A multi-agent-driven robotic AI chemist enabling autonomous chemical research on demand

Tao Song[1,2,†], Man Luo[1,†], Linjiang Chen[1,3,†,*], Yan Huang[1], Qing Zhu[1,4], Daobin Liu[1], Baicheng Zhang[1], Gang Zou[1], Fei Zhang[2,*], Weiwei Shang[2,*], Jun Jiang[1,5*], & Yi Luo[1,5*]

[1] Key Laboratory of Precision and Intelligent Chemistry, Hefei National Research Center for Physical Sciences at the Microscale, School of Chemistry and Materials Science, University of Science and Technology of China, Hefei, China
[2] School of Information Science and Technology, University of Science and Technology of China, Hefei, China
[3] School of Chemistry and School of Computer Science, University of Birmingham, Birmingham, UK
[4] Institute of Intelligent Innovation, Henan Academy of Sciences, Zhengzhou, China
[5] Hefei National Laboratory, University of Science and Technology of China, Hefei, China

† These authors contributed equally: T.S., M.L., L.C.
E-mails: linjiangchen@ustc.edu.cn (L.C.); zfei@ustc.edu.cn (F.Z.); wwshang@ustc.edu.cn (W.S.); jiangj1@ustc.edu.cn (J.J.); yiluo@ustc.edu.cn (Y.L.)

Abstract

The successful integration of large language models (LLMs) into laboratory workflows has demonstrated robust capabilities in natural language processing, autonomous task execution, and collaborative problem-solving.[1-4] This offers an exciting opportunity to realize the dream of autonomous chemical research on demand. Here, we report a robotic AI chemist powered by a hierarchical multi-agent system, ChemAgents, based on an on-board Llama-3-70B LLM, capable of executing complex, multi-step experiments with minimal human intervention. It operates through a Task Manager agent that interacts with human researchers and coordinates four role-specific agents—Literature Reader, Experiment Designer, Computation Performer, and Robot Operator—each leveraging one of four foundational resources: a comprehensive Literature Database, an extensive Protocol Library, a versatile Model Library, and a state-of-the-art Automated Lab. We demonstrate its versatility and efficacy through six experimental tasks of varying complexity, ranging from straightforward synthesis and characterization to more complex exploration and screening of experimental parameters, culminating in the discovery and optimization of functional materials. Our multi-agent-driven robotic AI chemist showcases the potential of on-demand autonomous chemical research to drive unprecedented efficiencies, accelerate discovery, and democratize access to advanced experimental capabilities across academic disciplines and industries.

Main

The chemistry laboratories of the future are poised to undergo a significant transformation, largely driven by the advent of large language models (LLMs) and their integration into everyday tasks and workflows. These models, with billions of parameters and pre-trained on extensive hybrid corpora, exhibit robust cognitive abilities in natural language processing. They demonstrate remarkable capabilities in semantic understanding, question answering, content generation, and code development across various tasks involving open human-machine interaction[5-10]. These capabilities are catalysing unprecedented opportunities for the intelligent transformation of chemical and materials research, giving rise to the development and deployment of a wide variety of LLM-powered research methods and tools[3,4,11-17]. One of the most exciting aspects where LLMs will have significant impact on chemistry and materials science is accelerating research and discovery by human-LLM collaboration. For example, GPT-4[10] has been integrated into an iterative process of discovering new metal-organic frameworks (MOFs)[4], operating through a cooperative workflow between GPT-4 and a human chemist. Through structured prompting of GPT-4 and in-text learning informed by human feedback, the human-LLM collaboration yielded the discovery of an isoreticular series of MOFs, each synthesized using distinct strategies and optimal conditions.

LLM-powered "agents" offer robust capabilities in natural language processing, autonomous task execution, and collaborative problem-solving[18-24]. Their integration with external tools and platforms enables them to perform a wide range of tasks across various domains, including chemistry, where they are revolutionizing research and discovery processes[2,25-28]. One such example is ChemCrow[2], a GPT-4-powered agent designed to streamline the reasoning process for various common chemical tasks, including organic synthesis, drug discovery, and materials design. It operates by sequentially prompting GPT-4 with instructions, guiding it to reason about the current state of the task, consider its relevance to the final goal, and plan the next steps accordingly. ChemCrow has proven to be an effective assistant to expert chemists, while also lowering the entry barrier for non-experts by offering a simple interface to access accurate chemical knowledge. More generally, LLM-powered agents are autonomous systems that leverage the capabilities of LLMs to interpret, process, and respond to textual inputs in a manner that mimics human-like understanding and reasoning. Their integration with external tools and platforms enables them to offer robust capabilities in autonomous task execution and collaborative problem-solving, thereby enhancing productivity, accelerating innovation, and democratizing access to advanced research tools.

Parallel to the rapid adoption of LLMs in chemical and materials research, robotics has also made significant strides in laboratory automation, providing a diverse array of systems tailored to meet the complex demands of autonomous chemistry research. These include, but are not limited to, automated platforms that handle high-throughput experimentation[29-37], benchtop fixed robotic arms that perform precise manipulations[38-

[42], and mobile robots that navigate laboratory environments to execute and coordinate complex tasks[43-45]. More sophisticated setups feature dual-arm robots capable of complex, coordinated tasks that mimic a human chemist's dexterity[46], as well as multi-robot systems that collaborate seamlessly to optimize workflow efficiency and output[47]. Naturally, the integration of automated robotic hardware with LLM techniques[48-54] has emerged as a new research direction in autonomous chemistry research[1,55,56]. For instance, Boiko et al. built Coscientist[1], a GPT-4-driven AI system capable of semi-autonomously designing, planning, and executing experiments through internet and document searches, code execution, and experimental automation.

Despite these recent advancements in autonomous chemical research driven by LLMs, when confronted with complex experimental scenarios, such as multi-step, multi-station, and/or multi-robot setups, the challenge arises: How can one effectively employ LLMs to construct an automated system capable of proficiently executing complex chemical tasks on demand? It is impractical to task LLMs with overly complex assignments in a single request. Likewise, in autonomous chemistry, it is impractical to prompt an LLM once or within a single conversation to perform a series of experimental tasks, which can involve interpreting the given problem, generating testing hypotheses, planning experimental procedures, implementing instructions for robots, adjusting hypotheses, making decisions for subsequent experiments, and much more.

To address this challenge, we have developed an approach featuring ChemAgents, a hierarchical LLM-based multi-agent system designed to meet diverse experimental demands. By enabling collaboration and coordination among agents, each with a unique role and associated abilities (Figure 1), ChemAgents amplifies the effectiveness of single LLM agents, allowing them to collectively tackle complex tasks and achieve defined objectives[57-61]. Using Llama-3-70B, these agents—Literature Reader, Experiment Designer, Computation Performer, and Robot Operator—receive textual instructions, perform specific operations (e.g., using external tools, writing code, etc.), and produce suitably formatted outputs that can be used directly as input instructions by other agents or interpreted by human researchers. A Task Manager agent coordinates the role-specific agents, integrating external software tools and robotic hardware to autonomously plan, manage, and execute complex chemical research tasks.

Underpinning ChemAgents are four foundational resources of the robotic AI chemist: a database of scientific literature, a library of experimental protocols, a library of machine learning models (many of which are pre-trained), and an autonomous chemistry lab with robots and automated experimental stations. Here, we demonstrate the versatility and performance of our multi-agent-driven robotic AI chemist in handling three categories of experimental tasks with increasing levels of complexity: (1) performing the synthesis of required compounds and/or performing required characterizations of given compounds (referred to as “make & measure”); (2) exploring and/or screening experimental parameters of desired functional materials (referred to as “exploration & screening”); and (3) combining literature mining, computational

modelling, and closed-loop optimization for the discovery of functional materials (referred to as "discovery & optimization"), as depicted in Figure 1.

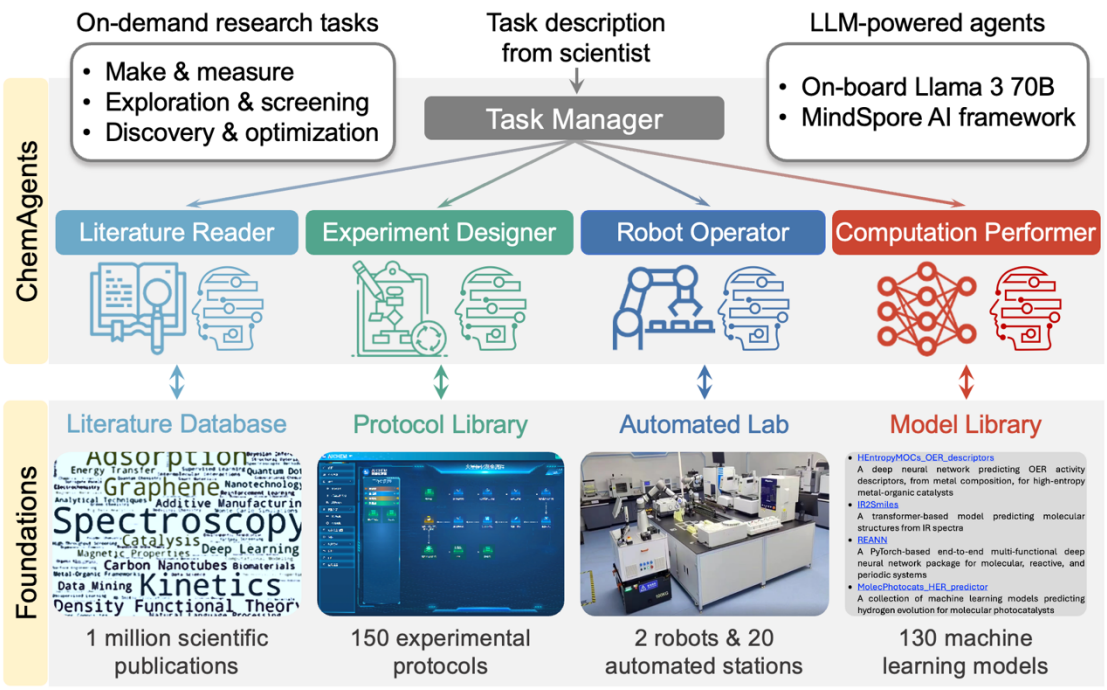

Figure 1: The multi-agent-driven robotic AI chemist's architecture. The LLM-based multi-agent system, ChemAgents, comprises a Task Manager agent, which manages four role-specific agents: Literature Reader, Experiment Designer, Computation Performer, and Robot Operator. Each role-specific agent uses built-in tools and operates on one of four foundational resources: Literature Database, Protocol Library, Model Library, and Automated Lab.

The ChemAgents architecture

We use the open-source Llama-3-70B as the LLM basis, implemented in Huawei's open-source AI framework MindSpore[62]. The ChemAgents architecture consists of a central coordinating entity, the Task Manager agent, that interacts with human researchers and manages four role-specific agents—Literature Reader, Experiment Designer, Computation Performer, and Robot Operator. The Task Manager has its own preset system prompt, within which we define its role, tasks, specific functions, and interface descriptions for all entities it can use (including the lower-ranking agents and their tools), as well as message formats for workflows and communications. By leveraging the capabilities of LLM and LangGraph, an extended library of LangChain, the Task Manager interacts with human researchers in natural language. It interprets the demand description, plans and decomposes tasks, and instructs and manages the role-specific agents to use the corresponding foundational resources to perform specified operations, facilitating an automated experimental workflow for the demand description.

Each role-specific agent has its own LLM instance(s) and built-in tools, operating on one of the four foundational resources of the system (Figure 1). The Literature Reader uses an external natural language processing (NLP) tool, detailed elsewhere by us[63], to

mine the Literature Database and accumulate knowledge pertinent to given experimental objectives. The Experiment Designer receives instructions, from the Task Manager, on intended experiments, plans experimental tasks, and generates step-by-step procedures by leveraging the Protocol Library. The Robot Operator translates experimental procedures from the Experiment Designer into codes and commands for operating the Automated Lab's robots and experimental stations, facilitating and completing the robotic execution of chemical experiments. The Computation Performer searches for suitable pre-trained machine learning models in the Model Library, using them as provided or enhancing them by expanding the model and incorporating additional training data. This agent also includes a Bayesian optimizer, which itself is an LLM-based agent managed by the Computation Performer and can be deployed for iterative optimization tasks.

Generating experimental designs from scientific literature

We constructed a local Literature Database containing titles and abstracts of approximately 1.2 million scientific publications in chemistry and materials science subjects, downloaded from Crossref. These publications cover a wide variety of research subjects, topics, and fields, as depicted by the word cloud in Figure 2a, with larger font sizes indicating more frequent occurrences of words in the database. To mine this Literature Database and analyse the resulting data to distil knowledge, we built the Literature Reader agent.

The Literature Reader uses two built-in tools: (1) LiteratureSearch, which performs searches in the Literature Database according to input keywords and returns relevant literature data; and (2) LiteratureMine, which employs an unsupervised syntactic distance analysis approach for mining literature data, capable of extracting information about chemical substances, physicochemical properties, chemical/material functions, and experimental conditions[63]. The Literature Reader functions in the following steps: It receives keywords from the Task Manager and then uses the LiteratureSearch tool to collect relevant publication titles and abstracts. These are subsequently passed to the LiteratureMine tool, which processes the data and generates statistics. These statistics can serve as prior knowledge for designing chemical experiments.

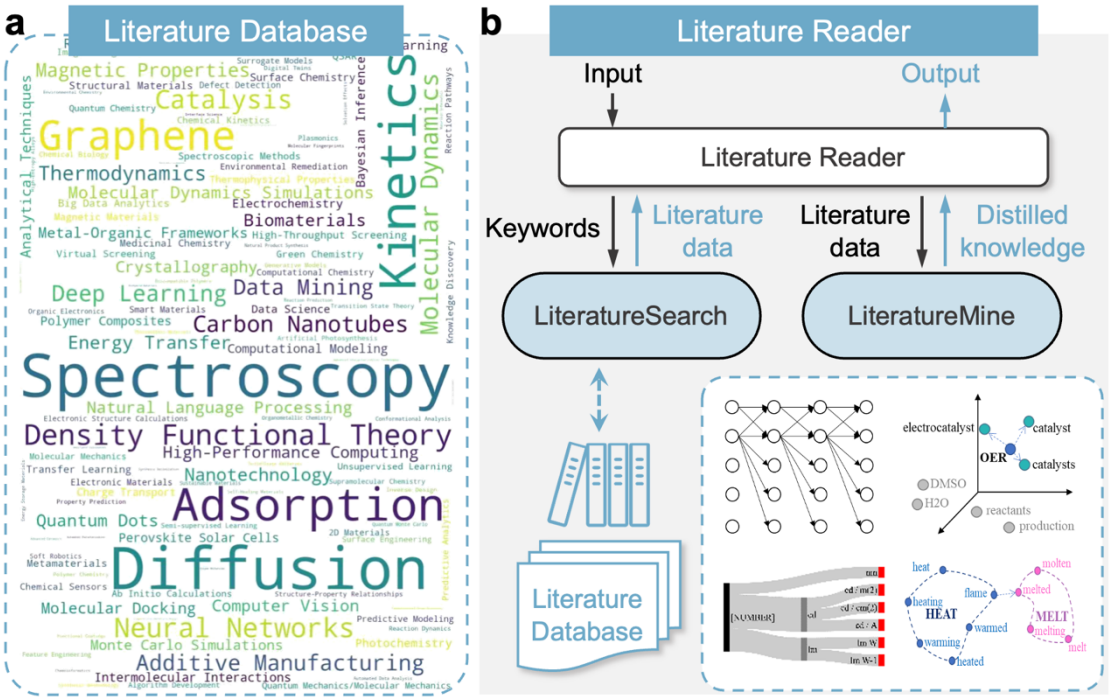

Figure 2: The robotic AI chemist's literature mining and analysing capabilities. a, Visualization of knowledge domains contained in the Literature Database, with larger text sizes indicating more frequent occurrences. **b,** Architecture of the Literature Reader agent: step 1 involves searching the scientific literature database, and step 2 entails the use of the literature-mining tool.

Generating experimental procedures from protocols

A key foundational resource for our robotic AI chemist is a Protocol Library, which has been accumulated over the past few years from experiments conducted in the Automated Lab. This library includes procedural templates for previous experiments as well as protocols derived from various configurations of the 20 experimental stations in the lab, even if not all configurations have yet been used. These protocols are stored in the XML format and describes experimental steps in natural language. The experimental stations encompass liquid dispensing, solid dispensing, magnetic stirring, drying, infrared spectroscopy, X-ray diffraction, photocatalysis, electrocatalysis, and more (see Supplementary Information for full details). Together, these protocols and stations define the experimental capabilities of the Automated Lab in chemical and material synthesis, characterization, and performance testing.

Figure 3a illustrates the architecture of the Experiment Designer agent. It operates as a multi-agent system itself, integrating two LLM-based agents: a Protocol writer and a Protocol critic. The Protocol writer functions through a pre-prompted LLM, using two built-in tools: (1) ProtocolSearch, which searches the Protocol Library based on keywords and returns the best-matching experimental template, and (2) StationQuery, which allows the Protocol writer to query detailed information about the available stations in the Automated Lab, such as specific functions and associated parameters. The Protocol critic is another LLM instance with its own preset prompts. These prompts

outline its specific tasks to check and correct experimental procedures against predefined expert rules, based on the operational constraints of various stations (see Supplementary Information for these expert rules).

The Experiment Designer operates as follows: First, the Protocol writer receives a description of the intended experiments from the Task Manager. It then invokes the ProtocolSearch tool to look for the best-matching template within the Protocol Library. If a match is found, the Protocol writer references the relevant template, adjusting the steps and parameters to formulate an experimental procedure tailored to the specific requirements. If no existing template applies, the Protocol writer invokes the StationQuery tool to get the list of available automated stations, plans the experimental workflow, and generates an experimental procedure relying solely on the capabilities of the LLM. Next, the Protocol critic engages in a reflection process to check the experimental procedure generated by the Protocol writer. It reviews, critiques, and improves the procedure in accordance with predefined expert rules. Finally, the Protocol critic formats the output to ensure the experimental procedure align with both the intended experiments and the lab's capabilities. Figure 3b shows part of a generated experimental procedure, which can be fed into the Robot Operator agent and compiled into robot-executable codes for conducting automated chemical experiments.

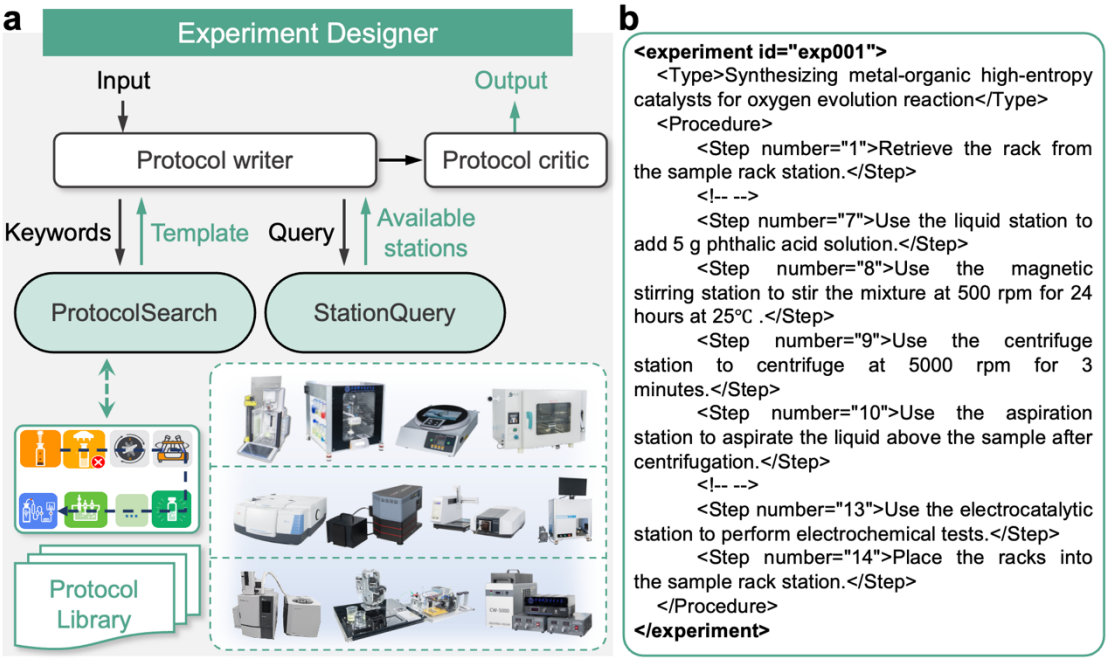

Figure 3: The robotic AI chemist's experiment design capabilities. a, Architecture of the Experiment Designer agent. **b,** Example of a generated experimental procedure.

From experimental procedures to automated execution of experiments

Our Automated Lab encompasses two robots and 20 automated stations with agent-machine interaction interfaces and a backend control system for all the hardware. Compared to our previous robotic AI chemist system[44,45], we now have two robots (Extended Data Figure 1): a fully mobile robot and a benchtop robotic arm. The mobile

robot moves freely within the laboratory, primarily handling tasks involving multiple instruments. The benchtop robot, placed on a platform at the centre of the lab, moves laterally on tracks and handles characterization and performance testing instruments on the platform, including Raman spectroscopy, X-ray diffraction, infrared spectroscopy, UV-vis spectroscopy, fluorescence spectroscopy, and electrochemical testing. The backend control system schedules tasks, manages data, and communicates with robots and stations via the HTTP protocol. To enhance scalability and flexibility, we have abstracted and encapsulated the operational methods of the robots and stations, establishing standard operation commands and a common interface specification. These commands encompass eight types of high-level Robot API interfaces in Python format (see Supplementary Information for details), which cover operations including visual positioning, motion planning, experimental operations, and instrument communication. These API interfaces enable the utilization of the hardware by the Robot Operator agent's codes, ensuring accurate and efficient execution of experimental operations.

Figure 4a illustrates the architecture of the Robot Operator agent. It is a multi-agent system, integrating three LLM-based agents—Code writer, Code critic, and Code proofreader—with the Automated Lab. The Code writer is a pre-prompted LLM instance equipped with a built-in tool: RobotAPIQuery, which is used to obtain the available high-level API functions for robots. The Code critic is another pre-prompted LLM instance, responsible for checking, correcting, critiquing, and improving the codes against predefined expert rules on robot codes (detailed in the Supplementary Information). The Code proofreader also utilizes the RobotAPIQuery tool and is tasked with performing final code checks for grammatical, spelling, and formatting errors according to the robot API interface specifications and Python code standards.

The Robot Operator functions as follows: First, the Code writer receives the experimental procedure generated by the Experiment Designer. It then invokes the RobotAPIQuery tool to obtain the list of available APIs for robots and writes a Python code leveraging the capabilities of the LLM. Next, the Code critic reviews and improves the code in accordance with predefined expert rules. The improved code is then passed to the Code proofreader, who again invokes the RobotAPIQuery tool to retrieve the list of available APIs for robots and proofreads the code for further enhancement. Finally, the Automated Lab receives the final code for the execution of automated chemical experiments. Figure 4b shows an example of a partial robot code for various operations at the magnetic stirring station, involving robot movement, visual positioning, and rack placement and retrieval (code block 3), as well as for sequential liquid dispensing operations (code block 6).

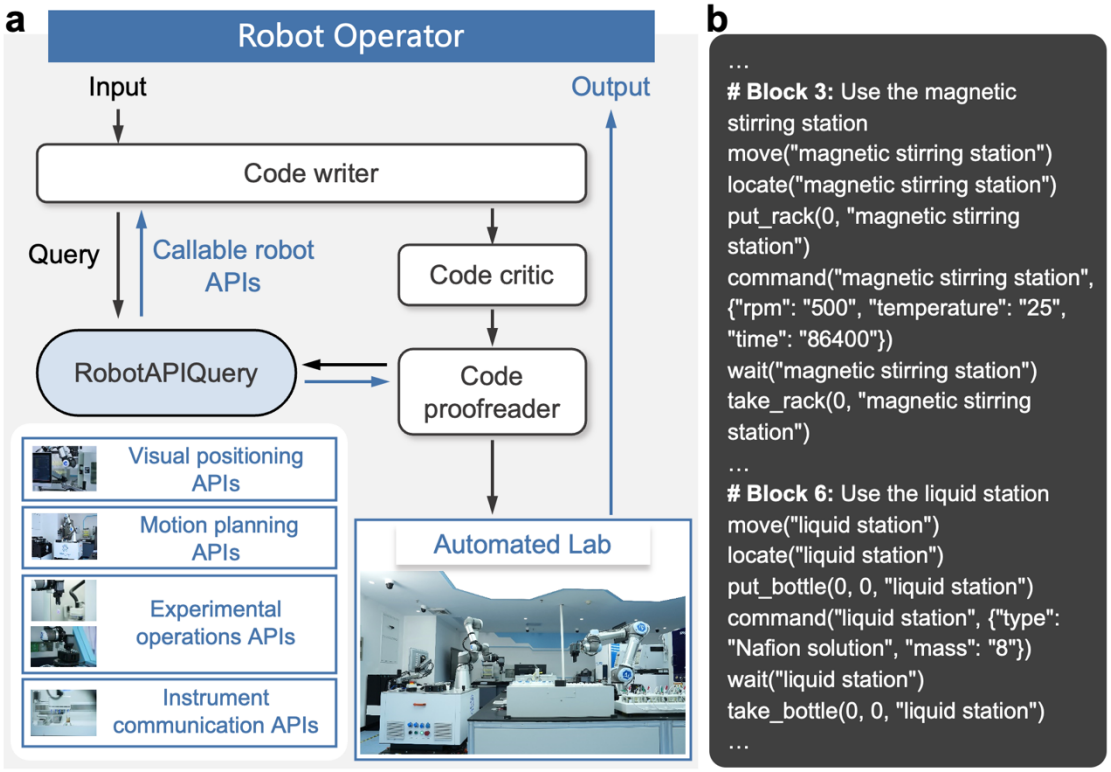

Figure 4: The robotic AI chemist's experiment execution capabilities. a, Architecture of the Robot Operator Agent. **b,** Example of code blocks for using the magnetic stirring and liquid dispensing stations.

Enabling discovery and optimization by leveraging computational capabilities

Most on-demand experimental tasks requested of our robotic AI chemist by human researchers follow a sequential make-measure-test fashion (e.g., measuring the infrared spectrum of a molecule) or involve exploring a prescribed design space (e.g., performing a full factorial experiment of certain variables). For such tasks, it is sufficient for the Task Manager to only call upon the Experiment Designer and the Robot Operator. However, research tasks that necessitate data-driven discovery or closed-loop iterative optimization (e.g., discovering high-performance catalysts in an expansive search space) require additional computational capabilities.

To address this need, we have compiled a library of machine learning models, most of which are pre-trained models from our previous publications or specifically developed by us for general use by the robotic AI chemist. Each pre-trained model provides auxiliary functions for processing input data, ensuring consistency in data preprocessing (such as normalization and handling missing values), and can be used for direct prediction. These models come with annotations detailing their technical specifications and origins, making them accessible for queries by human researchers or LLM-based agents. Additionally, the library includes open-source models that can be trained on-the-fly with experimental data collected by the robotic AI chemist, upon requests by specific research tasks.

Figure 6 illustrates the architecture of the Computation Performer agent. It integrates a pre-prompted LLM with two built-in tools, a Bayesian optimizer, and a deep learning computing platform. The two built-in tools are: (1) ModelSearch, which performs searches in the Model Library using keywords and returns a matching model, and (2) ModelFuse, which expands a pre-trained neural-network model by appending additional hidden layers to create a fused model. The Bayesian optimizer is an LLM-based agent, pre-prompted to write Bayesian optimization code using the task description provided to the Task Manager by the human researcher; the completion of its tasks relies solely on the capabilities of the LLM. The deep learning computing platform provides the Computation Performer with a complete Python environment necessary for running machine learning codes and obtaining results.

The Computation Performer operates as follows: First, the Computational task manager receives input information, including experimental task keywords, a detailed task description, and (if applicable) robotic experimental results. Then, it invokes the ModelSearch tool to obtain a suitable pre-trained model. Based on the specific task, it determines the subsequent actions to be taken and executes them sequentially to generate machine learning codes. This process may involve the use of the ModelFuse tool and the Bayesian optimizer, depending on the task specification. Finally, prediction results are obtained by executing the generated code on the deep learning computing platform.

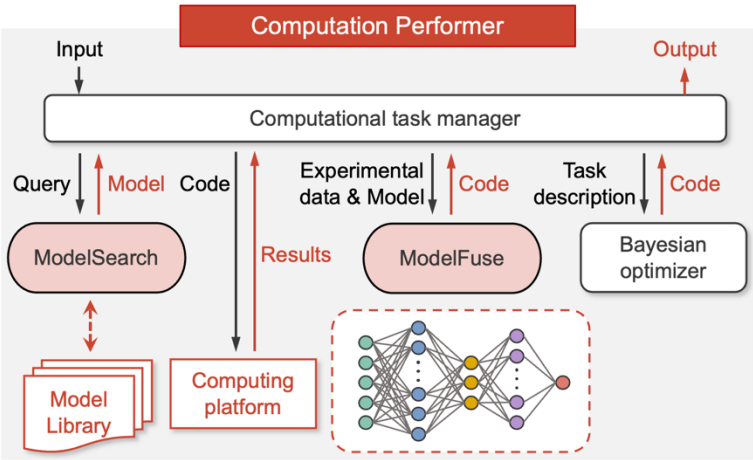

Figure 5: The robotic AI chemist's computational capabilities. Architecture of the Computation Performer agent.

On-demand autonomous chemical research

Our multi-agent-driven robotic AI chemist can perform a variety of experimental tasks on demand, with the primary limitation being the availability and capabilities of automated experimental stations in the laboratory. It generates suitable workflows upon understanding task descriptions provided by human researchers, judiciously selecting and combining only the necessary agents and stations according to the needs of specific workflows. To demonstrate this, we instructed the robotic AI chemist to carry out six

experimental tasks with varied levels of sophistication and complexity (Figure 6a–f), all employing the Experiment Designer and Robot Operator agents, with the last and most complex task also employing the Literature Reader and Computation Performer agents. The results are detailed below.

"Make & measure" tasks. These are the most straightforward type of experimental tasks for the robotic AI chemist. Upon receiving a task description, the Task Manager interprets and understands the instructions before passing on the necessary information to the Experiment Designer, which then identifies and uses an appropriate experimental protocol to generate a suitable experimental procedure. This procedure is translated into the required code-based instructions by the Robot Operator to execute robotic synthesis and testing in the Automated Lab, obtaining experimental results that are returned to the Task Manager. Figure 6a–c reports experimental results of three "make & measure" tasks.

Task 1 involved characterizing three azobenzene molecules using Fourier-transform infrared (FT-IR) spectroscopy. Azobenzenes are a class of organic molecules characterized by the presence of an N=N double bond linking two phenyl rings. They are known for their photoswitchable properties, which allow them to change their molecular structures and chemical properties upon irradiation. The Task Manager was instructed to measure the IR spectra of three azobenzene molecules, using the following prompt:

"*You have been assigned a scientific research task to measure the infrared spectra of azobenzene molecules. You need to conduct one measurement for each of the following solid powders: azobenzene, 1-(phenylazo)naphthalen-2-amine, and 4,4'-dihydroxymethylazobenzene.*"

Upon receiving this task description, the robotic AI chemist successfully planned, coordinated, and collected high-quality FT-IR spectra for the molecules. The results showed the expected characteristic peaks of stretching vibrations of N=N in the range of 1600–1500 cm^{-1}, as well as the stretching vibrations of C=C and C–H of the benzene ring in the range of 1500–1400 cm^{-1} and 3100–3000 cm^{-1}, respectively (Figure 6a).

Task 2 involved synthesizing researcher-specified metal oxides and then characterizing them using powder X-ray diffraction (PXRD). The robotic AI chemist synthesized six requested metal oxides, namely ZrO_2, ZnO, WO_3, Mn_3O_4, CuO, and Fe_2O_3, before collecting their PXRD patterns (Figure 6b). The successful syntheses of the metal oxides were confirmed by comparing experimentally collected PXRD patterns with their corresponding PDF cards (Supplementary Fig. S3). Task 3 involved using lead halide perovskite ($APbX_3$) quantum dots (PQDs) to achieve a wide gamut of high-purity colours through their tuneable emission wavelengths and sharp photoluminescence profiles. This was done by combining different types of $A^+Pb^{2+}(X^-)_3$ (A = CH_3NH_3, $(NH_2)_2CH$, or Cs; X = Cl, Br, or I). Per the task description, the robotic

AI chemist successfully prepared PQD inks of four colours (i.e., blue, green, yellow, and red) following predefined recipes of $APbX_3$ mixtures. These inks were then cast into thin films, and the fluorescence emission spectra of the films were measured, displaying the intended high-purity colours (Figure 6c).

"Exploration & screening" tasks. These tasks necessitate more advanced LLM capabilities compared to the "make & measure" ones. For instance, the Task Manager must have the ability to interpret task descriptions that require a deeper understanding of domain-specific knowledge, such as comprehending and executing instructions to "perform a full factorial experiment". This involves recognizing the concept of a factorial experiment and understanding the importance of systematically varying multiple factors to evaluate their effects. Similarly, the Experiment Designer must demonstrate flexibility and creativity in adapting existing protocols. This can involve assigning different values to variables based on experimental requirements and customizing protocols by expanding or combining them to accommodate more complex experimental designs. For example, it might need to integrate multiple experimental steps from different protocols or adjust procedural details to align with specific research goals. Tasks 4 and 5 fall within this category of on-demand research tasks.

Task 4 involved the robotic AI chemist performing a full factorial experiment to synthesize graphitic carbon nitrides (g-C_3N_4) and test them for hydrogen evolution reaction (HER) performance. For demonstration, the robotic AI chemist was instructed to consider three heating temperatures (500, 550, and 600 °C) and three heating durations (3, 4, and 5 hours). Measuring the HER performance of the resulting g-C_3N_4 materials served to illustrate the different sample qualities produced under the various synthesis conditions, highlighting the importance of screening these synthesis conditions to optimize the HER performance of g-C_3N_4. As shown in Figure 6d, the results indicate degradation or over-sintering of g-C_3N_4 at higher temperatures across all heating durations, particularly evident at 600 °C. Additionally, it is evident that increasing the heating time negatively affected the HER performance at all three temperatures.

Task 5 explored the photocatalytic degradation of organic pollutants in water using bismuth oxyhalides (BiOX, where X = Cl, Br, or I). These compounds have abundant availability, low toxicity, and suitable band gaps for visible light absorption, making them promising photocatalysts for various applications. Specifically, the robotic AI chemist was instructed to investigate the effect of halogen atom species in bismuth halide on the photocatalytic degradation of tetracycline (TC). Figure 6e shows that BiOBr exhibited the fastest photocatalytic degradation of TC, followed by BiOCl and then BiOI. The superior performance of BiOBr in degrading TC can be attributed to its suitable band gap (~2.7 to 3.1 eV), which is narrower than that of BiOCl (~3.5 eV), enabling it to absorb light more efficiently. Despite having the narrowest band gap (~1.7 to 1.9 eV), BiOI exhibited the lowest photocatalytic activity, likely due to a faster recombination of electron-hole pairs.

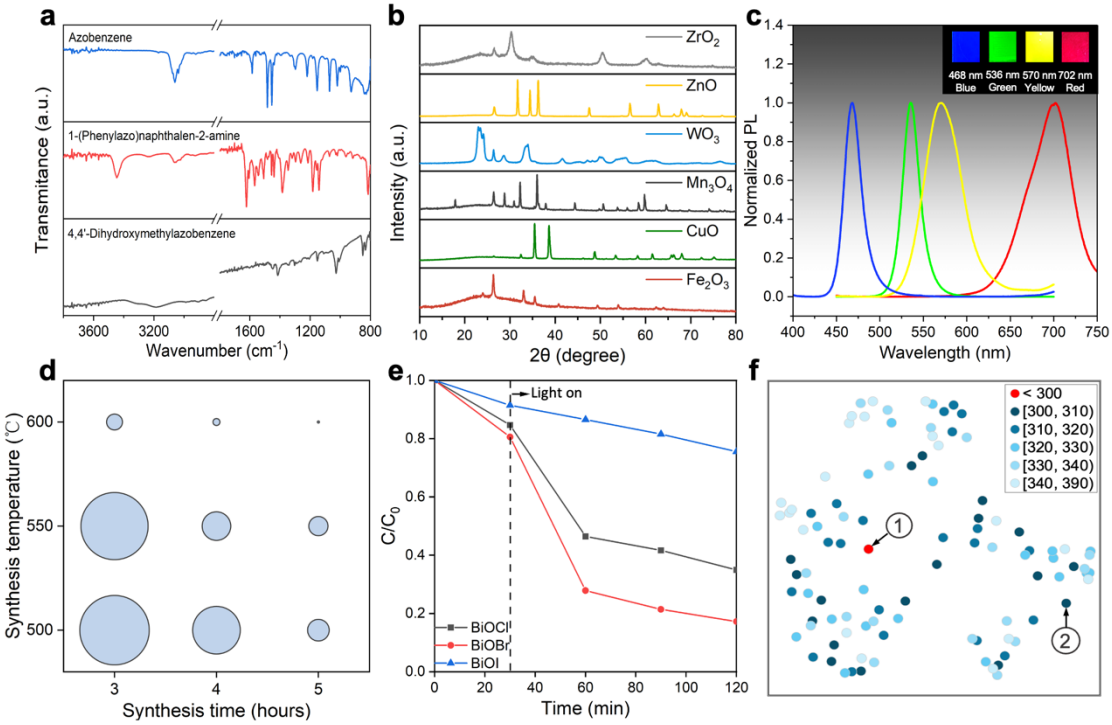

Figure 6: On-demand autonomous chemical research. a, Task 1: measured FT-IR spectra of three azobenzene molecules. **b,** Task 2: measured PXRD patterns of six metal oxides. **c,** Task 3: normalized fluorescence emission spectra of perovskite quantum dots films displaying different colours, shown as insets. **d,** Task 4: measured HER performances (ranging from 9.28 × 10^{-5} to 2.10 × 10^{-3} mmol/g) for g-C_3N_4; the symbol sizes are proportional to the corresponding HER performances. **e,** Task 5: measured photocatalytic degradation of tetracycline in water by bismuth oxyhalides. **f,** Task 6: 2D UMAP embedding of the 101 MO-HECs' compositional space, with symbol colours indicating the value ranges to which their measured overpotentials (in mV) belong; labels 1 and 2 indicate the BO-discovered catalyst and the best catalyst from the random sampling, respectively.

"Discovery & optimization" tasks. These are the most advanced research tasks that the robotic AI chemist can currently accomplish, requiring the use of all four role-specific agents and their corresponding foundational resources, as illustrated in Figure 1. A typical workflow involves (1) the Literature Reader generating experimental designs from the Literature Database, (2) the Experiment Designer creating experimental procedures from the Protocol Library, (3) the Robot Operator generating codes and commands for execution of experiments in the Automated Lab, and (4) the Computation Performer using pre-trained models from the Model Library to analyse data and generate optimization targets. As an example, we tasked the robotic AI chemist with discovering metal-organic high-entropy catalysts (MO-HECs) for the oxygen evolution reaction (OER). The process began with a single prompt to the Task Manager:

"*A scientific research task has been assigned to you to discover high-performance metal-organic high-entropy catalysts (MO-HECs), comprising five metallic elements, for the oxygen evolution reaction (OER). Please use the Literature Reader to identify prevalent metals for OER in the literature and select the top five recommendations. The automated laboratory is equipped with the necessary precursor solutions and other*

required chemicals. Please select 100 quinary metal compositions by random sampling, with each metal's proportion ranging from 5% to 35%, and conduct their syntheses and overpotential measurements. After obtaining the experimental data from the 100 automated experiments, use the Computation Performer to optimize the experiments. Use the compositions and measured overpotentials of these 100 experiments, together with a pre-trained model, to build a fused model that predicts overpotentials from compositions. Then, use this fused model in conjunction with Bayesian optimization to search for the composition that gives the lowest predicted overpotential. The constraints for Bayesian optimization are that the metal proportions must range from 5% to 35% and sum up to 100%. After the Computation Performer provides the results, your final experiment is to synthesize and test this optimal composition experimentally."

When instructed by the Task Manager to query the Literature Database for the most prevalent metallic elements related to OERs, the Literature Reader identified cobalt (Co), nickel (Ni), iron (Fe), manganese (Mn), and copper (Cu) as the top five recommendations. The Experiment Designer then identified an appropriate experimental protocol and generated a detailed procedure. This procedure included the preparation of required metal compositions for MO-HECs, their synthesis, separation, and sample preparation for electrochemical testing. Extended Data Figure 1b illustrates the key robotic operations throughout this experimental procedure.

The robotic AI chemist synthesized 100 MO-HECs with randomly chosen metal compositions and measured their overpotentials, all of which were above 300 mV at 10 mA cm^{-2}. Following the task description quoted above, the Computation Performer was then deployed to select a suitable pre-trained model from the Model Library, which predicts OER activity descriptors from metal composition. The Computation Performer generated a "fused" model by appending one hidden layer and a new output layer to the neural network architecture of the pre-trained model. This fused model was trained using the 100 experimental samples (compositions with measured overpotentials), tuning only the weights of the new hidden layer. As a result, the fused model predicted experimentally measured overpotentials for given metal compositions by correlating OER activity descriptors with compositions and overpotentials.

The Computation Performer used this trained fused model, in conjunction with its Bayesian optimizer, to conduct virtual experiments to identify metal compositions imparting reduced overpotentials to MO-HECs. Specifically, the Computation Performer ran a Bayesian optimization: when a specific composition was evaluated, the fused model acted as an "oracle" to provide its predicted overpotential. After the optimal composition with the lowest overpotential was found, it was passed on to the Experiment Designer and Robot Operator for experimental synthesis and testing. The MO-HEC with the optimal metal composition exhibited an overpotential of 266.1 mV at 10 mA cm^{-2}, a significant improvement over the 100 MO-HECs with randomly chosen metal compositions.

Figure 6f depicts a two-dimensional (2D) UMAP projection of the metal-composition space formed by the 101 experimentally measured MO-HECs. In this plot, points that are closer together represent compositions that are more similar based on Euclidean distance. This visualization highlights that the composition of the optimal catalyst identified leveraging the robotic AI chemist's computational capabilities is significantly different from that of the best catalyst among the 100 randomly selected samples. Additionally, the 2D UMAP embedding reveals that catalysts with similar compositions (i.e., those close to each other in the plot) can exhibit vastly different overpotential values. This lack of a strong composition-activity correlation underscores the challenge of discovering high-performance catalysts in this chemical space while highlighting the efficacy of our approach in identifying such catalysts. Furthermore, Supplementary Fig. S9 confirms that the optimal MO-HEC remained highly stable, showing less than a 2% reduction in performance over 500 hours of continuous operation.

Outlook

In this work, we present our latest advancement in the development and implementation of a multi-agent-driven robotic AI chemist, leveraging the capabilities of LLMs, specifically Llama-3-70B, and ChemAgents, a hierarchical multi-agent system. ChemAgents integrates various specialized LLM-based agents capable of performing complex research tasks both individually and collaboratively. It comprises an Task Manager agent that interacts with human researchers and manages four role-specific agents—Literature Reader, Experiment Designer, Computation Performer, and Robot Operator. This architecture enhances the effectiveness of single agents through collaboration and coordination, allowing them to collectively tackle complex tasks and achieve defined objectives. Underpinning ChemAgents are four foundational resources: a comprehensive Literature Database, an extensive Protocol Library, a versatile Model Library, and a state-of-the-art Automated Lab. These resources enable the robotic AI chemist to perform a wide range of tasks from literature mining and experimental procedure generation to computational modelling and robotic execution of experiments.

The robotic AI chemist demonstrated its versatility in handling a variety of experimental tasks with varying levels of complexity. Six experimental tasks were conducted, encompassing "make & measure," "exploration & screening," and "discovery & optimization" workflows. These experiments highlighted ChemAgents's adaptability and precision in generating experimental procedures, executing robotic experiments, and utilizing computational models for data-driven discovery and optimization. Notably, the robotic AI chemist successfully navigated the vast chemical space of possible quinary compositions to identify high-performance metal-organic high-entropy catalysts (MO-HECs) for electrochemical oxygen evolution reactions. Through literature mining, computational modelling, and Bayesian optimization, the system identified and experimentally validated an optimal MO-HEC composition, achieving a substantial improvement in overpotential performance.

The development of the multi-agent-driven robotic AI chemist marks a significant step towards fully autonomous chemical research. This advancement enables the execution of complex, multi-step experiments with minimal human intervention, thereby opening new avenues for scientific inquiry. As autonomous chemistry evolves from isolated autonomous labs[43,44] to coordinated, cloud-based systems[64], and, in the future, to advanced networks of Intelligent Scientist Systems[65], our multi-agent-driven robotic AI chemist aligns closely with this trajectory and will help accelerate these transitions. The system's ability to autonomously plan, coordinate, manage, and execute chemical research tasks promises to drive unprecedented efficiencies and innovations in research. Additionally, it democratizes the capability to conduct autonomous chemical research, making advanced robotic experimental methodologies and capabilities accessible to a diverse and broad range of researchers. This democratization not only broadens participation in autonomous chemical research but also fosters a more inclusive and collaborative scientific community, paving the way for accelerated discoveries and technological advancements.

Acknowledgements

Y.L., J.J., and W.S. acknowledge the Strategic Priority Research Program of the Chinese Academy of Sciences for funding (Grant XDB0450302). Y.L. acknowledges the Innovation Program for Quantum Science and Technology for funding (Grant 2021ZD0303303). J.J. acknowledges the National Natural Science Foundation of China (Grants 22025304, 22033007) and the CAS Project for Young Scientists in Basic Research (Grant YSBR-005) for funding. W.S. and Y.H. acknowledge the National Natural Science Foundation of China for funding through Grants U22A2056 and 22303091, respectively. Q.Z. acknowledges the National Natural Science Foundation of China (Grant 22103076) and Anhui Provincial Natural Science Foundation (Grant 2108085QB63) for funding. We gratefully acknowledge the USTC Center for Micro- and Nanoscale Research and Fabrication for providing experimental resources and the USTC supercomputing center, as well as Hefei Advanced Computing Center, for providing computational resources.

Contributions

Y.L., J.J., W.S., and L.C. conceived the project, conceptualized the design of the multi-agent-driven robotic AI chemist, and supervised the project. T.S. carried out the development of the multi-agent system and its integration with the robotic system and automated experimental stations, which was overseen by J.J., W.S., F.Z., and L.C. M.L., Q.Z., and D.L. carried out the experimental work, which was overseen by J.J., L.C., D.L., Q.Z., and G.Z. Y.H. carried out the theoretical modelling and calculations. B.Z. built the literature database. J.J., Y.L., L.C., and W.S. led the preparation of the manuscript, with contributions from all other authors.

Data availability

The authors declare that the data supporting the findings of this study are available within the paper and its Supplementary Information files.

Code availability

All Python scripts, experimental protocols, and pre-trained machine learning models used in this study are available at https://aichem.ustc.edu.cn/Jiang_Group.USTC, within the codes_for_paper, warehouse_for_experimental_protocols, and warehouse_for_pretrained_models repositories, respectively.

Competing interests

The authors declare no competing interests.

References

1 Boiko, D. A., MacKnight, R., Kline, B. & Gomes, G. Autonomous chemical research with large language models. *Nature* **624**, 570-578 (2023).

2 M. Bran, A. *et al.* Augmenting large language models with chemistry tools. *Nat. Mach. Intell.* **6**, 525-535 (2024).

3 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).

4 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).

5 Zhao, W. X. *et al.* A survey of large language models. Preprint at https://arxiv.org/abs/2303.18223 (2023).

6 Kaddour, J. *et al.* Challenges and applications of large language models. Preprint at https://arxiv.org/abs/2307.10169 (2023).

7 Thoppilan, R. *et al.* LaMDA: language models for dialog applications. Preprint at https://arxiv.org/abs/2201.08239 (2022).

8 Chowdhery, A. *et al.* PaLm: scaling language modeling with pathways. *J. Mach. Learn. Res.* **24**, 1-113 (2023).

9 Touvron, H. *et al.* LLaMA: open and efficient foundation language models. Preprint at https://arxiv.org/abs/2302.13971 (2023).

10 Achiam, J. *et al.* GPT-4 technical report. Preprint at https://arxiv.org/abs/2303.08774 (2023).

11 Jablonka, K. M., Schwaller, P., Ortega-Guerrero, A. & Smit, B. Leveraging large language models for predictive chemistry. *Nat. Mach. Intell.* **6**, 161-169 (2024).

12 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).

13 Beveridge, C. *et al.* Comprehensive lipidomic automation workflow using large language models. Preprint at https://arxiv.org/abs/2403.15076 (2024).

14 Polak, M. P. & Morgan, D. Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nat. Commun.* **15**, 1569 (2024).

15 Hocky, G. M. Connecting molecular properties with plain language. *Nat. Mach. Intell.* **6**, 249-250 (2024).
16 Li, J. *et al.* Empowering molecule discovery for molecule-caption translation with large language models: a ChatGPT perspective. Preprint at https://arxiv.org/abs/2306.06615 (2023).
17 Zhang, D. *et al.* ChemLLM: a chemical large language model. Preprint at https://arxiv.org/abs/2402.06852 (2024).
18 Wang, L. *et al.* A survey on large language model based autonomous agents. *Front. Comput. Sci.* **18**, 186345 (2024).
19 Xi, Z. *et al.* The rise and potential of large language model based agents: a survey. Preprint at https://arxiv.org/abs/2309.07864 (2023).
20 Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K. & Yao, S. Reflexion: language agents with verbal reinforcement learning. *Adv. Neural Inf. Process. Syst.* **36**, 8634-8652 (2023).
21 Patil, S. G., Zhang, T., Wang, X. & Gonzalez, J. E. Gorilla: large language model connected with massive APIs. Preprint at https://arxiv.org/abs/2305.15334 (2023).
22 Song, Y. *et al.* RestGPT: connecting large language models with real-world RESTful APIs. Preprint at https://arxiv.org/abs/2306.06624 (2023).
23 Yang, Z. *et al.* MM-REACT: prompting ChatGPT for multimodal reasoning and action. Preprint at https://arxiv.org/abs/2303.11381 (2023).
24 Shen, Y. *et al.* HuggingGPT: solving AI tasks with ChatGPT and its friends in Hugging Face. *Adv. Neural Inf. Process. Syst.* **36**, 38154-38180 (2023).
25 Chen, K. *et al.* Chemist-X: large language model-empowered agent for reaction condition recommendation in chemical synthesis. Preprint at https://arxiv.org/abs/2311.10776 (2023).
26 McNaughton, A. D. *et al.* CACTUS: chemistry agent connecting tool-usage to science. Preprint at https://arxiv.org/abs/2405.00972 (2024).
27 Ansari, M. & Moosavi, S. M. Agent-based learning of materials datasets from scientific literature. Preprint at https://arxiv.org/abs/2312.11690 (2023).
28 Chen, K. *et al.* An autonomous large language model agent for chemical literature data mining. Preprint at https://arxiv.org/abs/2402.12993 (2024).
29 Manzano, J. S. *et al.* An autonomous portable platform for universal chemical synthesis. *Nat. Chem.* **14**, 1311-1318 (2022).
30 Greenaway, R. *et al.* High-throughput discovery of organic cages and catenanes using computational screening fused with robotic synthesis. *Nat. Commun.* **9**, 2849 (2018).
31 Li, J. *et al.* Synthesis of many different types of organic small molecules using one automated process. *Science* **347**, 1221-1226 (2015).
32 Mehr, S. H. M., Craven, M., Leonov, A. I., Keenan, G. & Cronin, L. A universal system for digitization and automatic execution of the chemical synthesis literature. *Science* **370**, 101-108 (2020).
33 Granda, J. M., Donina, L., Dragone, V., Long, D.-L. & Cronin, L. Controlling an organic synthesis robot with machine learning to search for new reactivity. *Nature* **559**, 377-381 (2018).
34 Blair, D. J. *et al.* Automated iterative C sp 3–C bond formation. *Nature* **604**, 92-97 (2022).
35 Rohrbach, S. *et al.* Digitization and validation of a chemical synthesis literature database in the ChemPU. *Science* **377**, 172-180 (2022).
36 Slattery, A. *et al.* Automated self-optimization, intensification, and scale-up of photocatalysis in flow. *Science* **383**, eadj1817 (2024).

37 Bennett, J. *et al.* Autonomous reaction Pareto-front mapping with a self-driving catalysis laboratory. *Nat. Chem. Eng.* **1**, 240-250 (2024).
38 MacLeod, B. P. *et al.* Self-driving laboratory for accelerated discovery of thin-film materials. *Sci. Adv.* **6**, eaaz8867 (2020).
39 Zhao, H. *et al.* A robotic platform for the synthesis of colloidal nanocrystals. *Nat. Synth.* **2**, 505-514 (2023).
40 Szymanski, N. J. *et al.* An autonomous laboratory for the accelerated synthesis of novel materials. *Nature* **624**, 86-91 (2023).
41 Chen, J. *et al.* Navigating phase diagram complexity to guide robotic inorganic materials synthesis. *Nat. Synth.* **3**, 606-614 (2024).
42 Jiang, Y. *et al.* Autonomous biomimetic solid dispensing using a dual-arm robotic manipulator. *Digit. Discov.* **2**, 1733-1744 (2023).
43 Burger, B. *et al.* A mobile robotic chemist. *Nature* **583**, 237-241 (2020).
44 Zhu, Q. *et al.* An all-round AI-Chemist with a scientific mind. *Natl. Sci. Rev.* **9**, nwac190 (2022).
45 Zhu, Q. *et al.* Automated synthesis of oxygen-producing catalysts from Martian meteorites by a robotic AI chemist. *Nat. Synth.* **3**, 319-328 (2024).
46 Walker, M., Pizzuto, G., Fakhruldeen, H. & Cooper, A. I. Go with the flow: deep learning methods for autonomous viscosity estimations. *Digit. Discov.* **2**, 1540-1547 (2023).
47 Lunt, A. M. *et al.* Modular, multi-robot integration of laboratories: an autonomous workflow for solid-state chemistry. *Chem. Sci.* **15**, 2456-2463 (2024).
48 Vemprala, S., Bonatti, R., Bucker, A. & Kapoor, A. ChatGPT for robotics: design principles and model abilities. *IEEE Access* **12**, 55682-55696 (2024).
49 Liang, J. *et al.* Code as policies: language model programs for embodied control. *IEEE International Conference on Robotics and Automation*, 9493-9500 (2023).
50 Singh, I. *et al.* ProgPrompt: generating situated robot task plans using large language models. *IEEE International Conference on Robotics and Automation*, 11523-11530 (2023).
51 Wu, J. *et al.* TidyBot: personalized robot assistance with large language models. *Auton. Robot.* **47**, 1087-1102 (2023).
52 Ahn, M. *et al.* Do as I can, not as I say: grounding language in robotic affordances. Preprint at https://arxiv.org/abs/2204.01691 (2022).
53 Driess, D. *et al.* PaLM-E: an embodied multimodal language model. *Proceedings of the 40th International Conference on Machine Learning* **202**, 8469-8488 (2023).
54 Huang, W. *et al.* VoxPoser: composable 3D value maps for robotic manipulation with language models. Preprint at https://arxiv.org/abs/2307.05973 (2023).
55 Yoshikawa, N. *et al.* Large language models for chemistry robotics. *Auton. Robot.* **47**, 1057-1086 (2023).
56 Darvish, K. *et al.* ORGANA: a robotic assistant for automated chemistry experimentation and characterization. Preprint at https://arxiv.org/abs/2401.06949 (2024).
57 Guo, T. *et al.* Large language model based multi-agents: a survey of progress and challenges. Preprint at https://arxiv.org/abs/2402.01680 (2024).
58 Han, S. *et al.* LLM multi-agent systems: challenges and open problems. Preprint at https://arxiv.org/abs/2402.03578 (2024).
59 Park, J. S. *et al.* Generative agents: interactive simulacra of human behavior. Preprint at https://arxiv.org/abs/2304.03442 (2023).

60 Qian, C. *et al.* ChatDev: communicative agents for software development. Preprint at https://arxiv.org/abs/2307.07924 (2023).

61 Wu, Q. *et al.* AutoGen: enabling next-gen LLM applications via multi-agent conversation. Preprint at https://arxiv.org/abs/2308.08155 (2023).

62 *MindSpore*, https://www.mindspore.cn/en (2024).

63 Zhang, B. *et al.* Label-free data mining of scientific literature by unsupervised syntactic distance analysis. *J. Phys. Chem. Lett.* **15**, 212-219 (2024).

64 Strieth-Kalthoff, F. *et al.* Delocalized, asynchronous, closed-loop discovery of organic laser emitters. *Science* **384**, eadk9227 (2024).

65 Jiang, X. *et al.* Artificial intelligence and automation to power the future of chemistry. *Cell Rep. Phys. Sci.*, 102049 (2024).

Extended Data Figure 1: The Automated Lab. a, Depiction of the two experiment-conducting robots: a fully mobile robot and a benchtop robotic arm. **b,** Examples of key robotic operations along the autonomous workflow for sample synthesis and electrochemical testing.