Review | Open Access

Journal of Materials Informatics

Zhu *et al. J. Mater. Inf.* **2026**, *6*, 32 **DOI:**10.20517/jmi.2026.07

# A survey of agentic materials science and engineering: where are we and where are we going?

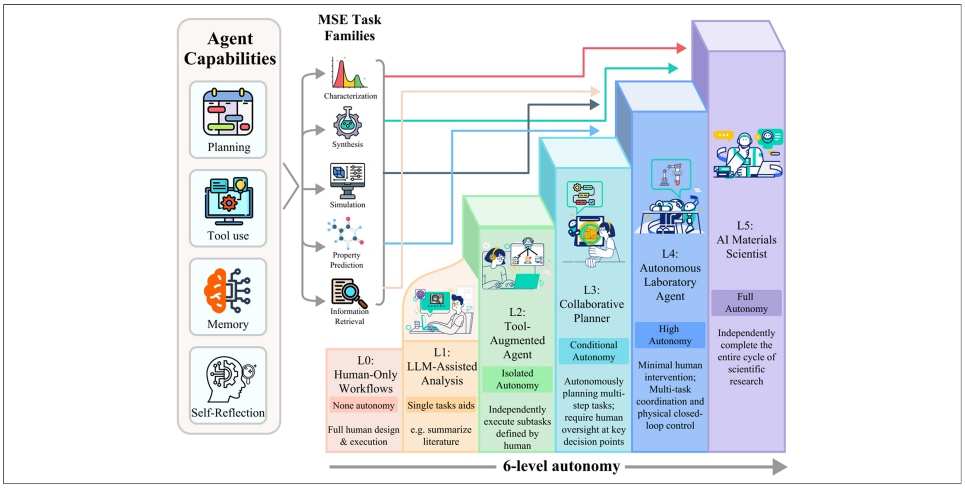
**Jiayi Zhu[1,#], Longhan Zhang[2,#], Yizhang Zhu[1], Xiaotian Lin[1], Yifan Wu[1], Shimin Di[3], Bang Liu[4], Yuyu Luo[1,*], Tongyi Zhang[2,*]**

**Keywords:**
Materials science and engineering, large language models, LLM-based agents, agentic materials science and engineering

**Citation:** Zhu, J.; Zhang, L.; Zhu, Y.; Lin, X.; Wu, Y.; Di, S.; Liu, B.; Luo, Y.; Zhang, T. A survey of agentic materials science and engineering: where are we and where are we going?. *J. Mater. Inf.* **2026**, *6*, 32. https://dx.doi.org/10.20517/jmi.2026.07

**Received:** 2 Mar 2026
**First Decision:** 24 Mar 2026
**Revised:** 7 Apr 2026
**Accepted:** 30 Apr 2026
**Published:** 26 May 2026

**Academic Editor:**
Xingjun Liu
**Copy Editor:**
Pei-Yun Wang
**Production Editor:**
Pei-Yun Wang




## Abstract

Agents, primarily built upon large language models (LLMs) and equipped with planning, tool use, memory, and self-reflection capabilities, are revolutionizing all aspects of materials science and engineering (MSE), from materials design and experimental execution to industrial manufacturing and deployment, thereby opening the age of agentic MSE. Rather than functioning as isolated artificial intelligence (AI) predictive models, these agents coordinate multi-step scientific workflows by retrieving and structuring knowledge, proposing and refining hypotheses, planning experiments, combining multimodal simulations and characterizations, and, when integrated with AI materials laboratories, closing the loop toward autonomous materials discovery. However, agentic systems exhibit varying degrees of autonomy, and their roles in materials research and development differ accordingly. To systematically examine the landscape of agentic MSE, this survey proposes a six-level autonomy framework (Levels 0-5) that characterizes the progression from human-only workflows to fully autonomous scientific agents. The framework aligns with

[1]Data Science and Analytics Thrust, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou 511453, Guangdong, China.
[2]Advanced Materials Thrust, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou 511453, Guangdong, China.
[3]School of Computer Science and Engineering, Southeast University, Nanjing 211189, Jiangsu, China.
[4]Department of Computer Science and Operations Research, University of Montreal, Montreal H3C 3J7, Canada.
[#]Authors contributed equally.

***Correspondence to:** Prof. Tongyi Zhang, Advanced Materials Thrust, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou 511453, Guangdong, China. E-mail: mezhangt@hkust-gz.edu.cn; Prof. Yuyu Luo, Data Science and Analytics Thrust, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou 511453, Guangdong, China. E-mail: yuyuluo@hkust-gz.edu.cn

key task families across the entire MSE pipeline, including information retrieval, property prediction, simulation, synthesis, and characterization. By reviewing recent advances in agentic MSE, we reveal uneven progress across these domains. Knowledge-centric capabilities often remain at an early stage, while experimental orchestration and characterization are starting to explore higher-level agent behaviors. Importantly, achieving mature autonomy requires coordinating multiple tasks rather than optimizing any single task in isolation. Collectively, these insights provide a structured roadmap for advancing agentic MSE toward higher levels of autonomy.

## INTRODUCTION

Materials science and engineering (MSE) is an important discipline at the intersection of physics, chemistry, and engineering, aiming to understand the complex relationships among a material's composition, structure, processing conditions, and resulting functional properties. In recent years, a wide range of data-driven and deep learning techniques have been explored across diverse materials contexts, focusing on specific research tasks that span the entire discovery pipeline. These efforts include natural language processing (NLP) for literature-based knowledge extraction, graph neural networks (GNNs)[1-6] for material representation learning and property prediction, generative models for inverse materials design[7,8], and optimization algorithms for process path optimization. Meanwhile, key research tools in MSE are undergoing an AI-driven paradigm shift toward higher-throughput data generation, encompassing atomic- and molecular-level computations[9-11], meso- and macro-scale simulations, autonomous characterization analysis, and autonomous high-throughput experimental platforms. Due to the diversity of material classes, such as metals, ceramics, polymers, semiconductors, and composites, as well as the inherently multimodal and multiscale nature of materials data, artificial intelligence (AI) for Materials Science and Engineering (AI4Mat) has emerged in a wide range of forms. Collectively, these developments converge toward a unified vision: the realization of an autonomous AI scientist deeply integrated into every stage of the materials research pipeline, enabling more efficient, accurate, and intelligent scientific discovery and materials innovation.

### Agentic MSE

The advent of large language models (LLMs) and LLM-based agents is the catalyst accelerating this vision. Unlike static predictive models, these agents are endowed with capabilities such as planning, memory, tool use, and self-reflection[12-14]. They can coordinate multi-step scientific workflows, including retrieving and structuring domain knowledge from literature, proposing and refining hypotheses, planning and parameterizing experiments, and invoking simulation or cheminformatics tools. When integrated with robotic platforms, agents can form a closed loop in material research by executing experiments in the physical world[15,16]. Early initiatives such as Coscientist[17] have demonstrated the autonomous design and execution of complex chemical tasks in both cloud-based and physical laboratory environments. Self-driving laboratories (SDLs), exemplified by A-Lab[18], are advancing toward higher levels of autonomy by employing active learning approaches to sustain long-term automated synthesis and discovery cycles. Collectively, these innovations signal a paradigm shift from traditional model-centric methodologies to comprehensive agentic systems that integrate data resources, computational tools, and experimental hardware within a cohesive framework for autonomous materials research[19].

We refer to agentic MSE as an emerging research paradigm in which LLM-based agents actively participate in the materials research and development cycle. In this paradigm, agents do not merely predict properties or extract information; instead, they exhibit the ability to perceive the environment, plan multi-step actions, invoke external computational or experimental tools, remember and refine prior outcomes, and autonomously pursue scientific objectives under human oversight. This agentic perspective transforms materials informatics from a data-analysis discipline into an integrated system of reasoning, experimentation,

and self-improvement. Consequently, agentic MSE, encompassing the design, evaluation, and governance of autonomous or semi-autonomous agents, presents considerable potential to accelerate discovery, ensure reproducibility, and facilitate collaboration with human scientists across all stages of materials research.

These trends motivate a fundamental shift from isolated, task-specific modeling to a workflow-oriented systems perspective for materials discovery and development. In this emerging paradigm, data resources, computational tools, experimental platforms, and control policies are no longer disparate elements but components integrated through unified agentic orchestration.

## A hierarchical framework for agentic MSE

However, transitioning to such integrated systems reveals a significant challenge, as the progression toward autonomy is highly uneven across the diverse landscape of MSE. This domain comprises distinct task families, ranging from purely informational tasks such as literature retrieval to physically demanding tasks including experimental synthesis. Each family presents unique barriers related to reasoning complexity, tool integration, and safety constraints. This leads to a landscape where AI capabilities vary drastically, extending from simple assistance in one domain to fully autonomous control in another.

To rigorously evaluate this heterogeneous progress, a simple catalog of individual models or a binary classification of automated *vs.* manual is insufficient. Such approaches fail to capture the nuance between a passive predictive model and an active reasoning agent. Therefore, we advocate examining the field through a hierarchical taxonomy that maps varying degrees of agent autonomy onto specific materials science tasks. Such a framework can provide a standardized metric for benchmarking progress. This methodology identifies not only where high autonomy has been achieved but also where critical gaps in reasoning and integration remain.

Therefore, to capture the progressive evolution of agentic MSE, we adopt the six-level hierarchy shown in Figure 1. Similar to the Society of Automotive Engineers (SAE) levels[20] of driving automation, this framework describes a progression from full human control to increasingly autonomous system behavior. Each level specifies a characteristic combination of agent capabilities, human roles, and agent responsibilities, together tracing the transition from human-only execution to fully autonomous scientific discovery.

• Level 0 - Human-Only. This is the baseline stage of traditional research, in which humans act as the sole executors, manually performing literature reviews, hypothesis formation, and experimentation. At this stage, agents are not yet involved.

• Level 1 - LLM-Assisted Analysis. Agents begin to play a purely supportive role in scientific workflows, essentially acting as intelligent research assistants or "copilots" for human scientists. At this stage, agents help retrieve information, summarize literature, and make simple predictions, but they do not take initiative or autonomously execute complex tasks. These systems excel at parsing scientific text and extracting structured knowledge[21]. However, their contributions remain advisory: they cannot yet plan multi-step experiments or make independent decisions, and any insights they provide still require human verification[22].

• Level 2 - Tool-Augmented Agent. Agents move beyond passive assistance and begin interacting with external tools to accomplish scientific tasks. At this stage, human researchers still define the overall goals, but agents can independently execute subtasks, such as retrieving data, running simulations, or invoking domain-specific libraries, without requiring step-by-step instructions. This tool-augmented paradigm enables agents to ground their reasoning in trusted computational resources, improving both reliability and scope. While humans remain responsible for high-level validation, agents can propose plausible synthesis

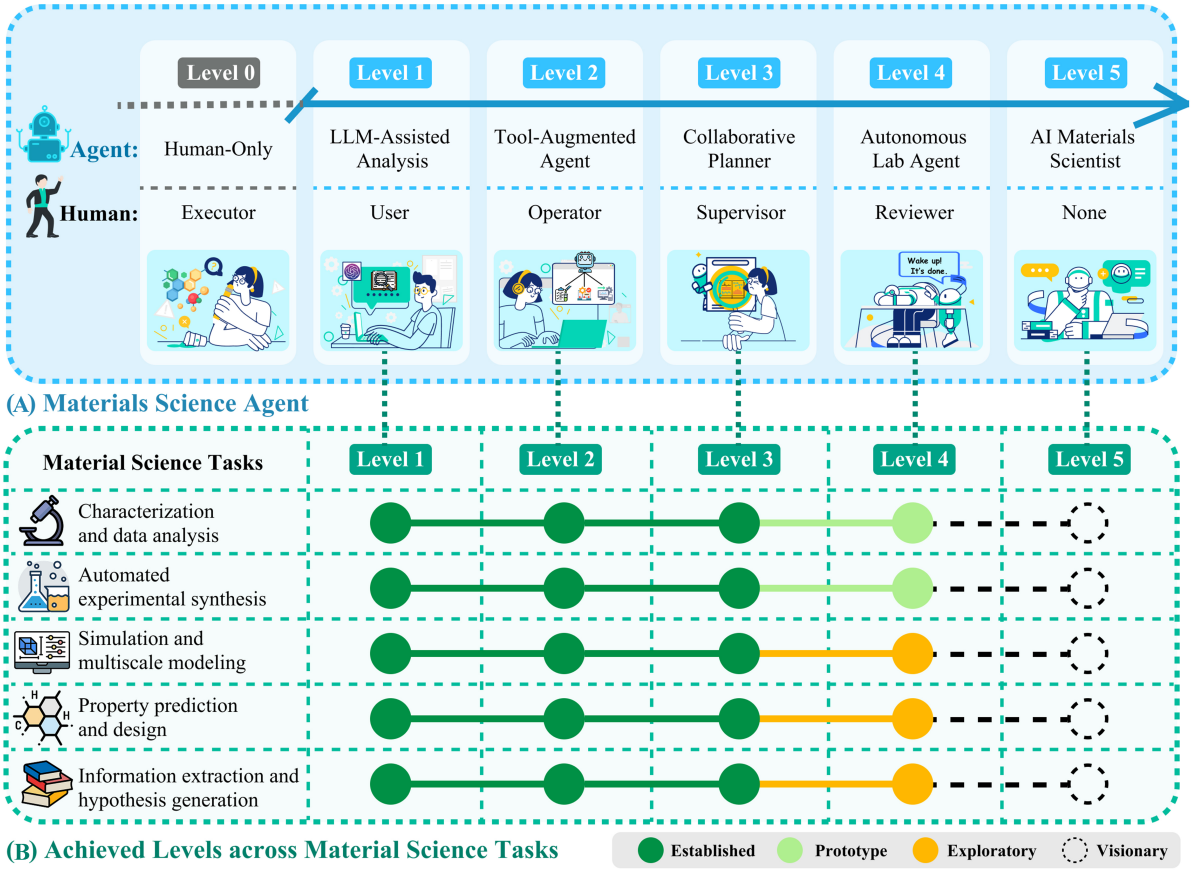



**Figure 1.** Overview of the six-level hierarchy for agentic MSE and its task mapping. (A) Progression from human-only workflows (Level 0) to fully autonomous AI materials scientists (Level 5); (B) Achieved autonomy levels across five core materials science tasks. Established: multiple independently published systems have demonstrated the capability with reproducible methods; Prototype: at least one published system has demonstrated the capability, but replication and generalization remain limited; Exploratory: the capability has been partially demonstrated in isolated or constrained settings, without full end-to-end validation; Visionary: no existing system has yet demonstrated the capability, representing a long-term research objective. MSE: Materials science and engineering; AI: artificial intelligence; LLM: large language model.

routes and predict material properties by drawing on databases and simulators.

• Level 3 - Collaborative Planner. Agents act as "conditionally automated" scientific assistants capable of autonomously planning multi-step tasks, although they still require human oversight at key decision points. Researchers provide high-level goals, and the agents use chain-of-thought reasoning[23], long-term memory[24], and tool invocation to autonomously decompose tasks and execute a series of actions.

• Level 4 - Autonomous Laboratory Agent. Agents at this level are not only capable of autonomous planning but can also operate for extended periods while interacting with real experimental environments. With minimal human intervention, they can complete the closed loop from experimental design to execution and data collection. The agents can continuously maintain working memory, adjust experimental plans based on intermediate results, and directly control laboratory instruments or invoke remote experimental platforms (e.g., A-Lab[18]; lab orchestration software such as ChemOS 2.0[25]). Humans primarily act as high-level supervisors, intervening only at milestone checkpoints or when anomalies occur.

• Level 5 - AI Materials Scientist. This stage represents the ultimate vision of a "fully autonomous" AI materials scientist. The agent can independently complete the entire scientific research cycle with virtually no human involvement, from formulating original hypotheses to planning research pathways, conducting physical experiments, and summarizing discoveries. Human input is limited to broad thematic directions, and research topics may even arise from the agent's intrinsic "curiosity".

Progress from Level 1 to Level 5 can be characterized by transformative transitions that mark distinct expansions in capability and responsibility:
• Level 1 → Level 2: Tool-Augmented Grounding. Agents advance from text-only analysis to grounded tool use, anchoring reasoning in materials databases, calculators, and simulators.
• Level 2 → Level 3: Multi-Step Planning with Memory. Agents adopt persistent contextual memory and decompose complex goals into executable plans to advance from Level 2 to Level 3 autonomy, often through multi-agent planner-executor or generator-critic structures.
• Level 3 → Level 4: Multi-Task Coordination and Physical Closed-Loop Control. Integration with robotics and instrumentation enables continuous operation across synthesis and characterization, pushing agents to Level 4 autonomy in materials research.
• Level 4 → Level 5: Self-Reflection and Hypothesis-Driven Science. The envisioned Level 5 "AI materials scientist" autonomously formulates testable hypotheses and produces verifiable reasoning chains under audit and governance[26-28].

As summarized in Figure 1B, development is uneven across tasks: All the five tasks have achieved up to Level 3 autonomy. Synthesis and characterization have reached the Level 4 prototype stage, as pioneering systems such as Coscientist[17] and AdaptiveXRD[29] have demonstrated closed-loop operation with real physical instrumentation and robotic hardware. In contrast, knowledge-centric tasks (information retrieval and property prediction) in systems such as AccelMat[30] and MARS[31] and simulation remain exploratory at Level 4, as current systems operate exclusively in the digital domain without verified physical execution. Figure 1 also serves as the organizational backbone of this survey: Section "THE HIERARCHY OF AUTONOMY IN MATERIALS SCIENCE AGENTS" follows this two-dimensional (2D) task–level matrix and discusses each task family vertically across autonomy levels.

Prior surveys emphasize chemistry-centric model catalogs and case studies of LLMs and agents[22,32-34]. Recent surveys provide complementary but different perspectives. The AI4MS survey[35] mainly offers an inventory-style overview of foundation models for materials science, featuring a task-driven taxonomy across six application areas and a broad summary of unimodal models, multimodal models, LLM agents, datasets, and tools. In parallel, Li *et al.* review the rise of AI agents in materials research, highlighting advances in knowledge processing, structure design, and property calculation, and discussing how tool use and experimental automation may support SDLs and, eventually, end-to-end autonomous materials creation[36]. In contrast, our survey focuses on operationalizing autonomy[20] in a materials-grounded way: we introduce a six-level autonomy framework and a 2D task–level map spanning the materials research workflow. We further specify per-level capability requirements and toolchains, which serve as practical design targets for building materials agents toward higher autonomy, with Level 5 as the long-term objective.

**Our Contributions.** We make the following contributions:

• A materials-science-grounded six-level autonomy framework that characterizes the progression from human-only workflows to highly autonomous scientific agents. By defining explicit capability criteria, the framework clarifies the evolving division of labor between human scientists and AI, laying the foundation for a future research paradigm defined by seamless human-agent collaboration.

• A structured background that establishes the research foundation for agentic materials science, comprising three components: a comparative analysis of traditional human-centered workflows and emerging agentic paradigms across the five core task families; a review of domain-specific foundation models and their development paradigms; and a synthesis of open-source agentic infrastructure. Together, these components define the current boundaries of what can be agentized in materials research.

• A task–level matrix aligning autonomy levels with core materials tasks, revealing uneven development across literature understanding, prediction and design, simulation, synthesis, and characterization, and identifying research directions.

• An analysis of key open challenges in agentic materials science, distinguishing cognition-centric limitations in digital reasoning tasks from execution-centric limitations in physical experimental workflows. Based on this analysis, we propose targeted research directions, including physically grounded reasoning, active perception for closed-loop experimentation, dynamic benchmarking, and safety and governance frameworks, providing a practical roadmap toward higher levels of autonomy.

**Paper Organization.** The remainder of this survey is organized as follows. Section "BACKGROUND: TASK OVERVIEW AND RESEARCH FOUNDATIONS" lays the research foundation for agentic MSE from three angles: a comparison between traditional human-centered and emerging agentic workflows across the five core task families; a review of domain-specific foundation models and their development paradigms; and a synthesis of the open-source agentic infrastructure that collectively defines the current boundaries of what can be agentized. Section "THE HIERARCHY OF AUTONOMY IN MATERIALS SCIENCE AGENTS" constitutes the analytical core of the survey. Guided by the six-level autonomy framework, it examines each of the five task families vertically across autonomy levels, revealing both the maturity and the remaining gaps in each domain. Cross-task agents that integrate multiple task families are discussed at the end of this section. Section "FUTURE WORK" identifies current challenges and proposes targeted research directions.

## BACKGROUND: TASK OVERVIEW AND RESEARCH FOUNDATIONS

### Key task families across MSE research

Before examining agentic systems at specific autonomy levels, we first establish the research foundations upon which they are built. This section introduces the five core task families that define the scope of MSE research, traces the workflow transformation from human-centered to agentic paradigms, and reviews the domain-specific foundation models and open-source infrastructure that collectively enable agentic behavior.

To establish consistent terminology for the subsequent analysis, we formalize five fundamental tasks that collectively represent the core of materials research. Each task occupies a distinct position in the data–model–experiment cycle and serves as a target for progressive agentic autonomy.

• **Information Extraction (IE) and Hypothesis Generation** focuses on extracting and structuring scientific knowledge from literature, patents, and databases. It converts unstructured textual, tabular, and graphical content into structured representations such as entities, relations, and process–property mappings. Based on the organized knowledge, agents generate scientifically grounded and testable hypotheses that guide downstream modeling and experimentation.

• **Property Prediction and Design** focuses on learning predictive relationships among composition, structure, processing conditions, and resulting material properties. It includes forward modeling for property estimation from known descriptors and inverse design for discovering new materials that meet specified performance objectives while ensuring thermodynamic stability and synthetic feasibility.

• **Simulation and Multiscale Modeling** integrates computational methods that operate across quantum, atomic, mesoscopic, and continuum scales. Its goal is to reproduce the physical, chemical, and mechanical behaviors of materials, connect phenomena across scales, and provide theoretical insights that complement and validate experimental results.

• **Automated Experimental Synthesis** addresses the autonomous planning, execution, and optimization of synthesis workflows using robotic, microfluidic, or high-throughput experimental systems. Agents select synthesis routes, control equipment, monitor reactions in real time[17], and adaptively adjust parameters through feedback from analytical measurements to achieve desired material outcomes with reproducibility and safety.

• **Characterization and Data Analysis** involves the acquisition, preprocessing, and interpretation of experimental data obtained from characterization instruments such as X-ray diffraction (XRD), X-ray photoelectron spectroscopy (XPS), scanning electron microscopy/transmission electron microscopy (SEM/TEM), and spectroscopy. It includes automated noise removal, feature extraction, and quantitative identification of structural, compositional, and electronic characteristics. Advanced systems further employ active learning to optimize measurement strategies for maximal information gain.

## From human-centered materials research to agentic workflows

Building upon the definitions above and the proposed autonomy framework, Figure 2 illustrates a fundamental transformation in MSE workflows. This schematic contrasts the traditional human-centric approach [all Subfigures (a) of Figure 2], characteristic of Level 0 and Level 1 autonomy, with the emerging agentic MSE research paradigm [all Subfigures (b) of Figure 2].

In prior MSE workflows, research tasks are mostly linear and handled separately. As illustrated in the left panel [all Subfigures (a) of Figure 2], the human scientist serves as the sole central processor who manually defines objectives, designs experiments, executes protocols, and interprets data. Feedback loops, such as redesigning synthesis routes or refining hypotheses, relied entirely on human intuition and manual intervention[37,38]. This bottleneck restricts discovery throughput and often disconnects high-level reasoning from low-level execution. Similarly, in computational domains such as simulation and characterization, data processing software operates as passive utilities that require continuous manual calibration and file manipulation.

Central to the agentic paradigm is the shift from fragmented manual steps to integrated reasoning and planning. In this framework, AI agents serve as orchestration hubs that actively perceive contextual information and formulate multi-stage strategies. As demonstrated in all subfigures (b) of Figure 2, the agent acts as a dynamic hub that seamlessly integrates external resources, ranging from internet application programming interfaces (APIs) and computational simulators to physical laboratory instruments[17,39]. This integration transforms previously independent software tools and hardware systems into active modules under agentic control. Most importantly, this architecture establishes autonomous feedback loops. Whether performing inverse design for property prediction or optimizing synthesis parameters in real time, the system iteratively refines its actions. By analyzing output data to automatically trigger redesigns or next-step suggestions, the agentic workflow closes the loop between decision making and execution, significantly reducing the need for continuous human oversight.

## Foundation models for MSE agents

While autonomous agents coordinate workflows, make decisions, and interact with tools, these actions ultimately depend on the expressive power and inductive biases of underlying models. These models encode scientific knowledge, structure–property relationships, and implicit physical constraints that shape how an

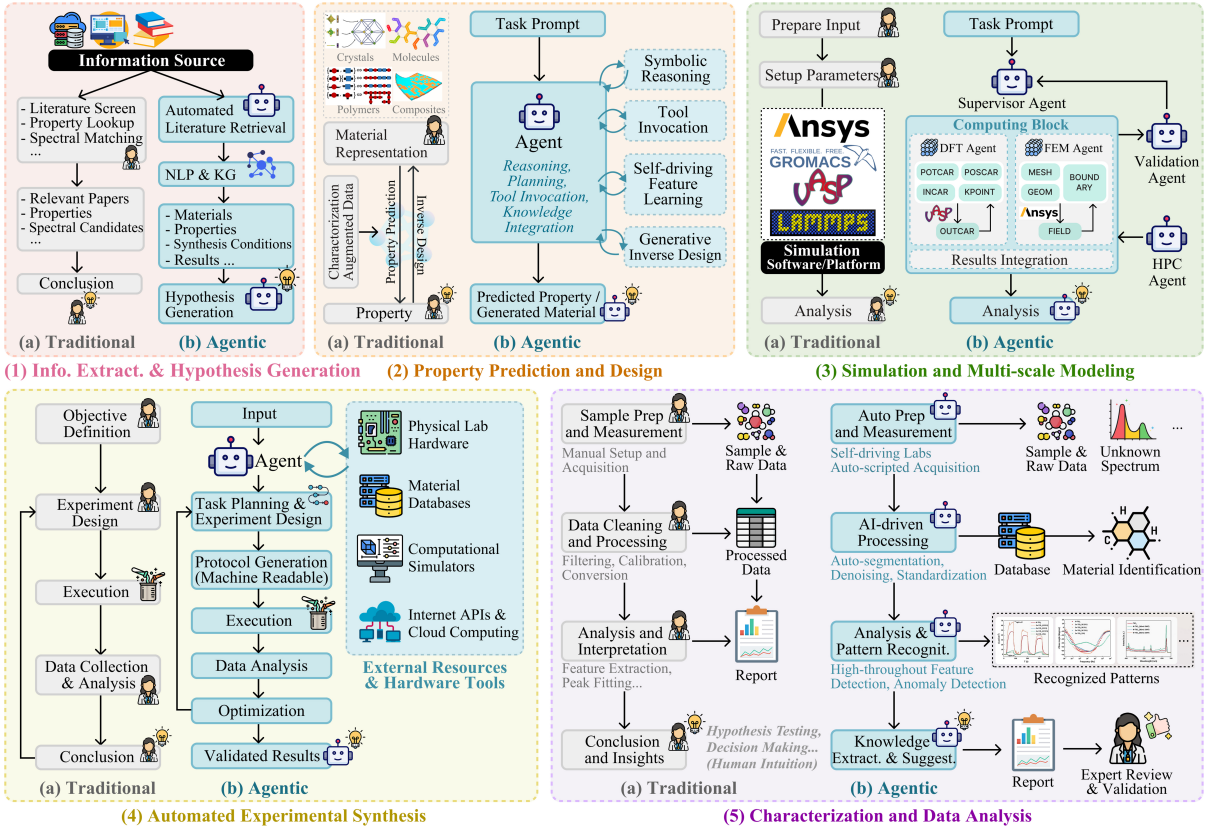



**Figure 2.** Comparative overview of materials research workflows: traditional methods *vs.* agentic approaches across five representative tasks: (1) IE and hypothesis generation, (2) property prediction and design, (3) simulation and multiscale modeling, (4) automated experimental synthesis, and (5) characterization and data analysis. In each panel, the left side shows a conventional human-driven workflow, while the right side illustrates an agentic alternative in which AI agents coordinate planning, knowledge integration, tool invocation, and iterative analysis. IE: Information extraction; AI: artificial intelligence; NLP: natural language processing; KG: knowledge graph; DFT: density functional theory; FEM: finite element method; HPC: high performance computing; APIs: application programming interfaces.

agent interprets information and takes actions. Before discussing agentic works within specific materials science tasks, we therefore review the foundation models upon which agentic systems are built. We first outline the distinct philosophies guiding the development of domain-specific LLMs for MSE, and then summarize representative models across key materials sub-domains.

## *Development paradigms of domain-specific LLMs*

Efforts to adapt LLMs to MSE can be grouped into several methodological families. Each family corresponds to a different view on how domain specialization should be integrated into general language models, and each involves different trade-offs in model capability, generalization, and computational cost.

### Continued pretraining

Current LLMs are pretrained on extensive and diverse text corpora. To further develop generalized scientific priors in a specific domain of interest, continued pretraining (CPT) is a key strategy. In this approach, established LLMs are further exposed to vast corpora of scientific texts, databases, and structured materials knowledge before being adapted to downstream tasks. Models such as MatSciBERT[40] and MatBERT[41] follow this path. Specifically, MatSciBERT is a materials-aware BERT model initialized from SciBERT and further pretrained on a large curated corpus (around 285M words) of MSE literature. Using RoBERTa-style[42] pretraining and domain-adaptive continuation, it achieves lower validation perplexity and

state-of-the-art performance across three downstream tasks. Compared to general-purpose pretrained LLMs, these studies[40,41] focus on scientific domains, particularly MSE. They typically require millions of domain-specific documents for models to learn statistical structures in materials knowledge, including composition-structure-property relationships, synthesis terminology, and common experimental or theoretical narratives. The learned prior knowledge is beneficial not only for NLP tasks but also for intelligent agent applications, where pretrained models serve as the core of decision making and are integrated into simulation or experimental workflows.

Supervised fine-tuning

Since CPT requires large-scale domain data, which may be unavailable in niche areas, supervised fine-tuning (SFT) provides an alternative strategy for domain adaptation. In this approach, curated labeled datasets targeting specific downstream tasks are constructed, such as property prediction[43], materials entity extraction[40], and reaction prediction[44]. These datasets are then used to fine-tune pretrained LLMs in a supervised manner. In the field of MSE, SFT is widely used to endow models with task-specific capabilities that cannot be effectively learned from unlabeled corpora alone. Task-centric SFT has been applied to models such as PolySea[45] and SteelBERT[46], enabling them to perform domain-specific tasks. However, these models risk lacking generalization capabilities beyond the scope of predefined tasks.

Reinforcement learning

Reinforcement learning (RL) has been prevailing in aligning LLMs with specific preferences or human feedback. It also provides a more dynamic paradigm for the decision making and iterative reasoning requirements of agentic materials workflows. In this framework, reward signals, including human preferences, rule-based evaluators, or self-consistency critics, will be constructed in the RL framework to evaluate LLM outputs and provide reward feedback for policy optimization. By rewarding coherent and actionable reasoning, RL enhances the ability of LLMs to meet the requirements of higher autonomy in agents operating in closed-loop computational or experimental environments.

### *Sub-domain specialized models*

Table 1 presents representative models created within various materials science subfields. These models demonstrate practical implementations of the described methodological paradigms and highlight methodological diversity across the discipline. They include tailored models that incorporate priors for various material classes, including molecules[47-49], polymers[45,50,51], alloys[46,52], perovskites[53], batteries[54], optical materials[55], and catalysts[56]. Unlike general-purpose LLMs designed to capture broad conceptual patterns, these models utilize domain-specific CPT or fine-tuning to capture chemical, structural, and processing features unique to each material category. Consequently, sub-domain LLMs provide more robust results compared to generalist models in downstream applications.

In agentic systems, these domain-specific priors make such models appropriate for incorporation into materials science workflows. Depending on their role, they may function as cognitive components for decision making, planning, and reasoning, or as execution tools such as property predictors, structure analyzers, or simulation surrogates. Accordingly, we categorize the models in Table 1 into "Core Models" and "Agentic Tools" to better reflect their functional roles within agentic systems.

## Agentic infrastructures for materials science

Beyond domain-specific models, recent progress in agentic MSE also depends on a broader infrastructure layer that supports information access, state tracking, planning, execution, and evaluation. Most of these open-source resources are not agentic in themselves. Instead, as shown in Table 2, they provide callable components that LLM-based agents can invoke and coordinate to build scientific workflows. Importantly,

**Table 1. Representative sub-domain specialized LLMs in MSE**

| Model | Sub-domain | Approach | Data scale | Modalities | Agentic usage |
|---|---|---|---|---|---|
| MatSciBERT[40] | General | CPT | 3.45B words | Text | Core models |
| MaterialBERT[6] | General | CPT | 8.4M entries | Text | Core models |
| MatterChat[57] | General | Pretrain + SFT | 142K samples | Text + molecular | Agentic tools |
| LLaMat-Chat[33] | General | CPT + SFT | 30B tokens | Text | Core models |
| LLaMat-CIF[33] | Crystal | CPT + SFT | 7M structures | Text + CIF | Agentic tools |
| CrystalLLM[58] | Crystal | Pretrain | 2.3M structures | Text + CIF | Agentic tools |
| Mol-LLM[47] | Molecules | SFT + RL | 3.3M samples | Text + molecular | Agentic tools |
| BindGPT[48] | Molecules | Pretrain + SFT + RL | 200M samples | Text + molecular | Agentic tools |
| ChemMLLM[49] | Molecules | SFT | 400K entries | Text + molecular + image | Agentic tools |
| BatGPT-Chem[59] | Molecules | SFT | 112K entries | Text + molecular | Core models |
| PolySea[45] | Polymers | SFT | 230K samples | Text + molecular | Core models |
| PolyBERT[50] | Polymers | Pretrain + SFT | 100M samples | Molecular | Agentic tools |
| TransPolymer[51] | Polymers | Pretrain + SFT | 5M samples | Text + molecular | Agentic tools |
| SteelBERT[46] | Alloys | Pretrain + SFT | 0.96B words | Text + tables | Core models |
| AlloyBERT[52] | Alloys | SFT | 1K samples | Text | Agentic tools |
| Perovskite-LLM[53] | Perovskites | SFT | 4.4M tokens | Text | Core models |
| BatteryBERT[54] | Battery materials | Pretrain + CPT + SFT | 3.3B tokens | Text | Core models |
| OpticalBERT[55] | Optical materials | Pretrain + CPT + SFT | 2.92B tokens | Text + tables | Core models |
| CatGPT[56] | Catalysts | Pretrain + SFT | 2M structures | Text-encoded structures | Agentic tools |

LLMs: Large language models; MSE: materials science and engineering; CPT: continued pretraining; SFT: supervised fine-tuning; RL: reinforcement learning; Core models: models that may participate in reasoning and decision making within agentic systems; Agentic tools: models that may be invoked to execute specific tasks.

these infrastructures do more than support engineering integration; they also instantiate the core mechanisms that make agentic scientific workflows possible, including retrieval and grounding, memory persistence, task decomposition, tool routing, execution control, and feedback-driven correction. In this sense, they externalize recurring steps in materials research into reusable computational modules and define much of what can be agentized with current infrastructure. Table 2 summarizes representative resources together with the scientific functions they support across knowledge, planning, execution, and evaluation.

### Knowledge and memory

Materials science research depends heavily on the scientific literature and diverse data resources including materials databases, simulation outputs, and experimental records. Researchers often need to review extensive papers, database entries, and experimental records to summarize existing findings and identify material systems and key parameters[115]. Agent-based systems require not only access to information, but also mechanisms for grounding their decisions in structured scientific evidence. In practice, this ability is enabled by a combination of retrieval-augmented generation (RAG), semantic similarity search, graph-based relation storage, and memory modules that preserve intermediate findings across workflow steps. For example, agents may retrieve structures, properties, phase stability data, or prior computational results from resources such as the Materials Project[60] and Materials Cloud[61], and then store relevant constraints or candidate information for later reasoning[62,63]. More broadly, vector databases[72-75], knowledge graphs (KGs)[64-66], ontology systems[69-71], and general-purpose data storage solutions such as MongoDB[67] and PostgreSQL[68] help connect scientific entities, conditions, and relationships across different sources. These mechanisms are important because materials workflows often require agents to accumulate evidence across multiple documents and data modalities rather than relying on a single query or static context. Other

**Table 2. Representative resources supporting core functional components of agentic MSE**

| Resource category | Representative tools or platforms | Scientific role in materials research |
|---|---|---|
| Knowledge/memory | Materials Project[60], Materials Cloud[61], NOMAD[62], AFLOW[63] | Knowledge retrieval & structured scientific data hub |
| | MatKG[64], MGED-KG[65], MeKG[66] | Materials KGs & Entity–relation storage |
| | MongoDB[67], PostgreSQL[68] | Data storage |
| | EMMO[69], ONTORULE[70], SLACKS[71] | Domain ontologies |
| | FAISS[72], Milvus[73], Qdrant[74], Weaviate[75] | Semantic storage & vector similarity search |
| | Neo4j[76], RDFLib[77], Letta (MemGPT)[78] | Graph-based knowledge storage |
| | LlamaIndex[79], Haystack[80], SerpAPI[81] | Retrieval-augmented document search |
| Decision/planning | LangChain[82], LangGraph[83], AutoGen[84], CrewAI[85] | Agentic orchestration & logic workflow construction |
| | ReAct[86], Reflexion[87] | Iterative reasoning and self-reflection |
| Execution/action | Fireworks[88], AiiDA[89], Simmate[90], Colmena[91], pymatgen[92] | Simulation workflow engines |
| | VASP[93], Quantum Espresso[94], ABINIT[95], GPAW[96] | Simulation engines & Python interfaces for DFT |
| | LAMMPS[97], GROMACS[98], OpenMM[99] | MD & interatomic models |
| | PyVISA[100], RoboRXN[101], PyLabRobot[102] | Automated experiment execution |
| | Pydantic-AI[103] | Workflow scheduling and automation |
| Evaluation/feedback | MatSciBench[104], MSQA[105], MatTools[106], ALDBench[107], MatBench Discovery[108], RxnBench[109], SDE[110], SFE[111] | Domain benchmarks |
| | LangSmith[112], OpenAI Evals[113], Ragas[114] | General evaluation frameworks and platforms |

MSE: Materials science and engineering; KGs: knowledge graphs; DFT: density functional theory; MD: molecular dynamics; AI: artificial intelligence.

commonly used tools[76-81] are summarized in Table 2.

### Decision and planning

In agent-driven materials research, a key challenge is to turn a broad scientific goal into a clear sequence of steps, and to keep the workflow consistent as new results appear. Agents must translate open-ended goals, such as identifying stable candidate materials or proposing synthesis conditions, into tractable substeps; maintain workflow state as new evidence appears; and revise plans when intermediate results are invalid, incomplete, or scientifically uninformative. These capabilities are commonly supported by mechanisms such as task decomposition, graph-structured workflow control, tool routing, reflection, and role-based multi-agent coordination. Frameworks such as LangChain[82] and LangGraph[83] support this process by providing building blocks to design multi-step agent workflows, connect external tools (such as databases, search engines, and code execution environments), and manage information flow between steps. They also help coordinate multiple roles or agents, enabling tasks such as literature search, data analysis, and result checking to be organized into a single pipeline. Beyond orchestration frameworks, the underlying reasoning algorithms are equally important. ReAct[86] enables agents to interleave reasoning and tool use across workflow steps, allowing them to iteratively refine hypotheses based on external evidence and execution feedback. Reflexion[87] further adds a self-reflection step, in which agents use feedback from earlier failures to improve subsequent decisions. This is particularly useful in multi-step materials workflows, where early errors in candidate selection, parameter setting, or intermediate result interpretation may affect later stages. Table 2 provides additional tools and frameworks[84,85].

### Execution and actions

After planning, agents also need reliable mechanisms to execute actions and obtain scientifically meaningful feedback from the environment. In materials workflows, execution is not limited to calling generic APIs; it often requires parameterizing simulations, launching structured workflows, managing intermediate artifacts, and coupling language-level reasoning with numerical or experimental engines. Workflow managers and

execution interfaces allow agents to organize computational jobs, pass structured inputs, monitor execution status, and collect outputs for downstream reasoning. They can also interface with common simulation codes[88-92], *ab initio* calculation packages[93-96], molecular dynamics (MD) tools[97-99], and automated experiment execution and workflow tools[100-103].

Evaluation and benchmarking

Existing evaluation practices for scientific LLMs and agents span a spectrum from static, capability-oriented benchmarks (e.g., domain QA[104-106], materials synthesis tasks[107]) to workflow-level assessments that test multi-step planning and tool use, and finally to realistic multimodal settings that require cross-document reasoning. Recently, several new benchmarks have further expanded this landscape, including scenario- and project-grounded evaluations for discovery workflows[110], hierarchical multimodal evaluations from localized perception to full-document synthesis[109,110], cognition-oriented multimodal evaluations that decompose scientific capability into perception, attribute understanding, and comparative reasoning across raw scientific data and multiple disciplines[111], and prospective, discovery-oriented evaluations for stability screening with task-relevant decision metrics[108]. These provide transferable principles for next-generation benchmarking in agentic MSE. General evaluation and observability frameworks are also increasingly used to assess LLM and agent systems beyond domain-specific benchmarks. LangSmith[112] provides tracing, observability, and experiment-level evaluation for LLM applications and AI agents. OpenAI Evals[113] offers a general framework for testing whether model outputs satisfy task-specific criteria, and is widely used for systematic evaluation and model comparison. Ragas[114] complements these platforms with metric-driven evaluation workflows, particularly for RAG and agentic applications, including reusable metrics and evaluation pipelines.

## THE HIERARCHY OF AUTONOMY IN MATERIALS SCIENCE AGENTS

As discussed in previous sections, we focus on five key tasks in MSE research: literature retrieval and hypothesis generation, property prediction and design, simulation and multiscale modeling, automated experimental synthesis, and characterization and data analysis. These tasks are selected to span the full spectrum of materials research, from abstract knowledge reasoning to concrete physical realization. Collectively, they capture how information flows through the research and development pipeline, where hypotheses are formed from literature, tested through simulations, verified in experiments, and interpreted through characterization and data analysis.

Each task offers a distinct perspective for examining the advancement of agentic MSE. Crucially, the progression of autonomy is not uniform across these domains. Within the same task, systems can operate at different levels of autonomy - from simple assistants that conceptually interact with humans (Level 1) to agents capable of multi-step planning or feedback-guided execution (Level 3-4). For example, in experimental synthesis, some systems still act as assistants that suggest procedures or parameters[116], while others already integrate planning with tool use and long-running execution, approaching higher autonomy through closed-loop operation[117]. Notably, as agents reach higher autonomy levels, the boundaries between tasks often become less clear. More advanced systems may span multiple tasks simultaneously and shift from a single-task view to a more system-level view. We discuss this in detail in the following sections.

This framework facilitates a dual analysis: horizontally across distinct tasks and vertically across levels of autonomy, revealing both the disparities and synergies in current progress. Tasks that are cognitively intensive yet computationally tractable - such as text mining or property prediction - have achieved greater maturity, whereas experiment-centric tasks continue to face bottlenecks regarding robotics integration and safety control. By intersecting these dimensions, we establish a task-level view of autonomy: for each domain,

**Table 3. Representative systems for IE and hypothesis generation**

| Methods | Year | Autonomy level | Multi-agent | Closed-loop | Equipment integration | Open source | Agentic system |
|---|---|---|---|---|---|---|---|
| ChemicalTagger[122] | 2011 | L1 | × | × | × | √ | × |
| ChemDataExtractor[123] | 2016 | L1 | × | × | × | △ | × |
| Mat2Vec[125] | 2019 | L1 | × | × | × | √ | × |
| Materials NER[126] | 2019 | L1 | × | × | △ | √ | × |
| Dunn *et al.*[130] | 2022 | L1 | × | × | × | × | × |
| Yan *et al.*[119] | 2022 | L1 | × | × | × | √ | × |
| Action_extractor[127] | 2023 | L1 | × | × | × | √ | × |
| SFBC[128] | 2023 | L1 | × | × | × | √ | × |
| MatKG[64] | 2024 | L1 | × | × | × | √ | × |
| MGED-KG[65] | 2024 | L1 | × | × | × | √ | × |
| MOF-KG[129] | 2024 | L1 | × | × | × | √ | × |
| Silva *et al.*[132] | 2024 | L1 | × | × | × | × | × |
| MaTableGPT[131] | 2025 | L1 | × | × | × | √ | × |
| HoneyBee[133] | 2023 | L2 | × | × | × | √ | × |
| HoneyComb[14] | 2024 | L2 | × | × | × | × | √ |
| Eunomia[12] | 2024 | L2 | × | △ | × | √ | √ |
| SciMON[134] | 2024 | L3 | √ | △ | × | √ | △ |
| Liu *et al.*[120] | 2025 | L3 | √ | √ | × | × | △ |
| SciAgents[121] | 2025 | L3 | √ | √ | △ | √ | √ |
| AccelMat[30] | 2025 | L4 | √ | √ | × | √ | √ |
| PriM[135] | 2025 | L4 | √ | √ | × | √ | √ |

Notes: √ = present; × = absent; △ = partial or simulated integration. IE: Information extraction; MatKG: Knowledge Graph of Materials Science; MGED-KG: Materials Genome Engineering Database Knowledge Graph; MOF-KG: Metal-Organic Framework Knowledge Graph.

we define its role, the current state of agentic systems, and the highest level of autonomy practically demonstrated to date. The following subsections discuss these tasks in detail.

## IE and hypothesis generation

MSE possesses a vast corpus of scientific literature, yet turning this unstructured text into structured datasets and actionable knowledge remains a significant challenge. IE tools are crucial for mining information on materials, properties, synthesis conditions, and performance metrics to build databases that accelerate materials design and understanding[118,119]. Recent advances in NLP, ranging from domain-tuned language models[40] to multi-agent systems (MAS)[120,121], are advancing IE beyond simple data gathering toward active hypothesis generation for new materials and experiments[30]. In this section, we examine the evolution of these capabilities, classifying systems by their level of autonomy in transforming raw text into novel scientific directions. Representative systems for this task are summarized in Table 3, including a brief comparison of their autonomy levels, multi-agent settings, closed-loop capabilities, equipment integration, and open-source availability. Similar summary tables are provided for each task in the following sections; therefore, this note is not repeated thereafter.

**Level 1.** Research at Level 1 focuses on automating specific, labor-intensive tasks within the scientific workflow, particularly IE and the construction of structured knowledge bases from unstructured literature. These systems act as intelligent assistants, parsing vast amounts of text to provide structured data that humans or downstream models can utilize. Early foundational efforts relied on rule-based systems and

statistical pipelines. Tools such as ChemicalTagger[122] employ grammar-based parsing to identify chemical action phrases, while ChemDataExtractor[123,124] combines part-of-speech tagging with rule-based logic to resolve interdependencies between text and tables for precise entity extraction. Subsequent approaches integrated deep learning to scale these capabilities. Mat2Vec[125] demonstrates that unsupervised word embeddings could capture latent chemical knowledge to predict future materials, a concept expanded by Materials NER[126] to mine inorganic materials from millions of abstracts. To address data scarcity in specialized domains such as superalloys, semi-supervised frameworks[119] including Action_extractor[127] have been developed, with SFBC[128] further refining accuracy by combining dynamic and static embeddings.

These extraction efforts evolve toward the construction of semantic KGs and the integration of LLMs. Systems such as MatKG (Knowledge Graph of Materials Science)[64] and MGED-KG (Materials Genome Engineering Database Knowledge Graph)[65] integrate entities into semantically linked networks, while MOF-KG (Metal-Organic Framework Knowledge Graph)[129] adds an LLM-powered interface for natural language querying. Concurrently, LLMs revolutionize extraction flexibility: Dunn *et al.* utilize fine-tuned models for joint entity-relation extraction[130], while MaTableGPT[131] and Silva *et al.*[132] leverage advanced serialization strategies to extract complex synthesis protocols and tabular data with high precision.

**Level 2.** Level 2 agents distinguish themselves by augmenting textual analysis with external tools and domain-specific knowledge, enabling robust, context-aware data curation. Unlike Level 1 systems that rely solely on pattern recognition within text, these agents can actively query databases, invoke APIs, or utilize specialized modules to validate and refine extracted information, thereby transforming static extraction into a dynamic and verified process. HoneyBee[133], an LLM progressively instruction-tuned for MSE, exemplifies this capability by generating trustworthy instruction data via MatSci-Instruct to execute domain-specific tasks with higher fidelity than general-purpose models. Building on this, HoneyComb[14] integrates a high-quality knowledge base (MatSciKB) with a sophisticated tool hub (ToolHub). It employs an inductive tool construction method to generate and refine API tools, allowing the agent to adaptively select and utilize appropriate tools for complex queries, thereby bridging the gap between static knowledge and dynamic tool execution. Furthermore, Eunomia[12] represents an agent-based framework where LLMs autonomously create structured datasets from literature and derive design guidelines. These systems showcase Level 2 autonomy by orchestrating the flow from raw text to actionable insights through tool use, though they remain single-agent planners.

**Level 3.** Level 3 agents transcend the execution of predefined workflows, exhibiting advanced capabilities in reasoning, planning, and generating novel scientific hypotheses. Operating as collaborative planners, they often employ multi-agent architectures to explore vast knowledge spaces. Liu *et al.* demonstrated that LLMs coupled with prompt engineering can generate valid materials design hypotheses that extend beyond the explicit knowledge of human designers[120]. By integrating diverse scientific principles, the model successfully proposed novel high-entropy alloys and halide solid electrolytes, which were subsequently validated in recent literature. Advancing the multi-agent paradigm, SciAgents[121] automates discovery through intelligent graph reasoning. It employs a suite of specialized agents (e.g., Ontologist, Scientist, Critic) that interact with an ontological KG to reveal hidden interdisciplinary relationships, generating hypotheses with precision that surpasses traditional methods. Furthermore, SciMON[134] ensures the novelty of generated hypotheses by retrieving "inspirations" from past literature and iteratively comparing generated ideas against prior work, addressing the common issue of low technical novelty in standard LLM outputs.

**Exploration of Level 4.** Research at Level 4 bridges the gap between digital hypothesis generation and physical execution, focusing on actionable experimental planning under real-world constraints. These agents are characterized by their ability to perform constraint-aware planning and integrate quantitative data to

**Table 4. Representative systems for property prediction and design**

| Methods | Year | Autonomy level | Multi-agent | Closed-loop | Equipment integration | Open source | Agentic system |
|---|---|---|---|---|---|---|---|
| AlloyBERT[52] | 2024 | L1 | × | × | × | √ | × |
| LLM-Prop[43] | 2025 | L1 | × | × | × | √ | × |
| MatterChat[57] | 2025 | L1 | × | × | × | × | × |
| LLM-Fusion[137] | 2025 | L1 | × | × | × | × | × |
| ChatGPT Material Explorer[138] | 2025 | L2 | × | × | × | △ | √ |
| Rep-CodeGen[155] | 2025 | L3 | √ | √ | × | √ | √ |
| SparksMatter[156] | 2025 | L3 | √ | √ | △ | √ | √ |
| MARS[31] | 2026 | L4 | √ | √ | √ | √ | √ |

Notes: √ = present; × = absent; △ = partial or simulated integration. LLM: Large language model.

refine feasibility. Several recent studies have begun to explore the transition toward Level 4 autonomy. AccelMat[30] introduced a goal-driven and constraint-guided LLM agent framework designed to generate viable hypotheses for materials discovery under specific real-world constraints. Utilizing a curated novel dataset from recent publications, which includes explicit design goals and constraints (e.g., cost, equipment availability), the framework presents effectiveness in planning synthesis routes and experimental procedures that are not only scientifically plausible but also practically feasible. This work moves beyond abstract hypothesis generation to actionable experimental planning. PriM[135] takes a further step by not only generating principle-guided hypotheses through multi-agent collaboration but also validating them within a surrogate model-based virtual laboratory. Although the experimental loop remains digital, this approach moves beyond static planning toward automated hypothesis-validation workflows that approximate physical closed-loop execution.

**Vision for Level 5.** While Level 4 agents demonstrate advanced planning and partial closed-loop capabilities, significant challenges remain in realizing fully autonomous discovery, positioning Level 5 primarily as a visionary goal. The primary barrier is the lack of continuous physical grounding. Current advanced agents operate predominantly within a digital hypothesis space and lack direct interfaces to control physical experiments or robotic platforms. Furthermore, while systems such as those proposed by AccelMat[30] and PriM[135] incorporate principle-guided reasoning and simulated validation, they operate without an active learning loop that autonomously requests experiments to resolve uncertainties. A true Level 5 "AI Scientist" would operate as a peer to human researchers, capable of identifying gaps in current theory, formulating original hypotheses, and managing the entire lifecycle of validation without human intervention. Future research should focus on integrating these reasoning engines with automated laboratory hardware (e.g., self-driving labs[37,136]) to create a truly closed-loop system where hypotheses are continuously tested and refined against physical reality.

## Property prediction and design

Advances in AI-driven property prediction and materials design are transforming how researchers discover and optimize new materials. Accurately predicting material properties or inversely designing materials with desired characteristics is crucial for accelerating the development of technologies in energy, electronics, catalysis, and related fields[57]. Traditionally, property prediction relied on experimental measurements or physics-based simulations, while inverse design was often a laborious trial-and-error process. Today, LLM-based agents are emerging as powerful tools to address these challenges. This section examines approaches for forward prediction and inverse design, structured according to ascending levels of agent autonomy, as summarized in Table 4.

**Level 1.** At the foundational Level 1, agents function as assistive tools for information retrieval and analysis in property prediction and design. Early systems such as ChemDataExtractor[123] utilized rule-based methods to parse scientific texts. More recent approaches, such as LLM-Prop[43], leverage natural language descriptions of crystals to predict properties, while AlloyBERT[52] predicts alloy properties from human-readable text. Advanced systems such as MatterChat[57] and LLM-Fusion[137] can integrate multimodal inputs (text, structure, fingerprints) to engage in complex dialogue and analysis. However, these systems remain passive, requiring users to direct all actions, and thus demonstrate Level 1 autonomy: they assist in property prediction and design tasks, but lack capabilities for autonomous planning, tool invocation, integration with experimental equipment, and closed-loop functionality.

**Level 2.** At Level 2, agents are expected to advance to acquire the ability to invoke external computational or simulation tools, evolving from passive assistants to active implementers. A representative example is ChatGPT Material Explorer[138], which can autonomously search materials databases and execute GNN-based property predictors in response to natural language queries. This Level 2 paradigm relies on a robust toolbox of specialized computational models for property prediction and design:

• Forward Prediction Tools: These tools span composition-input predictors such as CrabNet[139], ElemNet[140], and Roost[141], structure-input models such as MoMa[142], Crystalformer[143], and other crystal-graph architectures[144-146], and emerging text-input predictors capable of inferring properties directly from natural language descriptions of materials such as PolyBERT[50]. Together, these established models provide a unified computational substrate for agents to rapidly evaluate candidates from structure or composition to target properties.

• Generative Design Tools: These tools include early generative models such as CrystalGAN[147] and MolGPT[148], which demonstrated the feasibility of generative modeling for crystalline and molecular systems, respectively, as well as highly advanced goal-conditioned generators[149] such as PLaID++[150] and MatterGEN[151], which uses preference optimization to generate stable crystals meeting specific criteria.

• *In-Silico* Optimization Loops: The most advanced Level 2 workflows chain these tools into autonomous computational loops. For example, the deep RL agent by Pan *et al.* autonomously explores chemical space in simulation to discover compounds[152]. Similarly, the "deep dreaming" approach for metal-organic frameworks (MOFs)[153] integrates a generator and predictor into a self-contained *in silico* closed loop to iteratively optimize structures. Sequential optimization strategies such as Bayesian optimization[154], when integrated with forward property predictors, offer an additional route to efficient *in silico* search by guiding exploration toward high-performing regions of vast design spaces.

These tool-augmented agents can execute complex multi-step computational tasks. While systems such as the RL agent exhibit a form of *in silico* closed-loop behavior, they remain at Level 2, as they operate as single-agent systems without physical equipment integration.

**Level 3.** Level 3 is characterized by autonomous orchestration, where multiple role-specialized agents collaborate as a planning team for property prediction and design. Rep-CodeGen[155] is a representative example: a team of LLM agents iteratively writes, tests, and refines Python code to generate new material representations, and this closed-loop collaboration can discover representation schemes that improve property prediction accuracy. This marks a clear shift from a single tool-using agent (Level 2) to multi-agent collaboration within longer workflow pipelines. Beyond code-centered pipelines, SparksMatter[156] further illustrates Level 3 behavior at the task level. When given a high-level goal such as "design a soft semiconductor material", its planning agent can identify that the request implies multiple sub-tasks and

**Table 5. Representative systems for simulation and multiscale modeling**

| Methods | Year | Autonomy level | Multi-agent | Closed-loop | Equipment integration | Open source | Agentic system |
|---|---|---|---|---|---|---|---|
| NequIP[161] | 2022 | L1 | × | × | × | √ | × |
| M3GNet[166] | 2022 | L1 | × | × | × | √ | × |
| MACE[164] | 2024 | L1 | × | × | × | √ | × |
| DPA-2[163] | 2024 | L1 | × | × | × | √ | × |
| MatterSim[165] | 2024 | L1 | × | × | × | √ | × |
| MDAgent[170] | 2025 | L2 | × | △ | × | √ | √ |
| MDCrow[169] | 2025 | L2 | × | △ | × | √ | √ |
| AtomAgents[173] | 2024 | L3 | √ | △ | × | √ | √ |
| El Agente[171] | 2025 | L3 | √ | △ | △ | △ | √ |
| DREAMS[159] | 2025 | L3 | √ | △ | △ | √ | √ |
| MooseAgent[172] | 2025 | L3 | √ | △ | △ | √ | √ |
| MatSciAgent[174] | 2025 | L3 | √ | △ | △ | √ | √ |

Notes: √ = present; × = absent; △ = partial or simulated integration.

organize the workflow accordingly. Moreover, during prediction, the agent does not only output a numeric value but also provide chain-of-thought reasoning explanations for why a prediction is made and which factors drive the result, thereby improving interpretability for downstream design decisions. Overall, Level 3 systems can form closed loops within the computational domain, but they remain purely *in silico* without direct integration with laboratory hardware for physical experimentation.

**Exploration of Level 4 and Level 5.** The latter stages, Level 4 and Level 5, require linking autonomous computational planning with physical experimentation, and eventually with scientific problem discovery and validation. A key step from Level 3 to Level 4 is moving beyond *in silico* optimization toward physical closed-loop control by integrating agents with robotic lab platforms such as A-Lab[18] or ChemOS 2.0[25]. In the property prediction and design setting, recent systems such as MARS[31] begin to explore this direction by combining knowledge-grounded reasoning (e.g., hybrid RAG over domain literature) with tool-based analysis and coordinated execution. This enables predictions and decisions to be updated using real experimental feedback rather than remaining purely digital. Looking toward Level 5, agents should be able to autonomously select appropriate design and prediction strategies and integrate them into full experiment–computation validation loops, so that inverse design goals can be solved end-to-end with minimal human input. Achieving this level will require robust integration across tools and instruments, reliable closed-loop lab control, and multi-step reasoning that remains stable under real-world uncertainty.

## Simulation and multiscale modeling

In MSE, simulation is a crucial tool for designing new materials, validating conceptual hypotheses, and understanding material behaviors. Typical material simulation workflows can be classified into quantum mechanical calculations, atomistic simulations, mesoscale simulations, and continuum simulations. Traditionally (Level 0), these workflows required extensive manual effort: researchers had to construct simulation models manually, define assumptions and boundary conditions, select and calibrate parameters, and repeatedly debug numerical instabilities or convergence failures. Furthermore, bridging multiple scales posed additional challenges, as quantum, atomistic, and continuum simulations operate under different physical assumptions, resolution limits, and computational costs, making their integration into a coherent pipeline both time-consuming and error-prone. In recent years, data-driven machine learning models and autonomous agents have begun to augment this paradigm, creating a new ecosystem of powerful computational tools and automated workflows. Representative systems for material simulations are summarized in Table 5.

**Level 1.** At Level 1 autonomy, agents assist researchers by automating routine preprocessing tasks upon human request. One of the foundations of agentic material simulations is automated material calculation frameworks, including the atomic simulation environment (ASE)[157], FireWorks[88], and AiiDA[89]. These frameworks lay the groundwork for workflow definition and automation in material computations, ranging from density functional theory (DFT) and MD, to finite element analysis (FEA). LLM-based agents then assist with basic steps such as generating input files[158], validating structural data, selecting relevant simulation parameters, and retrieving prior results from databases - thus reducing the cognitive load associated with manual model setup[159]. By handling these preparatory and post-processing steps, Level 1 agents reduce the manual drudgery and cognitive burden associated with model setup, allowing researchers to focus on higher-level scientific questions.

Another foundational direction in applying AI to materials simulations is the use of deep learning models as surrogate solvers to replace computationally expensive physical calculations. Across quantum, atomistic, mesoscale, and continuum regimes, these models learn high-fidelity approximations of energies, forces, or field solutions, enabling orders-of-magnitude acceleration compared with first-principles or conventional numerical solvers. A representative example is the development of machine learning interatomic potentials (MLIPs), which bridge quantum and atomistic simulations by learning energy and force mappings from electronic-structure data. The predictive performance of MLIPs has been significantly improved by explicitly incorporating physical symmetries and constraints[160-162]. Moving toward more general-purpose large-scale atomistic modeling, universal potentials have been proposed to cover broader chemical and physical domains[163-166]. Another example is the application of physics-informed neural networks (PINNs) and neural partial differential equation (PDE) models for solving governing equations in mesoscale and continuum simulations[167,168].

While these works substantially improve computational efficiency and scalability, they remain passive components within the simulation pipeline, leaving the choice of simulation boundaries and scales to human researchers. Therefore, these approaches are primarily categorized as Level 1 automation that automates or accelerates specific execution, which lays the foundation for subsequent agentic systems.

**Level 2.** At Level 2, agents transition from passive analysis to active engagement with the scientific toolkit. While humans still define the overarching goals, these agents can independently invoke external tools such as electronic structure codes, MD engines, materials databases, and analysis libraries to execute intermediate tasks. LLMs at this level act as a control layer that interprets high-level scientific intent and dynamically decides which computational tools to invoke, in what sequence, and with which inputs. This "Tool-Augmented" paradigm grounds the reasoning of LLMs in rigorous computational engines, overcoming the hallucination limitations of pure language models.

For example, MDCrow[169] operates at Level 2 autonomy by enabling an LLM to dynamically select and sequence MD-related tool calls, such as solvation, OpenMM execution, and MDTraj analyses, according to high-level user objectives. These fundamental MD simulation workflows are encapsulated within an agentic toolset designed to autonomously execute MD simulations for exploration of the biochemical design space. The LLM serves as a conversational interface for coordinating, automating, and summarizing simulation steps. Similarly, MDAgents[170] employs a fine-tuned LLM to generate, validate, and execute MD simulation scripts, together with simple feedback loops that enable the agent to iteratively correct syntax errors or adjust simulation settings based on runtime feedback. Despite these advances, decision making at Level 2 remains task-oriented. This limitation motivates the transition to Level 3 autonomy, where agents begin to plan and orchestrate multi-step simulation campaigns with minimal human intervention.

**Level 3.** Level 3 marks the emergence of autonomous orchestration, where MASs plan and execute complex simulation workflows with minimal human intervention. DREAMS[159] exemplifies this paradigm through a hierarchical multi-agent framework that autonomously carries out sequences of DFT calculations. Similarly, El Agente[171] and MooseAgent[172] leverage cooperating LLM agents to translate high-level natural language goals into concrete quantum chemistry or multiphysics simulation tasks while handling execution and error monitoring. AtomAgents[173] introduced a physics-aware, multimodal, multi-agent architecture tailored for alloy design and discovery, in which multiple specialized agents collaboratively orchestrate atomistic simulations, code execution, and multimodal result analysis. By functioning as collaborative *in silico* planners, these systems decompose complex computational goals and coordinate specialized agents to accomplish them, often operating in a closed loop within the simulation environment.

Conceptually, tool usage under autonomous orchestration can further span across different simulation scales rather than relying solely on individual simulation tools. The development of AI-accelerated cross-scale simulation methods, as mentioned in Level 1, can be integrated into a practical simulation-based pipeline for materials design and engineering. In this context, MatSciAgent[174] represents a distinctive realization of Level 3 autonomy by unifying cross-scale materials simulation tasks within a modular multi-agent orchestration framework. Unlike systems that focus on coordinating a single class of simulation tools, MatSciAgent adopts a master–worker architecture in which a central agent interprets high-level natural-language requests, identifies the underlying task type, and delegates execution to specialized task-specific agents.

As a result, Level 3 agentic systems for simulation tasks are capable of executing typical multiscale materials simulation workflows in response to high-level user requests by coordinating tool usage and making adaptive decisions. However, despite their advanced planning capabilities, these systems remain confined to the digital realm and lack a direct interface with the physical world.

**Exploration of Level 4 and Level 5.** Although Level 3 systems mainly operate within digital simulation environments, several recent works have begun to explore Level 4 behaviors. As discussed above, systems such as DREAMS[159] and AtomAgents[173] not only plan multi-step simulation workflows, but also begin to handle long-running execution with monitoring and iterative adjustment of workflows based on intermediate results.

The key step toward Level 4 in simulation and multiscale modeling is therefore not only improved planning, but also closing the loop with the physical world. This requires coupling simulation agents with laboratory automation so that experimental measurements can be used to update model assumptions, parameters, and even the selection of simulation methods, and next rounds of simulations can be scheduled and executed with minimal human intervention. In such a setting, the agent becomes the controller of an experiment-simulation loop rather than merely a simulation planner[159,173]. Level 5 extends this concept to fully autonomous scientific discovery, in which the agent can connect multiscale modeling with experimental evidence at the level of scientific reasoning[175]. When persistent gaps appear between predictions and observations, a Level 5 agent should be able to propose plausible physical explanations, design integrated simulation-experiment campaigns to test them, and revise its models and hypotheses based on the resulting outcomes. While this capability remains a long-term goal, it represents the ultimate fusion of computation and experimentation for autonomous materials discovery.

## Automated experimental synthesis

Experimental synthesis and characterization of new materials is the most critical step in the MSE research pipeline. It is also the most important method to validate previous property prediction and simulation results. The advent of SDLs is redefining how materials are experimentally discovered and tested, moving

**Table 6. Representative systems for automated experimental synthesis**

| Methods | Year | Autonomy level | Multi-agent | Closed-loop | Equipment integration | Open source | Agentic system |
|---|---|---|---|---|---|---|---|
| MatScIE[177] | 2021 | L1 | × | × | × | √ | × |
| MatNexus[178] | 2023 | L1 | × | × | × | √ | × |
| CRISPR-GPT[189] | 2025 | L2 | × | △ | × | √ | √ |
| SciAgents[121] | 2024 | L3 | √ | × | × | √ | √ |
| ChatGPTResearchGroup[181] | 2023 | L3 | √ | √ | × | × | √ |
| LABMATE[190] | 2025 | L3 | √ | △ | × | × | √ |
| MOSAIC[191] | 2026 | L3 | √ | △ | △ | √ | △ |
| AlphaFlow[117] | 2023 | L4 | × | √ | √ | √ | × |
| Coscientist[17] | 2023 | L4 | √ | √ | △ | √ | √ |
| AutoMEX[192] | 2025 | L4 | √ | √ | √ | × | √ |

Notes: √ = present; × = absent; △ = partial or simulated integration.

experimentation from traditional manual work (Level 0) to autonomous closed-loop operation[176]. By integrating robotics, these platforms have the potential to iteratively plan and execute experiments with minimal human input, accelerating discovery and enhancing reproducibility. Recent developments in intelligent AI systems integrated with automated fabrication platforms enable closed-loop optimization of synthesis conditions and materials processing. Representative systems and their capabilities in experimental synthesis are summarized in Table 6.

**Level 1.** Level 1 autonomy focuses on establishing knowledge-grounded AI systems to assist physical synthesis, enabling human users to interact with the systems and acquire guidance on experimental design and optimization. This stage involves LLM-assisted analysis, where text-mining systems such as MatScIE[177] and MatNexus[178] extract critical data from scientific literature and technical manuals. These systems convert unstructured descriptive text into structured, queryable databases containing materials information and synthesis protocols. Similarly, LLMs have also been integrated into electronic experimental notebooks[179] to facilitate the digitization of routine materials experiments. Some preliminary studies have further utilized the generative capabilities of LLMs to directly predict synthesis pathways for inorganic[180] and organic[59] materials. These efforts reduce the burden on experimentalists by enabling rapid retrieval of protocols, identification of relevant control variables, and generation of hypothesis-driven suggestions for synthesis conditions. They also support the transformation of experience-oriented experimental knowledge into machine-readable and reusable representations, thereby improving efficiency and reproducibility. However, Level 1 systems remain advisory and lack independent planning capabilities or the ability to invoke external tools.

**Level 2.** Building upon knowledge-grounded LLM systems, Level 2 agents function as "digital chemists" by invoking external computational tools for experimental planning. At this level of autonomy, LLM-based agentic systems not only rely on prior knowledge acquired during pretraining and prompt-based reasoning, but also invoke external information sources and computational capabilities to support experimental design for materials fabrication and synthesis optimization. For example, active learning-based methods can function as external tools for optimizing material fabrication parameters[181] or can be combined with Bayesian optimization and LLMs to enhance contextual optimization[182-185]. For synthesis pathway prediction, advanced LLMs have demonstrated significant competence in predicting reaction outcomes and retrosynthetic routes[186]. External knowledge systems can further enhance synthesis route prediction[187]. This capability is further strengthened by specialized models, such as the conditional graph logic network (GLN)

for retrosynthesis[116] and graph-based networks for predicting solid-state synthesis routes[188]. Additionally, data-driven models can emulate human decision making to recommend precursors[180], while AI co-pilots such as CRISPR-GPT[189] can automate complex experimental designs. Level 2 agents are thus capable of complex computational planning and analysis, yet they remain single-agent systems without multi-step planning capabilities or closed-loop hardware control.

**Level 3.** Level 3 advances to multi-agent coordination, where specialized agents manage *in silico* workflows akin to those of a human research group. These works reflect the decomposition of complex experimental design tasks into subtasks that might involve previous tasks. SciAgents[121] conceptualizes this framework as a "team of AI agents", in which networks of agents collaborate autonomously through iterative cycles of hypothesis generation and computational validation. ChatGPTResearchGroup[181] organized multiple role-specialized LLM agents to collaboratively conduct closed-loop *in silico* planning and Bayesian optimization for materials synthesis. LABMATE[190] applied this paradigm to catalysis research by orchestrating agents responsible for literature review, simulation, data analysis, and hypothesis generation within a human-in-the-loop computational copilot framework. MOSAIC[191] trained 2,498 specialized chemistry experts and successfully guided the synthesis of over 35 novel compounds across areas such as pharmaceuticals, materials science, and agrochemicals. Although these multi-agent planners can execute complex multi-step computational workflows, they remain disconnected from equipment integration and require human involvement for experimental execution.

**Level 4.** The transition to Level 4 represents the current frontier of research, characterized by exploratory efforts to bridge the gap between digital planning, computation, and the physical execution of experiments. In this phase, pioneering systems act as autonomous planners that directly control robotic laboratories in a closed loop. This emerging capability is exemplified by LLM-driven prototypes such as Coscientist[17] and AutoMEX[192], which demonstrate the feasibility of using AI to autonomously issue commands to cloud-based lab robots and three-dimensional (3D) printers. Other exploratory platforms leverage algorithmic optimization to guide laboratory hardware, such as the microfluidic systems for nanoparticle synthesis developed by Tao *et al.*[193] and Sadeghi *et al.*[194]. More advanced implementations, such as AlphaFlow[117], utilize RL to control modular reactors and have successfully identified novel synthesis routes that outperform human-designed processes. Perhaps the most comprehensive demonstration of hardware flexibility is the mobile robotic chemist described by Burger *et al.*[195], which navigates a standard lab to execute closed-loop optimization. Collectively, these platforms represent early iterations of Level 4 intelligence. Although they are primarily proof-of-concept systems often orchestrated by a central planner, they successfully validate the core criteria for autonomous physical execution.

**Vision for Level 5.** Beyond these emerging implementations lies the aspirational goal of Level 5, which envisions a fully autonomous agent capable of independently formulating broad hypotheses, designing novel research directions, and executing experiments without constraints. However, such a system remains a distant prospect. In the foreseeable future, human experts will remain indispensable. The consensus in the field suggests that the evolution of SDLs will likely stabilize at a human-in-the-loop hybrid model rather than achieve complete human replacement, as argued by Hysmith *et al.*[196]. Ideally, this synergy between human intuition and transparent, interpretable AI will transform the laboratory into an engine of innovation, enabling scientists to tackle challenges on a previously unimaginable scale.

## Characterization and data analysis

Modern MSE relies on advanced characterization techniques that produce vast and complex datasets. These measurements probe material structures and properties across multiple length and time scales from complementary perspectives, posing significant challenges for consistent, knowledge-grounded

**Table 7. Representative systems for characterization and data analysis**

| Methods | Year | Autonomy level | Multi-agent | Closed-loop | Equipment integration | Open source | Agentic system |
|---|---|---|---|---|---|---|---|
| MatQnA[197] | 2025 | L1 | × | × | × | √ | × |
| MicroscopyGPT[204] | 2025 | L1 | × | × | × | × | × |
| S1-MMAlign[198] | 2026 | L1 | × | × | × | √ | × |
| Chen *et al.*[205] | 2025 | L2 | × | × | × | × | √ |
| Dara[206] | 2025 | L2 | × | × | △ | √ | × |
| Drug Discovery Agent[207] | 2025 | L2 | △ | × | √ | √ | √ |
| MatAgent[208] | 2025 | L2 | √ | √ | × | √ | √ |
| AutoMat[210] | 2025 | L3 | × | √ | × | √ | √ |
| IR-Agent[211] | 2025 | L3 | √ | × | √ | √ | √ |
| Multicrossmodal Agent[15] | 2025 | L3 | √ | √ | △ | √ | √ |
| SciLink[209] | 2025 | L3 | √ | √ | △ | √ | √ |
| AdaptiveXRD[29] | 2024 | L4 | × | √ | √ | √ | △ |
| AILA[212] | 2025 | L4 | √ | √ | √ | √ | √ |
| ORGANA[213] | 2025 | L4 | √ | △ | √ | √ | √ |

Notes: √ = present; × = absent; △= partial or simulated integration.

interpretation. To overcome these challenges, relevant agents have been introduced across workflows ranging from data collection to interpretation, shifting the field from traditional manual interpretation (Level 0) toward autonomous workflows, as summarized in Table 7.

**Level 1.** Level 1 serves as a materials characterization knowledge base and analytical assistant. For example, MatQnA[197] established a large multimodal benchmark for materials characterization techniques, including XPS, XRD, SEM, and TEM. S1-MMAlign[198] collected large-scale multimodal scientific image interpretation datasets, including materials science data. In addition, several AI systems have also been developed for the operation and interpretation of specific characterization techniques such as TEM[199], XRD[200,201], SEM[202], and spectroscopy[203], thereby reducing the expertise barrier and subjectivity inherent in manual analysis and laying the groundwork for higher-level autonomy as agentic tools. Notably, MicroscopyGPT[204] is a vision-language model (VLM) that solves the difficult problem of reconstructing full 3D atomic structures from 2D scanning transmission electron microscopy (STEM) images by directly mapping images to structured text. These systems fall within Level 1 as they facilitate AI-assisted IE from multimodal materials characterization data, but function as passive resources without autonomous planning or execution capabilities.

**Level 2.** These data and knowledge resources fuel a new generation of Level 2 (Tool-Augmented) agents, which function as automated analysts for specific tasks. Chen *et al.* proposed an LLM-driven multimodal framework for detecting scale bars and extracting related information from SEM images[205]. The framework uses a You Only Look Once (YOLO)-based detector to localize the scale bar and a hybrid optical character recognition (OCR) system to recognize the numeric value and unit. For diffraction analysis, Dara automates multiple-hypothesis phase identification and refinement from powder XRD data by searching candidate phase combinations and programmatically performing peak matching and Rietveld refinement (accelerated via parallel execution), while using domain-aware criteria to prune candidates and decide when to stop[206]. Drug Discovery Agent[207] can follow high-level prompts to detect and classify drug–cell phenotypes from microscopy images/videos by coordinating vision modules, thereby enabling scalable and near real-time screening. In addition, general-purpose agentic frameworks such as MatAgent[208] show that an LLM-based

MAS can run end-to-end experimental data analysis, ranging from exploratory statistics to modeling, visualization, and report generation.

**Level 3.** At Level 3, agents can cope with the complexity and diversity of characterization data by using multi-step planning and agent-based architectures to take an active role in the research process. Systems such as SciLink[209] and AutoMat[210] show how agents can break down complex goals and support end-to-end automated analysis. SciLink[209] can turn raw characterization data into scientific hypotheses, and then assess these claims against published literature. AutoMat[210] uses a "plan-then-execute" design and integrates multiple tools to transform STEM image inputs into reconstructed atomic crystal structures. In spectroscopy, IR-Agent[211] mimics the reasoning process of human experts and employs a team of agents for feature extraction, database retrieval, and final inference of molecular structures. For multimodal data from different sources, Bazgir *et al.* proposed a multi-agent framework with a dynamic gating mechanism[15]. This framework can analyze microscopy images and simulation videos while also retrieving relevant papers and web resources to provide contextual support and improve accuracy. Overall, Level 3 systems enable data-driven inverse reasoning and greatly improve both the efficiency and reliability of extracting scientific insights from raw characterization data.

**Level 4.** At Level 4, agents take on direct operational control of the physical laboratory: they can operate equipment, sustain long-running experiments, and autonomously determine how procedures should proceed during execution. AdaptiveXRD[29] is an autonomous and adaptive XRD system that enables agent-driven, real-time control of physical hardware and can autonomously adjust the scan step size and scan range during measurements. For complex and precise instruments such as atomic force microscopy (AFM), Artificially Intelligent Lab Assistant (AILA)[212] shows strong multi-agent collaboration and supports long-duration autonomous operation. Moreover, ORGANA[213] is a highly integrated automation platform that uses natural language interaction to automate complex chemistry experiments from end to end. It can translate high-level research goals into physical operation commands, marking a shift at Level 4 from single-instrument automation toward more system-level laboratory automation. While these platforms demonstrate promising Level 4 autonomy, they remain early prototypes rather than widely adopted and reliable systems.

**Vision for Level 5.** Level 5 represents a long-term goal: a characterization agent that can work with minimal human input. Beyond operating instruments, such an agent would be able to pose meaningful research questions, choose suitable characterization methods, and integrate evidence from multiple instruments to build a complete view of new materials. This would shift the focus from simply collecting measurements to understanding what the results imply. Reaching this level, however, will require major progress in linking different instruments, standardizing data and metadata, and improving multi-step reasoning. In the near term, Level 5 should be treated as a reference point, while most practical work should focus on strengthening the human–AI collaboration patterns seen in Level 3 and Level 4.

## Cross-task MSE agents

The preceding sections examine AI agents within specific, isolated research tasks such as synthesis planning, characterization, property prediction, and simulation. While these task-specific systems have demonstrated significant capabilities, more advanced systems are now emerging that transcend single-task boundaries, integrating diverse capabilities into cohesive research pipelines[39,214]. These cross-task agents represent a shift from specialized tools to holistic research orchestrators capable of managing the full research cycle, from hypothesis generation to experimental validation.

### *Pioneering autonomous laboratories*

Early Level 4 systems demonstrated that robotic platforms could leverage closed-loop machine learning to accelerate discovery. Pioneering examples such as adaptive rapid experimentation and spectroscopy (ARES)[215] (for carbon nanotubes) and Ada[136] (for thin films) showed that algorithms could design, execute, and analyze experiments faster than human researchers. This paradigm was significantly advanced by A-Lab[18], which integrates computation, literature mining, and robotics to autonomously discover 41 new inorganic materials within 17 days. Similarly, full-process *in silico* frameworks, such as the Level 3 system for perovskite solar cells developed by Ye *et al.*[216], demonstrate how agents can process heterogeneous data spanning materials, fabrication, and performance to uncover complex patterns, even without physical automation. These works highlight the power of integrating diverse data streams and operational modules into a cohesive research engine.

### *Unifying computational planning and physical execution*

Recent advances have demonstrated the integration of multi-agent planning with physical experimentation. LLM-RDF[217], a framework employing specialized agents to coordinate a complete reaction development cycle - from literature search and experiment design to hardware control and spectral analysis - successfully guided the development of a novel oxidation reaction. In the computational domain, TopoMAS[218] orchestrates literature search, hypothesis generation, and DFT simulations in an *in silico* closed loop, identifying novel topological quantum materials. AGAPI-Agents[219] also unifies open-source LLMs with more than 20 materials APIs to autonomously run multi-step, tool-grounded workflows for reproducible and accelerated materials design. Furthermore, ChemAgents[220] seamlessly integrates robotic experimentation, quantum simulations, and ML-driven spectral analysis to investigate azobenzene isomerization, uncovering new mechanistic insights with minimal human intervention. This framework represents a blueprint for autonomous molecular discovery, where agents manage the full "design-make-test-analyze" cycle across both digital and physical realms. To support such cross-domain reasoning, multimodal frameworks such as MatterChat[57] enable agents to process both textual knowledge and structural data, bridging the gap between literature understanding and atomic-level design.

## FUTURE WORK

### Current challenges and inherent limitations

Our analysis through the six-level framework reveals that the current research gaps in agentic MSE fall into two distinct categories, defined by the nature of the tasks involved.

### *Cognition-centric challenges*

These challenges primarily emerge in tasks such as information retrieval, property prediction, and simulation, which operate in the digital domain and rely on the reasoning capabilities of LLMs. Despite rapid progress, current systems are constrained by the intrinsic limitations of LLMs when applied to scientific domains. MSE data are often sparse, heterogeneous, and highly structured, yet LLMs typically process such data as ungrounded text. Consequently, retrieval agents may miss critical context, property predictors may extrapolate beyond physical validity, and simulation planners may generate workflows that are linguistically coherent but numerically unstable. Fundamentally, these systems often lack robust mechanisms for enforcing physical laws, estimating uncertainty, and recovering from failures, hindering their progression to higher levels of autonomous reasoning.

### *Execution-centric challenges*

This second category primarily appears in experimental synthesis and materials characterization, where the dominant difficulty shifts from language-based reasoning to physical interaction and real-world control. In materials science research, the execution bottleneck is driven by the heterogeneous and non-standardized

nature of laboratory environments, including diverse software–hardware interfaces, inconsistent data formats, and the intrinsic variability of material samples. These factors introduce substantial noise and uncertainty into experimental processes, making reliable execution significantly more challenging than in purely digital settings. Such execution-centric settings also expose reliability problems in instruction adherence. Recent AFM automation studies have shown that LLM agents can take extra actions beyond the given protocol, sometimes acting as if they rely on prior context or memory rather than the current instruction - a behavior referred to as "sleepwalking"[212]. This behavior can appear as risky physical actions beyond authorized limits or as functional code that exceeds the specified requirements, reflecting instruction drift during execution. Such behavior raises clear concerns regarding operational safety and the validity of closed-loop experiments.

Recent advances in collaborative robotics and automated laboratories have led to the development of middleware frameworks, hardware standardization efforts, and communication protocols (e.g., SiLA[221], ChemOS[25,222], Robot Operating System[223]), which provide an important technical pathway for device coordination and standardization in materials science labs. However, integrating agentic systems into these infrastructures remains non-trivial, as there still exists a gap between agent-level reasoning and device-level communication. Looking forward, an additional challenge lies in enabling effective human–agent collaboration, as future laboratory environments are likely to involve hybrid workflows in which autonomous systems and human operators must co-adapt, share context, and coordinate decisions under uncertainty.

Uncertainty is another fundamental challenge that permeates all aspects of agentic MSE[224]. At the single-agent level, uncertainty arises from the stochastic nature of LLM outputs, irreducible noise in experimental measurements, and approximation errors in computational simulations[225]. In MASs, these uncertainties do not remain local; instead, they can propagate across agents and even be amplified in a cascading manner. For example, an incorrect assumption introduced during IE may bias downstream property prediction and ultimately lead to suboptimal or even incorrect synthesis decisions[226]. To build trustworthy agentic MSE systems, uncertainty quantification should therefore evolve from a passive monitoring signal into an active control signal. In this context, the agentic uncertainty quantification (AUQ) framework[227] offers a promising direction. Inspired by dual-process theories of human cognition, AUQ converts uncertainty into a closed-loop behavioral signal through uncertainty-aware memory and uncertainty-aware reflection, aiming to mitigate hallucination cascades in long-horizon agent trajectories. More broadly, uncertainty in agentic MSE should not be treated solely as a property of model outputs, but as a system-level quantity that governs whether an agent should continue execution, request additional evidence, trigger self-correction, or defer to human oversight. This issue will become even more important as agentic MSE moves from laboratory prototypes toward industrial deployment, where robustness, reliability, and governance under uncertainty are essential.

Collectively, these challenges reveal that higher autonomy cannot be achieved by optimizing individual components in isolation. Instead, it requires tightly integrated systems that are knowledge-grounded for cognitive reasoning, perception-aware for physical execution, and equipped with principled mechanisms to quantify, propagate, and act on uncertainty at both the agent and system levels.

### Strategic directions for future research

To overcome these hurdles, future research must pivot from purely data-driven approaches toward the development of physically grounded and robustly embodied agents.

### *Physically grounded intelligence*

A critical step in addressing cognitive limitations is the development of Hybrid Neuro-Symbolic Reasoning systems[228]. By constraining the generative fluency of LLMs with thermodynamic verifiers and physics-informed logic, agents can ensure their hypotheses are not only novel but also physically viable. This approach entails training agents on "negative data" and physics-informed datasets to instill a form of scientific "common sense", effectively preventing the proposal of chemically unreasonable candidates.

### *Closing the physical execution gap*

To address execution-centric challenges, future systems must move beyond simple API calls to incorporate Active Perception, empowering agents to monitor experiments via computer vision and multimodal sensor feedback. This sensorimotor integration is essential for agents to adaptively correct errors in real time - such as detecting precipitation failures or blocked needles - rather than proceeding blindly. This capability is the foundation for creating truly adaptive and resilient autonomous laboratories.

### *Dynamic evaluation and benchmarking*

Establishing robust metrics is essential for quantifying progress across the proposed six-level autonomy hierarchy. Existing benchmarks, such as MatSciBench[104] and MSQA[105], mainly evaluate static reasoning or isolated property prediction, and therefore provide limited coverage of agentic behavior in long-horizon scientific workflows. However, evaluating an autonomous agent is fundamentally different from evaluating a static LLM: beyond final-answer correctness, it also requires assessing the quality of the reasoning trajectory, including multi-step planning, tool selection, feedback utilization, error recovery, and avoidance of unproductive loops. Recent benchmark efforts have begun to move in this direction. For example, SciAgentGym[229] explicitly evaluates long-horizon scientific tool use and analyzes process-level behaviors such as adaptation to execution errors, parameter tuning, strategic switching, loop escape, and recovery dynamics across interaction steps. Likewise, SGI-Bench[230] frames evaluation around scientist-aligned workflows, covering deep research, idea generation, dry/wet experiments, and experimental reasoning, and further introduces an agent-based evaluation framework to support multi-dimensional assessment.

Future benchmarking efforts should therefore move beyond outcome-only scoring and incorporate trajectory-level criteria that capture whether an agent can sustain coherent multi-step reasoning, recover from failures, and interact reliably with tools, data, and experimental systems. Such dynamic evaluation testbeds, ideally coupled with realistic noise, hardware constraints, and failure modes, will be essential for assessing agentic resilience in MSE and for charting progress toward fully autonomous AI materials scientists.

### *Safety and governance*

Finally, as agents evolve from advisory roles at lower levels to synthesis planning and direct physical execution at higher levels, their dual-use risks become increasingly serious, making safety and governance more critical. The development of autonomous systems requires robust and deterministic safety guardrails, together with specialized safety assessment tools and governance frameworks throughout the agent development lifecycle. To address dual-use risks, a range of advanced red-teaming methods for scientific agents has recently emerged[231]. In addition, the development of standardized safety benchmarks for toxicity screening is a necessary step toward measuring progress[232]. Relevant protocols should verify every chemical instruction against strict safety databases to ensure that the pursuit of autonomous discovery never compromises laboratory safety[175]. In the long term, trustworthy deployment will also require uncertainty-aware governance, in which quantified uncertainty is used not only for post hoc diagnosis, but also for real-time control, escalation, and safety intervention.

### *Ecosystem integration and real-world deployment*

Future agentic MSE systems will need to operate within a broader ecosystem that extends beyond the scientific workflow itself. Materials research is closely tied to supply chains for precursors, consumables, instruments, and software, as well as to funding mechanisms, certification procedures, and downstream industrial deployment. These external factors may strongly constrain what an agent can realistically propose or execute. A scientifically valid plan may still fail in practice due to unavailable materials, incompatible equipment, restricted software access, limited project budgets, or unmet regulatory requirements. At the same time, this broader integration opens an important opportunity: agentic MSE could evolve from optimizing isolated scientific tasks to coordinating science with operations. This includes resource-aware planning, procurement-aware experiment scheduling, traceable documentation for certification, and decision support for technology transfer into industrial settings. Accordingly, a major future direction is to develop ecosystem-aware agents that can reason not only over materials knowledge and laboratory feedback, but also over the logistical, economic, and regulatory contexts in which materials innovation actually unfolds. In this sense, higher autonomy can also be defined by the ability to remain actionable under real-world supply, budgetary, and regulatory constraints.

## CONCLUSIONS

This survey reviewed the fast-growing landscape of agentic MSE from a systems view, where agents connect data resources, computational tools, and (in some cases) experimental hardware into unified workflows. To describe this transition in a consistent way, we proposed a six-level autonomy framework and mapped it to five core task families in MSE. This task–level map helps move beyond "model lists" and instead shows what an agent can actually accomplish, what components it must integrate, and where key gaps remain.

A central finding is that progress is uneven across tasks because each task family faces different limitations in reasoning, tool integration, and safety constraints. This unevenness also reflects a broader workflow shift: traditional MSE work has often been linear and organized into separate steps, with humans manually linking high-level reasoning to low-level execution; agentic workflows aim to close this gap through unified reasoning and planning. Importantly, as autonomy increases, task boundaries become less clear: higher-level agents tend to combine multiple tasks into a unified process.

A brief cross-task comparison at Level 3 shows why a task–level lens is necessary. Although multi-agent coordination emerges across tasks, the bottlenecks differ: simulation tasks focus on the stable orchestration of long tool chains, information tasks emphasize evidence grounding and scientific validity, and synthesis/characterization tasks are limited mainly by hardware interfaces, sensing capabilities, and experimental variability.

Looking ahead, the long-term goal is Level 5 autonomy, but the path forward is not only "more capable models". It requires several system-level advances, including physically grounded intelligence, stronger active perception and embodied interaction, improved evaluation methods for long-horizon autonomy, and clearer safety and governance rules (including equity of access). In this sense, the six-level framework and the task–level map serve as practical guides: they make progress measurable, clarify what "higher autonomy" demands in each task family, and support a systematic transition from isolated tools to reliable human–AI collaboration in real-world MSE workflows.

## DECLARATIONS

### Acknowledgments

The authors would like to acknowledge Flaticon (https://www.flaticon.com/) and IconPark (https://iconpark.oceanengine.com/) for providing the icons and graphical assets used in the figures of this manuscript.

## Authors' contributions

Conceptualization and design of the review: Luo, Y.; Zhang, T.; Zhu, J.; Zhang, L.
Writing - manuscript: Zhu, J.; Zhang, L.; Zhu, Y.
Writing - review and editing: Lin, X.; Wu, Y.; Di, S.; Liu, B.
Supervision: Luo, Y.; Zhang, T.; Di, S.; Liu, B.
All authors reviewed and approved the final version of the manuscript.

## Availability of data and materials

Not applicable.

## AI and AI-assisted tools statement

Not applicable.

## Financial support and sponsorship

This work is supported by National Key R&D Program of China (No. 2025ZD0619400) and by the Guangzhou-HKUST (GZ) Joint Funding Program (No. 2023A03J0003).

## Conflicts of interest

Zhang, T. is the Editor-in-Chief of *Journal of Materials Informatics*, but was not involved in any stage of the editorial process, notably including reviewer selection, manuscript handling, or decision making. The other authors declared that there are no conflicts of interest.

## Ethical approval and consent to participate

Not applicable.

## Consent for publication

Not applicable.



## REFERENCES

1. Gao, X.; Tan, R.; Li, G. Research on text mining of material science based on natural language processing. *IOP. Conf. Ser. Mater. Sci. Eng.* **2020**, *768*, 072094. DOI
2. Li, Y.; Gupta, V.; Kilic, M. N. T.; et al. Hybrid-LLM-GNN: integrating large language models and graph neural networks for enhanced materials property prediction. *Digit. Discov.* **2025**, *4*, 376-83. DOI
3. Olivetti, E. A.; Cole, J. M.; Kim, E.; et al. Data-driven materials research enabled by natural language processing and information extraction. *Appl. Phys. Rev.* **2020**, *7*, 041317. DOI
4. Reiser, P.; Neubert, M.; Eberhard, A.; et al. Graph neural networks for materials science and chemistry. *Commun. Mater.* **2022**, *3*, 93. DOI PubMed PMC
5. Venugopal, V.; Sahoo, S.; Zaki, M.; Agarwal, M.; Gosvami, N. N.; Krishnan, N. M. A. Looking through glass: knowledge discovery from materials science literature using natural language processing. *Patterns* **2021**, *2*, 100290. DOI PubMed PMC
6. Yoshitake, M.; Sato, F.; Kawano, H.; Teraoka, H. MaterialBERT for natural language processing of materials science texts. *Sci. Technol. Adv. Mater. Methods.* **2022**, *2*, 372-80. DOI
7. Kim, K.; Kang, S.; Yoo, J.; et al. Deep-learning-based inverse design model for intelligent discovery of organic molecules. *npj. Comput. Mater.* **2018**, *4*, 67. DOI
8. Liu, Z.; Zhu, D.; Rodrigues, S. P.; Lee, K.; Cai, W. Generative model for the inverse design of metasurfaces. *Nano. Lett.* **2018**, *18*, 6570-6. DOI
9. Curtarolo, S.; Hart, G. L. W.; Nardelli, M. B.; Mingo, N.; Sanvito, S.; Levy, O. The high-throughput highway to computational materials design. *Nat. Mater.* **2013**, *12*, 191-201. DOI PubMed
10. Hachmann, J.; Olivares-Amaya, R.; Atahan-Evrenk, S.; et al. The Harvard Clean Energy Project: large-scale computational screening and design of organic photovoltaics on the world community grid. *J. Phys. Chem. Lett.* **2011**, *2*, 2241-51. DOI
11. Kanal, I. Y.; Owens, S. G.; Bechtel, J. S.; Hutchison, G. R. Efficient computational screening of organic polymer photovoltaics. *J. Phys. Chem. Lett.* **2013**, *4*, 1613-23. DOI PubMed

12. Ansari, M.; Moosavi, S. M. Agent-based learning of materials datasets from the scientific literature. *Digit. Discov.* **2024**, *3*, 2607-17. DOI
13. Ghafarollahi, A.; Buehler, M. J. Rapid and automated alloy design with graph neural network-powered large language model-driven multi-agent AI. *MRS. Bull.* **2025**, *50*, 1309-24. DOI
14. Zhang, H.; Song, Y.; Hou, Z.; Miret, S.; Liu, B. HoneyComb: a flexible LLM-based agent system for materials science. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024 Oct; Miami, USA. Association for Computational Linguistics; 2024. pp. 3369-82. DOI
15. Bazgir, A.; Praneeth Madugula, R.; Zhang, Y. Multicrossmodal automated agent for integrating diverse materials science data. *arXiv* **2025**, arXiv:2505.15132. https://doi.org/10.48550/arXiv.2505.15132. (accessed 2026-05-19).
16. Zhou, L.; Ling, H.; Yan, K.; et al. Toward greater autonomy in materials discovery agents: unifying planning, physics, and scientists. *arXiv* **2025**, arXiv:2506.05616. https://doi.org/10.48550/arXiv.2506.05616. (accessed 2026-05-19).
17. Boiko, D. A.; Macknight, R.; Kline, B.; Gomes, G. Autonomous chemical research with large language models. *Nature* **2023**, *624*, 570-8. DOI PubMed PMC
18. Szymanski, N. J.; Rendy, B.; Fei, Y.; et al. An autonomous laboratory for the accelerated synthesis of inorganic materials. *Nature* **2023**, *624*, 86-91. DOI PubMed PMC
19. Zhang, Y.; Khan, S. A.; Mahmud, A.; et al. Exploring the role of large language models in the scientific method: from hypothesis to discovery. *npj. Artif. Intell.* **2025**, *1*, 14. DOI
20. SAE Standard. J3016_202104 - Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. 2021. DOI
21. Schilling-Wilhelmi, M.; Ríos-García, M.; Shabih, S.; et al. From text to insight: large language models for chemical data extraction. *Chem. Soc. Rev.* **2025**, *54*, 1125-50. DOI
22. Ramos, M. C.; Collison, C. J.; White, A. D. A review of large language models and autonomous agents in chemistry. *Chem. Sci.* **2025**, *16*, 2514-72. DOI PubMed PMC
23. Wei, J.; Wang, X.; Schuurmans, D.; et al. Chain-of-thought prompting elicits reasoning in large language models. *arXiv* **2022**, arXiv:2201.11903. https://doi.org/10.48550/arXiv.2201.11903. (accessed 2026-05-19).
24. Zhang, Z.; Dai, Q.; Bo, X.; et al. A survey on the memory mechanism of large language model-based agents. *ACM. Trans. Inf. Syst.* **2025**, *43*, 1-47. DOI
25. Sim, M.; Vakili, M. G.; Strieth-Kalthoff, F.; et al. ChemOS 2.0: an orchestration architecture for chemical self-driving laboratories. *Matter* **2024**, *7*, 2959-77. DOI
26. Schmidgall, S.; Su, Y.; Wang, Z.; et al. Agent laboratory: using LLM agents as research assistants. *arXiv* **2025**, arXiv:2501.04227. https://doi.org/10.48550/arXiv.2501.04227. (accessed 2026-05-19).
27. Kapoor, S.; Stroebl, B.; Siegel, Z. S.; Nadgir, N.; Narayanan, A. AI agents that matter. *arXiv* **2024**, arXiv:2407.01502. https://doi.org/10.48550/arXiv.2407.01502. (accessed 2026-05-19).
28. Lu, C.; Lu, C.; Lange, R. T.; Foerster, J. N.; Clune, J.; Ha, D. The AI scientist: towards fully automated open-ended scientific discovery. *arXiv* **2024**, arXiv:2408.06292. https://doi.org/10.48550/arXiv.2408.06292. (accessed 2026-05-19).
29. Szymanski, N. J.; Bartel, C. J.; Zeng, Y.; Diallo, M.; Kim, H.; Ceder, G. Adaptively driven X-ray diffraction guided by machine learning for autonomous phase identification. *npj. Comput. Mater.* **2023**, *9*, 31. DOI
30. Kumbhar, S.; Mishra, V.; Coutinho, K.; Handa, D.; Iquebal, A.; Baral, C. Hypothesis generation for materials discovery and design using goal-driven and constraint-guided LLM agents. In *Findings of the Association for Computational Linguistics: NAACL 2025*, 2025 Mar; Albuquerque, New Mexico. Association for Computational Linguistics; 2025. pp. 7524-55. DOI
31. Shi, T.; Li, Y.; Wang, Z.; et al. Knowledge-driven autonomous materials research via collaborative multi-agent and robotic system. *Matter* **2026**, *9*, 102577. DOI
32. Pyzer-Knapp, E. O.; Manica, M.; Staar, P.; et al. Foundation models for materials discovery - current state and future directions. *npj. Comput. Mater.* **2025**, *11*, 61. DOI
33. Mishra, V.; Singh, S.; Ahlawat, D.; et al. Foundational large language models for materials research. *arXiv* **2024**, arXiv:2412.09560. https://doi.org/10.48550/arXiv.2412.09560. (accessed 2026-05-19).
34. Choi, J.; Nam, G.; Choi, J.; Jung, Y. A perspective on foundation models in chemistry. *JACS. Au.* **2025**, *5*, 1499-518. DOI PubMed PMC
35. Van, M. H.; Verma, P.; Zhao, C.; Wu, X. A survey of AI for materials science: foundation models, LLM agents, datasets, and tools. *arXiv* **2025**, arXiv:2506.20743. https://doi.org/10.48550/arXiv.2506.20743. (accessed 2026-05-19).
36. Li, C.; Ran, N.; Liu, J. Agentic material science. *J. Mater. Inf.* **2026**, *6*, 10. DOI
37. 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

38. Anstine, D. M.; Isayev, O. Generative models as an emerging paradigm in the chemical sciences. *J. Am. Chem. Soc.* **2023**, *145*, 8736-50. DOI PubMed PMC
39. Bran, A. M.; Cox, S.; Schilter, O.; Baldassari, C.; White, A. D.; Schwaller, P. Augmenting large language models with chemistry tools. *Nat. Mach. Intell.* **2024**, *6*, 525-35. DOI
40. Gupta, T.; Zaki, M.; Krishnan, N. M. A.; Mausam. MatSciBERT: a materials domain language model for text mining and information extraction. *npj. Comput. Mater.* **2022**, *8*, 102. DOI
41. Trewartha, A.; Walker, N.; Huo, H.; et al. Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science. *Patterns* **2022**, *3*, 100488. DOI PubMed PMC
42. Liu, Y.; Ott, M.; Goyal, N.; et al. RoBERTa: a robustly optimized BERT pretraining approach. *arXiv* **2019**, arXiv:1907.11692. https://doi.org/10.48550/arXiv.1907.11692. (accessed 2026-05-19).
43. Niyongabo Rubungo, A.; Arnold, C.; Rand, B. P.; Dieng, A. B. LLM-Prop: predicting the properties of crystalline materials using large language models. *npj. Comput. Mater.* **2025**, *11*, 186. DOI
44. Yu, B.; Baker, F. N.; Chen, Z.; Ning, X.; Sun, H. LlaSMol: advancing large language models for chemistry with a large-scale, comprehensive, high-quality instruction tuning dataset. *arXiv* **2024**, arXiv:2402.09391. https://doi.org/10.48550/arXiv.2402.09391. (accessed 2026-05-19).
45. Qiu, H.; Zhao, J.; Jing, E.; et al. Introducing PolySea: an LLM-based polymer smart evolution agent. *ChemRxiv* **2025**. DOI
46. Tian, S.; Jiang, X.; Wang, W.; et al. Steel design based on a large language model. *Acta. Mater.* **2025**, *285*, 120663. DOI
47. Yang, Z.; Lv, K.; Shu, J.; Li, Z.; Xiao, P. Incorporating molecular knowledge in large language models via multimodal modeling. *IEEE. Trans. Comput. Soc. Syst.* **2025**, *12*, 3660-70. DOI
48. Zholus, A.; Kuznetsov, M.; Schutski, R.; et al. BindGPT: a scalable framework for 3D molecular design via language modeling and reinforcement learning. *arXiv* **2024**, arXiv:2406.03686. https://doi.org/10.48550/arXiv.2406.03686. (accessed 2026-05-19).
49. Tan, Q.; Zhou, D.; Xia, P.; et al. ChemMLLM: chemical multimodal large language model. *arXiv* **2025**, arXiv:2505.16326. https://doi.org/10.48550/arXiv.2505.16326. (accessed 2026-05-19).
50. Kuenneth, C.; Ramprasad, R. polyBERT: a chemical language model to enable fully machine-driven ultrafast polymer informatics. *Nat. Commun.* **2023**, *14*, 4099. DOI PubMed PMC
51. Xu, C.; Wang, Y.; Barati Farimani, A. TransPolymer: a Transformer-based language model for polymer property predictions. *npj. Comput. Mater.* **2023**, *9*, 64. DOI
52. Chaudhari, A.; Guntuboina, C.; Huang, H.; Farimani, A. B. AlloyBERT: alloy property prediction with large language models. *Comput. Mater. Sci.* **2024**, *244*, 113256. DOI
53. Liu, X.; Sun, P.; Chen, S.; et al. Perovskite-LLM: knowledge-enhanced large language models for perovskite solar cell research. *arXiv* **2025**, arXiv:2502.12669. https://doi.org/10.48550/arXiv.2502.12669. (accessed 2026-05-19).
54. Huang, S.; Cole, J. M. BatteryBERT: a pretrained language model for battery database enhancement. *J. Chem. Inf. Model.* **2022**, *62*, 6365-77. DOI PubMed PMC
55. Zhao, J.; Huang, S.; Cole, J. M. OpticalBERT and OpticalTable-SQA: text- and table-based language models for the optical-materials domain. *J. Chem. Inf. Model.* **2023**, *63*, 1961-81. DOI PubMed PMC
56. Mok, D. H.; Back, S. Generative pretrained transformer for heterogeneous catalysts. *J. Am. Chem. Soc.* **2024**, *146*, 33712-22. DOI PubMed
57. Tang, Y.; Xu, W.; Cao, J.; et al. MatterChat: a multi-modal LLM for material science. *arXiv* **2025**, arXiv:2502.13107. https://doi.org/10.48550/arXiv.2502.13107. (accessed 2026-05-19).
58. Antunes, L. M.; Butler, K. T.; Grau-Crespo, R. Crystal structure generation with autoregressive large language modeling. *Nat. Commun.* **2024**, *15*, 10570. DOI PubMed PMC
59. Yang, Y.; Shi, R.; Li, Z.; et al. BatGPT-Chem: a foundation large model for chemical engineering. *Research* **2025**, *8*, 0827. DOI PubMed PMC
60. Jain, A.; Ong, S. P.; Hautier, G.; et al. Commentary: The Materials Project: a materials genome approach to accelerating materials innovation. *APL. Mater.* **2013**, *1*, 011002. DOI
61. Talirz, L.; Kumbhar, S.; Passaro, E.; et al. Materials Cloud, a platform for open computational science. *Sci. Data.* **2020**, *7*, 299. DOI PubMed PMC
62. Scheidgen, M.; Himanen, L.; Ladines, A. N.; et al. NOMAD: a distributed web-based platform for managing materials science research data. *J. Open. Source. Softw.* **2023**, *8*, 5388. DOI
63. Esters, M.; Oses, C.; Divilov, S.; et al. aflow.org: a web ecosystem of databases, software and tools. *Comput. Mater. Sci.* **2023**, *216*, 111808. DOI
64. Venugopal, V.; Olivetti, E. MatKG: an autonomously generated knowledge graph in Material Science. *Sci. Data.* **2024**, *11*, 217. DOI PubMed PMC

65. Zhang, Y.; Chen, F.; Liu, Z.; et al. A materials terminology knowledge graph automatically constructed from text corpus. *Sci. Data.* **2024**, *11*, 600. DOI PubMed PMC
66. Statt, M. J.; Rohr, B. A.; Guevarra, D.; Breeden, J.; Suram, S. K.; Gregoire, J. M. The materials experiment knowledge graph. *Digit. Discov.* **2023**, *2*, 909-14. DOI
67. MongoDB Inc. MongoDB: the world's leading modern data platform. https://www.mongodb.com/. (accessed 2026-05-18).
68. PostgreSQL: the world's most advanced open source relational database. https://www.postgresql.org/. (accessed 2026-05-18).
69. Elementary Multiperspective Material Ontology (EMMO). 2025. https://github.com/emmo-repo/EMMO. (accessed 2026-05-18).
70. de Sainte Marie, C.; Iglesias Escudero, M.; Rosina, P. The ONTORULE Project: where ontology meets business rules. In *Web Reasoning and Rule Systems*. RR 2011. Lecture Notes in Computer Science, vol 6902; Springer, Berlin, Heidelberg: 2011. pp. 24-9. DOI
71. Premkumar, V.; Krishnamurty, S.; Wileden, J. C.; Grosse, I. R. A semantic knowledge management system for laminated composites. *Adv. Eng. Inform.* **2014**, *28*, 91-101. DOI
72. Douze, M.; Guzhva, A.; Deng, C.; et al. The Faiss library. *IEEE. Trans. Big. Data.* **2026**, *12*, 346-61. DOI
73. Wang, J.; Yi, X.; Guo, R.; et al. Milvus: a purpose-built vector data management system. In *Proceedings of the 2021 International Conference on Management of Data*. Association for Computing Machinery; 2021. pp. 2614-27. DOI
74. Qdrant. 2025. https://github.com/qdrant/qdrant. (accessed 2026-05-19).
75. Weaviate. 2025. https://github.com/weaviate/weaviate. (accessed 2026-05-19).
76. Neo4j. Graph Intelligence Platform. 2026. https://neo4j.com/. (accessed 2026-05-19).
77. Krech, D.; Grimnes, G. A.; Higgins, G.; et al. RDFLib. 2023. DOI
78. Packer, C.; Wooders, S.; Lin, K.; et al. MemGPT: towards LLMs as operating systems. *arXiv* **2023**, arXiv:2310.08560. https://doi.org/1.48550/arXiv.2310.08560. (accessed 2026-05-19).
79. LlamaIndex. 2026. https://www.llamaindex.ai/. (accessed 2026-05-19).
80. Haystack. 2019. https://github.com/deepset-ai/haystack. (accessed 2026-05-19).
81. SerpApi: Google Search API. https://serpapi.com/. (accessed 2026-05-19).
82. LangChain. 2022. https://github.com/langchain-ai/langchain. (accessed 2026-05-19).
83. Langgraph. 2025. https://github.com/langchain-ai/langgraph. (accessed 2026-05-19).
84. Wu, Q.; Bansal, G.; Zhang, J.; et al. AutoGen: enabling next-gen LLM applications via multi-agent conversation. *arXiv* **2023**, arXiv:2308.08155. https://doi.org/10.48550/arXiv.2308.08155. (accessed 2026-05-19).
85. crewAI. 2025. https://github.com/crewAIInc/crewAI. (accessed 2026-05-19).
86. Yao, S.; Zhao, J.; Yu, D.; et al. ReAct: synergizing reasoning and acting in language models. *arXiv* **2022**, arXiv:2210.03629. https://doi.org/10.48550/arXiv.2210.03629. (accessed 2026-05-19).
87. Shinn, N.; Cassano, F.; Berman, E.; Gopinath, A.; Narasimhan, K.; Yao, S. Reflexion: language agents with verbal reinforcement learning. *arXiv* **2023**, arXiv:2303.11366. https://doi.org/10.48550/arXiv.2303.11366. (accessed 2026-05-19).
88. Jain, A.; Ong, S. P.; Chen, W.; et al. FireWorks: a dynamic workflow system designed for high-throughput applications. *Concurr. Comput. Pract. Exp.* **2015**, *27*, 5037-59. DOI
89. Huber, S. P.; Zoupanos, S.; Uhrin, M.; et al. AiiDA 1.0, a scalable computational infrastructure for automated reproducible workflows and data provenance. *Sci. Data.* **2020**, *7*, 300. DOI PubMed PMC
90. Sundberg, J. D.; Benjamin, S. S.; Mcrae, L. M.; Warren, S. C. Simmate: a framework for materials science. *J. Open. Source. Softw.* **2022**, *7*, 4364. DOI
91. Ward, L.; Pauloski, J. G.; Hayot-Sasson, V.; et al. Employing artificial intelligence to steer exascale workflows with colmena. *Int. J. High. Perform. Comput. Appl.* **2024**, *39*, 52-64. DOI
92. Ong, S. P.; Richards, W. D.; Jain, A.; et al. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **2013**, *68*, 314-9. DOI
93. Hafner, J.; Kresse, G. The Vienna AB-Initio Simulation Program VASP: an efficient and versatile tool for studying the structural, dynamic, and electronic properties of materials. In: Gonis, A.; Meike, A.; Turchi, P. E. A.; Editors. *Properties of Complex Inorganic Solids*. Springer US; 1997. pp. 69-82. DOI
94. Giannozzi, P.; Baroni, S.; Bonini, N.; et al. QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials. *J. Phys. Condens. Matter.* **2009**, *21*, 395502. DOI
95. Gonze, X.; Amadon, B.; Anglade, P.; et al. ABINIT: first-principles approach to material and nanosystem properties. *Comput. Phys. Commun.* **2009**, *180*, 2582-615. DOI
96. Mortensen, J. J.; Larsen, A. H.; Kuisma, M.; et al. GPAW: an open Python package for electronic structure calculations. *J. Chem. Phys.* **2024**, *160*, 092503. DOI
97. Thompson, A. P.; Aktulga, H. M.; Berger, R.; et al. LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Comput. Phys. Commun.* **2022**, *271*, 108171. DOI

98. Abraham, M. J.; Murtola, T.; Schulz, R.; et al. GROMACS: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1-2*, 19-25. DOI
99. Eastman, P.; Galvelis, R.; Peláez, R. P.; et al. OpenMM 8: molecular dynamics simulation with machine learning potentials. *J. Phys. Chem. B*. **2023**, *128*, 109-16. DOI PubMed PMC
100. Grecco, H. E.; Dartiailh, M. C.; Thalhammer-Thurner, G.; Bronger, T.; Bauer, F. PyVISA: the Python instrumentation package. *J. Open. Source. Softw.* **2023**, *8*, 5304. DOI
101. Pyzer-Knapp, E. O.; Pitera, J. W.; Staar, P. W. J.; et al. Accelerating materials discovery using artificial intelligence, high performance computing and robotics. *npj. Comput. Mater.* **2022**, *8*, 84. DOI
102. Wierenga, R. P.; Golas, S. M.; Ho, W.; Coley, C. W.; Esvelt, K. M. PyLabRobot: an open-source, hardware agnostic interface for liquid-handling robots and accessories. *Device* **2023**, *1*, 100111. DOI
103. Pydantic-Ai. 2025. https://github.com/pydantic/pydantic-ai. (accessed 2026-05-19).
104. Zhang, J.; Gan, J.; Wang, X.; et al. MatSciBench: benchmarking the reasoning ability of large language models in materials science. *arXiv* **2025**, arXiv:2510.12171. https://doi.org/10.48550/arXiv.2510.12171. (accessed 2026-05-19).
105. Cheung, J. J.; Shen, S.; Zhuang, Y.; Li, Y.; Ramprasad, R.; Zhang, C. MSQA: benchmarking LLMs on graduate-level materials science reasoning and knowledge. *arXiv* **2025**, arXiv:2505.23982. https://doi.org/10.48550/arXiv.2505.23982. (accessed 2026-05-19).
106. Liu, S.; Hu, B.; Ye, B.; Xu, J.; Srolovitz, D. J.; Wen, T. MatTools: benchmarking large language models for materials science tools. *arXiv* **2025**, arXiv:2505.10852. https://doi.org/10.48550/arXiv.2505.10852. (accessed 2026-05-19).
107. Yanguas-Gil, A.; Dearing, M. T.; Elam, J. W.; et al. Benchmarking large language models for materials synthesis: the case of atomic layer deposition. *arXiv* **2024**, arXiv:2412.10477. https://doi.org/10.48550/arXiv.2412.10477. (accessed 2026-05-19).
108. Riebesell, J.; Goodall, R. E. A.; Benner, P.; et al. A framework to evaluate machine learning crystal stability predictions. *Nat. Mach. Intell.* **2025**, *7*, 836-47. DOI
109. Li, H.; Fang, X.; Li, Y.; et al. RxnBench: a multimodal benchmark for evaluating large language models on chemical reaction understanding from scientific literature. *arXiv* **2025**, arXiv:2512.23565. https://doi.org/10.48550/arXiv.2512.23565. (accessed 2026-05-19).
110. Song, Z.; Lu, J.; Du, Y.; et al. Evaluating large language models in scientific discovery. *arXiv* **2025**, arXiv:2512.15567. https://doi.org/1.48550/arXiv.2512.15567. (accessed 2026-05-19).
111. Zhou, Y.; Wang, Y.; He, X.; et al. Scientists' first exam: probing cognitive abilities of MLLM via perception, understanding, and reasoning. *arXiv* **2025**, arXiv:2506.10521. https://doi.org/10.48550/arXiv.2506.10521. (accessed 2026-05-19).
112. LangSmith docs. https://docs.langchain.com/langsmith/home. (accessed 2026-05-19).
113. Evals. https://github.com/openai/evals. (accessed 2026-05-19).
114. Es, S.; James, J.; Espinosa-Anke, L.; Schockaert, S. Ragas: automated evaluation of retrieval augmented generation. *arXiv* **2023**, arXiv:2309.15217. https://doi.org/10.48550/arXiv.2309.15217. (accessed 2026-05-19).
115. Kononova, O.; He, T.; Huo, H.; Trewartha, A.; Olivetti, E. A.; Ceder, G. Opportunities and challenges of text mining in materials research. *iScience* **2021**, *24*, 102155. DOI PubMed PMC
116. Dai, H.; Li, C.; Coley, C. W.; Dai, B.; Song, L. Retrosynthesis prediction with conditional graph logic network. *arXiv* **2020**, arXiv:2001.01408. https://doi.org/10.48550/arXiv.2001.01408. (accessed 2026-05-19).
117. Volk, A. A.; Epps, R. W.; Yonemoto, D. T.; et al. AlphaFlow: autonomous discovery and optimization of multi-step chemistry using a self-driven fluidic lab guided by reinforcement learning. *Nat. Commun.* **2023**, *14*, 1403. DOI PubMed PMC
118. Huang, S.; Cole, J. M. A database of battery materials auto-generated using ChemDataExtractor. *Sci. Data.* **2020**, *7*, 260. DOI PubMed PMC
119. Yan, R.; Jiang, X.; Wang, W.; Dang, D.; Su, Y. Materials information extraction via automatically generated corpus. *Sci. Data.* **2022**, *9*, 401. DOI PubMed PMC
120. Liu, Q.; Polak, M. P.; Kim, S. Y.; et al. Beyond designer's knowledge: generating materials design hypotheses via large language models. *Acta. Mater.* **2025**, *297*, 121307. DOI
121. Ghafarollahi, A.; Buehler, M. J. SciAgents: automating scientific discovery through multi-agent intelligent graph reasoning. *arXiv* **2024**, arXiv:2409.05556. https://doi.org/10.48550/arXiv.2409.05556. (accessed 2026-05-19).
122. Hawizy, L.; Jessop, D. M.; Adams, N.; Murray-Rust, P. ChemicalTagger: a tool for semantic text-mining in chemistry. *J. Cheminform.* **2011**, *3*, 17. DOI PubMed PMC
123. Swain, M. C.; Cole, J. M. ChemDataExtractor: a toolkit for automated extraction of chemical information from the scientific literature. *J. Chem. Inf. Model.* **2016**, *56*, 1894-904. DOI PubMed
124. Kumar, P.; Kabra, S.; Cole, J. M. A database of stress-strain properties auto-generated from the scientific literature using ChemDataExtractor. *Sci. Data.* **2024**, *11*, 1273. DOI PubMed PMC

125. Tshitoyan, V.; Dagdelen, J.; Weston, L.; et al. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **2019**, *571*, 95-8. DOI PubMed
126. Weston, L.; Tshitoyan, V.; Dagdelen, J.; et al. Named entity recognition and normalization applied to large-scale information extraction from the materials science literature. *J. Chem. Inf. Model.* **2019**, *59*, 3692-702. DOI
127. Wang, W.; Jiang, X.; Tian, S.; et al. Alloy synthesis and processing by semi-supervised text mining. *npj. Comput. Mater.* **2023**, *9*, 183. DOI
128. Zhang, R.; Zhang, J.; Chen, Q.; et al. A literature-mining method of integrating text and table extraction for materials science publications. *Computa. Mater. Sci.* **2023**, *230*, 112441. DOI
129. An, Y.; Greenberg, J.; Kalinowski, A.; et al. Knowledge graph question answering for materials science (KGQA4MAT): developing natural language interface for metal-organic frameworks knowledge graph (MOF-KG) using LLM. *arXiv* **2023**, arXiv:2309.11361. https://doi.org/10.48550/arXiv.2309.11361. (accessed 2026-05-19).
130. Dagdelen, J.; Dunn, A.; Lee, S.; et al. Structured information extraction from complex scientific text with fine-tuned large language models. *Nat. Commun.* **2024**, *15*, 1418. DOI
131. Yi, G. H.; Choi, J.; Song, H.; et al. MaTableGPT: GPT-based table data extractor from materials science literature. *Adv. Sci.* **2025**, *12*, 2408221. DOI PubMed PMC
132. da Silva, V. T.; Rademaker, A.; Lionti, K.; et al. Automated, LLM enabled extraction of synthesis details for reticular materials from scientific literature. *arXiv* **2024**, arXiv:2411.03484. https://doi.org/10.48550/arXiv.2411.03484. (accessed 2026-05-19).
133. Song, Y.; Miret, S.; Zhang, H.; Liu, B. HoneyBee: progressive instruction finetuning of large language models for materials science. *arXiv* **2023**, arXiv:2310.08511. https://doi.org/10.48550/arXiv.2310.08511. (accessed 2026-05-19).
134. Wang, Q.; Downey, D.; Ji, H.; Hope, T. SciMON: scientific inspiration machines optimized for novelty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; 2024 Jul; Bangkok, Thailand. Association for Computational Linguistics; 2024. pp. 279-99. DOI
135. Lai, Z.; Pu, Y. PriM: principle-inspired material discovery through multi-agent collaboration. *arXiv* **2025**, arXiv:2504.08810. https://doi.org/10.48550/arXiv.2504.08810. (accessed 2026-05-19).
136. Macleod, B. P.; Parlane, F. G. L.; Morrissey, T. D.; et al. Self-driving laboratory for accelerated discovery of thin-film materials. *Sci. Adv.* **2020**, *6*, eaaz8867. DOI PubMed PMC
137. Boyar, O.; Priyadarsini, I.; Takeda, S.; Hamada, L. LLM-fusion: a novel multimodal fusion model for accelerated material discovery. *arXiv* **2025**, arXiv:2503.01022. https://doi.org/10.48550/arXiv.2503.01022. (accessed 2026-05-19).
138. Choudhary, K. ChatGPT Material Explorer: design and implementation of a custom GPT assistant for materials science applications. *Integr. Mater. Manuf. Innov.* **2025**, *14*, 276-83. DOI
139. Wang, A. Y.; Kauwe, S. K.; Murdock, R. J.; Sparks, T. D. Compositionally restricted attention-based network for materials property predictions. *npj. Comput. Mater.* **2021**, *7*, 77. DOI
140. Jha, D.; Ward, L.; Paul, A.; et al. ElemNet: deep learning the chemistry of materials from only elemental composition. *Sci. Rep.* **2018**, *8*, 17593. DOI PubMed PMC
141. Goodall, R. E. A.; Lee, A. A. Predicting materials properties without crystal structure: deep representation learning from stoichiometry. *Nat. Commun.* **2020**, *11*, 6280. DOI PubMed PMC
142. Wang, B.; Ouyang, Y.; Li, Y.; et al. MoMa: a modular deep learning framework for material property prediction. *arXiv* **2025**, arXiv:2502.15483. https://doi.org/10.48550/arXiv.2502.15483. (accessed 2026-05-19).
143. Taniai, T.; Igarashi, R.; Suzuki, Y.; et al. Crystalformer: infinitely connected attention for periodic structure encoding. *arXiv* **2024**, arXiv:2403.11686. https://doi.org/10.48550/arXiv.2403.11686. (accessed 2026-05-19).
144. Choudhary, K.; Decost, B. Atomistic line graph neural network for improved materials property predictions. *npj. Comput. Mater.* **2021**, *7*, 185. DOI
145. Xie, T.; Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* **2018**, *120*, 145301. DOI PubMed
146. Yan, K.; Liu, Y.; Lin, Y.; Ji, S. Periodic graph transformers for crystal material property prediction. *arXiv* **2022**, arXiv:2209.11807. https://doi.org/10.48550/arXiv.2209.11807. (accessed 2026-05-19).
147. Nouira, A.; Sokolovska, N.; Crivello, J. C. CrystalGAN: learning to discover crystallographic structures with generative adversarial networks. *arXiv* **2018**, arXiv:1810.11203. https://doi.org/10.48550/arXiv.1810.11203. (accessed 2026-05-19).
148. Bagal, V.; Aggarwal, R.; Vinod, P. K.; Priyakumar, U. D. MolGPT: molecular generation using a transformer-decoder model. *J. Chem. Inf. Model.* **2021**, *62*, 2064-76. DOI
149. Li, Z.; Cao, B.; Jiao, R.; et al. Materials generation in the era of artificial intelligence: a comprehensive survey. *arXiv* **2025**, arXiv:2505.16379. https://doi.org/10.48550/arXiv.2505.16379. (accessed 2026-05-19).

150. Xu, A.; Desai, R.; Wang, L.; Hope, G.; Ritz, E. PLaID++: a preference aligned language model for targeted inorganic materials design. *arXiv* **2025**, arXiv:2509.07150. https://doi.org/10.48550/arXiv.2509.07150. (accessed 2026-05-19).
151. Zeni, C.; Pinsler, R.; Zügner, D.; et al. A generative model for inorganic materials design. *Nature* **2025**, *639*, 624-32. DOI PubMed PMC
152. Karpovich, C.; Pan, E.; Olivetti, E. A. Deep reinforcement learning for inverse inorganic materials design. *Npj. Comput. Mater.* **2024**, *10*, 287. DOI
153. Cleeton, C.; Sarkisov, L. Inverse design of metal-organic frameworks using deep dreaming approaches. *Nat. Commun.* **2025**, *16*, 4806. DOI PubMed PMC
154. Zuo, Y.; Qin, M.; Chen, C.; et al. Accelerating materials discovery with Bayesian optimization and graph deep learning. *Mater. Today.* **2021**, *51*, 126-35. DOI
155. Huang, J.; Xing, Q.; Ji, J.; Yang, B. Code-generated graph representations using multiple LLM agents for material properties prediction. In *Proceedings of the 42nd International Conference on Machine Learning*. PMLR; 2025. pp. 25972-86. https://proceedings.mlr.press/v267/huang25an.html. (accessed 2026-05-19).
156. Ghafarollahi, A.; Buehler, M. J. Autonomous inorganic materials discovery via multi-agent physics-aware scientific reasoning. *arXiv* **2025**, arXiv:2508.02956. https://doi.org/10.48550/arXiv.2508.02956. (accessed 2026-05-19).
157. Hjorth Larsen, A.; Jørgen Mortensen, J.; Blomqvist, J.; et al. The atomic simulation environment - a Python library for working with atoms. *J. Phys. Condens. Matter.* **2017**, *29*, 273002. DOI
158. Chandrasekhar, A.; Farimani, A. B. Automating MD simulations for proteins using large language models: NAMD-agent. *arXiv* **2025**, arXiv:2507.07887. https://doi.org/10.48550/arXiv.2507.07887. (accessed 2026-05-19).
159. Wang, Z.; Huang, H.; Zhao, H.; et al. DREAMS: density functional theory based research engine for agentic materials simulation. *arXiv* **2025**, arXiv:2507.14267. https://doi.org/10.48550/arXiv.2507.14267. (accessed 2026-05-19).
160. Deng, B.; Zhong, P.; Jun, K.; et al. CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nat. Mach. Intell.* **2023**, *5*, 1031-41. DOI
161. Batzner, S.; Musaelian, A.; Sun, L.; et al. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat. Commun.* **2022**, *13*, 2453. DOI
162. Musaelian, A.; Batzner, S.; Johansson, A.; et al. Learning local equivariant representations for large-scale atomistic dynamics. *Nat. Commun.* **2023**, *14*, 579. DOI PubMed PMC
163. Zhang, D.; Liu, X.; Zhang, X.; et al. DPA-2: a large atomic model as a multi-task learner. *npj. Comput. Mater.* **2024**, *10*, 293. DOI PubMed PMC
164. Batatia, I.; Benner, P.; Chiang, Y.; et al. A foundation model for atomistic materials chemistry. *J. Chem. Phys.* **2025**, *163*, 184110. DOI
165. Yang, H.; Hu, C.; Zhou, Y.; et al. MatterSim: a deep learning atomistic model across elements, temperatures and pressures. *arXiv* **2024**, arXiv:2405.04967. https://doi.org/10.48550/arXiv.2405.04967. (accessed 2026-05-19).
166. Chen, C.; Ong, S. P. A universal graph deep learning interatomic potential for the periodic table. *Nat. Comput. Sci.* **2022**, *2*, 718-28. DOI PubMed
167. Li, W.; Bazant, M. Z.; Zhu, J. Phase-Field DeepONet: Physics-informed deep operator neural network for fast simulations of pattern formation governed by gradient flows of free-energy functionals. *Comput. Methods. Appl. Mech. Eng.* **2023**, *416*, 116299. DOI
168. Haghighat, E.; Raissi, M.; Moure, A.; Gomez, H.; Juanes, R. A physics-informed deep learning framework for inversion and surrogate modeling in solid mechanics. *Comput. Methods. Appl. Mech. Eng.* **2021**, *379*, 113741. DOI
169. Campbell, Q.; Cox, S.; Medina, J.; Watterson, B.; White, A. D. MDCrow: automating molecular dynamics workflows with large language models. *arXiv* **2025**, arXiv:2502.09565. https://doi.org/10.48550/arXiv.2502.09565. (accessed 2026-05-19).
170. Shi, Z.; Xin, C.; Huo, T.; et al. A fine-tuned large language model based molecular dynamics agent for code generation to obtain material thermodynamic parameters. *Sci. Rep.* **2025**, *15*, 10295. DOI PubMed PMC
171. Zou, Y.; Cheng, A. H.; Aldossary, A.; et al. El Agente: an autonomous agent for quantum chemistry. *Matter* **2025**, *8*, 102263. DOI
172. Zhang, T.; Liu, Z.; Xin, Y.; Jiao, Y. MooseAgent: a LLM based multi-agent framework for automating moose simulation. *arXiv* **2025**, arXiv:2504.08621. https://doi.org/10.48550/arXiv.2504.08621. (accessed 2026-05-19).
173. Ghafarollahi, A.; Buehler, M. J. AtomAgents: alloy design and discovery through physics-aware multi-modal multi-agent artificial intelligence. *arXiv* **2024**, arXiv:2407.10022. https://doi.org/10.48550/arXiv.2407.10022. (accessed 2026-05-19).
174. Chaudhari, A.; Ock, J.; Barati Farimani, A. Modular large language model agents for multi-task computational materials science. *Commun. Mater.* **2026**, *7*, 131. DOI
175. Gottweis, J.; Weng, W. H.; Daryin, A.; et al. Towards an AI co-scientist. *arXiv* **2025**, arXiv:2502.18864. https://doi.org/10.48550/arXiv.2502.18864. (accessed 2026-05-19).
176. Abolhasani, M.; Kumacheva, E. The rise of self-driving labs in chemical and materials sciences. *Nat. Synth.* **2023**, *2*, 483-92. DOI
177. Guha, S.; Mullick, A.; Agrawal, J.; et al. MatScIE: an automated tool for the generation of databases of methods and parameters used in the computational materials science literature. *Comput. Mater. Sci.* **2021**, *192*, 110325. DOI

178. Zhang, L.; Stricker, M. MatNexus: a comprehensive text mining and analysis suite for materials discovery. *SoftwareX* **2024**, *26*, 101654. DOI
179. Jalali, M.; Luo, Y.; Caulfield, L.; Sauter, E.; Nefedov, A.; Wöll, C. Large language models in electronic laboratory notebooks: transforming materials science research workflows. *Mater. Today. Commun.* **2024**, *40*, 109801. DOI
180. He, T.; Huo, H.; Bartel, C. J.; Wang, Z.; Cruse, K.; Ceder, G. Precursor recommendation for inorganic synthesis by machine learning materials similarity from scientific literature. *Sci. Adv.* **2023**, *9*, eadg8180. DOI PubMed PMC
181. Zheng, Z.; Zhang, O.; Nguyen, H. L.; et al. ChatGPT Research Group for Optimizing the Crystallinity of MOFs and COFs. *ACS. Cent. Sci.* **2023**, *9*, 2161-70. DOI PubMed PMC
182. Cissé, A.; Evangelopoulos, X.; Gusev, V. V.; Cooper, A. I. Language-based Bayesian optimization research assistant (BORA). In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligen*, Aug 16-22, 2025; Montreal, Canada. California: International Joint Conferences on Artificial Intelligence Organization; 2025. pp. 4967-75. DOI
183. Liu, T.; Astorga, N.; Seedat, N.; van der Schaar, M. Large language models to enhance bayesian optimization. *arXiv* **2024**, arXiv:2402.03921. https://doi.org/10.48550/arXiv.2402.03921. (accessed 2026-05-19).
184. Chang, C. Y.; Azvar, M.; Okwudire, C.; Kontar, R. A. LLINBO: trustworthy LLM-in-the-loop bayesian optimization. *arXiv* **2025**, arXiv:2505.14756. https://doi.org/10.48550/arXiv.2505.14756. (accessed 2026-05-19).
185. Yang, Z.; Wang, D.; Ge, L.; Wang, B.; Fu, T.; Li, Y. Reasoning BO: enhancing Bayesian optimization with long-context reasoning power of LLMs. *arXiv* **2025**, arXiv:2505.12833. https://doi.org/10.48550/arXiv.2505.12833. (accessed 2026-05-19).
186. Guo, T.; Guo, K.; Nan, B.; et al. What can large language models do in chemistry? A comprehensive benchmark on eight tasks. *arXiv* **2023**, arXiv:2305.18365. https://doi.org/10.48550/arXiv.2305.18365. (accessed 2026-05-19).
187. Yang, Y.; Liu, Z.; Wu, W.; et al. MaterialBrain: high-performance material synthesis extraction via human–AI-curated few-shot large language models. *J. Chem. Inf. Model.* **2025**, *66*, 228-45. DOI PubMed
188. Mcdermott, M. J.; Dwaraknath, S. S.; Persson, K. A. A graph-based network for predicting chemical reaction pathways in solid-state materials synthesis. *Nat. Commun.* **2021**, *12*, 3097. DOI PubMed PMC
189. Qu, Y.; Huang, K.; Yin, M.; et al. CRISPR-GPT for agentic automation of gene-editing experiments. *Nat. Biomed. Eng.* **2025**, *10*, 245-58. DOI PubMed PMC
190. Acharya, A.; Sharma, A. K.; Parker, D.; et al. LABMATE: language model based multi-agent system to accelerate catalysis experiments. In *Proceedings of the SC '25 Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis, Association for Computing Machinery*; 2025. pp. 607-15. DOI
191. Li, H.; Sarkar, S.; Lu, W.; et al. Collective intelligence for AI-assisted chemical synthesis. *Nature* **2026**, *651*, 107-15. DOI PubMed
192. Fan, H.; Huang, J.; Xu, J.; et al. AutoMEX: streamlining material extrusion with AI agents powered by large language models and knowledge graphs. *Mater. Design.* **2025**, *251*, 113644. DOI
193. Tao, H.; Wu, T.; Kheiri, S.; Aldeghi, M.; Aspuru-Guzik, A.; Kumacheva, E. Self-driving platform for metal nanoparticle synthesis: combining microfluidics and machine learning. *Adv. Funct. Mater.* **2021**, *31*, 2106725. DOI
194. Sadeghi, S.; Bateni, F.; Kim, T.; et al. Autonomous nanomanufacturing of lead-free metal halide perovskite nanocrystals using a self-driving fluidic lab. *Nanoscale* **2024**, *16*, 580-91. DOI PubMed
195. Burger, B.; Maffettone, P. M.; Gusev, V. V.; et al. A mobile robotic chemist. *Nature* **2020**, *583*, 237-41. DOI PubMed
196. Hysmith, H.; Foadian, E.; Padhy, S. P.; et al. The future of self-driving laboratories: from human in the loop interactive AI to gamification. *Digital. Discov.* **2024**, *3*, 621-36. DOI
197. Weng, Y.; Gao, L.; Zhu, L.; Huang, J. MatQnA: a benchmark dataset for multi-modal large language models in materials characterization and analysis. *arXiv* **2025**, arXiv:2509.11335. https://doi.org/10.48550/arXiv.2509.11335. (accessed 2026-05-19).
198. Wang, H.; Guo, L.; Huo, P.; et al. S1-MMAlign: a large-scale, multi-disciplinary dataset for scientific figure-text understanding. *arXiv* **2026**, arXiv:2601.00264. https://doi.org/10.48550/arXiv.2601.00264. (accessed 2026-05-19).
199. Botifoll, M.; Pinto-Huguet, I.; Rotunno, E.; et al. Artificial intelligence-assisted workflow for transmission electron microscopy: from data analysis automation to materials knowledge unveiling. *Adv. Mater.* **2025**, e06785. DOI PubMed
200. Davel, C.; Bassiri-Gharb, N.; Correa-Baena, J. Machine learning in X-ray diffraction for materials discovery and characterization. *Matter* **2025**, *8*, 102272. DOI
201. Cao, B.; Zheng, Z.; Liu, Y.; et al. XQueryer: an intelligent crystal structure identifier for powder X-ray diffraction. *Natl. Sci. Rev.* **2025**, *12*, nwaf421. DOI PubMed PMC
202. Li, C.; Han, X.; Yao, C.; Ban, X. MatSAM: efficient extraction of microstructures of materials via visual large model. *arXiv* **2024**, arXiv:2401.05638. https://doi.org/10.48550/arXiv.2401.05638. (accessed 2026-05-19).
203. Anker, A. S.; Butler, K. T.; Selvan, R.; Jensen, KMØ. Machine learning for analysis of experimental scattering and spectroscopy data in materials chemistry. *Chem. Sci.* **2023**, *14*, 14003-19. DOI PubMed PMC

204. Choudhary, K. MicroscopyGPT: generating atomic-structure captions from microscopy images of 2D materials with vision-language transformers. *J. Phys. Chem. Lett.* **2025**, *16*, 7028-35. DOI
205. Chen, Y.; Yang, R.; Zhang, Z.; Ahmed, M.; Wang, Y. A large-language-model assisted automated scale bar detection and extraction framework for scanning electron microscopic images. *arXiv* **2025**, arXiv:2510.11260. https://doi.org/10.48550/arXiv.2510.11260. (accessed 2026-05-19).
206. Fei, Y.; McDermott, M. J.; Rom, C. L.; Wang, S.; Ceder, G. Dara: automated multiple-hypothesis phase identification and refinement from powder X-ray diffraction. *Chem. Mater.* **2026**, *38*, 1364-76. DOI
207. Bazgir, A.; Zhang, Y. Drug discovery agent: an automated vision detection system for drug-cell interactions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, CVPR Workshops 2025, 2025; pp. 4308-16. https://openaccess.thecvf.com/content/CVPR2025W/CVDD/html/Bazgir_Drug_Discovery_Agent_An_Automated_Vision_Detection_System_for_Drug-Cell_CVPRW_2025_paper.html. (accessed 2026-05-19).
208. Bazgir, A.; Zhang, Y. Matagent: a human-in-the-loop multi-agent LLM framework for accelerating the material science discovery cycle. In *AI for Accelerated Materials Design-ICLR 2025*, 2025. https://openreview.net/pdf/7c4f3b61beb01e7f4740671b78cd1f777bd0e60a.pdf . (accessed 2026-05-19).
209. Yao, L.; Samantray, S.; Ghosh, A.; et al. Operationalizing serendipity: multi-agent AI workflows for enhanced materials characterization with theory-in-the-loop. *arXiv* **2025**, arXiv:2508.06569. https://doi.org/10.48550/arXiv.2508.06569. (accessed 2026-05-19).
210. Yang, Y.; Tang, Y.; Chen, Y.; et al. AutoMat: enabling automated crystal structure reconstruction from microscopy via agentic tool use. *arXiv* **2025**, arXiv:2505.12650. https://doi.org/10.48550/arXiv.2505.12650. (accessed 2026-05-19).
211. Noh, H.; Lee, N.; Na, G. S.; Kim, K.; Park, C. IR-agent: expert-inspired LLM agents for structure elucidation from infrared spectra. *arXiv* **2025**, arXiv:2508.16112. https://doi.org/10.48550/arXiv.2508.16112. (accessed 2026-05-19).
212. Mandal, I.; Soni, J.; Zaki, M.; et al. Evaluating large language model agents for automation of atomic force microscopy. *Nat. Commun.* **2025**, *16*, 9104. DOI
213. Darvish, K.; Skreta, M.; Zhao, Y.; et al. ORGANA: a robotic assistant for automated chemistry experimentation and characterization. *Matter* **2025**, *8*, 101897. DOI
214. Jia, S.; Zhang, C.; Fung, V. LLMatDesign: autonomous materials discovery with large language models. *arXiv* **2024**, arXiv:2406.13163. https://doi.org/10.48550/arXiv.2406.13163. (accessed 2026-05-19).
215. Nikolaev, P.; Hooper, D.; Webber, F.; et al. Autonomy in materials research: a case study in carbon nanotube growth. *npj. Comput. Mater.* **2016**, *2*, 16031. DOI
216. Ye, X.; Yuan, W.; Fu, P.; et al. A full-process artificial intelligence framework for perovskite solar cells. *Sci. China. Mater.* **2025**, *68*, 2526-35. DOI
217. Ruan, Y.; Lu, C.; Xu, N.; et al. An automatic end-to-end chemical synthesis development platform powered by large language models. *Nat. Commun.* **2024**, *15*, 10160. DOI PubMed PMC
218. Zhang, B.; Li, X.; Xu, H.; Jin, Z.; Wu, Q.; Li, C. TopoMAS: large language model driven topological materials multiagent system. *arXiv* **2025**, arXiv:2507.04053. https://doi.org/10.48550/arXiv.2507.04053. (accessed 2026-05-19).
219. Lee, J.; Ely, J.; Zhang, K.; Ajith, A.; Campbell, C. R.; Choudhary, K. AGAPI-agents: an open-access agentic AI Platform For Accelerated Materials Design on AtomGPT.org. *arXiv* **2025**, arXiv:2512.11935. https://doi.org/10.48550/arXiv.2512.11935. (accessed 2026-05-19).
220. Shen, Y.; Wang, L.; Huang, Y.; et al. Unlocking azobenzene isomerization mechanisms *via* an LLM agent-driven workflow integrating simulation, experiment, and machine learning. *Chem. Sci.* **2026**. DOI
221. Juchli, D. SiLA 2: the next generation lab automation standard. In *Smart Biolabs of the Future*, Beutel, S., Lenk, F. Eds.; Springer International Publishing, 2022; pp. 147–74. DOI
222. Roch, L. M.; Häse, F.; Aspuru-Guzik, A. Chapter 16: ChemOS: an orchestration software to democratize autonomous discovery. In: Brown, N.; Editors. *Artificial Intelligence In Drug Discovery*. Cambridge: Royal Society of Chemistry; 2020. pp. 349-88. DOI
223. Quigley, M.; Conley, K.; Gerkey, B.; et al. ROS: an open-source Robot Operating System. In *ICRA workshop on open source software*, 2009. http://lars.mec.ua.pt/public/LAR%20Projects/BinPicking/2016_RodrigoSalgueiro/LIB/ROS/icraoss09-ROS.pdf. (accessed 2026-05-19).
224. Fawzy, S. M.; M. Ali MK, Allam NK. Artificial intelligence-driven materials design for next-generation sustainable energy technologies. *ACS. Sustain. Chem. Eng.* **2026**, *14*, 4745-61. DOI
225. Bouchard, D.; Chauhan, M. S.; Skarbrevik, D.; Ra, H. K.; Bajaj, V.; Ahmad, Z. UQLM: a Python package for uncertainty quantification in large language models. *arXiv* **2025**, arXiv:2507.06196. https://doi.org/10.48550/arXiv.2507.06196. (accessed 2026-05-19).
226. Krasecki, V. K.; Sharma, A.; Cavell, A. C.; et al. The role of experimental noise in a hybrid classical-molecular computer to solve combinatorial optimization problems. *ACS. Cent. Sci.* **2023**, *9*, 1453-65. DOI PubMed PMC

227. Zhang, J.; Choubey, P. K.; Huang, K. H.; Xiong, C.; Wu, C. S. Agentic uncertainty quantification. *arXiv* **2026**, arXiv:2601.15703. https://doi.org/10.48550/arXiv.2601.15703. (accessed 2026-05-19).
228. Bougzime, O.; Jabbar, S.; Cruz, C.; Demoly, F. Unlocking the potential of generative AI through neuro-symbolic architectures: benefits and limitations. *arXiv* **2025**, arXiv:2502.11269. https://doi.org/10.48550/arXiv.2502.11269. (accessed 2026-05-19).
229. Shen, Y.; Yang, Y.; Xi, Z.; et al. SciAgentGym: benchmarking multi-step scientific tool-use in LLM agents. *arXiv* **2026**, arXiv:2602.12984. https://doi.org/10.48550/arXiv.2602.12984. (accessed 2026-05-19).
230. Xu, W.; Zhou, Y.; Zhou, Y.; et al. Probing scientific general intelligence of LLMs with scientist-aligned workflows. *arXiv* **2025**, arXiv:2512.16969. https://doi.org/10.48550/arXiv.2512.16969. (accessed 2026-05-19).
231. Chaturvedi, S. S.; Bergerson, J.; Mallick, T. Toward reliable, safe, and secure LLMs for scientific applications. *arXiv* **2026**, arXiv:2603.18235. https://doi.org/10.48550/arXiv.2603.18235. (accessed 2026-05-19).
232. Zhao, H.; Tang, X.; Yang, Z.; et al. ChemSafetyBench: benchmarking LLM safety on chemistry domain. *arXiv* **2024**, arXiv:2411.16736. https://doi.org/10.48550/arXiv.2411.16736. (accessed 2026-05-19).