Large-Model-Driven, Human-Machine Collaborative Robot Chemist Cloud Infrastructure Postprint

Authors: Chong Yuanyuan, Feng Shuo, Wang Song, Jiang Jun

Date: 2024-03-27T00:00:00+00:00

Abstract

In the current wave of technology driven by artificial intelligence, chemical scientific research is facing unprecedented opportunities and challenges. To promote the transformation of chemical research paradigms, this article proposes a construction plan for a machine chemist cloud facility. The system collects multichannel data to construct databases, develops chemistry knowledge-enhanced scientific large models, builds robot facility clusters, and constructs intelligent management and decision-making systems, thereby practicing a new paradigm of scientific research, substantially improving research efficiency, and solving scientific problems in end-user applications. This infrastructure is expected to propel paradigm shifts in scientific research and achieve major scientific breakthroughs in the field of chemistry.

Full Text

Large Model-Driven, Human-Computer Collaborative Robotic AI-Chemist Cloud Facility

Chong Yuanyuan, Feng Shuo, Wang Song, Jiang Jun*

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

Abstract

In the current wave of technological transformation driven by artificial intelligence, chemical science research is facing unprecedented opportunities and

challenges. To promote a paradigm shift in chemical research, this paper proposes a construction plan for a robotic AI-chemist cloud facility. This system implements a new scientific research paradigm by collecting multi-channel data to build databases, developing scientific large models enhanced with chemical knowledge, constructing clusters of robotic facilities, and establishing an intelligent management and decision-making system. This approach will dramatically improve research efficiency and solve scientific problems in end-user applications. This infrastructure is expected to drive paradigm transformation in scientific research and achieve major scientific breakthroughs in chemistry.

Keywords: paradigm shift in chemical research, robotic AI-chemist cloud facility, artificial intelligence, automated experiments, human-computer collaboration

---

1. Paradigm Transformation in Chemical Research: Challenges, Opportunities, and Trends

Chemistry, as a fundamental science, focuses on studying the composition, structure, and properties of matter, as well as its behavior under different conditions and interactions with other substances. Experimental and theoretical research methods complement each other and jointly advance chemical science, which holds significant importance and broad applications in developing new materials, exploring new energy sources, and improving biomedical technologies [**?**]. However, contemporary chemical research faces formidable challenges as research objects become increasingly complex and high-dimensional. Current mainstream research methods rely on exhaustive trial-and-error and variable complexity reduction, whose inefficiency and limitations are becoming ever more apparent [**?**]. From the atomic and molecular scale to macroscopic applied materials, material properties are influenced by various reaction conditions and interactions [**?**], making accurate prediction and description difficult. The growing desire to decode the underlying laws of complex systems from micro to macro scales will provide crucial guidance for automated synthesis optimization [**?**], on-demand inverse design of materials [**?**], and precise control of biomedical processes [**?**]. Nevertheless, the bottom-up evolution from fundamental principles—physical constants, Schrödinger equations, and the periodic table—to complex applications involves immense complexity and diversity [**?**, **?**], creating a disconnect between real-world problems and structure-performance relationships. Chemical synthesis still depends heavily on expert experience, falling far short of intelligent optimization goals. Incomplete data and unclear structure-property relationships remain major obstacles to inverse design of customized materials, while the lack of evolutionary information about central dogma processes in biology limits human understanding of disease mechanisms and the essence of life.

To actively address these challenges in chemical science, we must innovate re-

search methods and transform research paradigms. The arrival of the big data era has ushered in a data-driven research paradigm. Artificial intelligence excels at exploring relationships among variables in high-dimensional, highly complex data [**?**, **?**], offering unprecedented opportunities to meet these challenges. Deep learning and large models, as representative AI technologies, possess capabilities such as learning, adaptability, autonomous decision-making, pattern recognition, and prediction, demonstrating advantages in intelligent decision-making that surpass human capabilities (Figure 1 [Figure 1: see original paper]). In 2016, DeepMind' s AlphaGo program employed deep reinforcement learning, combining deep neural networks with reinforcement learning algorithms to efficiently search and make precise judgments about Go strategies, surpassing traditional human heuristic search methods [**?**]. This human-machine competition in Go became a milestone in AI history, first demonstrating AI' s potential in complex decision-making domains. In 2021, the protein structure prediction program AlphaFold2 achieved high-precision prediction of protein 3D structures by training on large-scale sequence data based on deep neural networks and self-attention mechanisms [**?**], representing a breakthrough with potential significance for drug design and disease diagnosis. In 2023, the conversational model ChatGPT captured global attention by generating language using Transformer architecture' s self-attention mechanisms and multi-layer neural networks, continuously iterating its language generation capabilities through unsupervised learning to enhance human-computer interaction [**?**]. This represents breakthrough progress in natural language processing, promising to help humans obtain information and make intelligent decisions, enabling the emergence of general cognitive intelligence.

Intelligent-driven robotic chemical research has also achieved a series of breakthroughs. In 2022, Cronin' s team at the University of Glasgow developed the automated robotic system Chemputer, which integrates literature reading, experimental protocol customization, compound synthesis, and characterization. It can convert synthesis steps from literature into machine-readable chemical description languages and store them in internal databases for automatic execution by robots [**?**]. Cooper' s team at the University of Liverpool developed a mobile robotic chemist that can efficiently conduct experiments and uses Bayesian algorithm optimization to analyze existing experimental data and improve experimental plans. However, Cooper noted that current robots lack a computational brain, do not utilize existing chemical knowledge, and cannot incorporate theories or physical models, making Bayesian optimization blind [**?**]. In 2022, Jiang Jun' s team at the University of Science and Technology of China developed an all-round AI-chemist driven by data intelligence, comprising machine reading, machine computing, and machine experimentation systems. It can learn from previous knowledge and wisdom, generate physical models and provide intelligent predictions, and conduct efficient experiments to generate full lifecycle data. This platform leverages the advantages of machine data being reproducible, trustworthy, traceable, and alignable, using precise experimental data to calibrate theoretical pre-trained models, achieving intelligent prediction

that integrates theory and practice [**?**].

1.3 Development Trends in Chemical Science Research

The international landscape has shifted dramatically since ChatGPT demonstrated the feasibility of general cognitive intelligence in early 2023. Within half a year, the United States, United Kingdom, Canada, Netherlands, Switzerland, and other countries accelerated investment in developing intelligent scientific large models as the "brain" for scientific equipment. In 2023, the U.S. updated its National Artificial Intelligence Research and Development Strategic Plan, investing substantial annual funding to support research in data science, AI, quantum information, and other fields [**?**]. Starting in 2023, the UK also invested in building intelligent innovation workshops that integrate large models, robotics, and intelligent alliances. In April 2023, Canada added 1.5 billion RMB in investment to the Acceleration Consortium for large-scale intelligent laboratory infrastructure. In July 2023, the Netherlands began building a robotic chemistry laboratory. In December 2022, Switzerland invested in creating public service facilities driven by large models. Machine scientists with chemical wisdom supporting industrial digitalization have become a reality: in 2022, 60% of Unilever' s annual R&D budget was used to purchase intelligent synthesis and testing services from the University of Liverpool' s Materials Innovation Factory.

In China, we currently lead locally in machine chemist systems covering the entire workflow of intelligent literature research, research planning, computation, experimentation, and optimization. However, we urgently need institutional project deployment in large-scale intelligent laboratories and chemical science large models to avoid the situation of "starting early but arriving late."China has a solid foundation for building robotic AI-chemist cloud facilities, including the "Zidong Taichu" all-modal large model developed by the Institute of Automation and Wuhan Artificial Intelligence Research Institute, the "Xinghuo" cognitive large model developed by iFLYTEK, and over 20 scientific data centers and intelligent computing centers deployed nationwide by the Chinese Academy of Sciences.

2. Robotic AI-Chemist Cloud Facility: A New Tool for Future Chemical Research

The rapid development of AI technology is bringing unprecedented opportunities and challenges to chemical science research. In the current wave of technological revolution and industrial transformation, developing new chemical research tools that integrate scientific data, AI algorithms, intelligent robots, and cloud platforms has become an urgent and necessary task. This tool is expected to solve the long-standing curse of dimensionality and black-box problems of complex giant systems in chemical science innovation, thereby driving disruptive breakthroughs in high-value chemicals, functional materials, and biochemical

medicine.

2.1 Concept of the Robotic AI-Chemist Cloud Facility

Traditional research workflows for human chemists typically consist of proposing requirements, reviewing literature, designing protocols, theoretical simulation and experimental validation, extracting theories, and solving practical problems. In contrast, the robotic AI-chemist cloud facility, encompassing databases, human-computer interaction, robotic experimenters, chemical workstations, and a chemical brain, can not only fully cover these processes but also customize solutions to specific problems through human-computer collaborative systems (Figure 2 [Figure 2: see original paper]).

Data is a crucial component of modern scientific research and is even more critical for the robotic AI-chemist cloud facility. By learning from vast amounts of chemical data in databases, the robotic AI-chemist can acquire previous knowledge and wisdom. Human-computer interaction raises scientific questions, which are then processed by the chemical brain that integrates scientific large models to establish physical models and provide intelligent predictions. The system subsequently generates research protocols, drives efficient robotic experimenters, chemical workstations, and intelligent computing servers to produce high-quality experimental and theoretical simulation data. This data feedback optimizes the scientific large models, forming vertical domain application models to solve specific scientific problems. Its unique advantage lies in efficiently integrating data knowledge, continuously adjusting theoretical and experimental designs, and achieving full-process intelligent deduction. Currently, the University of Science and Technology of China has successfully developed the world' s first all-round AI-chemist driven by data intelligence, providing a technical foundation for aggregating multi-disciplinary methodologies, integrating multi-domain knowledge logic, coupling the wisdom of chemical scientist communities, and reducing experimental workload.

2.2 Significance of Building the Robotic AI-Chemist Cloud Facility

The robotic AI-chemist cloud facility integrates deep learning and scientific large models into experimental robotic hardware, providing a technical foundation that will accelerate experimental design and data analysis processes, enhancing the efficiency and accuracy of chemical science research. Currently, multiple countries including the US and UK are accelerating investment in developing robotic research tools equipped with scientific large models. The intelligent field is characterized by "winner-takes-all"dynamics with almost no late-mover advantage; only by seizing the initiative and mastering advanced research tools first can China avoid being constrained in the new round of technological revolution. Therefore, capitalizing on China' s leading position in self-developed robotic chemists and building the robotic AI-chemist cloud facility can prevent China from being "choked" in basic research tools for the new paradigm of intelligent chemical research and help seize advantageous positions in the field.

The facility will also generate positive spillover effects across society, promoting industrial digitalization and productivity gains, and potentially catalyzing a new round of industrial transformation. Overall, the robotic AI-chemist cloud facility holds tremendous and far-reaching significance for enhancing China' s competitiveness in scientific and technological innovation and ensuring leadership in emerging technology fields, positioning China to achieve greater development and breakthroughs in the global new round of technological revolution.

3. Robotic AI-Chemist Cloud Facility: Layered Architecture

Through the mutual confrontation and collaborative evolution of scientific large model predictions and intelligent robot empirical validation, we can create a robotic AI-chemist cloud facility with chemical scientific intelligence that drives paradigm transformation and produces major scientific breakthroughs.

3.1 Chemical Science Database

In the data-driven research paradigm, effective integration and utilization of scientific data are the core drivers of innovation. However, current scientific data generally suffer from non-uniform standards, uneven quality, and relative isolation of multi-source data, limiting data-based chemical science research. Therefore, there is an urgent need to break data silos and integrate theoretical and experimental data from different sources to build an AI-enabled chemical science database with multi-disciplinary knowledge and multi-modal data. This will provide a solid data foundation for intelligent development in chemical science.

The chemical science database will embed AI models and integrate literature data, theoretical data, and experimental data, including four key aspects: (1) **Chemical science domain data aggregation**: Integrate data resources from various institutions, utilizing multi-modal data from scientific literature (text, tables, images) and large amounts of fundamental physicochemical data on chemical molecules and materials generated by first-principles simulations. Simultaneously, establish experimental data collection channels and national standards to enable automatic collection and rapid aggregation of standardized data. (2) **Scientific literature machine reading tool construction**: Through cleaning, screening, and annotation of corpus data from scientific journals, textbooks, and question banks, obtain high-value general-domain pre-training corpora and chemical science domain pre-training corpora. Use deep mining technology to extract computational and experimental data from literature content. (3) **Data curation and high-quality database construction**: Annotate pre-training corpora, compile computational and experimental data from literature, and conduct data classification and quality assessment. Develop data discrimination and quality scoring techniques based on interpretable models to intelligently clean data. (4) **Knowledge embedding and knowledge graph**

construction: Use mapping relationship analysis to build association models and establish chemical science knowledge graphs including structure, properties, and evolutionary correlations. Guide multi-modal data fusion through knowledge graphs to build unified, efficient, scalable, and clearly structured data storage formats. Use pre-trained models and other tools to embed knowledge graphs into chemical science large models, enhancing knowledge utilization efficiency.

3.2 Scientific Large Model

Current neural network-based large models suffer from core issues such as low reliability in predictions, insufficient depth in logical reasoning and semantic understanding, and weak interpretability and debuggability, resulting in poor performance in chemistry applications where high accuracy is required. To address these problems, we need to develop scientific large models based on mathematical logic, deeply integrating data-driven neural network models with knowledge-driven symbolic logic reasoning engines for application in intelligent science fields including mathematics, chemistry, and physics.

The scientific large model framework proposed in this study focuses on developing a knowledge-driven reasoning engine built upon domain ontologies and knowledge bases, connected to databases and potential databases to simulate human cognitive reasoning and decision-making capabilities, thereby compensating for defects in reliability, interpretability, and debuggability of large models. By integrating knowledge graphs and chemistry-aware knowledge enhancement algorithms, the scientific large model incorporates expert chemical knowledge and understanding, utilizes characteristic chemical descriptors, and creates clear AI algorithms based on chemical principles to solve complex challenges such as large-scale screening and strategy optimization, ultimately building a machine scientist brain with "chemical wisdom." Based on user requirements, it designs experimental protocols and operational workflows, analyzes experimental data in real time, adjusts intelligent models, and continuously feeds back to optimize experimental plans, achieving automatic decision-making and optimization of experimental protocols and processes.

3.3 Robotics Platform

The robotics platform will provide efficient and precise experimental and data processing solutions, including four specific aspects: (1) **Fully automated high-throughput research system based on microchannel continuous flow**: The system aims to conduct important organic chemical reactions and key functional material synthesis with precision, automation, and high throughput, requiring solutions to multi-domain technical problems and integration of multiple key functional subsystems including multi-channel reactant automatic switching, microchannel continuous flow reaction, product collection and post-processing, online detection and automatic sampling, chromatographic interfaces, reaction temperature control, master control, and human-computer interaction systems. (2) **Function expansion of fully automated high-**

throughput research systems: To ensure reliability of high-throughput experimental results, each subsystem is equipped with redundant sensors combined with visual recognition technology for real-time feedback and automatic screening of abnormal data. Researchers only need to prepare reactant libraries and input reaction matrices, and the system will complete experiments, post-processing, and detection, batch-outputting data. Future expansion of research scope can be achieved by adding subsystems and functional modules to complete more complex post-processing and detection analysis. (3) **Mobile manipulation robots for fully autonomous experiments**: Design hardware and software integration of six-degree-of-freedom robotic arms with omnidirectional mobile chassis; develop visual perception algorithms for laboratory environments, as well as dexterous control methods with high-precision visual guidance and real-time force feedback; research high-precision positioning and mapping methods using multi-modal data, and develop dynamic obstacle avoidance algorithms and task management systems to achieve fully autonomous experiments with mobile manipulation robots. (4) **Full-process intelligent chemistry laboratory**: Develop proprietary automatic encapsulation machines, liquid automatic dispensing workstations, and electrochemical automation testing workstations, while designing fully autonomous mobile manipulation robots, high-throughput experimental platforms, collaborative control systems for experimental equipment, and full-process task scheduling systems to build a full-process intelligent chemistry laboratory integrating chemical synthesis, spectroscopic characterization, and performance testing (Figure 3 [Figure 3: see original paper]), achieving full-scenario coverage of chemical research.

3.4 Intelligent Management Decision System

The intelligent management decision system is the intelligent chemistry cloud platform, comprising the robotic AI-chemist instruction set, operating system, and federated learning algorithm system, enabling the robotic AI-chemist to conduct transfer learning across different experimental tasks and laboratories, ultimately building a standardized intelligent chemistry laboratory at the cloud platform level (Figure 4 [Figure 4: see original paper]).

The instruction set includes standardization of four components: interface functions, communication protocols, equipment specifications, and data standards, to support alignment of data from different sources and achieve interconnectivity. The operating system features a friendly human-computer interaction interface, clear business workflows, and intuitive data visualization, helping researchers break free from physical space constraints to conduct experiments, simulations, and data analysis remotely, while facilitating overall system experimental task scheduling and resource allocation. The core of the federated learning algorithm system is "data stays, models move," enabling data sharing between different users and laboratories while ensuring data privacy and security. By publishing standard specifications for intelligent chemistry laboratories, the system enables cloud-based sharing of databases and AI models. This sys-

tem aims to achieve intelligent management and decision-making, promoting operational efficiency across different laboratories.

4. Conclusion: Transforming Chemical Research Through the Robotic AI-Chemist Cloud Facility

The first two "carbon-based" industrial revolutions represented by steam engines and electric motors helped humans break through "physical" limitations, while the third "silicon-based" information technology revolution represented by computers helped humans break through "computational" limitations. The intelligent era has arrived, and the fourth general intelligent industrial revolution that breaks through human "cognitive" limitations is imminent. In response to this era, the chemical science database, scientific large model, robotics platform, and intelligent management decision system of the robotic AI-chemist cloud facility will collectively liberate human researchers from limitations of "memory, physical capacity, computational power, and cognitive ability," breaking knowledge barriers, spatial constraints, and disciplinary boundaries in the research process. By intelligently connecting individual researchers and substantially enhancing their research capabilities, the facility will comprehensively transform China' s chemical science and even entire material science research paradigms.

References

1. Li, W., Li, Z., Zhang, H., et al. Efficient catalysts of surface hydrophobic Cu-BTC with coordinatively unsaturated Cu(I) sites for the direct oxidation of methane. *PNAS*, 2023, 120(10): e2206619120.
2. Dereka, B., Yu, Q., Lewis, N. H. C., et al. Crossover from hydrogen to chemical bonding. *Science*, 2021, 371: 160-164.
3. Majerle, A., Hadži, S., Aupič, J., et al. A nanobody toolbox targeting dimeric coiled-coil modules for functionalization of designed protein origami structures. *PNAS*, 2021, 118(17): e2021899118.
4. Mou, T., Pillai, H. S., Wang, S., et al. Bridging the complexity gap in computational heterogeneous catalysis with machine learning. *Nature Catalysis*, 2023, 6(2): 122-136.
5. Yang, J., Huang, Y., Qi, H., et al. Modulating the strong metal-support interaction of single-atom catalysts via vicinal structure decoration. *Nature Communications*, 2022, 13(1): 7845.
6. Sun, Y., Dai, S. High-entropy materials for catalysis: A new frontier. *Science Advances*, 2021, 7(20): eabg1600.
7. Rai, S. K., Khanna, R., Avni, A., et al. Heterotypic electrostatic interactions control complex phase separation of tau and prion into multiphasic

condensates and co-aggregates. *PNAS*, 2023, 120(2): e2216338120.

8. Li, Y. M., Yuan, J., Ren, H., et al. Fine-tuning the micro-environment to optimize the catalytic activity of enzymes immobilized in multivariate metal-organic frameworks. *Journal of the American Chemical Society*, 2021, 143(37): 15463-15473.
9. Tang, T. C., An, B., Huang, Y., et al. Materials design by synthetic biology. *Nature Reviews Materials*, 2021, 6(4): 332-348.
10. Hirschi, S., Ward, T. R., Meier, W. P., et al. Synthetic biology: Bottom-up assembly of molecular systems. *Chemical Reviews*, 2022, 122(21): 16294-16328.
11. Jousset, A., Eisenhauer, N., Merker, M., et al. High functional diversity stimulates diversification in experimental microbial communities. *Science Advances*, 2016, 2(6): e1600124.
12. Sun, Y., Latora, V. The evolution of knowledge within and across fields in modern physics. *Scientific Reports*, 2020, 10(1): 12097.
13. E, W. N. The Dawning of a new era in applied mathematics. *Notices of the American Mathematical Society*, 2021, 68(4): 565-566.
14. Schmidt, M., Lipson, H. Distilling free-form natural laws from experimental data. *Science*, 2009, 324: 81-85.
15. Silver, D., Huang, A., Maddison, C. J., et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 2016, 529: 484-489.
16. Jumper, J., Evans, R., Pritzel, A., et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 2021, 596: 583-589.
17. OpenAI. GPT-4 Technical Report. arXiv:2303.08774, 2023.
18. Steiner, S., Wolf, J., Glatzel, A., et al. Organic synthesis in a modular robotic system driven by a chemical programming language. *Science*, 2019, 363: eaav2211.
19. Burger, B., Maffettone, P. M., Gusev, V. V., et al. A mobile robotic chemist. *Nature*, 2020, 583: 237-241.
20. Zhu, Q., Zhang, F., Huang, Y., et al. An all-round AI-chemist with a scientific mind. *National Science Review*, 2022, 9(10): nwac190.
21. National Artificial Intelligence Research and Development Strategic Plan: 2023 Update. National Science and Technology Council, 2023.

Note: Figure translations are in progress. See original paper for figures.