



From execution loop to reasoning loop in autonomous materials research

Zhizhi Tan^{1,2}, Li Qian^{1,2}, Ying Wang^{1,2}, Nian Ran^{3,4}

Citation: Tan, Z.; Qian, L.; Wang, Y.; Ran, N. From execution loop to reasoning loop in autonomous materials research. *AI Agent* 2026, 2, 13. <https://dx.doi.org/10.20517/aiagent.2026.13>

Received: 23 Apr 2026

First Decision: 21 May 2026

Revised: 18 Jun 2026

Accepted: 29 Jun 2026

Published: 14 Jul 2026

Academic Editor:

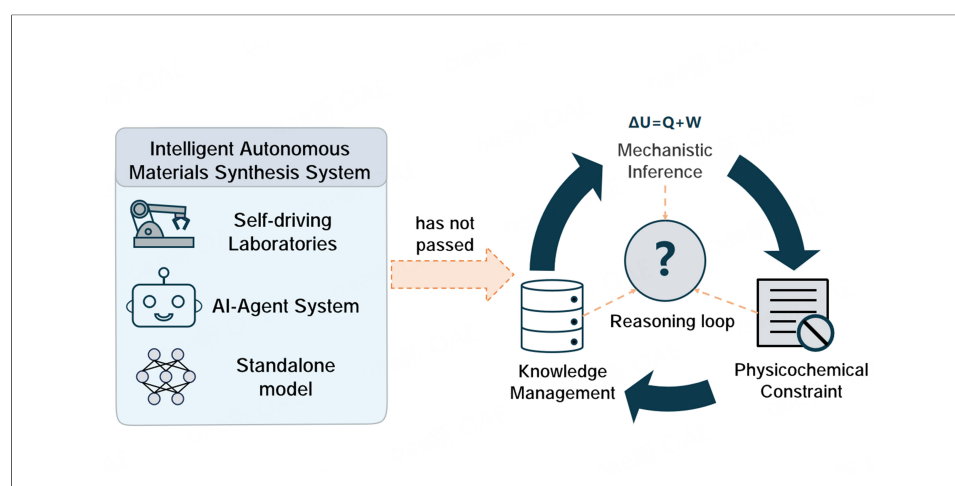
Hao Li

Copy Editor:

Tong Wang

Production Editor:

Tong Wang



INTRODUCTION

Over the past five years, materials research and development have achieved genuine and reproducible progress: multi-agent systems coordinate the execution of multi-step tasks, large language models mine the literature to propose initial synthesis routes, robots carry out experiments, Bayesian optimization searches the parameter space, and the optimized results provide feedback^[1,2]. However, during the same period, the narrative of “autonomous materials research” painted a vague picture of its ultimate goal. It points to a fully autonomous platform that can generalize across material systems, infer synthesis pathways from target properties, and generate mechanistic hypotheses from experimental observations. What has been accomplished and what that endpoint demands are not answers to the same question. The former accelerates iteration within a predefined search space; the latter requires mechanistic reasoning extracted from experimental results and guides the next scientific decision.



¹National Science Library, Chinese Academy of Sciences, Beijing 100190, China.

²Department of Information Resources Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190, China.

³State Key Laboratory of High Performance Ceramics, Shanghai Institute of Ceramics, Chinese Academy of Sciences, Shanghai 200050, China.

⁴Center of Materials Science and Optoelectronics Engineering, University of Chinese Academy of Sciences, Beijing 100049, China.

Correspondence to: Prof. Nian Ran, State Key Laboratory of High Performance Ceramics, Shanghai Institute of Ceramics, Chinese Academy of Sciences, Shanghai 200050, China. E-mail: rannian@mail.sic.ac.cn; Prof. Li Qian, Dr. Ying Wang, National Science Library, Chinese Academy of Sciences, Beijing 100190, China. E-mail: qianl@mail.las.ac.cn; wangy@mail.las.ac.cn

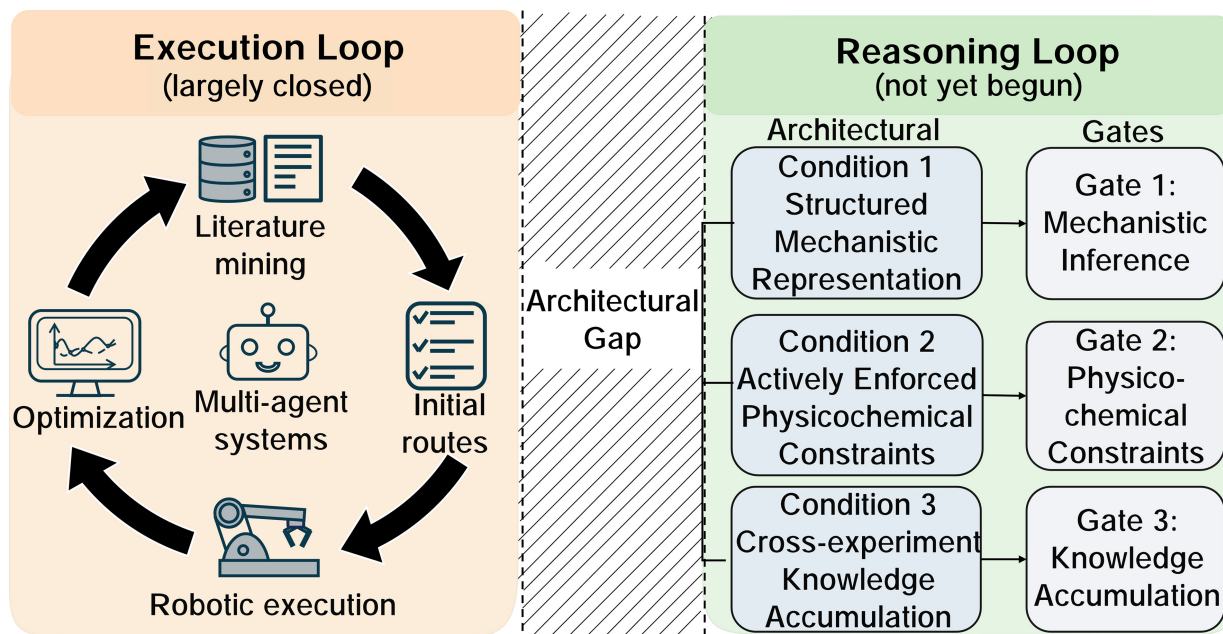


Figure 1. Architectural distinction between the Execution Loop and the Reasoning Loop. The Execution Loop (already feasible) forms a closed multi-agent cycle integrating four peripheral subsystems. An architectural gap separates it from the Reasoning Loop (aspirational), which requires three structural conditions to be realized.

These two categories of progress have long shared the same vocabulary, with no criteria to distinguish them. As a result, execution loop success is often mistaken for progress toward the overarching goal while the architectural gap in the reasoning loop is underestimated. This is not a progress gap; however, it's an architectural gap [Figure 1]. An intelligent autonomous materials research system is neither merely a faster robot nor a more knowledgeable literature reader. It is a system that combines execution loop and reasoning loop. The execution loop starts from an initial experimental scheme, runs high-throughput automated experiments, and refines them through optimization, enabling large-scale trial-and-error at machine speed without mechanistic understanding. By contrast, the reasoning loop is a cyclical inference process in which experimental observations are converted into updated mechanistic hypotheses, which in turn guide subsequent experimental decisions. At each iteration, the system compares new results with its current hypotheses, revises them upon discrepancy, and stores the refined understanding. The execution loop has largely been closed. The reasoning loop has not yet begun.

THE EXECUTION LOOP IS CLOSING

Self-driving laboratories (SDLs) have achieved a genuine engineering milestone over the past five years. Tom *et al.* map SDL autonomy along the software and hardware axes, finding most systems at level 2 to 3, namely closed-loop multi-task optimization^[1]. Huang *et al.* converted natural language to unit-operation code, discovering new Mn-W polyoxometalate clusters through 65 rounds of human and machine interaction^[3]. The gains in throughput, parallelism, and reproducibility are substantial.

But the execution loop is fundamentally a parameter optimization engine: it finds better points within a parameter space predefined by humans, knowing which combination of parameters performs better, but not why. The boundaries of the search space still need to be defined by researchers, rather than being derived by the system from new hypotheses generated through the previous round of experiments. ChemAgents is a

hierarchical multi-agent system built on Llama-3.1-70B that coordinates four specialized sub-agents across literature, experiment design, computation, and robotic operation, demonstrating the execution capability of a hierarchical multi-agent system across seven categories of chemical tasks^[4], yet the objective of each task was specified by humans. The achievements of the execution loop rest precisely within this scope.

HOW THE CONFLATION OF EXECUTION PROGRESS WITH REASONING PROGRESS OCCURS

Large language models create surface equivalence: systems that read literature, draft protocols, decompose tasks, call tools, and communicate in natural language appear to reason. These capabilities are real, but they belong to the execution loop. The confusion centers on the word “reasoning” itself. In one sense, it means stepwise inference of reactants from a product through retrosynthetic pathway decomposition and structured multi-step prediction^[5–7]. In another, it means generating mechanistic hypotheses from experimental characterization data and using them to constrain subsequent decisions. The former applies established rules; the latter uses mechanisms to explain phenomena. SUPERChem, a benchmark for expert-level chemical reasoning with process-level scoring, offers a sobering calibration: humans reach 40.3% accuracy, while the strongest current model attains only 38.5%^[8].

This pattern is also reflected in recent AI-agent discovery frameworks. DigHyd^[9] is an agentic workflow for hydrogen storage materials discovery that combines large language model (LLM) retrieval from domain-specific databases with machine learning property prediction. Its guidance comes from dataset-derived soft constraints and built-in LLM knowledge, not from iterative mechanistic refinement. SciToolAgent^[10] is a knowledge-graph-driven scientific agent that translates research task descriptions into multi-step tool-calling sequences across chemistry and materials science. Its core capability is task decomposition and tool orchestration rather than mechanistic understanding. Such frameworks use existing knowledge effectively, but their LLM reasoning processes remain opaque and vulnerable to hallucination.

The confusion also persists because dominant SDL frameworks classify execution autonomy rather than reasoning capability. Tom *et al.* proposed a five-level SDL classification along software and hardware autonomy^[1]. At Level 5, software proposes experiments within an algorithmically generated search space and hardware executes diverse experiments without human intervention, a state described as the “ultimate goal, not yet achieved”. This framework clarifies operational autonomy, but it does not explain what architecture would allow a system to infer mechanisms and use them to redirect inquiry.

THREE ARCHITECTURAL CONDITIONS THE REASONING LOOP REQUIRES THAT THE EXECUTION LOOP LACKS

The reasoning loop requires three things that differ fundamentally from what the execution loop provides [Figure 2].

The first is structured mechanism representation. For a target system to generalize across material systems, knowledge accumulated in system A must be transferable to system B. A structured mechanistic representation is any formalism that encodes the underlying regularities of a system in a form that can be communicated across components, transferred between pipeline stages, reasoned over to generate and revise mechanistic hypotheses, and persistently stored and updated, whether it takes the shape of causal graphs, symbolic rules, natural-language hypotheses, or other equivalent formalisms. This means characterization data, such as anomalous X-ray diffraction (XRD) peaks and fourier-transform infrared spectroscopy (FTIR) spectral shifts, must serve as the starting point for reasoning. At each iteration, the hypotheses are encoded by a set of descriptors, which are revised to reflect any updates to the hypotheses they represent. Current research planning systems have not reached this point. Material synthesis planning (MSP) LLM takes a target

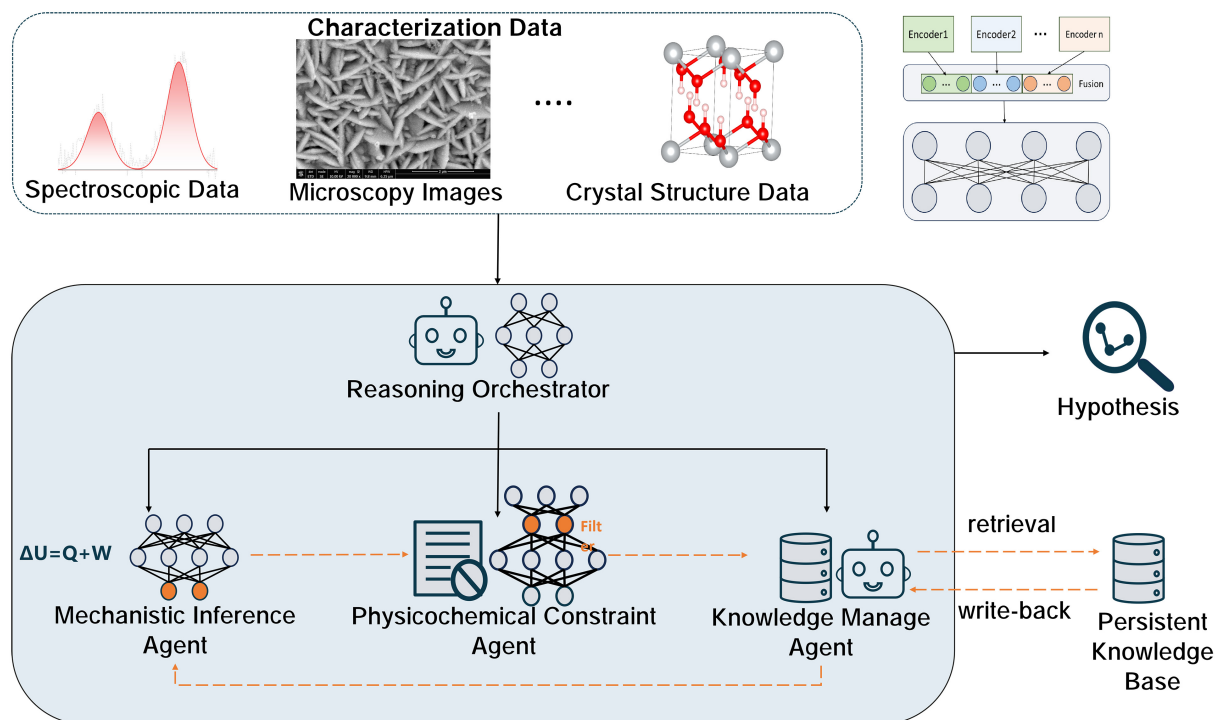


Figure 2. Proposed multi-agent architecture for the Reasoning Loop. Multi-modal characterization data are encoded and fused (already feasible), then routed to a central reasoning orchestrator. Three specialized agents handle mechanistic inference (aspirational), physicochemical constraint enforcement (aspirational), and knowledge management (partly feasible). Validated conclusions feed back iteratively to refine inference, closing the loop.

material as input and plans synthesis by predicting precursors and operations through a chemically consistent decision chain. It outperforms prior methods in precursor and operation sequence prediction^[11], but produces steps rather than mechanisms, and does not include characterization data in reasoning. MOFReasoner, a domain-specific LLM fine-tuned via knowledge distillation and chain-of-thought reasoning to think like a scientist^[12], it mainly distills established knowledge from historical literature and cannot generate mechanistic hypotheses from unseen characterization observations in new experiments.

The second is actively enforced physicochemical constraints. If physicochemical constraints do not intervene actively in the generation process, large language models will produce thermodynamically infeasible or charge-imbalanced proposals with high confidence. HalluMatDetector^[13] documents this pattern by detecting hallucinations in LLM-generated materials responses through self-consistency checks, external retrieval, and contradiction graph analysis. The key issue is not tool variety but how tools are invoked. In ChemAgents, density functional theory (DFT) calculations are called as needed, forming passive queries^[4], whose results may be overridden by later text generation, so the constraints do not truly shape the proposal. FlowER is a deep generative model that predicts reaction mechanisms by treating reaction prediction as electron redistribution. It uses a bond-electron matrix to enforce charge conservation and shows this approach works for organic reaction mechanism prediction^[14]. However, it covers only electron conservation in organic reactions, and the constraint system required for crystal synthesis is far more complex. Furthermore, FlowER has not yet been incorporated as a reasoning component in a closed-loop system that iteratively updates mechanistic hypotheses from experimental outcomes.

The third is cross-experiment knowledge accumulation. The target system requires the capacity for sustained inquiry: not merely completing individual tasks, but growing through accumulation. Current systems do the opposite. Each task starts with fresh retrieval of literature and protocol libraries, and each

experimental round remains isolated. This applies to ChemAgents^[4] and Bayesian optimization closed loops^[2], which both lack structured records of failed approaches.

A concrete example is the reported NiFe layered double hydroxide (LDH) system modified by CeO₂ and intercalated with (W₂O₇)²⁻ for large current seawater electrooxidation. After synthesis, it shows lower-angle XRD shifts. In an execution loop, this signal is usually recorded as phase characterization and linked to improved activity or stability, but the peak shift is not reused for reasoning. In a reasoning loop, the same anomaly becomes the start of mechanistic inference: the agent encodes it as a structural signal and uses the relation between diffraction angle and interplanar spacing to infer that (W₂O₇)²⁻ enters the LDH interlayer and expands layer spacing. This forms a transferable mechanism chain: intercalated (W₂O₇)²⁻ expands the galleries, modulates the local ionic environment, improves OH⁻ accessibility, suppresses Cl⁻ interference, and enhances seawater electrooxidation. The physicochemical constraint agent then uses this mechanism for later design. Candidate intercalants are constrained by charge, size, and interlayer compatibility rather than treated as arbitrary additives, allowing extension from tungstate to molybdate or broader polyoxometalate anions. The validated hypothesis and descriptors are stored in a persistent knowledge base, so similar XRD or spectral signals in another LDH system can retrieve this conclusion to guide design. Thus, a reasoning loop turns an XRD anomaly into a structured mechanism, a generative constraint, and transferable knowledge, whereas an execution loop can only identify better formulations.

THREE GATES: ACTIONABLE EVALUATION CRITERIA

The three architectural conditions outlined above can be translated into three actionable diagnostic questions, derived directly from the capability requirements of the target system and its current gaps.

Gate 1: Mechanistic inference. The reasoning loop should autonomously identify plausible mechanistic hypotheses from characterization data and represent them through appropriate, transferable forms that can be passed to other modules.

Gate 2: Physicochemical constraints. The reasoning loop should autonomously select and enforce relevant physicochemical constraints throughout its reasoning process, such that the outputs are guaranteed to satisfy those constraints.

Gate 3: Knowledge accumulation. The reasoning loop should autonomously retain task-relevant knowledge and adaptively retrieve and apply it in novel cross-task contexts.

Gate 1 enables mechanistic transfer across systems, Gate 2 maintains physical consistency in open search spaces, and Gate 3 ensures that experiments accumulate knowledge rather than repeat prior labor. By architectural completeness, existing systems fall into three categories: SDLs with wet-lab closed loops, AI-agent platforms without physical experimentation, and standalone predictive models. Representative systems from each category are evaluated against the three gates [Table 1].

None of the systems address more than one reasoning gap at once. For Gate 1, SDL systems support data acquisition and multimodal fusion but do not generate mechanistic hypotheses. Agent platforms mainly retrieve literature. EMSeek is closest because it infers plausible crystal structures from electron microscopy images. Most standalone prediction models lack mechanistic inference, except FlowER, which captures textbook reactivity trends through electron flow. For Gate 2, SDL systems provide partial constraint enforcement through preset model libraries or knowledge-embedded latent search, but neither offers hard physicochemical guarantees. Agent platforms apply LLM-embedded chemical knowledge after proposal

Table 1. Evaluation of representative AI systems against the three gates

Type	System	Description	Gate 1	Gate 2	Gate 3
SDL	ChemAgents ^[4]	Hierarchical multi-agent system coordinating literature, design, computation, and robotics	Fail×: Experimental designs drawn from literature abstracts; no hypothesis generated from observations	Partly Passed Δ: Selects from 130 preset ML models; DFT invoked passively and overridable	Partly Passed Δ: Historical protocol templates available; no automated write-back loop
	RoboChem-Flex ^[15]	Low-cost modular SDL combining customizable hardware with multi-objective Bayesian optimization	Partly PassedΔ: Data acquisition and analysis module present; no mechanistic hypothesis generated	Fail×: Search space manually selected from presets	Fail×: No cross-task knowledge accumulation
	CRESt ^[16]	Multimodal robotic platform integrating large multimodal models, Bayesian optimization, and robotics	Partly PassedΔ: Fuses SEM, composition, and literature into unified vectors; VLM generates anomaly-correction hypotheses but not mechanistic understanding	Partly Passed Δ: Knowledge-embedded latent-space search reduces candidate space; no hard physicochemical guarantee	Fail×: No cross-task knowledge retention
	DigHyd ^[9]	RAG-driven LLM agent with iterative ML feedback for hydrogen storage material discovery	Fail×: Retrieves candidate compositions from literature; no mechanistic hypothesis generated	Fail×: Relies on LLM built-in chemistry knowledge; unreliable and applied post-hoc	Fail×: No knowledge written back or retrieved across tasks
	EMSeek ^[17]	Modular multi-agent platform connecting electron microscopy to atomic-scale structural and property analysis	Nearly Passed √: Infers plausible crystal structures from EM images	Fail×: Physical consistency checks via LLM post-hoc; not enforced during proposal generation	Fail×: No cross-task knowledge retention
AI-agent	SciToolAgent ^[10]	LLM agent using a scientific tool knowledge graph for automated multi-tool integration	Fail×: No mechanistic representation or hypothesis generation	Partly Passed Δ: Proactively selects tool chains based on task context and knowledge graph	Partly Passed Δ: Stores task outputs and conversation history; no structured mechanistic accumulation
Standalone model	Cat-Advisor ^[18]	Multi-agent system combining LLM-curated databases and ML models for MgH ₂ catalyst design	Fail×: No mechanistic representation or hypothesis generation	Partly Passed Δ: Task-specific ML model predictions provide domain-constrained design guidance	Fail×: No cross-task knowledge retention
	MOFReasoner ^[12]	Domain-specific LLM fine-tuned via knowledge distillation and chain-of-thought reasoning	Fail×: Knowledge distilled from historical literature; cannot generate hypotheses from novel observations	Partly Passed Δ: Structured step-by-step reasoning applies chemical knowledge; selection and reliability not guaranteed	Fail×: No knowledge accumulation mechanism
	FlowER ^[14]	Deep generative model predicting reaction mechanisms via bond-electron matrix representation	Partly Passed Δ: Represents reactions through electron flow; captures textbook reactivity trends	Nearly Passed √: Charge conservation encoded as a hard architectural constraint; limited to electron balance, not generalizable	Fail×: No knowledge accumulation mechanism
	Chemma ^[19]	LLM fine-tuned with 1.28 million reaction Q&A pairs for organic synthesis route prediction and yield prediction	Fail×: No mechanistic understanding or explanation; yield prediction and retrosynthesis only	Fail×: No physicochemical constraint enforcement	Fail×: No knowledge accumulation mechanism

SDL: Self-driving laboratory; AI: artificial intelligence; SEM: scanning electron microscopy; VLM: vision-language model; DFT: density functional theory; ML: machine learning; LLM: large language model.

generation, making constraints unreliable and non-binding. FlowER is the strongest Gate 2 case because it encodes charge conservation as a hard architectural constraint, but it is limited to electron balance and is not integrated into a closed-loop workflow. Gate 3 is nearly absent. SDL systems treat outcomes as numerical fitness signals. Agent platforms store logs and outputs rather than structured mechanistic knowledge. Standalone models have no accumulation mechanism.

CONCLUSIONS AND OUTLOOK

AI-driven materials research has largely advanced the execution loop, making synthesis faster and more reproducible, while the reasoning loop needed for truly intelligent autonomy still lacks the required architecture. Without clear criteria separating these two forms of progress, execution successes are easily mistaken for movement toward mechanistic discovery. The starting point for the next phase is therefore clear: the field must recognize that structured mechanism representation, actively enforced physicochemical constraints, and cross-experiment knowledge accumulation remain unmet prerequisites. Only then can autonomous systems move from running experiments to advancing science.

DECLARATIONS

Authors' contributions

Contributed to the writing and preparation of the manuscript: Tan, Z.; Ran, N.; Qian, L.; Wang, Y.

Availability of data and materials

Not applicable.

AI and AI-assisted tools statement

Not applicable.

Financial support and sponsorship

This work is supported by Advanced Materials-National Science and Technology Major Project (2025ZD0619500) and The National Social Science Fund of China (24BTQ043).

Conflicts of interest

All authors declared that there are no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Copyright

© The Author(s) 2026.

REFERENCES

1. Tom, G.; Schmid, S. P.; Baird, S. G.; et al. Self-driving laboratories for chemistry and materials science. *Chem. Rev.* **2024**, *124*, 9633-732. DOI PubMed PMC
2. Slattery, A.; Wen, Z.; Tenblad, P.; et al. Automated self-optimization, intensification, and scale-up of photocatalysis in flow. *Science* **2024**, *383*, eadj1817. DOI PubMed
3. Huang, L.; Zhang, C.; Fu, Y.; et al. Natural-language-interfaced robotic synthesis for AI-copilot-assisted exploration of inorganic materials. *J. Am. Chem. Soc.* **2025**, *147*, 23014-25. DOI PubMed
4. Song, T.; Luo, M.; Zhang, X.; et al. a multiagent-driven robotic AI chemist enabling autonomous chemical research on demand. *J. Am. Chem. Soc.* **2025**, *147*, 12534-45. DOI PubMed
5. Li, X.; Wang, S.; Lin, Y.; Wu, Y. Retro-expert: collaborative reasoning for interpretable retrosynthesis. *arXiv* **2025**, arXiv:2508.10967. Available online: <https://doi.org/10.48550/arXiv.2508.10967> (accessed 8 July 2026).
6. Zhang, S.; Li, H.; Chen, L.; et al. Reasoning-driven retrosynthesis prediction with large language models via reinforcement learning. *arXiv* **2025**, arXiv:2507.17448. Available online: <https://doi.org/10.48550/arXiv.2507.17448> (accessed 8 July 2026).
7. Zhao, G.; Lu, Z.; Ge, Y.; et al. MolReasoner: toward effective and interpretable reasoning for molecular LLMs. *arXiv* **2025**, arXiv:2508.02066. Available online: <https://doi.org/10.48550/arXiv.2508.02066> (accessed 8 July 2026).
8. Zhao, Z.; Huang, Z.; Li, J.; et al. SUPERChem: a multimodal reasoning benchmark in chemistry. *arXiv* **2025**, arXiv:2512.01274. Available online: <https://doi.org/10.48550/arXiv.2512.01274> (accessed 8 July 2026).

9. Zhang, D.; Jia, X.; Tran, H. B.; et al. "DIVE" into hydrogen storage materials discovery with AI agents. *Chem. Sci.* **2026**, *17*, 3031-42. DOI PubMed PMC
10. Ding, K.; Yu, J.; Huang, J.; Yang, Y.; Zhang, Q.; Chen, H. SciToolAgent: a knowledge-graph-driven scientific agent for multitool integration. *Nat. Comput. Sci.* **2025**, *5*, 962-72. DOI PubMed
11. Noh, H.; Na, G. S.; Lee, N.; Park, C. MSP-LLM: a unified large language model framework for complete material synthesis planning. *arXiv* **2026**, arXiv:2602.07543. Available online: <https://doi.org/10.48550/arXiv.2602.07543> (accessed 8 July 2026).
12. Bai, X.; Zheng, Z.; Zhang, X.; Wang, H.; Yang, R.; Li, J. MOFReasoner: think like a scientist - a reasoning large language model via knowledge distillation. *Digit. Discov.* **2026**, *5*, 869-77. DOI
13. Vangala, B. P.; Mahmud, S.; Neupane, P.; Selvaraj, J.; Cheng, J. HalluMat: detecting hallucinations in LLM-generated materials science content through multi-stage verification. *arXiv* **2025**, arXiv:2512.22396. Available online: <https://doi.org/10.48550/arXiv.2512.22396> (accessed 8 July 2026).
14. Joung, J. F.; Fong, M. H.; Casetti, N.; Liles, J. P.; Dassanayake, N. S.; Coley, C. W. Electron flow matching for generative reaction mechanism prediction. *Nature* **2025**, *645*, 115-23. DOI PubMed
15. Pilon, S.; Savino, E.; Bayley, O. M.; et al. A flexible and affordable self-driving laboratory for automated reaction optimization. *Nat. Synth.* **2026**. DOI
16. Zhang, Z.; Ren, Z.; Hsu, C. W.; et al. A multimodal robotic platform for multi-element electrocatalyst discovery. *Nature* **2025**, *647*, 390-6. DOI PubMed
17. Chen, G.; Yuan, W.; You, F. Bridging electron microscopy and materials analysis with an autonomous agentic platform. *Sci. Adv.* **2026**, *12*, ead0583. DOI PubMed PMC
18. Yao, T.; Yang, Y.; Cai, J.; et al. From LLM to agent: a large-language-model-driven machine learning framework for catalyst design of MgH₂ dehydrogenation. *J. Magnes. Alloys.* **2026**, *14*, 101858. DOI
19. Zhang, Y.; Han, Y.; Chen, S.; et al. Large language models to accelerate organic chemistry synthesis. *Nat. Mach. Intell.* **2025**, *7*, 1010-22. DOI

Disclaimer/Publisher's Note: All statements, opinions, and data contained in this publication are solely those of the individual author(s) and contributor(s) and do not necessarily reflect those of OAE and/or the editor(s). OAE and/or the editor(s) disclaim any responsibility for harm to persons or property resulting from the use of any ideas, methods, instructions, or products mentioned in the content.



© The Author(s) 2026. Open Access This article is licensed under a Creative Commons Attribution 4.0 International License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, sharing, adaptation, distribution and reproduction in any medium or format, for any purpose, even commercially, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.