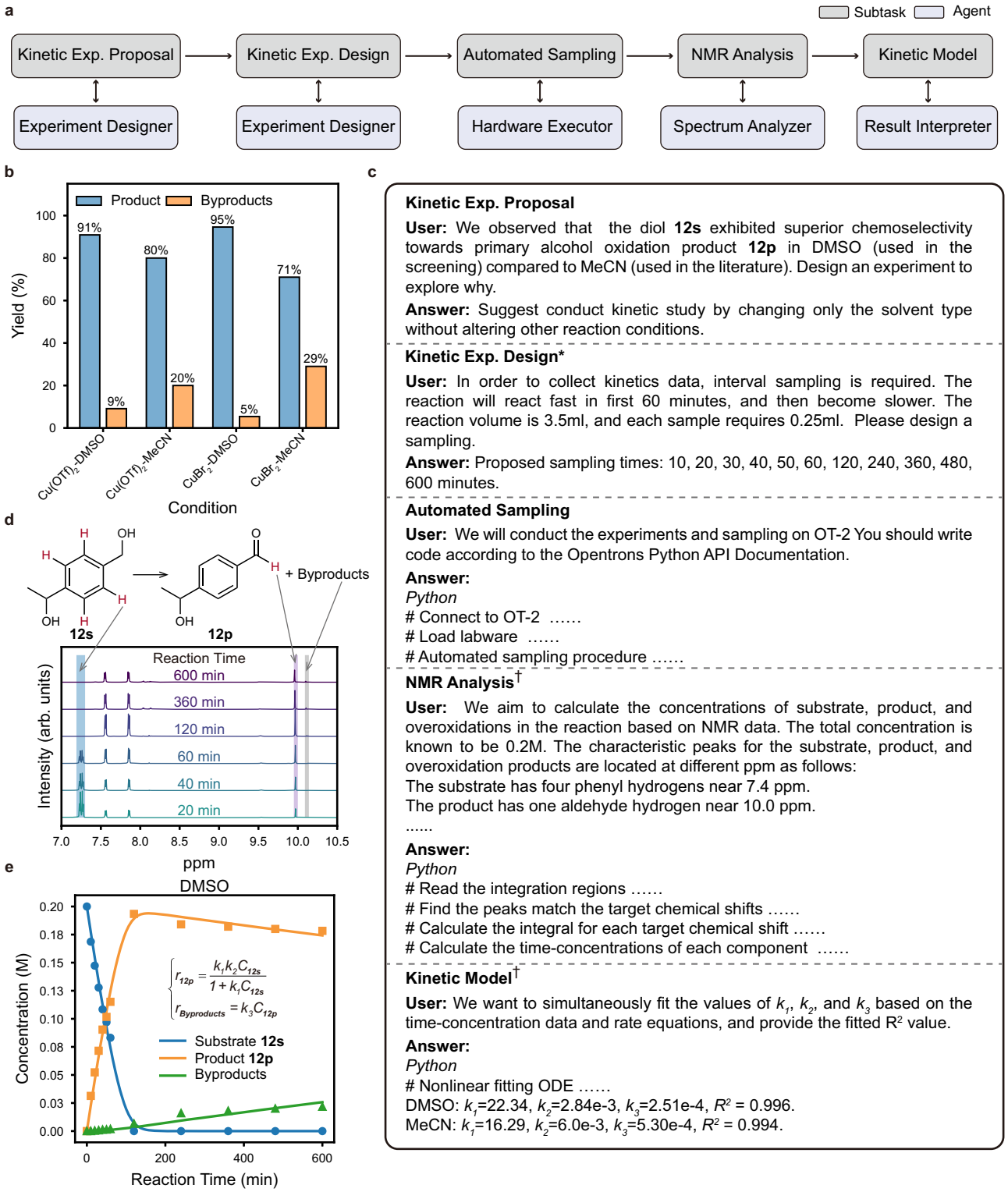

Fig. 5 | LLM-based agents facilitated reaction kinetics study. a Workflow for reaction kinetic study copiloted by Experiment Designer, Hardware Executor, Spectrum Analyzer, and Result Interpreter agents. **b** Comparison of the yields for products and overoxidation byproducts of diol **12 s** in different solvents (MeCN from the literature[48] and DMSO in this work) and copper catalysts. **c** The interaction between human chemists with large language model (LLM)-based agents for reaction kinetics study (see details in Supplementary Table 19-24). The dagger symbol indicates that the numerical results were generated by the agents' code interpreter, while the asterisk denotes that the numerical results were directly provided by the LLM. **d** Characteristic proton nuclear magnetic resonance (^{1}H NMR) peaks identified by Spectrum Analyzer for calculations of reaction samples' compositions. **e** The time-course concentration profile in DMSO solvent, and the fitted reaction kinetic curves given by Result Interpreter, with rate constants $k_1 = 22.34\ \mathrm{M}^{-1}$, $k_2 = 2.84 \times 10^{-3}\ \mathrm{M{\bullet}min}^{-1}$, $k_3 = 2.51 \times 10^{-4}\ \mathrm{min}^{-1}$, and the coefficient of determination $R^2 = 0.996$.

Reaction condition optimization

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)[12,14], and the mixed-integer nonlinear program (MINLP) algorithm[59], 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.

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 (Supplementary Information Section 5.1 and 5.4). 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. 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[41,60]. Thus, we chose to use BO as the optimizer in this work.

To demonstrate LLM-based agents copiloted reaction optimization workflow (Fig. 6a), we conducted the condition optimization for the selective oxidation of diol (**12s**) to the corresponding mono-oxidized 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**. 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) CuX_2/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 work-up procedure description [Add 0.75 mL HEDP.] into standardized JSON procedure steps (Supplementary Table 37) for display on the web application (Supplementary Fig. 44). Hardware Executor generated code templates based on these JSON procedure steps to define the automated synthesis platform operation workflows. 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: $CuCl_2$, $CuBr_2$, $Cu(OTf)_2$, $Cu(BF_4)_2$; 4. Base type: NMI, DBU.] into JSON format (Supplementary Table 38) that was used as inputs for the Bayesian optimizer (Supplementary Fig. 46). At last, users reviewed the entire experimental plan before running the reaction optimization on the automation hardware (Supplementary Fig. 47-50).

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** (Fig. 6b). Multiple high-yield reaction conditions were identified within the design space (Supplementary Table 39). 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[61], 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 (Fig. 6c), based on which the optimal conditions should be confidently identified. In comparison, Result Interpreter was used to determine the appropriate stopping point for the optimization task using the concept of balancing exploration and exploitation for black-box function optimization (Supplementary Table 40). During the exploitation of $CuBr_2$-DBU combination (after 12 experiments), Result Interpreter indicated that the yield was sufficiently high to consider stopping optimization, however, it still recommended further exploration in copper catalysts based on exploration considerations. Then, BO continued to explore two more catalysts (i.e., $Cu(BF_4)_2$ and $Cu(OTf)_2$). After several small condition adjustments proposed by BO near the high-yield conditions, the reaction yield did not increase significantly, and a yield decrease was observed in the 22nd experiment. Result Interpreter once again suggested considering the cessation of the optimization. After the 26th experiment, Result Interpreter assessed the reaction yield as sufficiently high and the exploration of the reaction space as comprehensively executed, explicitly recommending the termination of further optimization (Fig. 6d). This comparison showed that the optimization stopping suggestions given by Result Interpreter agent were more intuitive and also required less experiments to identify high-yield reaction conditions compared to PI stopping criterion. Unlike the PI stopping criterion relying on human experience to pre-define the stopping threshold (improper selection may lead to poor optimization results or excessive number of optimization experiments), utilizing Result Interpreter to terminate optimization offers better flexibility and adaptability.

Reaction scale-up and product purification

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[62]. 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 (Fig. 7a).

Among various high-yield (≥94.5%) conditions during the condition optimization of diol oxidation, Experiment Designer selected the condition used in 35th experiment for scaling up (Fig. 7b, Supplementary Table 41). 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% (Fig. 7c). 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 (Fig. 7b and Supplementary Table 42). 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 (Supplementary Table 43-44). We then conducted the scale-up experiment according to the parameters proposed by Experiment Designer.

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

https://doi.org/10.1038/s41467-024-54457-x

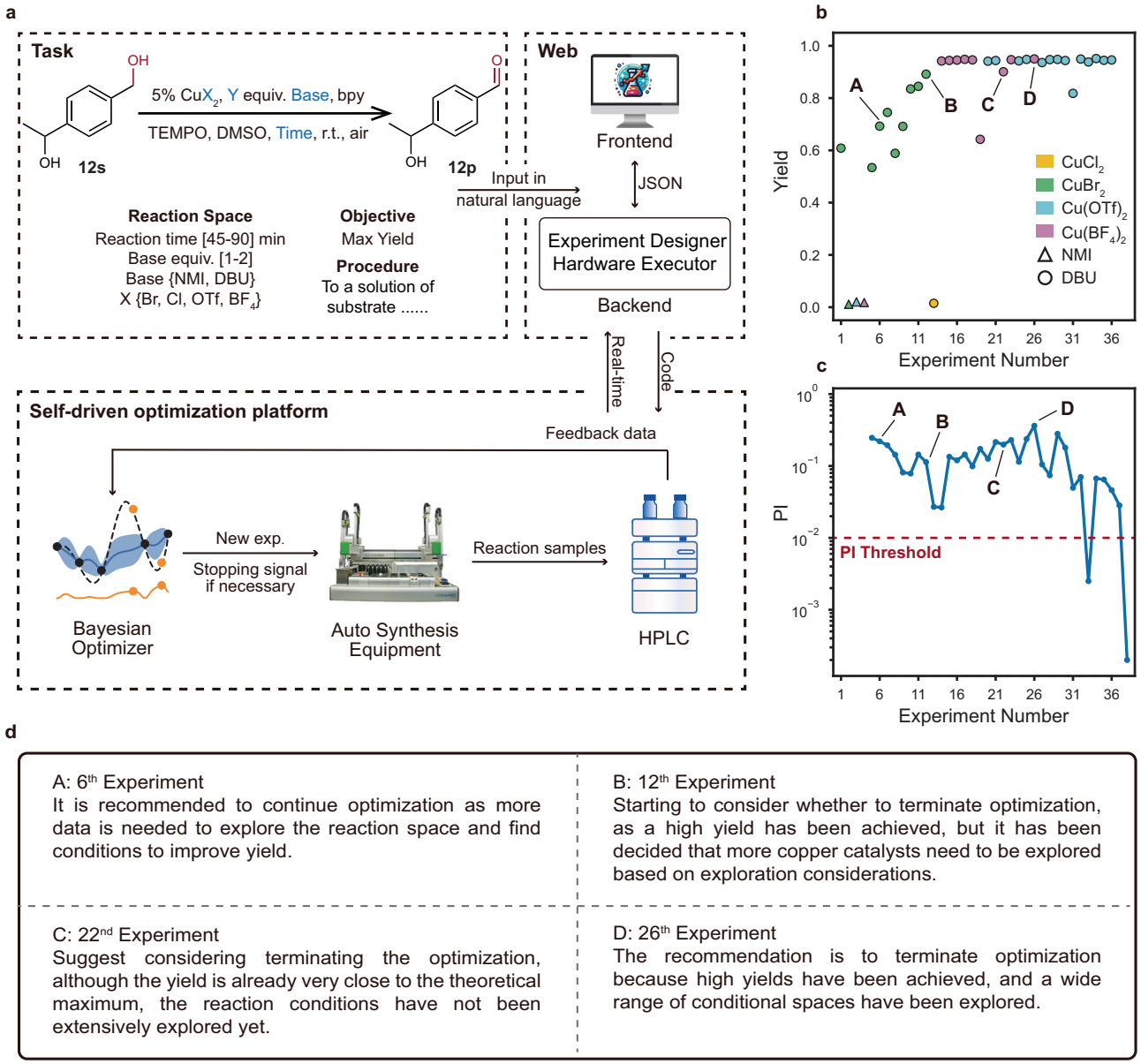

Fig. 6 | LLM-based agents facilitated self-driven reaction condition optimization. a The large language model (LLM)-based agents copiloted self-driven reaction optimization system. Users interface with the hardware system via natural language through the web application with Experiment Designer and Hardware Executor as the backend. The exact transcript of the natural language description of the task is provided in Supplementary Table 37-38. The automated reaction optimization platform, driven by Bayesian optimization (BO) algorithm, performed closed-loop reaction and analysis using automated Unchained synthesis platform and high-performance liquid chromatography (HPLC), respectively. The image of Unchained Labs Big Kahuna synthesis platform was obtained from the Unchained Labs website (https://unchained-labs.cn). The evolution profile of (**b**) yield and (**c**) probability of improvement (PI) value during the closed-loop reaction optimization process. **d** Result Interpreter's recommendations on whether reaction optimization should be terminated at 6th, 12th, 22nd, and 26th experiment (see detailed interaction dialogs in Supplementary Table 40).

target compound. A recent publication has applied machine learning model to predict the R_f value of a given compound structure in different eluent compositions[63]. 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. 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. 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**. 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 Rf value for the impurity, providing a sufficiently large difference for effective separation (Fig. 7b and Supplementary Table 45). 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% (Fig. 7d).

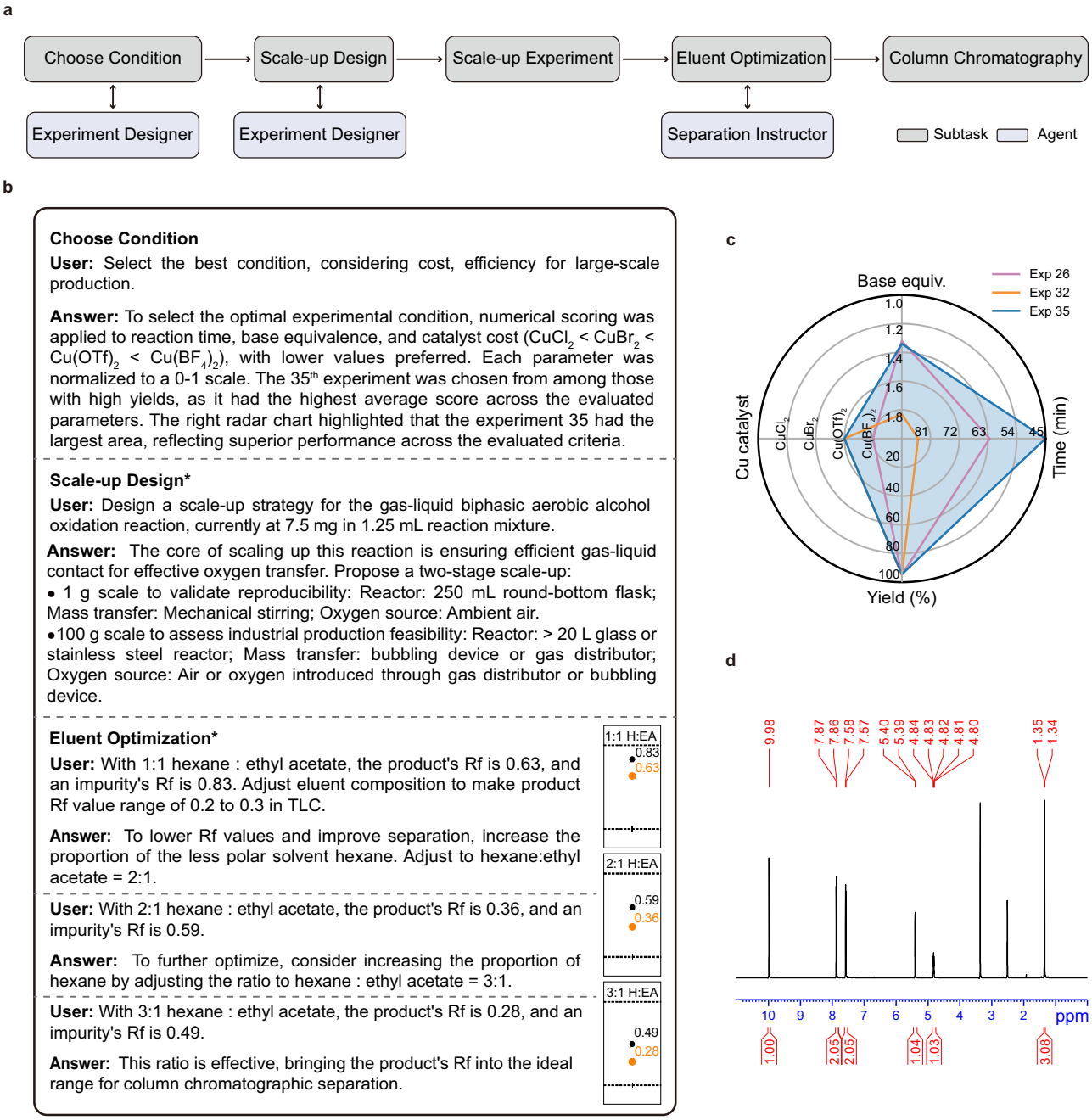

Fig. 7 | LLM-based agents facilitated reaction scale-up and product purification. **a** Workflow for reaction scale-up and product purification copiloted by Experiment Designer and Separation Instructor agents. **b** The interaction between human chemists with Experiment Designer for reaction scale-up and Separation Instructor for finding the optimal eluent composition (see detailed interaction dialogs in Supplementary Table 41-45). The asterisk denotes that the numerical results were directly provided by the large language model (LLM). **c** Radar chart for comparing three high-yield reaction conditions obtained during self-driven reaction optimization (experiments 26, 32, and 35). **d** Proton nuclear magnetic resonance (^{1}H NMR) spectrum of the purified target product (**12p**) in DMSO-d6 (See complete spectral information in Supplementary Information Section 6.7).

Applications

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.

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[64], 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**) [65] (Fig. 8a). Experiment Designer proposed a sampling schedule over a 2-h reaction period (Fig. 8b and Supplementary Table 46), based on which Hardware Executor generated the OT-2 running code to automate the reaction sampling process (Supplementary Table 47). Subsequently, the obtained kinetic data were supplied to Result Interpreter, which identified $r_{15} = 0.2C_{13}C_{14}^2$ was the best-fit kinetic model ($R^2 = 0.995$) among various possible kinetic

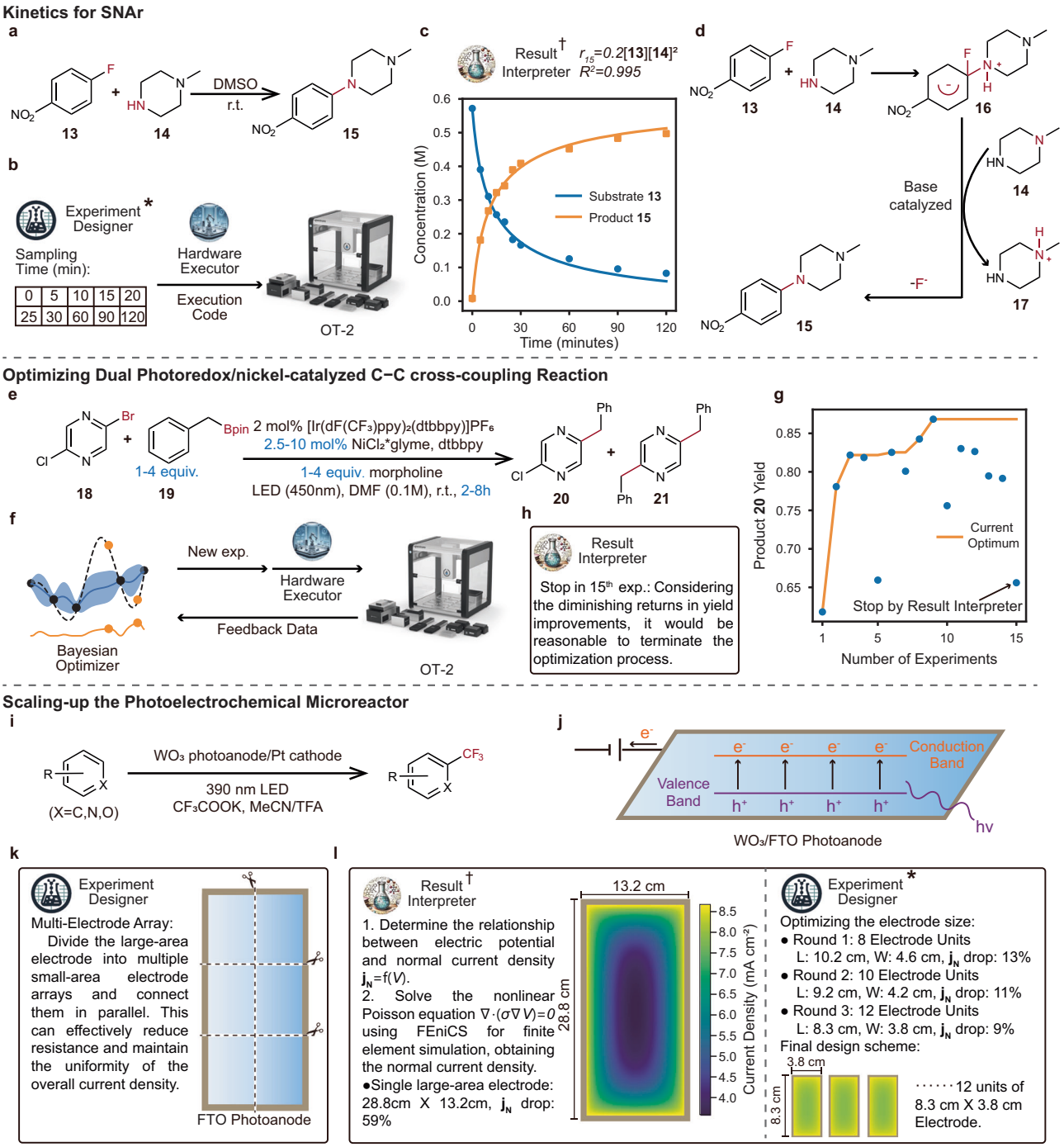

Fig. 8 | The applications of LLM-RDF in the chemical synthesis development. **a** Nucleophilic aromatic substitution (S_NAr) reaction. **b** Experiment Designer proposed a 2-hour sampling schedule for S_NAr kinetic experiments, and Hardware Executor generated liquid handler OT-2 code to automate the sampling process. **c** The kinetic model of the S_NAr reaction identified by Result Interpreter as the most suitable through analysis of kinetic data among various possible models. **d** Mechanism of the S_NAr reaction. **e** Photocatalytic cross-coupling reaction via amino radical transfer (ART) strategy. **f** Workflow for automated photocatalytic reaction optimization, in which Hardware Executor generated OT-2 running code, OT-2 executed the experiments, and Bayesian optimization (BO) algorithm suggested next-round trials. **g** Yield of **19** during the photocatalytic reaction optimization process driven by BO. **h** Result Interpreter's recommendation at 15th experiment to stop reaction optimization. **i** Photoelectrochemical decarboxylative trifluoromethylation. **j** Photoelectrochemical reaction mechanism in tungsten trioxide (WO_3) fluorine-doped tin oxide (FTO) glass photoanode: The incident photon (hv) excites the photoanode, generating electron-hole pairs (e^-/h^+) pairs. Electrons flow to the circuit via FTO, while holes drive the oxidation of trifluoroacetate. **k** Multi-electrode array approach proposed by Experiment Designer. The right part illustrates the division of the photoanode into smaller sections. **l** The left part shows how Result Interpreter constructed a finite-element conductivity (FEC) model for the current distribution simulation in FTO photoanode, revealing a 59% edge-to-center current density drop in a single large-area electrode. The right part illustrates the optimization process of FTO photoanode dimensions by Experiment Designer. 12 parallel-connected array photoanodes, each measuring 3.8 cm × 8.3 cm, meeting the design requirements. The heatmap plots of current distribution for the FTO photoanode in panel l share the color bar. The dagger symbol indicates numerical results from LLM-based agents' code interpreter, while the asterisk denotes those provided by the LLM directly. The images of Opentrons OT-2 liquid handler in Fig. 8b and f were obtained from the Opentrons website (www.opentrons.com.cn).

models (Fig. 8c and Supplementary Table 48). In terms of mechanistic explanation, Result Interpreter inferred that the second-order dependence on the concentration of N-methylpiperazine (C_{14}) indicated the bifunctional roles of **14** in S_NAr reaction besides being a nucleophile (Supplementary Table 48). 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 (Fig. 8d)[65].

The recently discovered amino radical transfer (ART) strategy enabled $C(sp^2)$-$C(sp^3)$ cross-coupling reactions between alkyl boronic esters and aryl halides under mild visible-light irradiation, representing an important advancement in the cross-coupling chemistry (Fig. 8e)[66]. 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. 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**). 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**. Hardware Executor generated the OT-2 running code based on the optimization task description for automating the execution of the experiments (Supplementary Table 49). The experimental outcome was fed into BO to suggest next-round trial (Fig. 8f). After three rounds of iteration, each consisting of five experiments (Fig. 8g), Result Interpreter concluded that the diminishing gains in yield improvement made it reasonable to terminate the optimization process (Fig. 8h and Supplementary Table 50). Under the optimal reaction condition of 1.38 equivalents of morpholine, 3 equivalents of **19**, and 10 mol% $NiCl_2$•glyme over 7.1 h, substrate **18** was fully consumed, yielding 87% of product **20** with almost no formation of byproduct **21** (Supplementary Table 51).

The recently emerging semiconductor-based heterogeneous photoelectrochemistry provides an unique approach to achieve single-electron transfer for radical generations[67]. However, due to the high sheet resistivity (~7 Ω/□) of the fluorine-doped tin oxide (FTO) glass for loading semiconductor photoelectrocatalysts (Fig. 8j), its nonuniform current distribution on the large-size FTO glass electrode creates significant challenges for scaling-up the synthesis throughput. To address this scale-up challenge, we attempted to employ LLM-RDF to propose a viable solution. Experiment Designer proposed the strategy of dividing the large electrode into an array of multiple small electrodes and connecting them in parallel, referred to as the multi-electrode array strategy, and suggested optimizing the size of the electrode units through finite element analysis (FEA) (Fig. 8k and Supplementary Table 52). Following this strategy, we sought to reproduce the photoelectrochemical microreactor (PEC-μReactor) design that was carefully engineered via COMSOL simulation and experimental validation by human researchers for decarboxylative trifluoromethylation reaction[68]. Here, we targeted a total 380 cm^2 photoanode size and <10% current distribution non-uniformity as reported. Result Interpreter first determined the relationship between electric potential and normal current density (Supplementary Table 53 and 54), and utilized the open-source FEA simulation package (FEniCS[69]) to construct a finite-element conductivity (FEC) model for the current distribution simulation in FTO photoanodes (Fig. 8i and Supplementary Table 55). The FEC model revealed that a single 380 cm^2 photoanode (width : length = 1 : 2.2) had a 59% edge-to-center current drop (Fig. 8l), resulting inefficient usage of photoanode. Experiment Designer followed the multi-electrode array strategy and identified that 12 small pieces of FTO photoanodes with 3.8 cm width and 8.3 cm length (fixed width-to-length ratio as the large-size photoanode) would suffice to keep the edge-to-center current drop within 10% (Fig. 8l and Supplementary Table 56). This photoanode scale-up design proposed by LLM agents was consistent with the solution originally reported[68].

Limitations and outlook

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.

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 (Supplementary Fig. 14-15, 62-63) before execution to avoid potential equipment damage or even personal injuries[70]. A recent study has demonstrated that introducing another LLM to automatically inspect and modify the responses from LLMs could partially mitigate unreliable response issues[71].

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 fine-tuning methods, significantly enhances their performance on chemistry-related tasks[38,72–77]. 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.

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 handle 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[78].

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 (Supplementary Information Section 4.6). 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 (Supplementary Fig. 40). 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.

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[79] 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.

Discussion

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. 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. 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. Although the LLM technology is still nascent in chemistry applications primarily due to the aforementioned limitations, we would envision that this work outlines a viable avenue to a deeper engagement of LLM technology in reaction development and relevant fields in the future.

Methods

Construction of LLM-based agents

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: (1) Literature Scouter: This agent was developed using Consensus[80] available from OpenAI's GPT store, which can access Semantic Scholar database for academic literatures. (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. (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. (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. (5) Separation Instructor: This agent instructs on identifying the appropriate TLC eluent composition to be used for subsequent flash column chromatography separation. (6) Result Interpreter: This agent interprets and concludes experiment results based on fundamental chemical knowledge.

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

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

OT-2 liquid handler platform

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. 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. 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.

Automated reaction optimization platform

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

Reaction optimization algorithm

The Bayesian optimization algorithm and the PI stopping criterion was developed and discussed in previous work[61]. In brief, it is composed of Gaussian process (GP) model and acquisition functions (AF). GP was a mixed kernel (Supplementary Equation (3)), combining the Matérn52 kernel (Supplementary Equation (1)) with the categorical kernel (Supplementary Equation (2)), to handle the reaction's design space, which includes both continuous and categorical variables. The new experiment candidates are proposed by maximizing the multi-points expected improvement (qEI) acquisition functions:

$$\begin{aligned}\left\{\mathbf{x}_{new}^{(k)}\right\}_{k=1}^{q} &= \mathrm{argmax}\,\mathrm{qEI}\left(\left\{\mathbf{x}^{(k)}\right\}_{k=1}^{q}\right)\\ &= \mathrm{argmax}\,\mathbb{E}_n\left(\mathrm{ReLu}\left(\max_{i=1,..,q}\mathrm{f}(\mathbf{x}_i)-\mathrm{f}_n(\mathbf{x}^+)\right)\right)\end{aligned} \tag{3}$$

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

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).

$$\mathrm{PI}(\mathbf{x}) = \mathbb{P}\left(\mathrm{f}(\mathbf{x}) \geq \mathrm{f}(\mathbf{x}^+) + \xi\right) = \Phi\left(\frac{\mu(\mathbf{x}) - \mathrm{f}(\mathbf{x}^+) - \xi}{\sigma(\mathbf{x})}\right) \tag{4}$$

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.

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[81]. 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[81] (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).

References

1. DiMasi, J. A., Grabowski, H. G. & Hansen, R. W. Innovation in the pharmaceutical industry: new estimates of R&D costs. *J. Health Econ.* **47**, 20–33 (2016).
2. Feng, F., Lai, L. & Pei, J. Computational chemical synthesis analysis and pathway design. *Front. Chem.* **6**, 199 (2018).
3. Molga, K., Szymkuć, S. & Grzybowski, B. A. Chemist ex machina: advanced synthesis planning by computers. *Acc. Chem. Res.* **54**, 1094–1106 (2021).
4. Andersson, S. et al. Making medicinal chemistry more effective—application of lean sigma to improve processes, speed and quality. *Drug Discov. Today* **14**, 598–604 (2009).
5. Struble, T. J. et al. Current and future roles of artificial intelligence in medicinal chemistry synthesis. *J. Med. Chem.* **63**, 8667–8682 (2020).
6. Griffin, D. J., Coley, C. W., Frank, S. A., Hawkins, J. M. & Jensen, K. F. Opportunities for machine learning and artificial intelligence to advance synthetic drug substance process development. *Org. Process Res. Dev.* **27**, 1868–1879 (2023).
7. Stokes, J. M. et al. A deep learning approach to antibiotic discovery. *Cell* **180**, 688–702.e13 (2020).
8. Wong, F. et al. Discovery of a structural class of antibiotics with explainable deep learning. *Nature* **626**, 177–185 (2024).
9. Zahrt, A. F. et al. Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning. *Science* **363**, eaau5631 (2019).
10. Coley, C. W. et al. A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science* **365**, eaax1566 (2019).
11. Mikulak-Klucznik, B. et al. Computational planning of the synthesis of complex natural products. *Nature* **588**, 83–88 (2020).
12. Shields, B. J. et al. Bayesian reaction optimization as a tool for chemical synthesis. *Nature* **590**, 89–96 (2021).
13. Wang, J. Y. et al. Identifying general reaction conditions by bandit optimization. *Nature* **626**, 1025–1033 (2024).
14. Slattery, A. et al. Automated self-optimization, intensification, and scale-up of photocatalysis in flow. *Science* **383**, eadj1817 (2024).
15. Steiner, S. et al. Organic synthesis in a modular robotic system driven by a chemical programming language. *Science* **363**, eaav2211 (2019).
16. ANTHROP\C. *Claude 3.5 Sonnet*. https://www.anthropic.com/news/claude-3-5-sonnet (2024).
17. Gemini Team et al. Gemini: A family of highly capable multimodal models. *arXiv* http://arxiv.org/abs/2312.11805 (2024).
18. Dubey, A. et al. The lama 3 herd of models. *arXiv* http://arxiv.org/abs/2407.21783 (2024).
19. Jiang, A. Q. et al. Mistral 7B. *arXiv* http://arxiv.org/abs/2310.06825 (2023).
20. Yang, A. et al. Qwen2 technical report. *arXiv* http://arxiv.org/abs/2407.10671 (2024).
21. Wang, L. et al. A survey on large language model based autonomous agents. *Front. Comput. Sci.* **18**, 186345 (2024).
22. Zheng, Z., Zhang, O., Borgs, C., Chayes, J. T. & Yaghi, O. M. ChatGPT chemistry assistant for text mining and the prediction of MOF synthesis. *J. Am. Chem. Soc.* **145**, 18048–18062 (2023).
23. Zhang, W. et al. Fine-tuning large language models for chemical text mining. *Chem. Sci.* **15**, 10600–10611 (2024).
24. Leong, S. X. Automated electrosynthesis reaction mining with multimodal large language models (MLLMs). *Chem. Sci.* https://doi.org/10.26434/chemrxiv-2024-7fwxv (2024).
25. Zheng, Z. et al. Image and data mining in reticular chemistry powered by GPT-4V. *Digit. Discov.* **3**, 491–501 (2024).
26. Chen, K. et al. Chemist-X: large language model-empowered agent for reaction condition recommendation in chemical synthesis. *arXiv* http://arxiv.org/abs/2311.10776 (2024).
27. M. Bran, A. et al. Augmenting large language models with chemistry tools. *Nat. Mach. Intell.* **6**, 525–535 (2024).
28. Zheng, Z. et al. Integrating machine learning and large language models to advance exploration of electrochemical reactions. *ChemRxiv* https://doi.org/10.26434/chemrxiv-2024-pk105-v2 (2024).
29. Song, T. et al. A multi-agent-driven robotic AI chemist enabling autonomous chemical research on demand. *ChemRxiv* https://doi.org/10.26434/chemrxiv-2024-w953h-v2 (2024).
30. Zheng, Z. et al. Shaping the water-harvesting behavior of metal–organic frameworks aided by fine-tuned GPT models. *J. Am. Chem. Soc.* **145**, 28284–28295 (2023).
31. Wang, H. et al. Efficient evolutionary search over chemical sace with large language models. *arXiv* http://arxiv.org/abs/2406.16976 (2024).
32. Parrilla-Gutiérrez, J. M. et al. Electron density-based GPT for optimization and suggestion of host–guest binders. *Nat. Comput. Sci.* **4**, 200–209 (2024).
33. Li, J. et al. Empowering molecule discovery for molecule-caption translation with large language models: a ChatGPT perspective. *IEEE Trans. Knowl. Data Eng.* **36**, 6071–6083 (2024).
34. Kang, Y. & Kim, J. ChatMOF: an artificial intelligence system for predicting and generating metal-organic frameworks using large language models. *Nat. Commun.* **15**, 4705 (2024).
35. Janakarajan, N., Erdmann, T., Swaminathan, S., Laino, T. & Born, J. Language models in molecular discovery. *arXiv* http://arxiv.org/abs/2309.16235 (2023).
36. McNaughton, A. D. et al. CACTUS: Chemistry agent connecting tool-usage to science. *arXiv* https://doi.org/10.48550/arXiv.2405.00972 (2024).
37. Sprueill, H. W. et al. ChemReasoner: Heuristic search over a large language model's knowledge space using quantum-chemical feedback. *arXiv* https://doi.org/10.48550/arXiv.2402.10980 (2024).
38. Livne, M. et al. nach0: multimodal natural and chemical languages foundation model. *Chem. Sci.* **15**, 8380–8389 (2024).
39. Zheng, Z. et al. A GPT-4 reticular chemist for guiding MOF discovery**. *Angew. Chem.* **135**, e202311983 (2023).
40. Zheng, Z. et al. ChatGPT research group for optimizing the crystallinity of MOFs and COFs. *ACS Cent. Sci.* **9**, 2161–2170 (2023).
41. Boiko, D. A., MacKnight, R., Kline, B. & Gomes, G. Autonomous chemical research with large language models. *Nature* **624**, 570–578 (2023).
42. Yoshikawa, N. et al. Large language models for chemistry robotics. *Auton. Robots* **47**, 1057–1086 (2023).
43. Darvish, K. et al. ORGANA: A robotic assistant for automated chemistry experimentation and characterization. *arXiv* http://arxiv.org/abs/2401.06949 (2024).
44. Wu, W. & Jiang, H. Palladium-catalyzed oxidation of unsaturated hydrocarbons using molecular oxygen. *Acc. Chem. Res.* **45**, 1736–1748 (2012).

45. OpenAI. GPT-4 technical report. *arXiv* https://doi.org/10.48550/arXiv.2303.08774 (2023).
46. Brown, T. B. et al. Language models are few-shot learners. *arXiv* https://doi.org/10.48550/arXiv.2005.14165 (2020).
47. Lewis, P. et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Adv. Neural Inform. Process. Syst.* **33**, 9459–9474 (2020).
48. Hoover, J. M. & Stahl, S. S. Highly practical copper(I)/TEMPO catalyst system for chemoselective aerobic oxidation of primary alcohols. *J. Am. Chem. Soc.* **133**, 16901–16910 (2011).
49. Huang, Z., Li, F., Chen, B. & Yuan, G. Sustainable catalytic oxidation of alcohols over the interface between air and water. *Green. Chem.* **17**, 2325–2329 (2015).
50. Kakiuchi, N., Maeda, Y., Nishimura, T. & Uemura, S. Pd(II)-hydrotalcite-catalyzed oxidation of alcohols to aldehydes and ketones using atmospheric pressure of air. *J. Org. Chem.* **66**, 6620–6625 (2001).
51. Nikitas, N. F., Tzaras, D. I., Triandafillidi, I. & Kokotos, C. G. Photochemical oxidation of benzylic primary and secondary alcohols utilizing air as the oxidant. *Green. Chem.* **22**, 471–477 (2020).
52. Ahneman, D. T., Estrada, J. G., Lin, S., Dreher, S. D. & Doyle, A. G. Predicting reaction performance in C–N cross-coupling using machine learning. *Science* **360**, 186–190 (2018).
53. Tu, Z., Stuyver, T. & Coley, C. W. Predictive chemistry: machine learning for reaction deployment, reaction development, and reaction discovery. *Chem. Sci.* **14**, 226–244 (2023).
54. Buitrago Santanilla, A. et al. Nanomole-scale high-throughput chemistry for the synthesis of complex molecules. *Science* **347**, 49–53 (2015).
55. Perera, D. et al. A platform for automated nanomole-scale reaction screening and micromole-scale synthesis in flow. *Science* **359**, 429–434 (2018).
56. Yin, G. Understanding the oxidative relationships of the metal oxo, hydroxo, and hydroperoxide intermediates with manganese(IV) complexes having bridged cyclams: correlation of the physicochemical properties with reactivity. *Acc. Chem. Res.* **46**, 483–492 (2013).
57. Christensen, M. et al. Development of an automated kinetic profiling system with online HPLC for reaction optimization. *React. Chem. Eng.* **4**, 1555–1558 (2019).
58. Hoover, J. M., Ryland, B. L. & Stahl, S. S. Mechanism of copper(I)/TEMPO-catalyzed aerobic alcohol oxidation. *J. Am. Chem. Soc.* **135**, 2357–2367 (2013).
59. Baumgartner, L. M., Coley, C. W., Reizman, B. J., Gao, K. W. & Jensen, K. F. Optimum catalyst selection over continuous and discrete process variables with a single droplet microfluidic reaction platform. *React. Chem. Eng.* **3**, 301–311 (2018).
60. Yang, C. et al. Large language models as optimizers. *arXiv* http://arxiv.org/abs/2309.03409 (2023).
61. Ruan, Y., Lin, S. & Mo, Y. AROPS: A framework of automated reaction optimization with parallelized scheduling. *J. Chem. Inf. Model.* **63**, 770–781 (2023).
62. Lovato, K., Fier, P. S. & Maloney, K. M. The application of modern reactions in large-scale synthesis. *Nat. Rev. Chem.* **5**, 546–563 (2021).
63. Xu, H. et al. High-throughput discovery of chemical structure-polarity relationships combining automation and machine-learning techniques. *Chem* **8**, 3202–3214 (2022).
64. Boström, J., Brown, D. G., Young, R. J. & Keserü, G. M. Expanding the medicinal chemistry synthetic toolbox. *Nat. Rev. Drug Discov.* **17**, 709–727 (2018).
65. Ashworth, I. W., Frodsham, L., Moore, P. & Ronson, T. O. Evidence of rate limiting proton transfer in an S_N Ar aminolysis in acetonitrile under synthetically relevant conditions. *J. Org. Chem.* **87**, 2111–2119 (2022).
66. Speckmeier, E. & Maier, T. C. ART—An amino radical transfer strategy for C(sp^2)–C(sp^3) coupling reactions, enabled by dual photo/nickel catalysis. *J. Am. Chem. Soc.* **144**, 9997–10005 (2022).
67. Okada, Y. Synthetic semiconductor photoelectrochemistry. *Chem. Rec.* **21**, 2223–2238 (2021).
68. Chen, Y. et al. Scalable decarboxylative trifluoromethylation by ion-shielding heterogeneous photoelectrocatalysis. *Science* **384**, 670–676 (2024).
69. Logg, A., Mardal, K.-A. & Wells, G. *Automated Solution of Differential Equations by the Finite Element Method: The FEniCS Book* 2012th edn, Vol 744 (Springer Science & Business Media, 2012).
70. Tang, X. et al. Prioritizing safeguarding over autonomy: risks of LLM agents for science. *arXiv* http://arxiv.org/abs/2402.04247 (2024).
71. Kirchner, J. H. et al. Prover-verifier games improve legibility of LLM outputs. *arXiv* http://arxiv.org/abs/2407.13692 (2024).
72. Zhang, C. et al. SynAsk: Unleashing the power of large language models in organic synthesis. *arXiv* http://arxiv.org/abs/2406.04593 (2024).
73. Zhao, Z. et al. ChemDFM: Dialogue foundation model for chemistry. arXiv http://arxiv.org/abs/2401.14818 (2024).
74. Zhang, D. et al. ChemLLM: A chemical large language model. *arXiv* https://doi.org/10.48550/arXiv.2402.06852 (2024).
75. Chen, L. et al. PharmaGPT: Domain-specific large language models for bio-pharmaceutical and chemistry. *arXiv* https://doi.org/10.48550/arXiv.2406.18045 (2024).
76. Chiang, Y., Hsieh, E., Chou, C.-H. & Riebesell, J. LLaMP: Large language model made powerful for high-fidelity materials knowledge retrieval and distillation. *arXiv* https://doi.org/10.48550/arXiv.2401.17244 (2024).
77. Jablonka, K. M., Schwaller, P., Ortega-Guerrero, A. & Smit, B. Leveraging large language models for predictive chemistry. *Nat. Mach. Intell.* **6**, 161–169 (2024).
78. Team, Q. *Introducing Qwen2-Math. Qwen* http://qwenlm.github.io/blog/qwen2-math/ (2024).
79. Wu, Q. et al. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. *arXiv* http://arxiv.org/abs/2308.08155 (2023).
80. ChatGPT - Consensus. *ChatGPT* https://chat.openai.com/g/g-bo0FiWLY7-consensus (2024).
81. Ruan-Yixiang. An automatic end-to-end chemical synthesis development platform powered by large language models. *Zenodo* https://doi.org/10.5281/zenodo.13440868 (2024).

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.