Review

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,*]

[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

ORCID: Tongyi Zhang (0000-0002-9646-9668), Yuyu Luo (0000-0001-9530-3327)

How to cite this article: 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:[Accept]. http://dx.doi.org/10.20517/jmi.2026.07

Received: 2 March 2026 | **Revised:** 7 April 2026 | **Accepted:** 30 April 2026

Abstract

Agents mostly built upon large language models (LLMs) with planning, tool-use,

memory, and self-reflection capabilities are revolutionizing all aspects of materials science and engineering (MSE), from materials design, experiment execution, industrial manufacture, to deployment, thereby opening the age of agentic MSE. Rather than isolated AI predictive models, these agents coordinate multi-step scientific workflows by retrieving and structuring knowledge, proposing and refining hypotheses, planning experiments, combining multi-modal simulations and characterizations, and, when integrated with AI materials laboratories, closing the loop toward autonomous discovery of materials. 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 key task families across all steps in MSE, including information retrieval, property prediction, simulation, synthesis, and characterization. By reviewing recent advances in agentic MSE, we reveal uneven progress. Knowledge-centric capabilities often remain early-stage, while experimental orchestration and characterization are starting to explore higher-level agent behaviors. Importantly, mature autonomy requires coordinating multiple tasks rather than optimizing a single one. Collectively, these insights provide a structured roadmap for advancing agentic MSE toward higher autonomy.

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

INTRODUCTION

Materials science and engineering 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 materials science and engineering 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, as well as autonomous and high-throughput experimental platforms. Due to the diversity of materials classes, such as metals, ceramics, polymers, semiconductors, and composites, and the inherently multi-modal and multi-scale nature of materials data, the emergence of AI for Materials Science and Engineering (AI4Mat) is motivated 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 materials science and engineering

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 planning, memory, tool use, and self-reflection capabilities[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 autonomous design and execution of intricate chemical tasks in both cloud-based and physical laboratory environments. Self-driving laboratories, exemplified by A-Lab[18], are advancing towards higher 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 unify data resources, computational tools, and experimental hardware within a cohesive framework for autonomous materials research[19].

We refer to *agentic materials science and engineering* 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 materials science and engineering*, which includes the design, evaluation, and governance of autonomous or semi-autonomous agents, presents considerable potential to accelerate discovery, ensure reproducibility, and collaborate 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 under unified agentic orchestration.

A hierarchical framework for agentic materials science and engineering

However, transitioning to such integrated systems reveals a significant challenge because the progression toward autonomy is highly uneven across the diverse landscape of materials science and engineering. 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 in 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 full autonomous control in another.

To rigorously evaluate this heterogeneous progress, a simple catalog of individual models or a binary classification of automated versus manual is insufficient. It fails to capture the nuance between a passive predictive model and an active, reasoning agent. Therefore, we advocate examining the field through a hierarchical taxonomy to map the varying degrees of agent autonomy against specific materials science tasks. Such a framework can provide a standardized metric to benchmark 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 materials science and engineering, we adopt the six-level hierarchy as shown in Figure 1. Similarly to the 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 roles*, together tracing the transition from human-only execution to fully autonomous scientific discovery.

- Level 0 - Human-Only. The baseline state of traditional research. The human acts as the sole executor, manually performing literature review, hypothesis formation, and experimentation. The agent's role is uninvolved yet.

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

- Level 2 - Tool-Augmented Agent. Agents move beyond passive assistance and begin to interact with external tools to accomplish scientific tasks. In this stage, the human researcher still defines the overall goal, but the agent can independently execute subtasks such as retrieving data, running simulations, or invoking domain-specific libraries without requiring step-by-step direction. This tool-augmented paradigm allows the agent to ground its reasoning in trusted computational resources, improving both reliability and scope. While humans remain responsible for high-level validation, the agent can propose plausible synthesis routes and predict 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, though they still require human oversight at key decision points. Researchers provide high-level goals, and the

agent uses chain-of-thought reasoning[23], long-term memory[24], and tool invocation to autonomously decompose the task 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 complete the closed loop from experimental design to execution and data collection. The agent 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, stepping in only at milestone checkpoints or when anomalies occur.

• Level 5 - AI Materials Scientist. This represents the ultimate vision of the "fully autonomous" AI materials scientist. The agent can independently complete the entire cycle of scientific research with virtually no human involvement: from formulating original hypotheses to planning 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 of 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 material 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 1 (b), development is uneven across tasks: All the 5 tasks have *established* up to Level 3. Synthesis and characterization have reached Level 4 *prototype* status, 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 3 follows exactly this two-dimensional task-level matrix, discussing each task family vertically across autonomy levels.

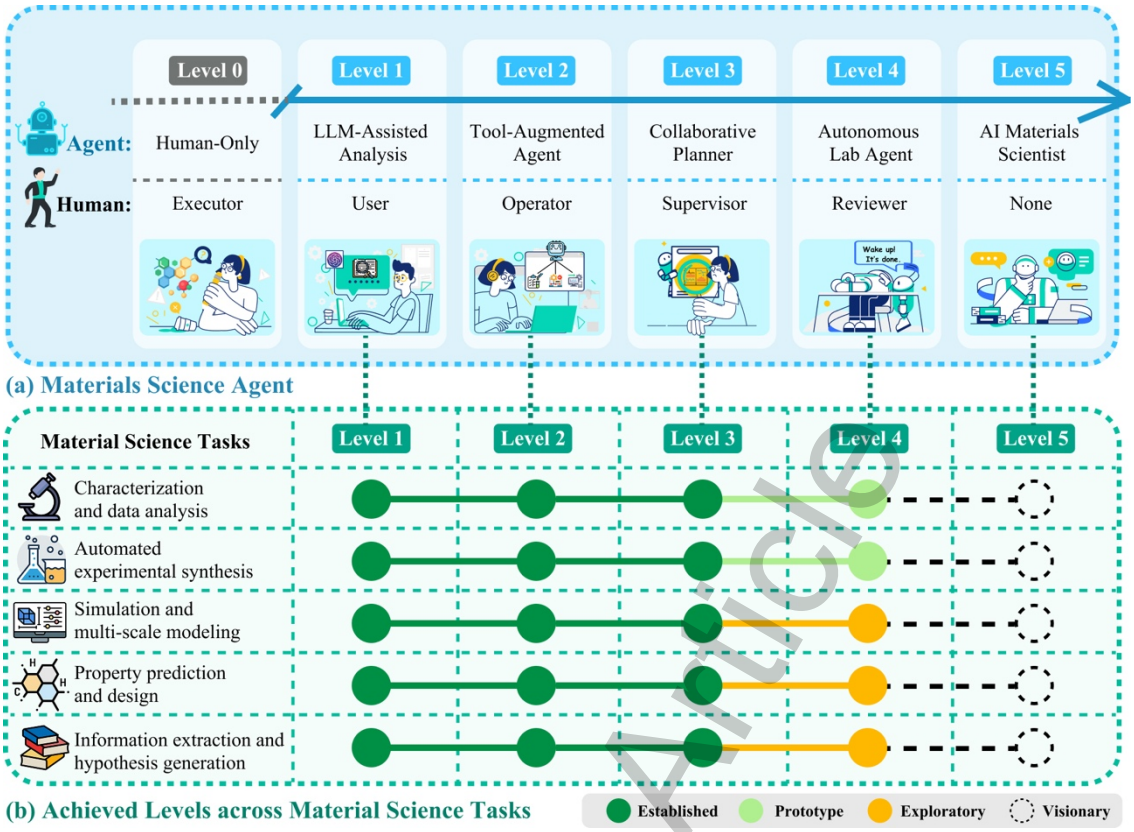

Figure 1. Overview of the six-level hierarchy for agentic materials science and engineering and its task mapping. (a) The progression from human-only workflows (Level 0) to fully autonomous AI materials scientists (Level 5). (b) Achieved autonomy levels across five core materials science tasks. Established: 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.

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, with 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.[36] 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 self-driving laboratories and, eventually, end-to-end autonomous materials creation. In contrast, our survey focuses on operationalizing autonomy[20] in a materials-grounded way: we introduce a six-level autonomy framework and a two-dimensional 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, it 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 lays 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 the open-source agentic infrastructure - that collectively defines the current boundary 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 between cognition-centric limitations in digital reasoning tasks and execution-centric limitations in physical experimental workflows. Based on this diagnosis, 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 2 lays the research foundation for agentic MSE from three angles: a comparison of 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 boundary of what can be agentized. Section 3 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 4 identifies current challenges and proposes targeted research directions.

BACKGROUND: TASK OVERVIEW AND RESEARCH FOUNDATIONS

Key task families across materials science and engineering 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 materials science and engineering 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 a consistent terminology for subsequent analysis, we formalize five fundamental tasks that together 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 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 the learning of 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 Multi-Scale 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 adjust parameters adaptively 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 XRD, XPS, 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 the fundamental transformation of materials science and engineering workflows. This schematic contrasts the human-centric traditional approach (all Subfigures (a) of Figure 2), characteristic of Level 0 and Level 1, with the emerging agentic materials science and engineering research (all Subfigures (b) of Figure 2).

In prior materials science and engineering workflows, research tasks were mostly linear and handled separately. As illustrated in the flow diagram on the left (all Subfigures (a) of Figure 2), the human scientist acted as the only central processor who manually defined the objectives, designed experiments, executed the protocols, and interpreted the data. Feedback loops, such as the redesign of a synthesis recipe or the refinement of a hypothesis, relied entirely on human intuition and manual intervention[37,38]. This bottleneck restricted the throughput of discovery and often disconnected high-level reasoning from low-level execution. Similarly, in computational domains like simulation and characterization, data processing software operated as passive utilities, requiring continuous manual calibration and file manipulation.

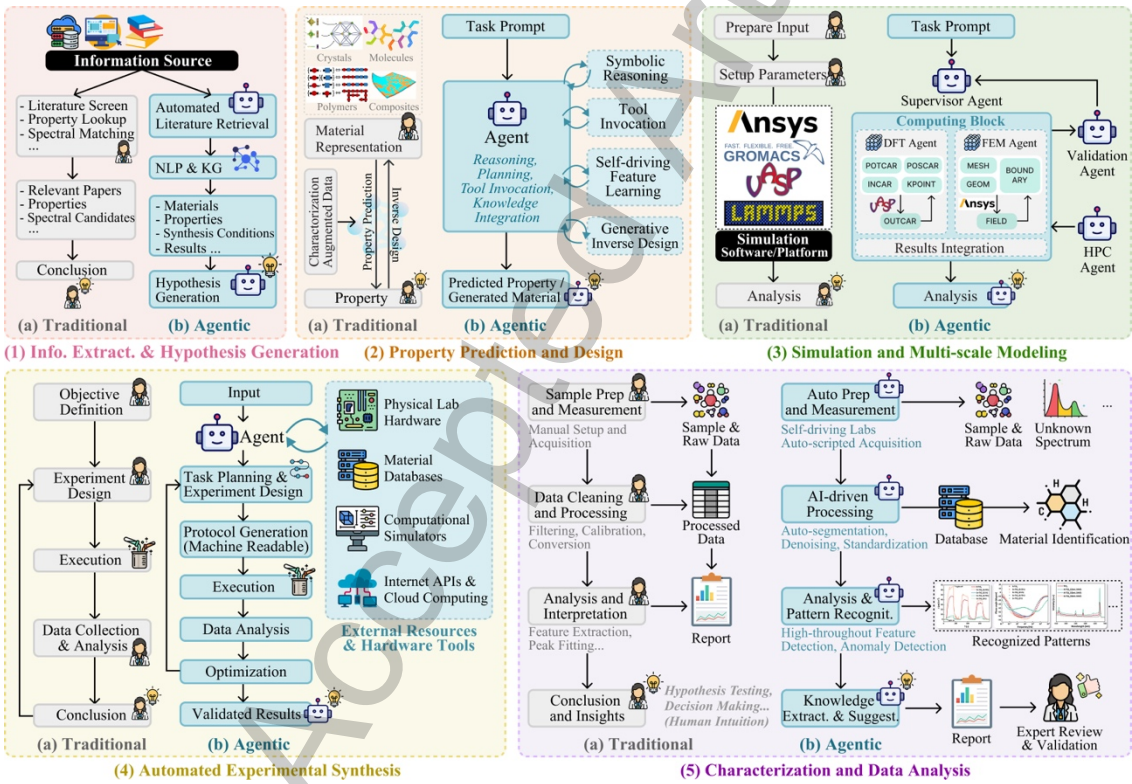

Figure 2. A comparative overview of materials research: traditional methods vs. agentic approaches across five representative materials research tasks: (1) information extraction and hypothesis generation, (2) property prediction and design, (3) simulation and multi-scale modeling, (4) automated experimental 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.

Central to the agentic paradigm is the shift from disjointed manual steps to unified reasoning and planning. Here, agents serve as the orchestration core, actively perceiving context to formulate multi-stage strategies. As demonstrated in the subfigures (b) of Figure 2, the agent acts as a dynamic hub that seamlessly integrates external resources ranging from Internet APIs and computational simulators to physical laboratory instruments[17,39]. This transforms software tools and hardware into active modules under agentic control. Most critically, this architecture establishes autonomous feedback loops. Whether performing inverse design for property prediction or optimizing synthesis parameters in real time, the system engages in iterative refinement. 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 constant human oversight.

Foundation models for materials science and engineering 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 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 materials science and engineering, and then summarize representative models across key materials sub-domains.

Development paradigms of domain-specific LLMs

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

Continued Pretraining (CPT). Current available large language models are pretrained on an extensive range of text corpora. To further develop generalized scientific priors in a specific domain of interest, continued pretraining is one key approach. The established

LLMs will be exposed to vast corpora of scientific texts, databases, and structured materials knowledge before being adapted to downstream tasks. Models like MatSciBERT[40] and MatBERT[41] follow this path. Specifically, MatSciBERT is a materials-aware BERT model pretrained on a large, curated corpus (around 285M words) of materials science and engineering literature and initialized from SciBERT. Using RoBERTa-style[42] pretraining and domain-adaptive continuation, it achieves lower validation perplexity and sets state-of-the-art results on three downstream tasks. Compared to general-purpose pre-trained large language models, these studies[40,41] focus on the fields of science, materials science, and engineering. They typically require millions of domain-specific documents for the models to learn the statistical structure of materials knowledge, including the composition-structure-property patterns, synthesis terminology, and common experimental or theoretical narratives. The learned prior knowledge is beneficial not only for natural language processing tasks but also for intelligent agent applications, where the pre-trained models serve as the core of decision-making and are integrated with simulation or experimental workflows.

Supervised Fine-Tuning (SFT). Since continuous pre-training methods require large amounts of data, and some niche domains may lack such data, another strategy for developing domain-specific large language models is supervised fine-tuning. Typically, a curated, labelled dataset defined on specific downstream tasks will be collected, such as property prediction[43], materials entity extraction[40], and reaction prediction[44]. Then, this dataset will be used to fine-tune existing LLMs in a supervised manner. In the field of materials science and engineering, supervised fine-tuning (SFT) is widely used to endow models with specific task capabilities that cannot be effectively learned solely from unlabeled corpora. Task-centric supervised fine-tuning 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 (RL). Reinforcement learning has been prevailing in aligning LLMs with specific preferences or human feedback. Also, it provides a more dynamic paradigm for the decision-making and iterative reasoning demands of agentic materials workflows. Specifically, the reward signals, including human preferences, rule-based

evaluators, or self-consistency critics, will be constructed in the RL framework to provide rewards to any LLM outputs. By rewarding coherent and actionable reasoning, RL fine-tuning helps LLMs to live up to the requirements of higher-level autonomy for agents operating in closed-loop computational or experimental environments.

Sub-Domain Specialized Models

Table 1 presents an overview of key models created within various materials science subfields. These models demonstrate the real-world applications of the described philosophies and emphasize the diversity of methods throughout the discipline. It included tailored models that incorporate priors for categories of materials, including molecules[47-49], polymers[45,50,51], alloys[46,52], perovskites[53], batteries[54], optical materials[55], and catalysts[56]. These models are different from generalized LLMs that seek to grasp a wide range of conceptual patterns but utilize specialized continued pretraining or fine-tuning to capture the chemical, structural, and processing features unique to each materials category. Consequently, these sub-domain LLMs provide more robust results compared to generalist models in downstream applications.

In the context of agentic systems, these domain-specific priors make them appropriate for incorporation into material science workflows. Depending on their functional role, they may operate either in cognitive functions like decision-making, planning and reasoning, or serving as execution tools like property predictors, structure analyzer, or simulation surrogates. Based on this distinction, we categorize the domain-specific LLMs in Table 1 into "Core Models" and "Agentic Tools" to better reflect their potential roles within agentic systems.

Table 1. Representative sub-domain specialized LLMs in materials science and engineering

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

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.

Agentic infrastructures for material science

Beyond domain-specific models, recent progress in agentic materials science and engineering 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 by themselves. Instead, as shown in Table 2, they provide callable components that LLM-based agents can invoke and coordinate to build scientific workflows. Importantly, 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 roles they support across knowledge, planning, execution, and evaluation.

Table 2. Representative resources supporting core functional components of agentic materials science and engineering

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 knowledge graphs & 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]	Role-based multi-agent planning
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]	Molecular Dynamics & 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 Discovey[108], RxnBench[109], SDE[110], SFE[111]	Domain benchmarks
	LangSmith[112], OpenAI Evals[113], Ragas[114]	General Evaluation Frameworks and Platforms

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 to 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, 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 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[64-66], and ontology systems[69-71], and general-purpose data storage solutions such as MongoDB[67] and PostgreSQL[68] help connect scientific entities, conditions, and relations across different sources. These mechanisms are important because materials workflows often require agents to accumulate evidence across multiple documents and data modalities rather than rely on a single query or static context. Other commonly used tools[76-81] are summarized in the 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 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, and code execution), and manage the flow of information between steps. They also help coordinate multiple roles or agents, so that tasks like literature search, data analysis, and result checking can be organized into a single pipeline. Beyond orchestration frameworks, the reasoning algorithms underlying these agents are equally important. ReAct[86] allows agents to interleave reasoning and tool use across workflow steps, so that they can gradually 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 later 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[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 jobs, pass structured inputs, monitor execution status, and collect outputs for downstream reasoning. They can also call common simulation codes[88-92], ab initio calculations[93-96], molecular dynamics[97-99], and material screening tools[100-103].

Evaluation and Benchmarking. Existing evaluation practices for scientific LLMs and agents span a spectrum from static, capability-unit benchmarks (e.g., domain QA[104-106], materials synthesis[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 evaluation for discovery workflows[110], hierarchical multimodal evaluation from localized perception to full document synthesis[109,110], cognition-oriented multimodal evaluation that decomposes scientific capability into perception, attribute understanding, and comparative reasoning across raw scientific data and multiple disciplines[111],and prospective, discovery oriented evaluation for stability screening with task relevant decision metrics[108], offering transferable principles for next generation benchmarking in agentic materials science and engineering. 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 materials science and engineering research: literature retrieval and hypothesis generation, property prediction and design, simulation and multi-scale 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 materials science and engineering. 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 at once 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, 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.

Information extraction and hypothesis generation

Materials science and engineering possess a vast corpus of scientific literature, yet turning this unstructured text into structured datasets and actionable knowledge remains a significant challenge. Information extraction (IE) tools are crucial for mining materials, properties, synthesis conditions, and performance metrics to build the databases that accelerate materials design and understanding[118,119]. Recent advances in NLP, from domain-tuned language models[40] to multi-agent systems[120,121], are pushing the progress of information extraction beyond simple data gathering towards 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. We summarize representative systems for this task in Table 3, including a brief comparison of their autonomy level, multi-agent setting, closed-loop capability, equipment integration, and open-source availability. Similar summary tables are provided for each task in the following sections, and we do not repeat this note thereafter.

Level 1. Research at Level 1 focuses on automating specific, labor-intensive tasks within the scientific workflow, particularly information extraction 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 like ChemicalTagger[122] utilize 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 like superalloys, semi-supervised frameworks[119] such as Action_extractor[127] are developed, with SFBC[128] further refining accuracy by combining dynamic and static embeddings.

These extraction efforts evolve into the construction of semantic Knowledge Graphs (KGs) and the integration of LLMs. Systems like MatKG[64] and MGED-KG[65] integrate entities into semantically linked networks, while MOF-KG[129] adds an LLM-powered interface for natural language querying. Concurrently, LLMs revolutionize extraction flexibility: Dunn et al.[130] utilize fine-tuned models for joint entity-relation extraction, 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, verified process. HoneyBee[133], an LLM progressively instruction-tuned for materials science and engineering, exemplifies this 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 insight 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 the generation of novel scientific hypotheses. Operating as collaborative planners, they often employ multi-agent architectures to explore vast knowledge spaces. Liu et al.[120] demonstrated that LLMs coupled with prompt engineering can generate valid materials design hypotheses that extend beyond the explicit knowledge of human designers. By integrating diverse scientific principles, the model successfully proposed novel high-entropy alloys and halide solid electrolytes, which were subsequently validated by 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 knowledge graph to reveal hidden interdisciplinary relationships, generating hypotheses with a precision that surpasses traditional methods. Furthermore, SciMON[134] ensures the novelty of these 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 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 that includes explicit design goals and constraints (e.g., cost, equipment availability), it presents effectiveness in planning synthesis routes and experimental procedures that are not only scientifically plausible but also practically feasible. This 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 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 like 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" operates 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.

Table 3. Representative Systems for Information Extraction and Hypothesis Generation Notes: ✓ = present; ✗ = absent; △= partial or simulated integration

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	✓	✓	✗	✓	✓

Property prediction and design

Advances in AI-driven property prediction and materials design are transforming how researchers discover and optimize new materials. Accurately predicting a material's 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 the ascending levels of agent autonomy they exhibit, 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 like 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 like 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, demonstrating Level 1 autonomy by assisting in property prediction and design tasks, but lack capabilities for autonomous planning, tool invocation, or integration to 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 like PolyBERT[50]. Together, these established models provide a unified computational substrate for agents to rapidly evaluate candidates from structure or composition to the property of interest.

- *Generative Design Tools*: These tools include early generative models like CrystalGAN[147] and MolGPT[148] that demonstrated the feasibility of generative

modeling for crystalline and molecular systems, respectively, and 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.[152] autonomously explores chemical space in simulation to discover compounds. Similarly, the "deep dreaming" approach for 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 like 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 work together 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 on a longer workflow. 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 planner agent can identify that the request implies multiple sub-tasks and organize the workflow accordingly. Moreover, during prediction, the agent does not only output a numeric value; it can also use chain-of-thought style reasoning to explain why a prediction is made and what factors drive the result, improving interpretability for downstream design decisions. Overall, Level 3 systems can form closed loops within the computational domain, but they still 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 finding and validation. A key step from Level 3 to Level 4 is to move beyond in silico optimization and achieve 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, so that predictions and decisions can be updated using real experimental feedback rather than remaining purely digital. Looking toward Level 5, agents should be able to select appropriate design and prediction strategies on their own and connect them with full experiment-compute 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.

Table 4. Representative Systems for Property Prediction and Design. Notes: ✓ = present; ✗ = absent; △= partial or simulated integration

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	✓	✓	✓	✓	✓

Simulation and multi-scale modeling

In materials science and engineering, simulation is a crucial tool for designing new materials, validating conceptual hypothesis 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 by hand, tune 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 prone to mistakes. 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 requests. One of the foundations of agentic material simulations is the automated material calculation frameworks including ASE[157], FireWorks[88], AiiDA[89]. These frameworks lay the groundwork for workflow definition and automation in material computations from Density Functional Theory (DFT), Molecular Dynamics (MD), to Finite Element Analysis (FEA). LLM-based agents then assisted with basic steps such as generating input files[158], validating structural data,

selecting relevant simulation parameters, or retrieving prior results from databases - thus reducing the cognitive load associated with manual model setup[159]. By handling these preparatory and post-processing steps, these Level 1 agents reduce the manual drudgery and cognitive load 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 to energies, forces, or field solutions, enabling orders-of-magnitude acceleration compared with first-principles or 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 span broader chemical and physical domains[163-166]. Another example is the application of physics-informed neural networks (PINNs) or neural PDE models in 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 the decision on simulation scale to human researchers. Therefore, these works are primarily categorized as Level 1 automation that automates or accelerates specific execution, which lay the foundation of 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, molecular dynamics 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 call, in what sequence, and with what inputs. This “Tool-Augmented” paradigm grounds the LLM's reasoning in rigorous

computational engines, overcoming the hallucination limitations of pure language models.

For example, MDCrow[169] operates on Level 2 autonomy by letting an LLM dynamically select and sequence MD tool calls like solvation, OpenMM execution, and MDTraj analyses under high-level user objectives. These MD-related basic simulation workflows were wrapped in an agentic toolset serving the objective in autonomously executing MD simulations to explore biochemical design space. The LLM serves as a chatting interface to coordinating, automating, and summarizing simulation steps. Similarly, MDAgents[170] employs a fine-tuned LLM to generate, validate, and execute MD simulation scripts, with simple feedback loops that allow 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 multi-agent systems plan and execute complex simulation workflows with minimal human intervention. DREAMS[159] exemplifies this with 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, handling execution and error monitoring. AtomAgents[173] introduced a physics-aware, multi-modal multi-agent architecture tailored for alloy design and discovery with multiple specialized agents collaboratively orchestrating 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 achieve them, often operating in a closed-loop within the simulation environment.

Conceptually, the tool usage can further span across different simulation scales instead of just using single simulation tools under the autonomous orchestration. The development of cross-scale simulation methods accelerated by AI, mentioned in Level 1,

can be invoked to form a practical simulation-based pipeline for material 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, agentic systems for simulation tasks at Level 3 will be capable of carrying out typical multi-scale material simulation workflows given high-level requests by coordinating tool usage and making decisive adjustments. However, despite their advanced planning capabilities, these systems remain confined to the digital realm, lacking a direct interface with the physical world.

Exploration of Level 4 and Level 5. Although Level 3 systems mainly operate inside the digital simulation environment, 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 the workflow based on intermediate results.

The key step toward Level 4 in simulation and multi-scale modeling is, therefore, not only better 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 choice of simulation method, and the next round of simulations can be scheduled and executed with minimal human intervention. In such a setting, the agent becomes a controller of an experiment--simulation loop, rather than a simulator-only planner[159,173]. Level 5 extends this idea to full autonomous scientific discovery, where the agent can connect multi-scale modeling with experimental evidence at the level of scientific reasoning[175]. When persistent gaps appear between prediction and observation, a Level 5 agent should be able to propose a plausible physical explanation, design an integrated simulation--experiment campaign to test it, and revise its models and hypotheses based on the outcomes. While this remains a long-term goal, it captures the ultimate fusion of computation and experimentation for autonomous materials discovery.

Table 5. Representative Systems for Simulation and Multi-scale Modeling. Notes: ✓ = present; ✗ = absent; △= partial or simulated integration

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	✓	△	✗	✓	✓
EI Agente[171]	2025	L3	✓	△	△	△	✓
DREAMS[159]	2025	L3	✓	△	△	✓	✓
MooseAgent[172]	2025	L3	✓	△	△	✓	✓
MatSciAgent[174]	2025	L3	✓	△	△	✓	✓

Automated experimental synthesis

The experimental synthesis and characterization of new materials is the most critical step in the material science and engineering research pipeline. It is also the most important method to validate the previous property prediction and simulation results. The advent of self-driving laboratories (SDLs) is redefining how materials are discovered and tested experimentally, moving experimentation from traditional manual work (Level 0) to autonomous, closed-loop operation[176]. By integrating with robotics, these platforms may 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 a knowledge-grounded AI system to assist physical synthesis that humans can interact with 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 or manuals. These systems convert unstructured descriptive text into structured, queryable databases of materials and synthesis protocols. Similarly, LLMs are also integrated into electronic experimental notebooks[179] to facilitate the digitization of everyday material experiments. Some preliminary work also utilizes the generative nature of LLMs to directly predict synthesis paths 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, and aiming for the transformation from experience-oriented experimental knowledge to machine-readable and reusable representations to improve efficiency and reproducibility. However, Level 1 level systems remain advisory and lack independent planning capabilities or the ability to invoke external tools.

Level 2. Building upon the knowledge-grounded LLM systems, Level 2 agents function as “digital chemists” by invoking external computational tools for experimental planning. In this level of autonomy, the LLM-based agentic systems not just rely on their prior knowledge in the pretraining step and prompts, but also invoke external information source or capabilities to assist the experiment design for material fabrication and

synthesis optimization. For example, active learning based methods can function as an external tool for the material fabrication parameter optimization[181] or combine Bayesian optimization with LLM to enhance contextual optimization[182-185]. For the synthesis path prediction, advanced LLMs demonstrate significant competence in predicting reaction outcomes and retrosynthetic routes[186]. External knowledge systems can be applied to enhance the synthesis route prediction[187]. This capability is further enhanced 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 like CRISPR-GPT[189] 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 or closed-loop hardware control.

Level 3. Level 3 advances to multi-agent coordination, where specialized agents manage in silico workflows akin to a human research group. These works signify the decomposition of a complex experimental design task into subtasks that might involve previous tasks. SciAgents[121] envision this as a "team of AIs" in which networks of agents collaborate to autonomously cycle through 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 literature review, simulation, data analysis, and hypothesis generation agents 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, and agrochemicals. Although these multi-agent planners can execute complex, multi-step computational workflows, they remain disconnected from equipment integration and require human-in-the-loop executions on experiments.

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 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 like Coscientist[17] and AutoMEX[192], which demonstrate the feasibility of using AI to autonomously issue commands to cloud lab robots and 3D printers. Other exploratory platforms leverage algorithmic optimization to guide hardware, such as the microfluidic systems for nanoparticle synthesis developed by Tao et al.[193] and Sadeghi et al.[194]. More advanced implementations, like AlphaFlow[117], utilize reinforcement learning to control modular reactors, successfully discovering 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 embody the early iterations of Level 4 intelligence. While they are primarily proof-of-concept systems often orchestrated by a central planner, they successfully validate the core criteria of 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 any experiment without constraints. However, such a system remains a distant prospect. In the foreseeable future, human experts will remain indispensable. The consensus within the field suggests that the evolution of SDLs will likely stabilize at a human-in-the-loop hybrid model rather than total replacement, as argued by Hysmith et al.[196]. Ideally, this synergy, guided by human intuition and transparent, interpretable AI, will transform the laboratory into an engine of innovation, allowing scientists to tackle challenges at a scale previously unimaginable.

Table 6. Representative Systems for Automated Experimental Synthesis. Notes: ✓ = present; ✗ = absent; △= partial or simulated integration

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	✓	✓	✓	✗	✓

Characterization and data analysis

Modern materials science and engineering rely on advanced characterization techniques that produce vast, complex datasets. These measurements probe material structures and properties across multiple length and time scales from complementary perspectives, posing significant challenges for consistent and knowledge-grounded interpretation. To overcome this challenge, relevant agents are introduced, 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 material characterization knowledge base and analytical assistants. For example, MatQnA[197] established a large multi-modal benchmark for material characterization techniques, including X-ray Photoelectron Spectroscopy (XPS),

X-ray Diffraction (XRD), Scanning Electron Microscopy (SEM), and Transmission Electron Microscopy (TEM). S1-MMAlign[198] collected large-scale multi-modal science image interpretation datasets, including material science data. Some AI systems were also developed for operation and interpretation of specific characterization techniques like 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 STEM images by mapping images directly to structured text. These systems fall into Level 1 as they facilitate AI-assisted information extraction from multi-modal material characterization data but function as passive resources without autonomous planning or execution capabilities.

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

published literature. AutoMat[210] uses a "plan then execute" design and integrates multiple tools to transform STEM image inputs into reconstructed atomic crystal structures. For spectroscopy, IR-Agent[211] follows human experts' reasoning and forms a team of agents for feature extraction, database retrieval, and final reasoning to infer molecular structures. For multimodal data from different sources, Bazgir et al.[15] propose a multi-agent framework with a dynamic gating mechanism. It can analyze microscopy images and simulation videos, and it also retrieves papers and web resources to provide context 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 decide how the procedure should proceed during execution. AdaptiveXRD[29] is an autonomous and adaptive XRD system. It enables agent-driven, real-time control of physical hardware and can make its own decisions to adjust the scan step size and scan range during measurement. For complex and precise instruments such as atomic force microscopy (AFM), 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 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 explore Level 4 autonomy, they remain early prototypes rather than widely adopted, 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, it would be able to pose useful research questions, choose suitable characterization methods, and combine evidence from multiple instruments to build a complete view of a new material. 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 is best treated as a reference point, while most

practical work should focus on strengthening the human-AI collaboration patterns seen in Levels 3 and 4.

Table 7. Representative Systems for Characterization and Data Analysis. Notes: ✓ = present; ✗ = absent; △= partial or simulated integration

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	✓	△	✓		✓	✓

Cross-task materials science and engineering 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 cycle from hypothesis generation to experimental validation.

Pioneering autonomous laboratories. Early Level 4 systems demonstrate that robotic platforms could leverage closed-loop machine learning to accelerate discovery. Pioneering examples like ARES[215] (for carbon nanotubes) and Ada[136] (for thin films) show that algorithms could design, execute, and analyze experiments faster than human researchers. This paradigm is significantly advanced by the A-Lab[18], which integrates computation, literature mining, and robotics to autonomously discover 41 new inorganic materials in 17 days. Similarly, full-process in silico frameworks, such as the Level 3 system for perovskite solar cells by Ye et al.[216], demonstrate how agents can digest 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 guiding 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 20+ materials APIs to autonomously run multi-step, tool-grounded workflows for reproducible, 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 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, multi-modal frameworks like 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 materials science and engineering 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. Materials science and engineering data is often sparse, heterogeneous, and highly structured, yet LLMs typically process it 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 failure, 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 to physical interaction and real-world control rather than pure language-based reasoning. 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 show that LLM agents can take extra actions that are outside 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 can appear both as risky physical actions beyond authorized limits and as functional code that exceeds the specified requirements, reflecting instruction drift during execution. Such behavior raises clear concerns for safety and for 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 important technical pathway for device coordination and standardization in material science labs. However, integrating agentic systems into these infrastructures remains non-trivial, as there still exists gap between the agent-level reasoning and device level communications. Looking forward, an additional challenge lies in enabling effective human-agent collaboration, as future laboratory environments are likely to involve hybrid workflows where 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 multi-agent systems (MAS), these uncertainties do not remain local. They can propagate across agents and even be amplified in a cascading manner: an incorrect assumption introduced during information extraction 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 only 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 demands 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 level.

Strategic directions for future research

To overcome these hurdles, future research must pivot from purely data-driven approaches toward developing 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 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 measuring the quality of the reasoning trajectory, including multi-step planning, tool selection, feedback utilization, error recovery, and avoidance of unproductive loops. Recent benchmark efforts begin to move in this direction. For example, SciAgentGym[229] explicitly evaluate long-horizon scientific tool-use and analyze 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 experiment, 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 materials science and engineering and for charting progress toward fully autonomous AI materials scientists.

Safety and governance. Finally, as agents move from advisory roles at lower levels to synthesis planning and direct physical execution at higher levels, their dual-use risks also become more serious, making safety and governance increasingly critical. Deploying 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, building 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 because of 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 context 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 materials science and engineering (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 do, what it must integrate, and where key gaps remain.

A central finding is that progress is uneven across tasks because each task family faces different limits in reasoning, tool integration, and safety constraints. This unevenness also reflects a broader workflow shift: traditional MSE work was often linear and handled as 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 grows, task boundaries become less clear: higher-level agents tend to combine multiple tasks into one connected process.

A brief cross-task comparison at Level 3 shows why a task-level lens is necessary. Multi-agent coordination emerges across tasks, but the bottlenecks differ: simulation focuses on stable orchestration of long tool chains, information tasks focus on evidence grounding and scientific validity, and synthesis/characterization are limited mainly by hardware interfaces, sensing, 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 advances, including physically grounded intelligence, stronger active perception and embodied interaction, better evaluation 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 serves as practical guides: they make progress measurable, clarify what “higher autonomy” demands in each task family, and support a disciplined move from isolated tools to reliable human-AI collaboration in real MSE workflows.

DECLARATIONS

Acknowledgement

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

Conceived and designed this 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

None.

Conflicts of interest

Tongyi Zhang is the Editor-in-Chief of the journal *Journal of Materials Informatics*, but was not involved in any steps of editorial processing, notably including reviewer selection, manuscript handling, and decision making, while the other authors have declared that they have 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 Conference Series: Materials Science and Engineering 2020, 768 (7), 072094. https://doi.org/10.1088/1757-899X/768/7/072094

2. Li, Y.; Gupta, V.; Kilic, M. N. T.; Choudhary, K.; Wines, D.; Liao, W.-k.; Choudhary, A.; Agrawal, A. Hybrid-LLM-GNN: integrating large language models and graph neural networks for enhanced materials property prediction. Digital Discovery 2025, 4 (2), 376-383. https://doi.org/10.1039/d4dd00199k

3. Olivetti, E. A.; Cole, J. M.; Kim, E.; Kononova, O.; Ceder, G.; Han, T. Y.-J.; Hiszpanski, A. M. Data-driven materials research enabled by natural language processing and information extraction. Applied Physics Reviews 2020, 7 (4). https://doi.org/10.1063/5.0021106

4. Reiser, P.; Neubert, M.; Eberhard, A.; Torresi, L.; Zhou, C.; Shao, C.; Metni, H.; van Hoesel, C.; Schopmans, H.; Sommer, T.; et al. Graph neural networks for materials science and chemistry. Commun Mater 2022, 3 (1), 93. https://doi.org/10.1038/s43246-022-00315-6

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 (N Y) 2021, 2, 100290. https://doi.org/10.1016/j.patter.2021.100290

6. Yoshitake, M.; Sato, F.; Kawano, H.; Teraoka, H. MaterialBERT for natural language processing of materials science texts. Science and Technology of Advanced Materials: Methods 2022, 2 (1), 372-380. https://doi.org/10.1080/27660400.2022.2124831

7. Kim, K.; Kang, S.; Yoo, J.; Kwon, Y.; Nam, Y.; Lee, D.; Kim, I.; Choi, Y.-S.; Jung, Y.; Kim, S.; et al. Deep-learning-based inverse design model for intelligent discovery of organic molecules. npj Computational Materials 2018, 4 (1). https://doi.org/10.1038/s41524-018-0128-1

8. Liu, Z.; Zhu, D.; Rodrigues, S. P.; Lee, K. T.; Cai, W. Generative Model for the Inverse Design of Metasurfaces. Nano Lett 2018, 18 (10), 6570-6576. https://doi.org/10.1021/acs.nanolett.8b03171

9. Curtarolo, S.; Hart, G. L.; Nardelli, M. B.; Mingo, N.; Sanvito, S.; Levy, O. The high-throughput highway to computational materials design. Nat Mater 2013, 12, 191-201. https://doi.org/10.1038/nmat3568

10. Hachmann, J.; Olivares-Amaya, R.; Atahan-Evrenk, S.; Amador-Bedolla, C.; Sánchez-Carrera, R. S.; Gold-Parker, A.; Vogt, L.; Brockway, A. M.; Aspuru-Guzik, A. The Harvard Clean Energy Project: Large-Scale Computational Screening and Design of Organic Photovoltaics on the World Community Grid. The Journal of Physical Chemistry

Letters 2011, 2 (17), 2241-2251. https://doi.org/10.1021/jz200866s
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 (10), 1613-1623. https://doi.org/10.1021/jz400215j
12. Ansari, M.; Moosavi, S. M. Agent-based learning of materials datasets from the scientific literature. Digital Discovery 2024, 3 (12), 2607-2617. https://doi.org/10.1039/d4dd00252k
13. Ghafarollahi, A.; Buehler, M. J. Rapid and automated alloy design with graph neural network-powered large language model-driven multi-agent AI. MRS Bulletin 2025, 50 (11), 1309-1324. https://doi.org/10.1557/s43577-025-00953-4
14. Zhang, H.; Song, Y.; Hou, Z.; Miret, S.; Liu, B. HoneyComb: A Flexible LLM-Based Agent System for Materials Science. Miami, Florida, USA, November, 2024; Association for Computational Linguistics: pp 3369-3382. https://doi.org/10.18653/v1/2024.findings-emnlp.192.
15. Bazgir, A.; Praneeth Madugula, R. c.; Zhang, Y. Multicrossmodal Automated Agent for Integrating Diverse Materials Science Data. arXiv 2025. https://doi.org/10.48550/arXiv.2505.15132.
16. Zhou, L.; Ling, H.; Yan, K.; Zhao, K.; Qian, X.; Arróyave, R.; Qian, X.; Ji, S. Toward Greater Autonomy in Materials Discovery Agents: Unifying Planning, Physics, and Scientists. CoRR 2025, abs/2506.05616. https://doi.org/10.48550/ARXIV.2506.05616
17. Boiko, D. A.; MacKnight, R.; Kline, B.; Gomes, G. Autonomous chemical research with large language models. Nature 2023, 624, 570-578. https://doi.org/10.1038/s41586-023-06792-0
18. Szymanski, N. J.; Rendy, B.; Fei, Y.; Kumar, R. E.; He, T.; Milsted, D.; McDermott, M. J.; Gallant, M.; Cubuk, E. D.; Merchant, A.; et al. An autonomous laboratory for the accelerated synthesis of novel materials. Nature 2023, 624, 86-91. https://doi.org/10.1038/s41586-023-06734-w
19. Zhang, Y.; Khan, S. A.; Mahmud, A.; Yang, H.; Lavin, A.; Levin, M.; Frey, J.; Dunnmon, J.; Evans, J.; Bundy, A.; et al. Exploring the role of large language models in the scientific method: from hypothesis to discovery. npj Artificial Intelligence 2025, 1 (1). https://doi.org/10.1038/s44387-025-00019-5
20. Practice, S. I. R. Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles. SAE Standard J3016_202104 2021. https://doi.org/10.4271/J3016_202104

21. Schilling-Wilhelmi, M.; Rios-Garcia, M.; Shabih, S.; Gil, M. V.; Miret, S.; Koch, C. T.; Marquez, J. A.; Jablonka, K. M. From text to insight: large language models for chemical data extraction. Chem Soc Rev 2025, 54 (3), 1125-1150. https://doi.org/10.1039/d4cs00913d

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, 2514-2572. https://doi.org/10.1039/d4sc03921a

23. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E. H.; Le, Q. V.; Zhou, D. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, 2022.

24. Zhang, Z.; Dai, Q.; Bo, X.; Ma, C.; Li, R.; Chen, X.; Zhu, J.; Dong, Z.; Wen, J.-R. A Survey on the Memory Mechanism of Large Language Model-based Agents. In ACM Trans. Inf. Syst., 11/, 2025; Vol. 43, pp 155:151-155:147. https://doi.org/10.1145/3748302.

25. Sim, M.; Vakili, M. G.; Strieth-Kalthoff, F.; Hao, H.; Hickman, R. J.; Miret, S.; Pablo-García, S.; Aspuru-Guzik, A. ChemOS 2.0: An orchestration architecture for chemical self-driving laboratories. Matter 2024, 7 (9), 2959-2977. https://doi.org/10.1016/j.matt.2024.04.022 (accessed 2026/04/25).

26. Schmidgall, S.; Su, Y.; Wang, Z.; Sun, X.; Wu, J.; Yu, X.; Liu, J.; Moor, M.; Liu, Z.; Barsoum, E. Agent laboratory: Using llm agents as research assistants. In Findings of the Association for Computational Linguistics: EMNLP 2025, 2025; pp 5977-6043.

27. Kapoor, S.; Stroebl, B.; Siegel, Z. S.; Nadgir, N.; Narayanan, A. AI Agents That Matter. Trans. Mach. Learn. Res. 2025, 2025.

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. CoRR 2024, abs/2408.06292. https://doi.org/10.48550/ARXIV.2408.06292

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 Computational Materials 2023, 9 (1), 31. https://doi.org/10.1038/s41524-023-00984-y

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; pp 7524-7555. https://doi.org/10.18653/V1/2025.FINDINGS-NAACL.420.

31. Shi, T.; Li, Y.; Wang, Z.; Xu, W.; Jiang, G.; Dai, D.; Zhou, J.; Huang, H.; He, R.; Ramakrishna, S.; et al. Knowledge-driven autonomous materials research via collaborative multi-agent and robotic system. Matter 2026, 9 (2). https://doi.org/10.1016/j.matt.2025.102577 (accessed 2026/04/26).

32. Pyzer-Knapp, E. O.; Manica, M.; Staar, P.; Morin, L.; Ruch, P.; Laino, T.; Smith, J. R.; Curioni, A. Foundation models for materials discovery - current state and future directions. npj Computational Materials 2025, 11 (1), 61. https://doi.org/10.1038/s41524-025-01538-0

33. Mishra, V.; Singh, S.; Ahlawat, D.; Zaki, M.; Bihani, V.; Grover, H. S.; Mishra, B.; Miret, S.; Mausam; Krishnan, N. M. A. Foundational Large Language Models for Materials Research. CoRR 2024, abs/2412.09560. https://doi.org/10.48550/ARXIV.2412.09560

34. Choi, J.; Nam, G.; Choi, J.; Jung, Y. A Perspective on Foundation Models in Chemistry. JACS Au 2025, 5 (4), 1499-1518. https://doi.org/10.1021/jacsau.4c01160

35. Van, M.-H.; Verma, P.; Zhao, C.; Wu, X. A Survey of AI for Materials Science: Foundation Models, LLM Agents, Datasets, and Tools. CoRR 2025, abs/2506.20743. https://doi.org/10.48550/ARXIV.2506.20743

36. Li, C.; Ran, N.; Liu, J. Agentic material science. Journal of Materials Informatics 2026, 6 (1), 10.

37. Tom, G.; Schmid, S. P.; Baird, S. G.; Cao, Y.; Darvish, K.; Hao, H.; Lo, S.; Pablo-García, S.; Rajaonson, E. M.; Skreta, M.; et al. Self-Driving Laboratories for Chemistry and Materials Science. Chemical Reviews 2024, 124 (16), 9633-9732. https://doi.org/10.1021/acs.chemrev.4c00055

38. Anstine, D. M.; Isayev, O. Generative Models as an Emerging Paradigm in the Chemical Sciences. Journal of the American Chemical Society 2023, 145 (16), 8736-8750. https://doi.org/10.1021/jacs.2c13467

39. M. Bran, A.; Cox, S.; Schilter, O.; Baldassari, C.; White, A. D.; Schwaller, P. Augmenting large language models with chemistry tools. Nature Machine Intelligence 2024, 6 (5), 525-535. https://doi.org/10.1038/s42256-024-00832-8

40. Gupta, T.; Zaki, M.; Krishnan, N. M. A.; Mausam. MatSciBERT: A materials domain

language model for text mining and information extraction. npj Computational Materials 2022, 8 (1). https://doi.org/10.1038/s41524-022-00784-w

41. Trewartha, A.; Walker, N.; Huo, H.; Lee, S.; Cruse, K.; Dagdelen, J.; Dunn, A.; Persson, K. A.; Ceder, G.; Jain, A. Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science. Patterns 2022, 3 (4). https://doi.org/10.1016/j.patter.2022.100488

42. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. CoRR 2019, abs/1907.11692.

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 Computational Materials 2025, 11 (1). https://doi.org/10.1038/s41524-025-01536-2

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. CoRR 2024, abs/2402.09391. https://doi.org/10.48550/ARXIV.2402.09391

45. Qiu, H.; Zhao, J.; Jing, E.; Hu, W.; Lv, Y.; Li, X.; Sun, Z.-Y. Introducing PolySea: An LLM-Based Polymer Smart Evolution Agent. ChemRxiv 2025, 2025 (0417). https://doi.org/doi:10.26434/chemrxiv-2025-zw65g

46. Tian, S.; Jiang, X.; Wang, W.; Jing, Z.; Zhang, C.; Zhang, C.; Lookman, T.; Su, Y. Steel design based on a large language model. Acta Materialia 2025, 285, 120663. https://doi.org/https://doi.org/10.1016/j.actamat.2024.120663

47. Yang, Z.; Lv, K.; Shu, J.; Li, Z.; Xiao, P. Incorporating Molecular Knowledge in Large Language Models via Multimodal Modeling. IEEE Transactions on Computational Social Systems 2025, 12 (5), 3660-3670. https://doi.org/10.1109/TCSS.2024.3506158

48. Zholus, A.; Kuznetsov, M.; Schutski, R.; Shayakhmetov, R.; Polykovskiy, D.; Chandar, S.; Zhavoronkov, A. BindGPT: A Scalable Framework for 3D Molecular Design via Language Modeling and Reinforcement Learning. In 2025.

49. Tan, Q.; Zhou, D.; Xia, P.; Liu, W.; Ouyang, W.; Bai, L.; Li, Y.; Fu, T. ChemMLLM: Chemical Multimodal Large Language Model. CoRR 2025, abs/2505.16326. https://doi.org/10.48550/ARXIV.2505.16326

50. Kuenneth, C.; Ramprasad, R. polyBERT: a chemical language model to enable fully machine-driven ultrafast polymer informatics. Nature Communications 2023, 14 (1),

4099. https://doi.org/10.1038/s41467-023-39868-6

51. Xu, C.; Wang, Y.; Barati Farimani, A. TransPolymer: a Transformer-based language model for polymer property predictions. npj Computational Materials 2023, 9 (1), 64. https://doi.org/10.1038/s41524-023-01016-5

52. Chaudhari, A.; Guntuboina, C.; Huang, H.; Farimani, A. B. AlloyBERT: Alloy Property Prediction with Large Language Models. CoRR 2024, abs/2403.19783. https://doi.org/10.48550/ARXIV.2403.19783

53. Liu, X.; Sun, P.; Chen, S.; Zhang, L.; Dong, P.; You, H.; Zhang, Y.; Yan, C.; Chu, X.; Zhang, T.-y. Perovskite-LLM: Knowledge-Enhanced Large Language Models for Perovskite Solar Cell Research. In Findings of the Association for Computational Linguistics: EMNLP 2025, 2025; pp 494-518.

54. Huang, S.; Cole, J. M. BatteryBERT: A Pretrained Language Model for Battery Database Enhancement. Journal of Chemical Information and Modeling 2022, 62 (24), 6365-6377. https://doi.org/10.1021/acs.jcim.2c00035

55. Zhao, J.; Huang, S.; Cole, J. M. OpticalBERT and OpticalTable-SQA: Text- and Table-Based Language Models for the Optical-Materials Domain. Journal of Chemical Information and Modeling 2023, 63 (7), 1961-1981. https://doi.org/10.1021/acs.jcim.2c01259

56. Mok, D. H.; Back, S. Generative Pretrained Transformer for Heterogeneous Catalysts. Journal of the American Chemical Society 2024, 146 (49), 33712-33722. https://doi.org/10.1021/jacs.4c11504

57. Tang, Y.; Xu, W.; Cao, J.; Ma, J.; Gao, W.; Farrell, S.; Erichson, N. B.; Mahoney, M. W.; Nonaka, A.; Yao, Z. MatterChat: A Multi-Modal LLM for Material Science. CoRR 2025, abs/2502.13107. https://doi.org/10.48550/ARXIV.2502.13107

58. Antunes, L. M.; Butler, K. T.; Grau-Crespo, R. Crystal structure generation with autoregressive large language modeling. Nature Communications 2024, 15 (1), 10570. https://doi.org/10.1038/s41467-024-54639-7

59. Yang, Y.; Shi, R.; Li, Z.; Jiang, S.; Lu, B.-L.; Zhao, Q.; Yang, Y.; Zhao, H. BatGPT-Chem: A Foundation Large Model for Chemical Engineering. Research 2025, 8, 0827. https://doi.org/doi:10.34133/research.0827

60. Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; et al. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. APL Materials 2013, 1 (1). https://doi.org/10.1063/1.4812323

61. Talirz, L.; Kumbhar, S.; Passaro, E.; Yakutovich, A. V.; Granata, V.; Gargiulo, F.; Borelli, M.; Uhrin, M.; Huber, S. P.; Zoupanos, S.; et al. Materials Cloud, a platform for open computational science. Sci Data 2020, 7, 299. https://doi.org/10.1038/s41597-020-00637-5

62. Scheidgen, M.; Himanen, L.; Ladines, A. N.; Sikter, D.; Nakhaee, M.; Fekete, Á.; Chang, T.; Golparvar, A.; Márquez, J. A.; Brockhauser, S.; et al. NOMAD: A distributed web-based platform for managing
materials science research data. Journal of Open Source Software 2023, 8 (90). https://doi.org/10.21105/joss.05388

63. Esters, M.; Oses, C.; Divilov, S.; Eckert, H.; Friedrich, R.; Hicks, D.; Mehl, M. J.; Rose, F.; Smolyanyuk, A.; Calzolari, A.; et al. aflow.org: A web ecosystem of databases, software and tools. Computational Materials Science 2023, 216. https://doi.org/10.1016/j.commatsci.2022.111808

64. Venugopal, V.; Olivetti, E. MatKG: An autonomously generated knowledge graph in Material Science. Sci Data 2024, 11, 217. https://doi.org/10.1038/s41597-024-03039-z

65. Zhang, Y.; Chen, F.; Liu, Z.; Ju, Y.; Cui, D.; Zhu, J.; Jiang, X.; Guo, X.; He, J.; Zhang, L.; et al. A materials terminology knowledge graph automatically constructed from text corpus. In Sci Data, Vol. 11; 2024; p 600. https://doi.org/10.1038/s41597-024-03448-0.

66. Statt, M. J.; Rohr, B. A.; Guevarra, D.; Breeden, J. N.; Suram, S. K.; Gregoire, J. M. The materials experiment knowledge graph. Digital Discovery 2023, 2 (4), 909-914. https://doi.org/10.1039/d3dd00067b

67. MongoDB Inc. MongoDB: The World's Leading Modern Data Platform. https://www.mongodb.com/ (accessed 2025-12-30).

68. The PostgreSQL Global Development Group. PostgreSQL: The World's Most Advanced Open Source Relational Database. https://www.postgresql.org/ (accessed 2025-12-30).

69. emmo-repo/EMMO. 2025. https://github.com/emmo-repo/EMMO (accessed 2025-12-30).

70. de Sainte Marie, C.; Iglesias Escudero, M.; Rosina, P. The ONTORULE Project : Where Ontology Meets Business Rules. Berlin, Heidelberg, 2011; Springer Berlin Heidelberg: pp 24-29.

71. Premkumar, V.; Krishnamurty, S.; Wileden, J. C.; Grosse, I. R. A semantic knowledge management system for laminated composites. Advanced Engineering Informatics 2014,

28 (1), 91-101. https://doi.org/https://doi.org/10.1016/j.aei.2013.12.004
72. Douze, M.; Guzhva, A.; Deng, C.; Johnson, J.; Szilvasy, G.; Mazaré, P.-E.; Lomeli, M.; Hosseini, L.; Jégou, H. The Faiss Library. IEEE Transactions on Big Data 2026, 12 (2), 346-361. https://doi.org/10.1109/tbdata.2025.3618474
73. Wang, J.; Yi, X.; Guo, R.; Jin, H.; Xu, P.; Li, S.; Wang, X.; Guo, X.; Li, C.; Xu, X.; et al. Milvus: A Purpose-Built Vector Data Management System. In 2021.
74. Qdrant/Qdrant. 2025. https://github.com/qdrant/qdrant (accessed 2025-12-30).
75. Weaviate. 2025. https://github.com/weaviate/weaviate (accessed 2025-12-30).
76. Neo4j Graph Intelligence Platform. 2026. https://neo4j.com/ (accessed 2026-02-28).
77. Krech, D., Gunnar AAstrand Grimnes, Graham Higgins, Jörn Hees, Iwan Aucamp, Niklas Lindström, Natanael Arndt, et al.,. RDFLib; 2023. https://doi.org/10.5281/zenodo.8206632.
78. Packer, C.; Fang, V.; Patil, S. G.; Lin, K.; Wooders, S.; Gonzalez, J. E. MemGPT: Towards LLMs as Operating Systems. CoRR 2023, abs/2310.08560. https://doi.org/10.48550/ARXIV.2310.08560
79. LlamaIndex. 2026. https://www.llamaindex.ai/ (accessed 2026-04-22).
80. Haystack. 2019. https://github.com/deepset-ai/haystack (accessed 2025-12-30).
81. SerpApi: Google Search API. https://serpapi.com/ (accessed 2025-12-30).
82. LangChain. 2022. https://github.com/langchain-ai/langchain (accessed 2025-12-30).
83. Langchain-Ai/Langgraph. 2025. https://github.com/langchain-ai/langgraph (accessed 2025-12-30).
84. Wu, Q.; Bansal, G.; Zhang, J.; Wu, Y.; Zhang, S.; Zhu, E.; Li, B.; Jiang, L.; Zhang, X.; Wang, C. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation Framework. CoRR 2023, abs/2308.08155. https://doi.org/10.48550/ARXIV.2308.08155
85. crewAIInc/crewAI. 2025. https://github.com/crewAIInc/crewAI (accessed 2025-12-30).
86. Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K. R.; Cao, Y. ReAct: Synergizing Reasoning and Acting in Language Models. In The Eleventh International Conference on Learning Representations, ICLR 2023, 2023.
87. Shinn, N.; Cassano, F.; Gopinath, A.; Narasimhan, K.; Yao, S. Reflexion: language agents with verbal reinforcement learning. In Advances in Neural Information Processing

Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, 2023.
88. Jain, A.; Ong, S. P.; Chen, W.; Medasani, B.; Qu, X.; Kocher, M.; Brafman, M.; Petretto, G.; Rignanese, G.-M.; Hautier, G.; et al. FireWorks: a dynamic workflow system designed for high-throughput applications. Concurrency and Computation: Practice and Experience 2015, 27 (17), 5037-5059. https://doi.org/https://doi.org/10.1002/cpe.3505
89. Huber, S. P.; Zoupanos, S.; Uhrin, M.; Talirz, L.; Kahle, L.; Häuselmann, R.; Gresch, D.; Müller, T.; Yakutovich, A. V.; Andersen, C. W.; et al. AiiDA 1.0, a scalable computational infrastructure for automated reproducible workflows and data provenance. Scientific Data 2020, 7 (1), 300. https://doi.org/10.1038/s41597-020-00638-4
90. Jack D. Sundberg, S. S. B., Lauren M. McRae,and Scott C. Warren,. Simmate: a framework for materials science. Journal of Open Source Software 2022, 7 (75), 4364. https://doi.org/10.21105/joss.04364
91. Ward, L. T.; Pauloski, J. G.; Hayot-Sasson, V.; Babuji, Y. N.; Brace, A.; Chard, R.; Chard, K.; Thakur, R.; Foster, I. T. Employing artificial intelligence to steer exascale workflows with colmena. Int. J. High Perform. Comput. Appl. 2025, 39 (1), 52-64. https://doi.org/10.1177/10943420241288242
92. Ong, S. P.; Richards, W. D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Chevrier, V. L.; Persson, K. A.; Ceder, G. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. Computational Materials Science 2013, 68, 314-319. https://doi.org/10.1016/j.commatsci.2012.10.028
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 Properties of Complex Inorganic Solids, Gonis, A., Meike, A., Turchi, P. E. A. Eds.; Springer US, 1997; pp 69-82. https://doi.org/10.1007/978-1-4615-5943-6_10.
94. Giannozzi, P.; Baroni, S.; Bonini, N.; Calandra, M.; Car, R.; Cavazzoni, C.; Ceresoli, D.; Chiarotti, G. L.; Cococcioni, M.; Dabo, I.; et al. QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials. Journal of Physics: Condensed Matter 2009, 21 (39), 395502. https://doi.org/10.1088/0953-8984/21/39/395502
95. Gonze, X.; Amadon, B.; Anglade, P. M.; Beuken, J. M.; Bottin, F.; Boulanger, P.; Bruneval, F.; Caliste, D.; Caracas, R.; Côté, M.; et al. ABINIT: First-principles approach to material and nanosystem properties. Computer Physics Communications 2009, 180

(12), 2582-2615. https://doi.org/https://doi.org/10.1016/j.cpc.2009.07.007
96. Mortensen, J. J.; Larsen, A. H.; Kuisma, M.; Ivanov, A. V.; Taghizadeh, A.; Peterson, A.; Haldar, A.; Dohn, A. O.; Schäfer, C.; Jónsson, E. Ö.; et al. GPAW: An open Python package for electronic structure calculations. The Journal of Chemical Physics 2024, 160 (9). https://doi.org/10.1063/5.0182685 (accessed 4/25/2026).
97. Thompson, A. P.; Aktulga, H. M.; Berger, R.; Bolintineanu, D. S.; Brown, W. M.; Crozier, P. S.; in 't Veld, P. J.; Kohlmeyer, A.; Moore, S. G.; Nguyen, T. D.; et al. LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. Computer Physics Communications 2022, 271, 108171. https://doi.org/https://doi.org/10.1016/j.cpc.2021.108171
98. Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX 2015, 1-2, 19-25. https://doi.org/https://doi.org/10.1016/j.softx.2015.06.001
99. Eastman, P.; Galvelis, R.; Peláez, R. P.; Abreu, C. R. A.; Farr, S. E.; Gallicchio, E.; Gorenko, A.; Henry, M. M.; Hu, F.; Huang, J.; et al. OpenMM 8: Molecular Dynamics Simulation with Machine Learning Potentials. The Journal of Physical Chemistry B 2024, 128 (1), 109-116. https://doi.org/10.1021/acs.jpcb.3c06662
100. Grecco, H. E.; Dartiailh, M. C.; Thalhammer-Thurner, G.; Bronger, T.; Bauer, F. PyVISA: the Python instrumentation package. Journal of Open Source Software 2023, 8 (84). https://doi.org/10.21105/joss.05304
101. Pyzer-Knapp, E. O.; Pitera, J. W.; Staar, P. W. J.; Takeda, S.; Laino, T.; Sanders, D. P.; Sexton, J.; Smith, J. R.; Curioni, A. Accelerating materials discovery using artificial intelligence, high performance computing and robotics. npj Computational Materials 2022, 8 (1). https://doi.org/10.1038/s41524-022-00765-z
102. Wierenga, R. P.; Golas, S.; Ho, W.; Coley, C.; Esvelt, K. M. PyLabRobot: An Open-Source, Hardware Agnostic Interface for Liquid-Handling Robots and Accessories. bioRxiv 2023. https://doi.org/10.1101/2023.07.10.547733
103. Pydantic/Pydantic-Ai. 2025. https://github.com/pydantic/pydantic-ai (accessed 2025-12-30).
104. Zhang, J.; Gan, J.; Wang, X.; Jia, Z.; Gu, C.; Chen, J.; Zhu, Y.; Ma, M. D.; Zhou, D.; Li, L.; et al. MatSciBench: Benchmarking the Reasoning Ability of Large Language Models in Materials Science. CoRR 2025, abs/2510.12171. https://doi.org/10.48550/ARXIV.2510.12171

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. CoRR 2025, abs/2505.23982. https://doi.org/10.48550/ARXIV.2505.23982
106. Liu, S.; Xu, J.; Ye, B.; Hu, B.; Srolovitz, D. J.; Wen, T. MatTools: Benchmarking Large Language Models for Materials Science Tools. CoRR 2025, abs/2505.10852. https://doi.org/10.48550/ARXIV.2505.10852
107. Yanguas-Gil, A.; Dearing, M. T.; Elam, J. W.; Jones, J. C.; Kim, S.; Mohammad, A.; Nguyen, C. T.; Sengupta, B. Benchmarking large language models for materials synthesis: the case of atomic layer deposition. CoRR 2024, abs/2412.10477. https://doi.org/10.48550/ARXIV.2412.10477
108. Riebesell, J.; Goodall, R. E. A.; Benner, P.; Chiang, Y.; Deng, B.; Ceder, G.; Asta, M.; Lee, A. A.; Jain, A.; Persson, K. A. A framework to evaluate machine learning crystal stability predictions. Nature Machine Intelligence 2025, 7 (6), 836-847. https://doi.org/10.1038/s42256-025-01055-1
109. Li, H.; Fang, X.; Li, Y.; Huang, C.; Wang, J.; Wang, X.; Bai, H.; Hao, B.; Lin, S.; Liang, H.; et al. RxnBench: A Multimodal Benchmark for Evaluating Large Language Models on Chemical Reaction Understanding from Scientific Literature. CoRR 2025, abs/2512.23565. https://doi.org/10.48550/ARXIV.2512.23565
110. Song, Z.; Lu, J.; Du, Y.; Yu, B.; Pruyn, T. M.; Huang, Y.; Guo, K.; Luo, X.; Qu, Y.; Qu, Y.; et al. Evaluating Large Language Models in Scientific Discovery. CoRR 2025, abs/2512.15567. https://doi.org/10.48550/ARXIV.2512.15567
111. Zhou, Y.; Wang, Y.; He, X.; Xiao, R.; Li, Z.; Feng, Q.; Guo, Z.; Yang, Y.; Wu, H.; Huang, W.; et al. Scientists' First Exam: Probing Cognitive Abilities of MLLM via Perception, Understanding, and Reasoning. CoRR 2025, abs/2506.10521. https://doi.org/10.48550/ARXIV.2506.10521
112. LangSmith docs. https://docs.langchain.com/langsmith/home (accessed 2026-04-07).
113. Openai/Evals. https://github.com/openai/evals (accessed 2025-12-30).
114. Es, S.; James, J.; Anke, L. E.; Schockaert, S. RAGAs: Automated Evaluation of Retrieval Augmented Generation. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - System Demonstrations, 2024; pp 150-158.
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 (3). https://doi.org/10.1016/j.isci.2021.102155 (accessed 2026/04/25).

116. Dai, H.; Li, C.; Coley, C. W.; Dai, B.; Song, L. Retrosynthesis Prediction with Conditional Graph Logic Network. In Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 2019; pp 8870-8880.

117. Volk, A. A.; Epps, R. W.; Yonemoto, D. T.; Masters, B. S.; Castellano, F. N.; Reyes, K. G.; Abolhasani, M. AlphaFlow: autonomous discovery and optimization of multi-step chemistry using a self-driven fluidic lab guided by reinforcement learning. Nature Communications 2023, 14 (1), 1403. https://doi.org/10.1038/s41467-023-37139-y

118. Huang, S.; Cole, J. M. A database of battery materials auto-generated using ChemDataExtractor. Scientific Data 2020, 7 (1), 260. https://doi.org/10.1038/s41597-020-00602-2

119. Yan, R.; Jiang, X.; Wang, W.; Dang, D.; Su, Y. Materials information extraction via automatically generated corpus. Scientific Data 2022, 9 (1), 401. https://doi.org/10.1038/s41597-022-01492-2

120. Liu, Q.; Polak, M. P.; Kim, S. Y.; Shuvo, M. D. A. A.; Deodhar, H. S.; Han, J.; Morgan, D.; Oh, H. Beyond designer's knowledge: Generating materials design hypotheses via large language models. CoRR 2024, abs/2409.06756. https://doi.org/10.48550/ARXIV.2409.06756

121. Ghafarollahi, A.; Buehler, M. J. SciAgents: Automating scientific discovery through multi-agent intelligent graph reasoning. CoRR 2024, abs/2409.05556. https://doi.org/10.48550/ARXIV.2409.05556

122. Hawizy, L.; Jessop, D. M.; Adams, N.; Murray-Rust, P. ChemicalTagger: A tool for semantic text-mining in chemistry. Journal of Cheminformatics 2011, 3 (1), 17. https://doi.org/10.1186/1758-2946-3-17

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 (10), 1894-1904. https://doi.org/10.1021/acs.jcim.6b00207

124. Kumar, P.; Kabra, S.; Cole, J. M. A Database of Stress-Strain Properties Auto-generated from the Scientific Literature using ChemDataExtractor. In Sci Data, Vol. 11; 2024; p 1273. https://doi.org/10.1038/s41597-024-03979-6.

125. Tshitoyan, V.; Dagdelen, J.; Weston, L.; Dunn, A.; Rong, Z.; Kononova, O.; Persson, K. A.; Ceder, G.; Jain, A. Unsupervised word embeddings capture latent knowledge from

materials science literature. In Nature, Vol. 571; 2019; pp 95-98. https://doi.org/10.1038/s41586-019-1335-8.

126. Weston, L.; Tshitoyan, V.; Dagdelen, J.; Kononova, O.; Trewartha, A.; Persson, K. A.; Ceder, G.; Jain, A. Named Entity Recognition and Normalization Applied to Large-Scale Information Extraction from the Materials Science Literature. J Chem Inf Model 2019, 59 (9), 3692-3702. https://doi.org/10.1021/acs.jcim.9b00470

127. Wang, W.; Jiang, X.; Tian, S.; Liu, P.; Lookman, T.; Su, Y.; Xie, J. Alloy synthesis and processing by semi-supervised text mining. npj Computational Materials 2023, 9 (1). https://doi.org/10.1038/s41524-023-01138-w

128. Zhang, R.; Zhang, J.; Chen, Q.; Wang, B.; Liu, Y.; Qian, Q.; Pan, D.; Xia, J.; Wang, Y.; Han, Y. A literature-mining method of integrating text and table extraction for materials science publications. Computational Materials Science 2023, 230. https://doi.org/10.1016/j.commatsci.2023.112441

129. An, Y.; Greenberg, J.; Kalinowski, A.; Zhao, X.; Hu, X.; Uribe-Romo, F. J.; Langlois, K.; Furst, J.; Gómez-Gualdrón, D. A. Knowledge Graph Question Answering for Materials Science (KGQA4MAT): Developing Natural Language Interface for Metal-Organic Frameworks Knowledge Graph (MOF-KG). CoRR 2023, abs/2309.11361. https://doi.org/10.48550/ARXIV.2309.11361

130. Dunn, A.; Dagdelen, J.; Walker, N.; Lee, S.; Rosen, A. S.; Ceder, G.; Persson, K. A.; Jain, A. Structured information extraction from complex scientific text with fine-tuned large language models. CoRR 2022, abs/2212.05238. https://doi.org/10.48550/ARXIV.2212.05238

131. Yi, G. H.; Choi, J.; Song, H.; Miano, O.; Choi, J.; Bang, K.; Lee, B.; Sohn, S. S.; Buttler, D.; Hiszpanski, A.; et al. MaTableGPT: GPT-Based Table Data Extractor from Materials Science Literature. In Adv Sci (Weinh), Vol. 12; 2025; p e2408221. https://doi.org/10.1002/advs.202408221.

132. Silva, V. T. d.; Rademaker, A.; Lionti, K.; Giro, R.; Lima, G.; Fiorini, S. R.; Archanjo, M.; Carvalho, B. W.; Neumann, R.; Souza, A.; et al. Automated, LLM enabled extraction of synthesis details for reticular materials from scientific literature. CoRR 2024, abs/2411.03484. https://doi.org/10.48550/ARXIV.2411.03484

133. Song, Y.; Miret, S.; Zhang, H.; Liu, B. HoneyBee: Progressive Instruction Finetuning of Large Language Models for Materials Science. In Findings of the Association for Computational Linguistics: EMNLP 2023, 2023; pp 5724-5739.

https://doi.org/10.18653/V1/2023.FINDINGS-EMNLP.380.
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), ACL 2024, 2024; pp 279-299. https://doi.org/10.18653/V1/2024.ACL-LONG.18.
135. Lai, Z.; Pu, Y. PriM: Principle-Inspired Material Discovery through Multi-Agent Collaboration. CoRR 2025, abs/2504.08810. https://doi.org/10.48550/ARXIV.2504.08810
136. MacLeod, B. P.; Parlane, F. G. L.; Morrissey, T. D.; Häse, F.; Roch, L. M.; Dettelbach, K. E.; Moreira, R.; Yunker, L. P. E.; Rooney, M. B.; Deeth, J. R.; et al. Self-driving laboratory for accelerated discovery of thin-film materials. Science Advances 2020, 6 (20), eaaz8867. https://doi.org/doi:10.1126/sciadv.aaz8867
137. Boyar, O.; Priyadarsini, I.; Takeda, S.; Hamada, L. LLM-Fusion: A Novel Multimodal Fusion Model for Accelerated Material Discovery. CoRR 2025, abs/2503.01022. https://doi.org/10.48550/ARXIV.2503.01022
138. Choudhary, K. ChatGPT Material Explorer: Design and Implementation of a Custom GPT Assistant for Materials Science Applications. Integrating Materials and Manufacturing Innovation 2025, 14 (3), 276-283. https://doi.org/10.1007/s40192-025-00410-9
139. Wang, A. Y.-T.; Kauwe, S. K.; Murdock, R. J.; Sparks, T. D. Compositionally restricted attention-based network for materials property predictions. npj Computational Materials 2021, 7 (1). https://doi.org/10.1038/s41524-021-00545-1
140. Jha, D.; Ward, L.; Paul, A.; Liao, W. K.; Choudhary, A.; Wolverton, C.; Agrawal, A. ElemNet: Deep Learning the Chemistry of Materials From Only Elemental Composition. In Sci Rep, Vol. 8; 2018; p 17593. https://doi.org/10.1038/s41598-018-35934-y.
141. Goodall, R. E. A.; Lee, A. A. Predicting materials properties without crystal structure: deep representation learning from stoichiometry. In Nat Commun, Vol. 11; 2020; p 6280. https://doi.org/10.1038/s41467-020-19964-7.
142. Wang, B.; Ouyang, Y.; Li, Y.; Wang, Y.; Cui, H.; Zhang, J.; Wang, X.; Ma, W.-Y.; Zhou, H. MoMa: A Modular Deep Learning Framework for Material Property Prediction. CoRR 2025, abs/2502.15483. https://doi.org/10.48550/ARXIV.2502.15483
143. Taniai, T.; Igarashi, R.; Suzuki, Y.; Chiba, N.; Saito, K.; Ushiku, Y.; Ono, K. Crystalformer: Infinitely Connected Attention for Periodic Structure Encoding. In The Twelfth International Conference on Learning Representations, ICLR 2024, 2024.

144. Choudhary, K.; DeCost, B. Atomistic Line Graph Neural Network for improved materials property predictions. npj Computational Materials 2021, 7 (1). https://doi.org/10.1038/s41524-021-00650-1

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 (14), 145301. https://doi.org/10.1103/PhysRevLett.120.145301

146. Yan, K.; Liu, Y.; Lin, Y.; Ji, S. Periodic Graph Transformers for Crystal Material Property Prediction. In Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, 2022.

147. Nouira, A.; Sokolovska, N.; Crivello, J.-C. CrystalGAN: Learning to Discover Crystallographic Structures with Generative Adversarial Networks. In Proceedings of the {AAAI} 2019 Spring Symposium on Combining Machine Learning with Knowledge Engineering (AAAI-MAKE 2019, 2019.

148. Bagal, V.; Aggarwal, R.; Vinod, P. K.; Priyakumar, U. D. MolGPT: Molecular Generation Using a Transformer-Decoder Model. J Chem Inf Model 2022, 62 (9), 2064-2076. https://doi.org/10.1021/acs.jcim.1c00600

149. Li, Z.; Cao, B.; Jiao, R.; Wang, L.; Wang, D.; Liu, Y.; Chen, D.; Li, J.; Liu, Q.; Rong, Y.; et al. Materials Generation in the Era of Artificial Intelligence: A Comprehensive Survey. CoRR 2025, abs/2505.16379. https://doi.org/10.48550/ARXIV.2505.16379

150. Xu, A.; Desai, R.; Wang, L.; Hope, G.; Ritz, E. PLaID++: A Preference Aligned Language Model for Targeted Inorganic Materials Design. CoRR 2025, abs/2509.07150. https://doi.org/10.48550/ARXIV.2509.07150

151. Zeni, C.; Pinsler, R.; Zügner, D.; Fowler, A.; Horton, M.; Fu, X.; Wang, Z.; Shysheya, A.; Crabbé, J.; Ueda, S.; et al. A generative model for inorganic materials design. Nature 2025, 639 (8055), 624-632. https://doi.org/10.1038/s41586-025-08628-5

152. Pan, E.; Karpovich, C.; Olivetti, E. A. Deep Reinforcement Learning for Inverse Inorganic Materials Design. CoRR 2022, abs/2210.11931. https://doi.org/10.48550/ARXIV.2210.11931

153. Cleeton, C.; Sarkisov, L. Inverse design of metal-organic frameworks using deep dreaming approaches. Nature Communications 2025, 16 (1), 4806. https://doi.org/10.1038/s41467-025-59952-3

154. Zuo, Y.; Qin, M.; Chen, C.; Ye, W.; Li, X.; Luo, J.; Ong, S. P. Accelerating materials discovery with Bayesian optimization and graph deep learning. Materials Today 2021,

51, 126-135. https://doi.org/https://doi.org/10.1016/j.mattod.2021.08.012
155. Huang, J.; Xing, Q.; Ji, J.; Yang, B. Code-Generated Graph Representations Using Multiple LLM Agents for Material Properties Prediction. In Forty-second International Conference on Machine Learning, ICML 2025, 2025.
156. Ghafarollahi, A.; Buehler, M. J. Autonomous Inorganic Materials Discovery via Multi-Agent Physics-Aware Scientific Reasoning. arXiv 2025. https://doi.org/10.48550/arXiv.2508.02956.
157. Hjorth Larsen, A.; Jørgen Mortensen, J.; Blomqvist, J.; Castelli, I. E.; Christensen, R.; Dułak, M.; Friis, J.; Groves, M. N.; Hammer, B.; Hargus, C.; et al. The atomic simulation environment - a Python library for working with atoms. Journal of Physics: Condensed Matter 2017, 29 (27), 273002. https://doi.org/10.1088/1361-648X/aa680e
158. Chandrasekhar, A.; Farimani, A. B. Automating MD simulations for Proteins using Large language Models: NAMD-Agent. CoRR 2025, abs/2507.07887. https://doi.org/10.48550/ARXIV.2507.07887
159. Wang, Z.; Huang, H.; Zhao, H.; Xu, C.; Zhu, S.; Janssen, J.; Viswanathan, V. DREAMS: Density Functional Theory Based Research Engine for Agentic Materials Simulation. CoRR 2025, abs/2507.14267. https://doi.org/10.48550/ARXIV.2507.14267
160. Deng, B.; Zhong, P.; Jun, K.; Riebesell, J.; Han, K.; Bartel, C. J.; Ceder, G. CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling. Nature Machine Intelligence 2023, 5 (9), 1031-1041. https://doi.org/10.1038/s42256-023-00716-3
161. Batzner, S.; Musaelian, A.; Sun, L.; Geiger, M.; Mailoa, J. P.; Kornbluth, M.; Molinari, N.; Smidt, T. E.; Kozinsky, B. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. Nature Communications 2022, 13 (1), 2453. https://doi.org/10.1038/s41467-022-29939-5
162. Musaelian, A.; Batzner, S.; Johansson, A.; Sun, L.; Owen, C. J.; Kornbluth, M.; Kozinsky, B. Learning local equivariant representations for large-scale atomistic dynamics. Nature Communications 2023, 14 (1), 579. https://doi.org/10.1038/s41467-023-36329-y
163. Zhang, D.; Liu, X.; Zhang, X.; Zhang, C.; Cai, C.; Bi, H.; Du, Y.; Qin, X.; Peng, A.; Huang, J.; et al. DPA-2: a large atomic model as a multi-task learner. npj Computational Materials 2024, 10 (1), 293. https://doi.org/10.1038/s41524-024-01493-2
164. Batatia, I.; Benner, P.; Chiang, Y.; Elena, A. M.; Kovács, D. P.; Riebesell, J.; Advincula, X. R.; Asta, M.; Avaylon, M.; Baldwin, W. J.; et al. A foundation model for

atomistic materials chemistry. The Journal of Chemical Physics 2025, 163 (18). https://doi.org/10.1063/5.0297006 (accessed 4/26/2026).
165. Yang, H.; Hu, C.; Zhou, Y.; Liu, X.; Shi, Y.; Li, J.; Li, G.; Chen, Z.; Chen, S.; Zeni, C.; et al. MatterSim: A Deep Learning Atomistic Model Across Elements, Temperatures and Pressures. arXiv 2024. https://doi.org/10.48550/arXiv.2405.04967.
166. Chen, C.; Ong, S. P. A universal graph deep learning interatomic potential for the periodic table. Nat. Comput. Sci. 2022, 2 (11), 718-728. https://doi.org/10.1038/S43588-022-00349-3
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. Computer Methods in Applied Mechanics and Engineering 2023, 416, 116299. https://doi.org/https://doi.org/10.1016/j.cma.2023.116299
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. Computer Methods in Applied Mechanics and Engineering 2021, 379, 113741. https://doi.org/https://doi.org/10.1016/j.cma.2021.113741
169. Campbell, Q. L.; Cox, S.; Medina, J.; Watterson, B.; White, A. D. MDCrow: Automating Molecular Dynamics Workflows with Large Language Models. CoRR 2025, abs/2502.09565. https://doi.org/10.48550/ARXIV.2502.09565
170. Shi, Z.; Xin, C.; Huo, T.; Jiang, Y.; Wu, B.; Chen, X.; Qin, W.; Ma, X.; Huang, G.; Wang, Z.; et al. A fine-tuned large language model based molecular dynamics agent for code generation to obtain material thermodynamic parameters. Scientific Reports 2025, 15 (1), 10295. https://doi.org/10.1038/s41598-025-92337-6
171. Zou, Y.; Cheng, A. H.; Aldossary, A.; Bai, J.; Leong, S. X.; Campos-Gonzalez-Angulo, J. A.; Choi, C.; Ser, C. T.; Tom, G.; Wang, A.; et al. El Agente: An autonomous agent for quantum chemistry. Matter 2025, 8 (7). https://doi.org/10.1016/j.matt.2025.102263 (accessed 2026/04/26).
172. Zhang, T.; Liu, Z.; Xin, Y.; Jiao, Y. MooseAgent: A LLM Based Multi-agent Framework for Automating Moose Simulation. CoRR 2025, abs/2504.08621. https://doi.org/10.48550/ARXIV.2504.08621
173. Ghafarollahi, A.; Buehler, M. J. AtomAgents: Alloy design and discovery through physics-aware multi-modal multi-agent artificial intelligence. CoRR 2024,

abs/2407.10022. https://doi.org/10.48550/ARXIV.2407.10022
174. Chaudhari, A.; Ock, J.; Barati Farimani, A. Modular large language model agents for multi-task computational materials science. Communications Materials 2026. https://doi.org/10.1038/s43246-025-00994-x
175. Gottweis, J.; Weng, W.-H.; Daryin, A. N.; Tu, T.; Palepu, A.; Sirkovic, P.; Myaskovsky, A.; Weissenberger, F.; Rong, K.; Tanno, R.; et al. Towards an AI co-scientist. CoRR 2025, abs/2502.18864. https://doi.org/10.48550/ARXIV.2502.18864
176. Abolhasani, M.; Kumacheva, E. The rise of self-driving labs in chemical and materials sciences. Nature Synthesis 2023, 2 (6), 483-492. https://doi.org/10.1038/s44160-022-00231-0
177. Guha, S.; Mullick, A.; Agrawal, J.; Ram, S.; Ghui, S.; Lee, S.-C.; Bhattacharjee, S.; Goyal, P. MatScIE: An automated tool for the generation of databases of methods and parameters used in the computational materials science literature. Computational Materials Science 2021, 192, 110325. https://doi.org/https://doi.org/10.1016/j.commatsci.2021.110325
178. Zhang, L.; Stricker, M. MatNexus: A comprehensive text mining and analysis suite for materials discovery. SoftwareX 2024, 26, 101654. https://doi.org/https://doi.org/10.1016/j.softx.2024.101654
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. Materials Today Communications 2024, 40, 109801. https://doi.org/https://doi.org/10.1016/j.mtcomm.2024.109801
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. Science Advances 2023, 9 (23), eadg8180. https://doi.org/doi:10.1126/sciadv.adg8180
181. Zheng, Z.; Zhang, O.; Nguyen, H. L.; Rampal, N.; Alawadhi, A. H.; Rong, Z.; Head-Gordon, T.; Borgs, C.; Chayes, J. T.; Yaghi, O. M. ChatGPT Research Group for Optimizing the Crystallinity of MOFs and COFs. ACS Central Science 2023, 9 (11), 2161-2170. https://doi.org/10.1021/acscentsci.3c01087
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 Intelligence, IJCAI 2025, 2025; pp 4967-4975. https://doi.org/10.24963/IJCAI.2025/553.

183. Liu, T.; Astorga, N.; Seedat, N.; Schaar, M. v. d. Large Language Models to Enhance Bayesian Optimization. In The Twelfth International Conference on Learning Representations, ICLR 2024, 2024.
184. Chang, C.-Y.; Azvar, M.; Okwudire, C. E.; Kontar, R. A. LLINBO: Trustworthy LLM-in-the-Loop Bayesian Optimization. CoRR 2025, abs/2505.14756. https://doi.org/10.48550/ARXIV.2505.14756
185. Yang, Z.; Ge, L.; Han, D.; Fu, T.; Li, Y. Reasoning BO: Enhancing Bayesian Optimization with Long-Context Reasoning Power of LLMs. CoRR 2025, abs/2505.12833. https://doi.org/10.48550/ARXIV.2505.12833
186. Guo, T.; Guo, K.; Nan, B.; Liang, Z.; Guo, Z.; Chawla, N. V.; Wiest, O.; Zhang, X. What can Large Language Models do in chemistry? A comprehensive benchmark on eight tasks. In Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, 2023.
187. Yang, Y.; Liu, Z.; Wu, W.; Zhang, Y.; Zhang, H.; Lin, J.; Wu, S.; Chen, Z.; Sun, Y.; Li, R.; et al. MaterialBrain: High-Performance Material Synthesis Extraction via Human-AI-Curated Few-Shot Large Language Models. Journal of Chemical Information and Modeling 2026, 66 (1), 228-245. https://doi.org/10.1021/acs.jcim.5c02299
188. McDermott, M. J.; Dwaraknath, S. S.; Persson, K. A. A graph-based network for predicting chemical reaction pathways in solid-state materials synthesis. Nature Communications 2021, 12 (1), 3097. https://doi.org/10.1038/s41467-021-23339-x
189. Qu, Y.; Huang, K.; Yin, M.; Zhan, K.; Liu, D.; Yin, D.; Cousins, H. C.; Johnson, W. A.; Wang, X.; Shah, M.; et al. CRISPR-GPT for agentic automation of gene-editing experiments. Nature Biomedical Engineering 2026, 10 (2), 245-258. https://doi.org/10.1038/s41551-025-01463-z
190. Acharya, A.; Sharma, A. K.; Parker, D.; Vega, T.; Ashraf, R. A.; Isenberg, N. M.; Strube, J.; Rallo, R. 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, SC Workshops 2025, 2025; pp 607-615. https://doi.org/10.1145/3731599.3767399.
191. Li, H.; Sarkar, S.; Lu, W.; Loftus, P. O.; Qiu, T.; Shee, Y.; Cuomo, A. E.; Webster, J.-P.; Kelly, H. R.; Manee, V.; et al. Collective intelligence for AI-assisted chemical synthesis. Nature 2026, 651 (8104), 107-115. https://doi.org/10.1038/s41586-026-

10131-4
192. Fan, H.; Huang, J.; Xu, J.; Zhou, Y.; Fuh, J. Y. H.; Lu, W. F.; Li, B. AutoMEX: Streamlining material extrusion with AI agents powered by large language models and knowledge graphs. Materials & Design 2025, 251, 113644. https://doi.org/https://doi.org/10.1016/j.matdes.2025.113644
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. Advanced Functional Materials 2021, 31 (51), 2106725. https://doi.org/https://doi.org/10.1002/adfm.202106725
194. Sadeghi, S.; Bateni, F.; Kim, T.; Son, D. Y.; Bennett, J. A.; Orouji, N.; Punati, V. S.; Stark, C.; Cerra, T. D.; Awad, R.; et al. Autonomous nanomanufacturing of lead-free metal halide perovskite nanocrystals using a self-driving fluidic lab. Nanoscale 2024, 16 (2), 580-591, 10.1039/D3NR05034C. https://doi.org/10.1039/D3NR05034C
195. Burger, B.; Maffettone, P. M.; Gusev, V. V.; Aitchison, C. M.; Bai, Y.; Wang, X.; Li, X.; Alston, B. M.; Li, B.; Clowes, R.; et al. A mobile robotic chemist. Nature 2020, 583 (7815), 237-241. https://doi.org/10.1038/s41586-020-2442-2
196. Hysmith, H.; Foadian, E.; Padhy, S. P.; Kalinin, S. V.; Moore, R. G.; Ovchinnikova, O. S.; Ahmadi, M. The future of self-driving laboratories: from human in the loop interactive AI to gamification. Digital Discovery 2024, 3 (4), 621-636, 10.1039/D4DD00040D. https://doi.org/10.1039/D4DD00040D
197. Weng, Y.; Gao, L.; Zhu, L.; Huang, J. MatQnA: A Benchmark Dataset for Multi-modal Large Language Models in Materials Characterization and Analysis. CoRR 2025, abs/2509.11335. https://doi.org/10.48550/ARXIV.2509.11335
198. Wang, H.; Guo, L.; Huo, P.; Lin, X.; Yuan, Y.; Jiang, J.; Liu, J. S1-MMAlign: A Large-Scale, Multi-Disciplinary Dataset for Scientific Figure-Text Understanding. CoRR 2026, abs/2601.00264. https://doi.org/10.48550/ARXIV.2601.00264
199. Botifoll, M.; Pinto-Huguet, I.; Rotunno, E.; Galvani, T.; Coll, C.; Kavkani, P. H.; Spadaro, M. C.; Niquet, Y.-M.; Eriksen, M. B.; Martí-Sánchez, S.; et al. Artificial Intelligence-Assisted Workflow for Transmission Electron Microscopy: From Data Analysis Automation to Materials Knowledge Unveiling. Advanced Materials n/a (n/a), e06785. https://doi.org/https://doi.org/10.1002/adma.202506785
200. Davel, C.; Bassiri-Gharb, N.; Correa-Baena, J.-P. Machine learning in X-ray diffraction for materials discovery and characterization. Matter 2025, 8 (9). https://doi.org/10.1016/j.matt.2025.102272 (accessed 2026/04/26).

201. Cao, B.; Zheng, Z.; Liu, Y.; Zhang, L.; Wong, L. W.-Y.; Weng, L.-T.; Li, J.; Li, H.; Zhang, T.-Y. XQueryer: an intelligent crystal structure identifier for powder X-ray diffraction. National Science Review 2025, 12 (12). https://doi.org/10.1093/nsr/nwaf421 (accessed 4/26/2026).
202. Li, C.; Han, X.; Yao, C.; Ban, X. MatSAM: Efficient Extraction of Microstructures of Materials via Visual Large Model. arXiv 2024. https://doi.org/10.48550/arXiv.2401.05638.
203. Anker, A. S.; Butler, K. T.; Selvan, R.; Jensen, K. M. Ø. Machine learning for analysis of experimental scattering and spectroscopy data in materials chemistry. Chemical Science 2023, 14 (48), 14003-14019, 10.1039/D3SC05081E. https://doi.org/10.1039/D3SC05081E
204. Choudhary, K. MicroscopyGPT: Generating Atomic-Structure Captions from Microscopy Images of 2D Materials with Vision-Language Transformers. The Journal of Physical Chemistry Letters 2025, 16 (27), 7028-7035. https://doi.org/10.1021/acs.jpclett.5c01257
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. CoRR 2025, abs/2510.11260. https://doi.org/10.48550/ARXIV.2510.11260
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. CoRR 2025, abs/2510.19667. https://doi.org/10.48550/ARXIV.2510.19667
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 4269-4277.
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.
209. Yao, L.; Samantray, S.; Ghosh, A.; Roccapriore, K.; Kovarik, L.; Allec, S. I.; Ziatdinov, M. A. Operationalizing Serendipity: Multi-Agent AI Workflows for Enhanced Materials Characterization with Theory-in-the-Loop. CoRR 2025, abs/2508.06569. https://doi.org/10.48550/ARXIV.2508.06569
210. Yang, Y.; Tang, Y.; Chen, Y.; Chen, X.; Qiu, J.; Xiong, H.; Yin, H.; Luo, Z.; Zhang,

Y.; Tao, S.; et al. AutoMat: Enabling Automated Crystal Structure Reconstruction from Microscopy via Agentic Tool Use. CoRR 2025, abs/2505.12650. https://doi.org/10.48550/ARXIV.2505.12650

211. Noh, H.; Lee, N.; Na, G. S.; Kim, K.; Park, C. IR-Agent: Expert-Inspired LLM Agents for Structure Elucidation from Infrared Spectra. CoRR 2025, abs/2508.16112. https://doi.org/10.48550/ARXIV.2508.16112

212. Mandal, I.; Soni, J.; Zaki, M.; Smedskjaer, M. M.; Wondraczek, K.; Wondraczek, L.; Gosvami, N. N.; Krishnan, N. A. Evaluating large language model agents for automation of atomic force microscopy. Nature Communications 2025, 16 (1), 9104.

213. Darvish, K.; Skreta, M.; Zhao, Y.; Yoshikawa, N.; Som, S.; Bogdanovic, M.; Cao, Y.; Hao, H.; Xu, H.; Aspuru-Guzik, A. ORGANA: A robotic assistant for automated chemistry experimentation and characterization. Matter 2025, 8 (2).

214. Jia, S.; Zhang, C.; Fung, V. LLMatDesign: Autonomous Materials Discovery with Large Language Models. CoRR 2024, abs/2406.13163. https://doi.org/10.48550/ARXIV.2406.13163

215. Nikolaev, P.; Hooper, D.; Webber, F.; Rao, R.; Decker, K.; Krein, M.; Poleski, J.; Barto, R.; Maruyama, B. Autonomy in materials research: a case study in carbon nanotube growth. npj Computational Materials 2016, 2 (1), 16031. https://doi.org/10.1038/npjcompumats.2016.31

216. Ye, X.; Yuan, W.; Fu, P.; Yang, X.; Chu, X.; Bai, Y.; Sun, Y.; Cheng, H.-M. A full-process artificial intelligence framework for perovskite solar cells. Science China Materials 2025, 68 (7), 2526-2535. https://doi.org/10.1007/s40843-025-3416-3

217. Ruan, Y.; Lu, C.; Xu, N.; He, Y.; Chen, Y.; Zhang, J.; Xuan, J.; Pan, J.; Fang, Q.; Gao, H.; et al. An automatic end-to-end chemical synthesis development platform powered by large language models. Nature Communications 2024, 15 (1), 10160. https://doi.org/10.1038/s41467-024-54457-x

218. Zhang, B.; Li, X.; Xu, H.; Jin, Z.; Wu, Q.; Li, C. TopoMAS: Large Language Model Driven Topological Materials Multiagent System. CoRR 2025, abs/2507.04053. https://doi.org/10.48550/ARXIV.2507.04053

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. CoRR 2025, abs/2512.11935. https://doi.org/10.48550/ARXIV.2512.11935

220. Shen, Y.; Wang, L.; Huang, Y.; Zhang, X.; Huang, M.; Li, H.; He, J.; Cai, A.; Wang, Y.; Smith, P. E. S.; et al. Unlocking azobenzene isomerization mechanisms via an LLM

agent-driven workflow integrating simulation, experiment, and machine learning. Chemical Science 2026, 10.1039/D5SC08794E. https://doi.org/10.1039/D5SC08794E

221. Babel, W.; Endo, I.; Enfors, S.; Fiechter, A.; Hoare, M.; Hu, W.; Mattiasson, B.; Nielsen, J.; Sahm, H.; Schügerl, K. Advances in biochemical engineering/biotechnology. Cell 2007, 107.

222. Roch, L. M.; Häse, F.; Aspuru-Guzik, A. ChemOS: An Orchestration Software to Democratize Autonomous Discovery. In Artificial Intelligence in Drug Discovery, Brown, N. Ed.; The Royal Society of Chemistry, 2020; p 0. https://doi.org/10.1039/9781788016841-00349.

223. Quigley, M.; Conley, K.; Gerkey, B.; Faust, J.; Foote, T.; Leibs, J.; Wheeler, R.; Ng, A. Y. ROS: an open-source Robot Operating System. In ICRA workshop on open source software, 2009; Kobe: Vol. 3, p 5.

224. Fawzy, S. M.; M. Ali, M. K.; Allam, N. K. Artificial Intelligence-Driven Materials Design for Next-Generation Sustainable Energy Technologies. ACS Sustainable Chemistry & Engineering 2026, 14 (10), 4745-4761. https://doi.org/10.1021/acssuschemeng.6c01084

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. J. Mach. Learn. Res. 2026, 27, 13:11-13:10.

226. Krasecki, V. K.; Sharma, A.; Cavell, A. C.; Forman, C.; Guo, S. Y.; Jensen, E. T.; Smith, M. A.; Czerwinski, R.; Friederich, P.; Hickman, R. J.; et al. The Role of Experimental Noise in a Hybrid Classical-Molecular Computer to Solve Combinatorial Optimization Problems. ACS Central Science 2023, 9 (7), 1453-1465. https://doi.org/10.1021/acscentsci.3c00515

227. Zhang, J.; Choubey, P. K.; Huang, K.-H.; Xiong, C.; Wu, C.-S. Agentic Uncertainty Quantification. CoRR 2026, abs/2601.15703. https://doi.org/10.48550/ARXIV.2601.15703

228. Bougzime, O.; Jabbar, S.; Cruz, C.; Demoly, F. Unlocking the Potential of Generative AI through Neuro-Symbolic Architectures: Benefits and Limitations. CoRR 2025, abs/2502.11269. https://doi.org/10.48550/ARXIV.2502.11269

229. Shen, Y.; Yang, Y.; Xi, Z.; Hu, B.; Sha, H.; Zhang, J.; Peng, Q.; Shang, J.; Huang, J.; Fan, Y.; et al. SciAgentGym: Benchmarking Multi-Step Scientific Tool-use in LLM Agents. CoRR 2026, abs/2602.12984. https://doi.org/10.48550/ARXIV.2602.12984

230. Laboratory, S. A. I. Probing Scientific General Intelligence of LLMs with Scientist-Aligned Workflows. CoRR 2025, abs/2512.16969. https://doi.org/10.48550/ARXIV.2512.16969

231. Chaturvedi, S. S.; Bergerson, J. D.; Mallick, T. Toward Reliable, Safe, and Secure LLMs for Scientific Applications. CoRR 2026, abs/2603.18235. https://doi.org/10.48550/ARXIV.2603.18235

232. Zhao, H.; Tang, X.; Yang, Z.; Han, X.; Feng, X.; Fan, Y.; Cheng, S.; Jin, D.; Zhao, Y.; Cohan, A.; et al. ChemSafetyBench: Benchmarking LLM Safety on Chemistry Domain. CoRR 2024, abs/2411.16736. https://doi.org/10.48550/ARXIV.2411.16736