nature reviews methods primers

https://doi.org/10.1038/s43586-026-00534-y

Primer

Check for updates

Agentic AI systems for scientific discovery

Bin Feng[1,2], Peter Walther[1], Leili Zhang[3], Reid A. Coyle[1], Seongmin Kim[4,5], Abhijoy Mandal[5,6], Zijian Zhang[5,6,7], Sven M. Papidocha[4,5], Anna Kelmanson[8], David Ming Segura[8,9], Saumya Thakur[8,9], Shyam Chand Pal[1], Varinia Bernales[4,6,10] ✉, Wenhao Gao[11] ✉, Xin Yang[12] ✉, Le Cong[13] ✉, Philippe Schwaller[8,9] ✉, Alán Aspuru-Guzik[4,5,6,7,10,14,15,16] ✉ & Zhiling Zheng[1,2] ✉

Abstract

Artificial intelligence (AI) agents are emerging in scientific research as systems that can transform the accumulated decisions of scientists into traceable workflows. In fact, modern scientific practice depends on continual coordination of knowledge dispersed across literature, data, theory and experiment, which, however, is often made through manual, case-by-case judgement. This Primer introduces agentic AI systems as a general method to structure this process through natural language interfaces, serving as a computational brain in the laboratory. Central to this framework are practical workflows for in-depth literature search and knowledge mining, orchestration of physics-based simulations, and integration with automated experimental platforms. We begin by describing the foundations of AI agents, including task decomposition, memory, tool use and communication between agents, and by distinguishing single-agent and multi-agent designs, as well as fully autonomous and human-in-the-loop configurations. Equally important are the means by which these systems are evaluated, documented and reported reproducibly. Finally, we examine major bottlenecks, including hallucination, tool misuse, limited interpretability and safety risks, and outline a future in which coordinated networks of agents, instruments and researchers transform scientific discovery from isolated tasks into continuous adaptive experimentation.

Sections

Introduction
Experimentation
Results
Applications
Reproducibility and data deposition
Limitations and optimizations
Outlook

A full list of affiliations appears at the end of the paper. ✉e-mail: varinia@bernales.org; whgao@seas.upenn.edu; xiny@ksu.edu; congle@stanford.edu; philippe.schwaller@epfl.ch; alan@aspuru.com; z.z@wustl.edu

Introduction

Advances in computer technology have enabled machine-assisted experimental design, instrument control and data analysis, thereby giving rise to high-throughput experimentation[1–4]. However, experimental discoveries in this data-intensive environment have largely hitherto arisen from predefined pipelines rather than from adaptive design. Furthermore, the rapid growth of datasets and digital tools has rendered information integration increasingly challenging[5,6]. With further developments in automation, the concept of self-driving laboratories was introduced to close the experimental loop through robotics and feedback[7–16]. Yet some systems face challenges in transferability due to fixed workflows[7,17,18]. Concurrently, machine learning models in experimental workflows are engineered for specialized tasks[19–25], with limited orchestration, whereas large language models (LLMs) demonstrate strong task comprehension in digital workflows[26–31] but are not yet broadly realized in physical experimental systems[26,32–35]. Taken together, these limitations point to the need for a framework capable of coordinating reasoning, tool use and experimental action within a single workflow[31,36–39].

Artificial intelligence (AI) agents and the workflows built around them have emerged as a new methodology for scientific discovery[37,40–42]. Here, an agentic AI system is defined as comprising one or more AI agents, the workflow that organizes their actions and the environment with which they interact in pursuit of a scientific objective, including digital resources, physical instruments and human participants. Given their diverse architectures and levels of autonomy, we view agency as a continuum shaped by discretion over actions, environmental impact, interaction horizon and persistence of goals and state. Nonetheless, this Primer focuses on LLM-based agents that act through natural language and tool interfaces in computational or physical environments[43–45]. Through role specialization and coordinated task execution, such systems can scale into agentic ecosystems. Here, an agentic ecosystem refers to a coordinated network of agents, researchers, data sources, computational resources and physical instruments linked through shared interfaces and feedback (Fig. 1a).

Typical scientific agentic systems involve multimodal frameworks that decompose tasks, invoke tools and refine strategies based on intermediate results, all centred on goals expressed in natural language. Across successive stages of development, three overlapping directions have acquired distinct current emphases: cognitive agents, embodied laboratory agents and collaborative agents (Fig. 1b). We distinguish these by their predominant mode of operation, namely digital cognition and computation, interaction with physical environments or coordination among agents and humans, while recognizing that real systems may span all three and that these capabilities are likely to converge as scientific agentic systems mature[45]. Structurally, they integrate reasoning models, tool use, memory and iterative planning to support complex experimental discovery workflows.

Compared with conventional machine learning pipelines or simple prompt-based LLM interactions, systems with greater degrees of agency can exercise broader control over actions and coordinate

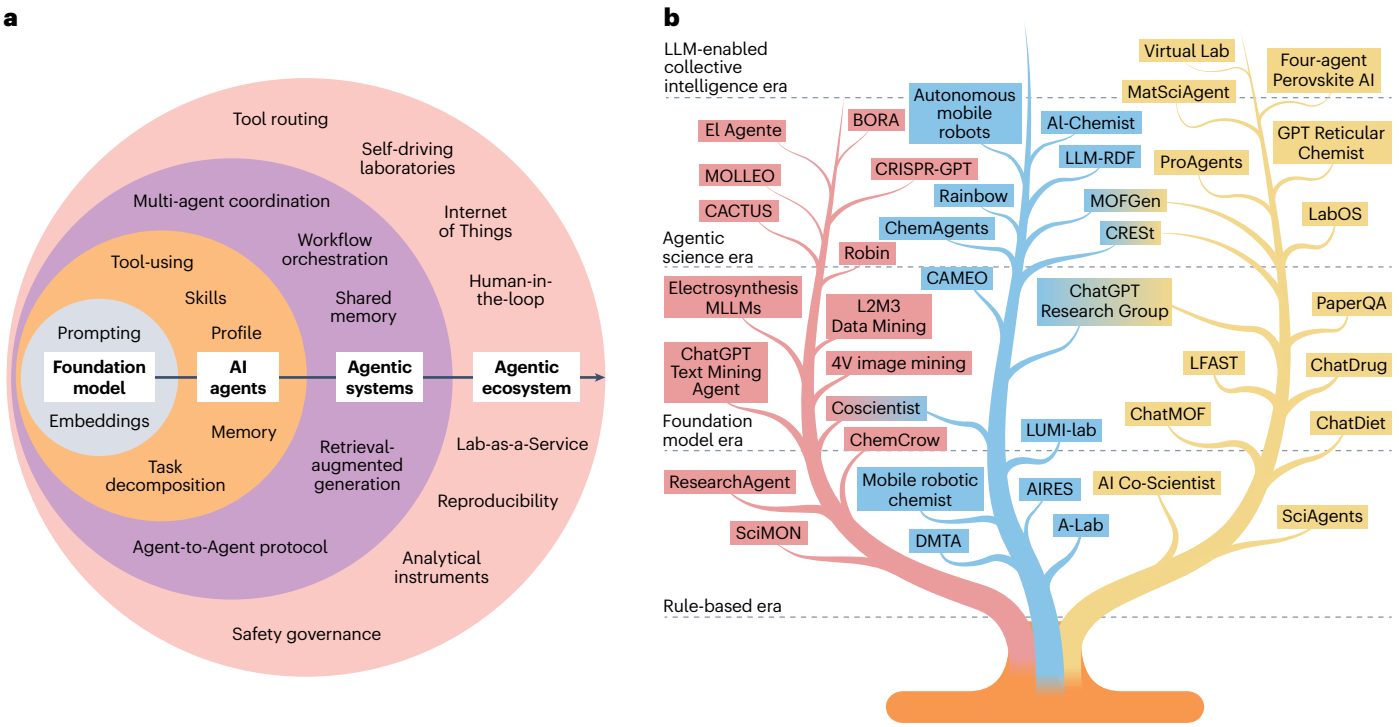


Fig. 1 | Overview of agentic experimental discovery. a, Progressive assembly from foundation models to artificial intelligence (AI) agents, agentic systems and agentic ecosystems. The main labels represent successive levels of system organization: models provide reasoning; agents add tool-enabled action; systems coordinate agents within workflows; and ecosystems connect agents with researchers, data, computation and instruments. **b**, Illustrative classification and development of scientific AI agents over time. Agents are classified by predominant function as cognitive (red), embodied (blue) or collaborative (yellow), corresponding to digital reasoning, physical laboratory interaction and coordinated human–AI or multi-agent operation, respectively. The four eras denote broad stages from rule-based systems and foundation models to agentic science with action capability and collective intelligence through coordinated systems. LLM, large language model; MLLM, multimodal large language model.

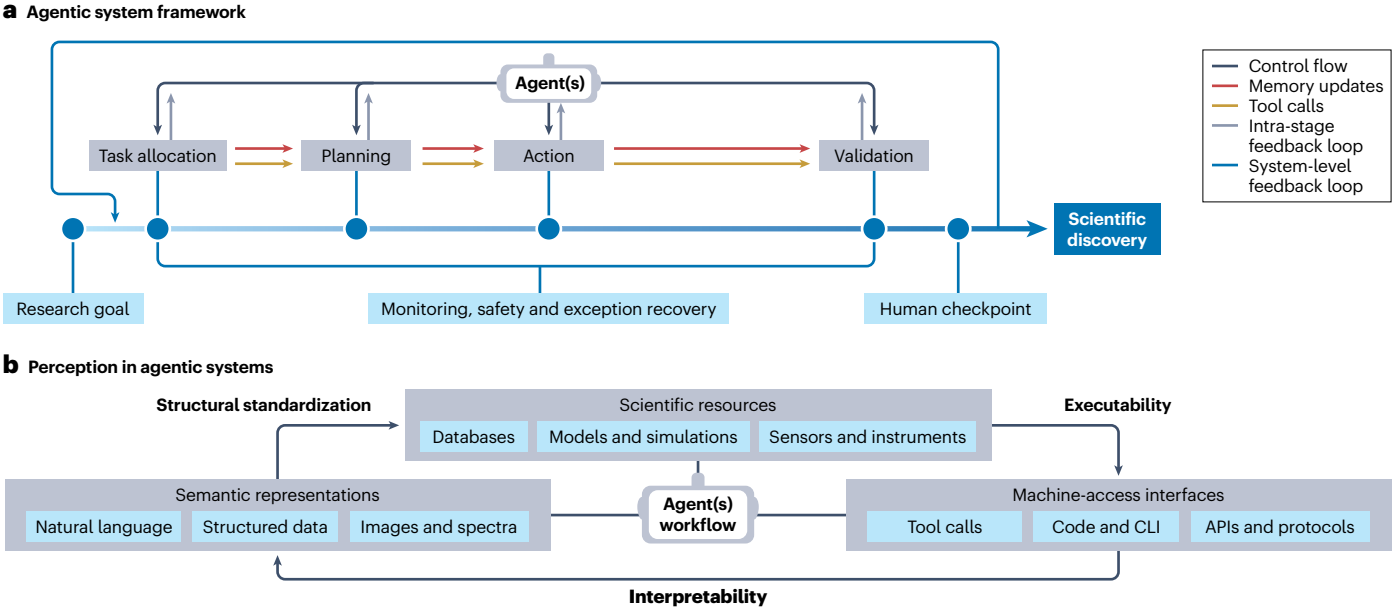


Fig. 2 | Operational architecture of agentic AI systems. a, Representative workflow of an agentic artificial intelligence (AI) system. A defined scientific problem is addressed through task allocation, planning, action and validation. Memory updates (red), tool use (yellow) and local feedback (grey) support adaptive revision, whereas a system-level feedback loop (blue), monitoring and human checkpoints guide progression towards scientific discovery. These functions may occur within a single agent or be distributed among a coordinator and role-specialized agents. **b**, Illustrative pathways through which scientific information is organized, accessed and interpreted by an agent. Scientific resources define where information resides, and their structural standardization reflects the consistency with which it is organized. Machine-access interfaces establish how agents reach and act on these resources, with their executability determining whether such interactions can be carried out programmatically. Semantic representations provide the means by which information is exchanged and processed, and their interpretability reflects how readily it can be understood and used in reasoning. API, application programming interface; CLI, Gemini command line interface.

reasoning with execution[36]. Moreover, by using natural language as a bridge between human researchers' intent, knowledge and experimental platforms, such agentic systems can support adaptive workflows that respond to intermediate results, potentially improving the flexibility, generalizability and scalability of scientific processes[46,47]. Nevertheless, important challenges remain, as long-horizon reasoning may lead to error accumulation and consequent deviation[48–50], whereas the expansion of system autonomy raises considerations of safety and ethics, and technical aspects, including memory design and multi-model orchestration, remain to be resolved.

This Primer introduces agentic AI systems as a general framework for scientific discovery. We outline their conceptual foundations and deployment methodologies. Following a systematic categorization of their core functionalities, we discuss their representative applications including literature mining, simulation orchestration and automated experimentation. We further discuss considerations for reproducibility and methodological reporting in such systems. Finally, we analyse current limitations and outline directions for the development of agentic systems capable of more reliable and autonomous experimental decision-making.

Experimentation

In the context of agentic AI systems, experimentation involves the construction and operation of adaptive workflows in which scientific objectives are translated into executable decisions. By virtue of this design, AI agents move beyond static text-based interaction and support sustained, multistep operation across literature, computation, instrumentation and human decision-making.

At a fundamental level, such systems are organized around a control framework that governs actions — most often a LLM — to coordinate workflow execution (Fig. 2). Prompts are iteratively processed into executable instructions that invoke tools, coordinate agents, retrieve information and trigger downstream experimental actions to reduce reliance on manual intervention. This section describes two principal elements of an agentic AI system for scientific discovery: construction and execution. The former defines the foundational components of the system, including profiles, perception, tools and memory, whereas the latter governs how these components interact through planning, action, coordination and adaptation.

Construction

Modern agentic systems are organized around one or more LLMs responsible for reasoning (Fig. 2a). An agent's behaviour is steered by a profile which spells out identity, objectives and operating limits. Capabilities beyond text generation are surfaced as callable tools, including interfaces to laboratory instruments, software, search engines and machine learning models. A memory subsystem maintains the task state and accumulates experience across turns. Agentic systems can be implemented through native application programming interface (API) integrations, custom orchestration loops or agent-development toolkits; illustrative examples include LangChain, Pydantic AI, OpenAI Agents SDK and LangGraph. In practice, these components can be organized within an agent harness, through which their interactions

Table 1 | Principal tool categories in agentic scientific systems

Category	Examples
Information access	Search engines such as Perplexity and Tavily, databases (PubMed, ChEMBL), input file generators[138,238]
Computation and modelling	Python executors, high-performance computing schedulers[67,239]
Data management	Natural language to SQL[240], terminal commands, electronic laboratory notebooks[241], laboratory information management systems[242]
Physical interaction and workflow	Instrument control[36,243], robotic platforms[244], subagent delegation[126], NL-to-XDL conversion[245], safety shutdowns[171,246]

Tools are classified by their predominant function, although many support more than one stage of a scientific workflow.

are coordinated during execution. Meanwhile agent frameworks such as Claude Code, OpenAI Codex, the Gemini command line interface (CLI), OpenClaw, Hermes Agent and OpenCode interact with heterogeneous external resources, including databases, software packages and laboratory instruments, each of which would require custom integration. The Model Context Protocol (MCP)[51] addresses this need for custom integration by providing a unified interface between agents and their environments. Agents can also connect via command line interfaces and interact with local file systems to support laboratory automation pipelines.

Profile. Profiles govern how agents behave during execution. They are often represented as versioned Markdown files (such as Claude Code and OpenClaw), thereby rendering system behaviour auditable and reproducible. In multi-agent systems, agents require external-facing capability descriptions in addition to their individual internal behaviour characteristics in order to discover and invoke other agents. The Agent-to-Agent protocol serves this purpose by providing machine-readable agent cards, which specify modalities, reusable capability modules, end points and authentication requirements. Whereas profiles steer an agent's internal reasoning and operating limits, agent cards specify what an agent can do and how it can be reached.

Perception. Perception defines how an agent acquires information from its environment (Fig. 2b). Text remains the dominant modality, including prompts, retrieved literature, tool outputs and execution traces. Increasingly, LLMs accept multimodal inputs[52,53], which allow agents to process scientific images[54–57] and edit molecular structures through vision[58]. Note that in scientific settings, domain-specific signals often require specialized preprocessing before model reasoning[59].

Tools. Tools provide the operational interface between reasoning and execution (Table 1). In principle, any computational component, from conventional algorithms to machine learning models, can serve as a tool if it exposes a callable interface with defined inputs and outputs. The most direct approach is to emit executable code (Python, Bash, JSON[60]) that an interpreter runs, exposing the full expressiveness of existing ecosystems including cheminformatics stacks[61] and laboratory hardware such as Python Protocol API Documentation. Evolutionary strategies can plan over large toolsets[62]. Standard-protocol approaches expose tools via JSON Schema, for example, Cheminformatics Microservice[63] and ChemCloud[64]. Standardized protocols such as the MCP provide a common interface for connecting agents to external tools and data sources, which, in essence, reduces the need for custom wrappers and improves interoperability across local and remote resources. Example MCP servers include PubChem and ChEMBL, as well as machine learning stacks such as Chemprop for molecular property prediction[65,66]. Additionally, tools need not be fixed at the time of design, as agents can synthesize, validate and reuse new tools on demand[67,68].

Memory. Memory preserves contextual continuity across long-horizon workflows[69]. Short-term memory is provided by the model context, which contains current instructions, observations and tool outputs[70]. Long-term memory persists beyond a single context and may take the form of text files, databases, vector stores, workflow logs, skills or knowledge graphs[71–73]. The choice of memory representation should depend on the scale and structure of the information: smaller or more ordered knowledge can be maintained in hierarchical Markdown-based formats, including the recently introduced Open Knowledge Format, whereas larger or less structured corpora may benefit from vector stores or graph-based representations. In experimental workflows, memory serves to preserve the state (such as which experiments have been attempted), support learning from prior outcomes (including failed conditions and successful protocols) and enable reproducibility by retaining provenance, decisions and intermediate data. Memories can also be supported by hierarchical, graph-based and network-based structures[71,72,74,75], which can be further modified by tools[76] (Hermes Agent) so that memory evolves with experience. For large-scale memory, vector-based retrieval-augmented generation represents stored information numerically and retrieves the entries most similar to the current query[77]. Retrieval alone, however, does not ensure that memory remains reliable over long workflows, as entries may become redundant, contradictory or obsolete. Active life-cycle management, including consolidation, conflict resolution and selective forgetting, is therefore necessary, and emerging approaches address this through agentic self-organization of memory entries[78] or layered architectures with explicit compression and update operations[79]. Recent graph-based retrieval and scientific-agent frameworks suggest that dynamic knowledge graphs can function as external memory and reasoning substrates, enabling multi-hop retrieval, provenance-aware context management and tool-updatable state evolution in long-horizon workflows[75].

Execution

Execution refers to the mechanisms that, given a defined agent architecture, collectively drive the system towards a high-level scientific goal, encompassing planning, reasoning, multi-agent coordination and direct engagement with the experimental environment (Fig. 2a). The fidelity of this coordination is a primary determinant of whether the system achieves reliable scientific outcomes[80–82]. Systems may be open loop, executing a fixed sequence of steps without feedback[44,83], or closed loop, cycling between planning, action and observation until a goal is reached[47,54,84,85]. In experimental science, closed-loop designs are considerably more practical, as outcomes are rarely predictable in advance. Consequently, these feedback-driven systems have been demonstrated across numerous domains such as autonomous electrochemistry[86], distributed chemical synthesis[84,87], autonomous biotechnology[88] and closed-loop organoid experimentation[89].

To illustrate how such systems can be constructed and used in practice, we present a lipid nanoparticle (LNP) formulation-discovery

task as a worked example (Fig. 3). The starting point is the scientific objective, which defines the capabilities that the agent requires and, in turn, its architecture. For a narrowly defined task, a single agent can make successive model, tool and API calls, whereas a broader workflow involving information retrieval, input preparation, simulation and evaluation can distribute these functions among specialized agents when the benefit warrants the additional coordination and computational cost[90]. Reusable scientific skills then provide instructions for accessing external tools and resources[41], whereas a repository-based workspace organizes project data and intermediate outputs. Transient instructions and current observations may remain in the prompt context, whereas prior attempts, intermediate results and provenance are retained persistently[91]. Tool outputs are checked for format and scientific plausibility before downstream use, with human approval reserved for irreversible, safety-critical or physical actions[92]. The resulting candidates can then be reviewed by researchers and, upon approval, passed to an experimental loop. Measurements are recorded in a results spreadsheet and returned to the workspace, where they provide the basis for subsequent candidate redesign.

Planning. Planning accounts for the cognitive layer of agentic systems. It determines how subtasks should be sequenced, which tools should be invoked and how the workflow should adapt as new information arrives. Single-path planning generates a linear chain-of-thought sequence, efficient yet brittle when early steps fail. Multi-path planning evaluates several candidate trajectories in parallel[93], well suited to situations where intermediate results must discriminate between competing protocols, as in the agentic tree search of AI Scientist v2 (ref. 94). Hierarchical planning separates high-level goals from low-level executable steps; ChemAgents implements this through a Task Manager coordinating four role-specific subagents covering literature retrieval, experimental design, computation and robotic operation[95]. Lastly, feedback-driven planning continuously revises the plan against real instrument outputs, as demonstrated in autonomous materials phase-diagram mapping[46] and closed-loop molecular discovery[47]. These paradigms are inherently complementary and are often combined.

Reasoning. Reasoning is the process by which an agent interprets observations, draws inferences from intermediate results and decides

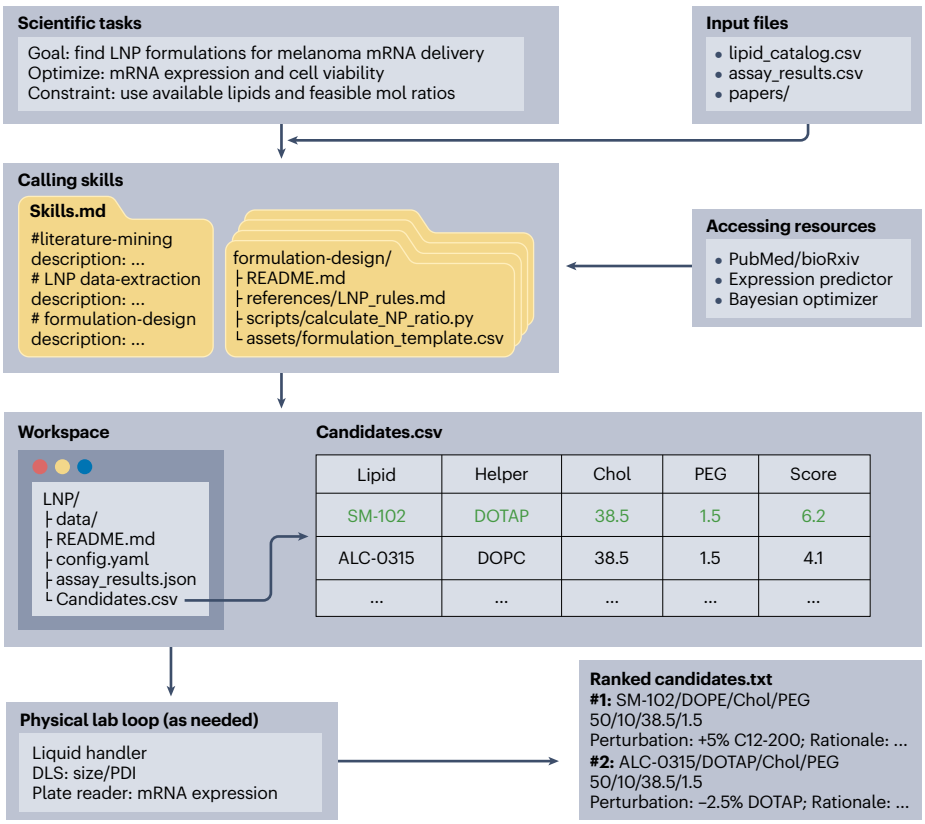


Fig. 3 | Practical construction of a LNP formulation-discovery agent. The scientific tasks are defined by researchers, with relevant background knowledge and supporting files supplied as inputs, including the catalogue of available lipids, prior assay measurements and literature. These inputs are incorporated through skills, each organized as a self-contained folder of instructions, reference rules, scripts and templates through which the agent is instructed on how to carry out the corresponding task. External resources can further augment the agent and are made accessible through application programming interfaces (APIs) or other callable interfaces. During operation, the agent queries these resources and writes outputs, including predictor-derived candidate scores, to files in the workspace, whereas task constraints restrict consideration to feasible formulations. Where needed, the agent can invoke a physical laboratory loop for experimental validation or additional measurements, including liquid handling, dynamic light scattering and plate-reader assays. The resulting intermediate outputs are evaluated by the agent against the stated objectives and constraints to produce a final ranking of candidates, together with proposed compositional perturbations and accompanying rationales. Chol, cholesterol; DLS, dynamic light scattering; DOPE, 1,2-dioleoyl-*sn*-glycerophosphoethanolamine; DOTAP, 1,2-dioleoyl-3-trimethylammonium propane; LNP, lipid nanoparticle; PDI, polydispersity index; PEG, polyethylene glycol.

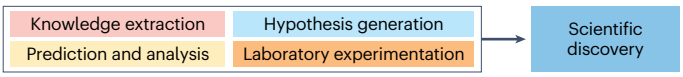


Fig. 4 | Overview of scientific result analysis in agentic systems. Domain knowledge is extracted and transformed into testable hypotheses, which are refined through prediction and analysis and subsequently examined by laboratory experimentation. The agentic character lies in an adaptive trajectory instead of a fixed linear workflow, in which one or more agents make autonomous decisions over subsequent actions and tool use, retain intermediate results as state and revisit earlier stages as new evidence emerges. Provenance and decisions are recorded throughout, which in essence allows the path to discovery to be traced while human involvement shifts towards defining goals and approving consequential actions.

whether its current plan remains appropriate. This is primarily realized using the ReAct mechanism[96], which interleaves explicit reasoning traces with tool calls and outcome observations, allowing the agent to revise its course in response to unexpected results. Each cycle comprises a written thought that defines what the agent needs to establish, an action through an appropriate tool and an observation of the result. The agent then considers the complete sequence and determines what to establish or do next, which initiates a new cycle. In practice, this spans both short-horizon tasks, such as parsing instrument outputs, and longer-horizon ones, such as adaptive replanning across a failed multistep synthesis. Agent frameworks have operationalized this pattern across domains of quantum computing[54], molecular dynamics[97] and gene editing[35].

Multi-agent systems. Multi-agent execution strategies form an important class of execution topologies in contemporary agentic systems and represent a major direction for future research. These strategies decompose scientific discovery into interacting roles, such as planners, hypothesis generators, experiment designers, data analysts and critics, coordinated through a shared state or workflow[98–100]. This division of labour enables agents to have simultaneous competence in subtasks of literature search, hypothesis generation and refinement, experiment design and execution (including on automated platforms) and iterative update of the models of the system under study. In drug discovery[101], physics[54], biology[102] and materials science[98,103], such architectures have begun to deliver end-to-end or lab-in-the-loop pipelines that demonstrate promising performance relative to human teams on tightly scoped case studies; however, a substantial gap remains between success in these bounded scenarios and true open-ended research competence.

Human-in-the-loop. Some decision points in experimental science, whether by virtue of their stakes or their dependence on tacit domain knowledge, remain beyond the reliable reach of automated reasoning. Human-in-the-loop (HITL) architectures distribute human checkpoints across the pipeline, from task allocation through to final validation, with exception handling at intermediate stages to catch anomalous results before they propagate. Human oversight catches compounding errors, prevents irreversible actions, preserves accountability and injects expertise where current models are weakest[104]. Full autonomy in polymer self-driving laboratories was shown to introduce algorithmic biases, supporting a hybrid model grounded in symbiotic autonomy[105]. Human checkpoints are intended to slow the workflow at consequential decisions, and this is by design[106]. The practical difficulty arises when the required expert is unavailable, as an unresolved checkpoint can stall the entire run. We therefore propose that workflows define explicit timeout, safe-pause and escalation rules, with a pending decision suspending the run rather than defaulting to execution. Looking ahead, adaptive HITL, whereby agents determine when to escalate a decision based on their own uncertainty, represents a compelling direction for sharpening the boundary between human and machine responsibility[107].

Results

The defining result of agentic AI systems is the generation of a ladder of outputs by which a scientific question is transformed into traceable executions. First, evidence is rendered into structured knowledge, and then knowledge is formulated into testable hypotheses. Hypotheses become candidates or protocols, and selected outputs are subjected to experimental validation. This section examines the representative results produced by agentic AI systems during experimental discovery and, more importantly, the principal considerations for their interpretation and validation (Fig. 4). Within this progression, cognitive agents primarily focus on knowledge synthesis, reasoning, planning and computational tool use. Embodied agents extend these capabilities to instruments and physical experiments, whereas collaborative agents, in turn, coordinate work among agents and researchers.

Output and analysis

AI systems for science have been historically evaluated by single-metric, isolated outputs such as classification and regression outputs[1,39,108,109]. By contrast, trajectory — as a linked succession of representations and actions that turns a question into a testable form — is closer to how experimentalists make discoveries. It yields four classes of output.

The first output class is extracted knowledge. Here, agents convert unstructured scientific corpora into representations amenable to downstream reasoning (Fig. 4). Early systems such as Robot Scientist Adam encoded biological observations into formal logical predicates to reason over metabolic pathways[9,110]. More recent systems instead leverage LLMs to extract structured knowledge directly from the literature. For example, in 2023, ChatGPT Chemistry Assistant[26] and LLM-NERRE[111] demonstrated the construction of JSON or tabulated synthesis datasets based on hundreds of publications describing metal–organic frameworks (MOFs), and later in the same year Coscientist[36] showed the conversion of textual synthesis procedures into robotic workflows. These outputs form the scaffolds that determine what the agent can later ask, vary and test.

The second class is ranked hypotheses. In early systems, this was handled through abductive inference over a constrained knowledge base[110,112,113]. Nowadays, agentic systems expand this idea by using literature retrieval, knowledge graphs and multi-agent critique. SciMON, for example, uses existing literature to maximize technical depth[114], whereas SciAgents searches ontological graphs for interdisciplinary links[98] and CoDHy uses graph-based evidence signals and literature retrieval from PubMed to rank drug combinations for biomarkers[115]. HypoChainer joins visualization and human expertise[116], SciMuse identifies non-obvious interdisciplinary insights from 58 million research papers[117] and AstroAgents analyses high-dimensional mass spectrometry datasets to propose evidence-based hypotheses[118]. The output list from this stage treats uncertainty as a first-class object and, instead of offering a single attractive answer, yields a ranked distribution of possibilities needed for adaptive experimentation.

The third class involves iterative refinement of scientific candidates through simulations, tool orchestration and automated feedback. For instance, RepurAgent combines a hierarchical multi-agent architecture with episodic memory and was assessed through pathway-level comparison with Google Co-Scientist outputs[119]. Agentomics improves protein-engineering performance while maintaining agreement between validation and hidden-test outcomes[120]. LARA-HPC[121] illustrates a validation-first mode for simulation execution in which dry-run primitives identify invalid inputs or physical configurations before formal execution. Thus, optimized outputs should be judged by both scientific merit and procedural soundness.

The last class, where the agentic pipeline is tested against the experiment, is empirically validated scientific knowledge. The analytical dimensions used here include reproducibility, efficiency, scientific validity, actionability and human oversight. Reproducibility refers to whether results or key conclusions can be regenerated and traced under documented conditions. Efficiency pertains to the time, computational, experimental and human resources required for each valid result. Scientific validity involves consistency across evidence, controls and domain constraints. Actionability refers to whether an output is sufficiently specified and reliable to support a defined downstream decision or experiment. Finally, human oversight refers to the adequacy, traceability and burden of human intervention, including when, why and how often experts intervene and how those interventions affect the outcome. In mathematical research, Aletheia uses a propose–verify cycle in which generators, verifiers and revisers refine candidate solutions before expert review[122]. Nothing in this loop is specific to mathematics, and the checking is done in ordinary prose rather than in a formal proof system, so the same mechanism should carry over to other fields. Applying the same principle to biomedical hypothesis generation, Google's AI Co-Scientist[123] nominated epigenetic modifiers for liver fibrosis, and the HDAC and BRD4 inhibitors among them were subsequently shown to suppress fibrogenesis in human hepatic organoids[124]. In materials science research, the GPT Reticular Chemist guided the discovery of four novel MOFs by identifying modulator effects on crystallinity improvement[125]. At this level, the result includes raw measurements, derived statistical analyses and the provenance connecting them to earlier agent decisions.

Validation strategies

As agentic systems produce heterogeneous outputs, validation is considered to be a staged process progressing from internal consistency checks towards computational and experimental confirmation[35,126,127] (Table 2). Internal validation is the first filter, which asks whether an output is coherent (internally consistent and stable across repeated runs), grounded (every substantive claim traceable to a source document, dataset or tool output) and well formed (containing all required fields and readable automatically by software)[126–129]. Within the context of knowledge extraction, this includes comparison against source documents, assessment of citation grounding, schema completeness, consistency across repeated runs and detection of unsupported claims[26,111,130,131]. For ranked hypotheses or protocols, this means testing stability across prompts, retrieval contexts and model variants, and inspecting whether subtasks, tool calls and intermediate states are propagated correctly[36,126,127]. For planning tasks, agreement with expert workflows or similarity between generated and reference task sequences may also be informative[132]. The above-mentioned checks are usually inexpensive and often sufficient to expose obvious failure modes[128,129], such as hallucinated facts, malformed outputs, omitted constraints or brittle rankings[127,128,132].

Computational confirmation provides the next filter for outputs intended to guide modelling or design[126,127,133]. Here, agent-prioritized molecules, materials, mechanisms or conditions can often be screened through benchmark datasets, surrogate models and physics-grounded simulations[126,127]. At this stage, it is helpful to distinguish testing strategies from the evaluation metrics rooted in them[127–129]. Testing strategies specify how an agentic system is challenged, for example through repeated runs, cross-domain transfer or robustness tests such as input perturbation and fault injection, in which errors or failures are deliberately introduced to test whether the system can detect and recover from them. Evaluation metrics specify what is measured during those tests, such as the task success rate, cross-run agreement and efficiency under realistic conditions.

Repeated runs are especially important because agentic systems are inherently stochastic and path-dependent[127]. Reporting only best-case trajectories may therefore substantially exaggerate capability[127–129], which can render reproducible performance difficult to distinguish from serendipitous success.

Wet-lab confirmation remains the strongest test for claims intended to affect real experiments[35,36,134]. Validation should follow the same standards expected for conventional experimental science, including appropriate controls, replication, predefined success criteria and transparent reporting of failures alongside successes. Human intervention should also be documented explicitly[135], as many systems operate in HITL settings, where researchers define design spaces, approve actions, correct failures or intervene at safety-critical points[35,129,134]. Consequently, validation should clarify which actions are autonomous and which are dependent on expert correction or approval. Additionally, negative results are highly informative in this context[127–129]. Failed experiments may reveal an unsupported literature synthesis, overfitting to a benchmark or a plan that was syntactically valid yet scientifically unsound. Because automated checks cannot judge mechanistic plausibility or novelty, expert post-run evaluation remains the decisive filter across all stages and should be reported explicitly.

Interpretation

Interpretation should begin by separating utility from evidence[128,129,135]. In essence, utility concerns what an output enables researchers to do,

Table 2 | Staged validation of outputs from scientific AI agents through internal checks, computational assessment and experimental confirmation, with complementary quantitative and qualitative methods

Category	Quantitative methods	Qualitative methods
Internal	Workflow feasibility Step success likelihood Tool-call success	Schema verification Logical consistency Expert overview
Computation	Benchmark performances Variance across seeds Rank correlation Success rate	Code/workflow inspection Analysis of failure modes Sensitivity analysis
Wet-lab	Experimental measurements Improvement over baseline Efficiency of the method	Observation of phenomena Mechanistic plausibility Expert assessments

AI, artificial intelligence.

whereas evidence concerns the scientific conclusions that the output can support. It should be remarked that this distinction matters, as an output may direct the course of discovery well before sufficient evidence exists to establish its scientific validity. First, factual outputs must be interpreted in the context of their source and origin. Summaries, mechanisms and high-scoring candidates are useful only to the extent that they are traceable, reproducible and properly bounded[36,126,134,135]. In this context, one can consider an output as being properly bounded when it makes clear what can and cannot be concluded from the available evidence, so that a finding specific to one system or set of conditions is not assumed to hold beyond them. Ungrounded outputs may still serve as testable hypotheses or heuristics, but they should be explicitly identified as unsupported rather than treated as evidence; where consequential, advancing ungrounded outputs to downstream testing should require human approval[26,111,130,131,135].

Tool use must be interpreted separately from the quality of the final outcome[36,126,135]. An agent may succeed by selecting the appropriate database, simulator, code environment or robotic interface[36,41,58,67,75,126,133,136–139], or fail because inputs were malformed, outputs were misread or recovery steps were absent[127–129]. The execution success rate, tool-selection accuracy, robustness under perturbation and completion time often reveal capabilities that are not captured by outcome-based measures alone[127,128,132]. Structural assessments of workflows can determine whether an agent has solved a task in a scientifically meaningful way or merely reached the correct answer by a brittle route[129,132].

Finally, output should be interpreted with explicit attention to the degree of autonomy achieved by the system. Many reported systems depend on human-defined objectives, curated tools, pre-specified search spaces, approval checkpoints and safety restrictions[35,36,134]. Their reported performance therefore reflects a mixture of model capability, architectural design and human scaffolding[35,134,135]. Reports should specify who defined the goal, selected tools, corrected failures, approved irreversible actions and chose outputs for validation[129,134,135]. Only then can agentic results be compared fairly and used without overstating what has been automated[126,127].

Applications

Agentic AI systems are particularly beneficial when used to convert scattered scientific information into verifiable actions. Their applications in experimental discovery can therefore be read as a progression from text to computation to the laboratory: reasoning over prior knowledge, generating and ranking candidate hypotheses, orchestrating computational workflows and closing experimental loops. This ordering defines how uncertainty is transformed from ambiguity in literature or datasets into ranked hypotheses, executable protocols and, eventually, empirical evidence (Fig. 5).

Literature mining and knowledge synthesis

A central bottleneck in scientific AI is the data desert created by unstructured literature. Although enormous quantities of literature reports exist, much of the available data remain embedded within prose, tables, schemes and supplementary files. Historically, converting these into machine-actionable databases required extensive standardization. LLM-based agents are changing this relationship between literature and data, as evidenced by early implementations using LLMs as the context searching agent and schema summarization agent to automate the data mining of experimental protocols[26,111,140,141] (Fig. 5a). A further step is multimodal extraction and analysis of the data distributed across text and plots. Through the direct encoding of images or the use of external tools for image conversion[131,142], the agentic system enables previously inaccessible experimental knowledge and otherwise overlooked properties reported within graphics into machine-readable datasets amenable to downstream reaction prediction and synthesis planning[143,144].

Hypothesis generation and ranking

Between knowledge synthesis and downstream computational or experimental execution, agentic systems can transform integrated evidence into testable candidate hypotheses and prioritize them according to evidential support, novelty, plausibility, feasibility and risk. SciAgents[98] uses bioinspired multi-agent graph reasoning over scientific ontological knowledge graphs to propose and refine hypotheses, whereas CoDHy[115] builds task-specific knowledge graphs from curated biomedical databases and literature to create and rank biomarker-guided drug combination hypotheses (Fig. 5b). Knowledge graph grounding can mitigate hallucination and improve the reliability and safety of hypothesis generation, particularly in biomedical contexts, but cannot eliminate errors. Independent expert review and computational or experimental validation therefore remain necessary.

Computational workflow orchestration

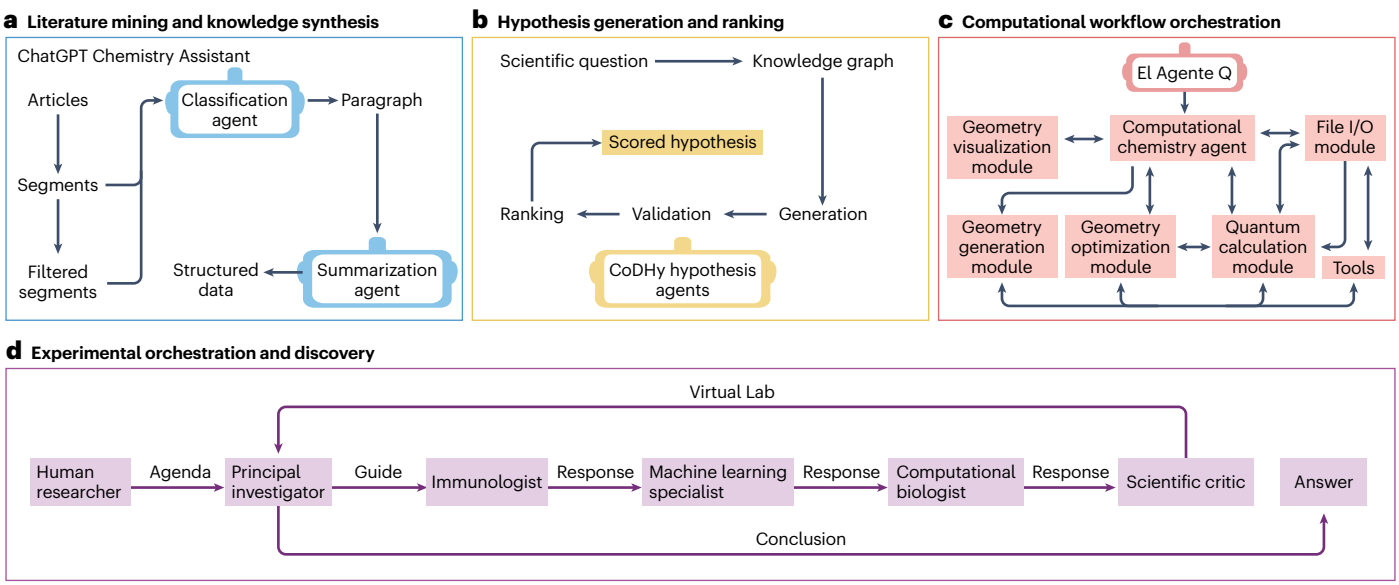
At present, in silico orchestration is arguably the most tractable domain for agentic AI because digital tools have explicit interfaces, failures are observable and repeated runs can test robustness[145]. A prominent example is the El Agente ecosystem[126], and its first member, El Agente Q, introduced a structured multi-agent architecture for autonomous quantum chemistry calculations (Fig. 5c). In this framework, agents select electronic-structure methods, prepare input files, invoke external packages, inspect intermediate results and revise calculations through explicit workflow graphs and provenance records. Similarly, in materials sciences, end-to-end structured simulations have also been performed by agents[146–148].

Complementary advances have emerged in life sciences. For instance, BioAgents applies the same principle to genomics, decomposing questions, retrieving domain knowledge and constructing multistep bioinformatics workflows, using retrieval-augmented generation over curated bioinformatics documentation to ground its recommendations[149]. Although current implementations focus primarily on conceptual reasoning and workflow guidance, these systems achieve performance comparable with human experts on a narrowly scoped set of conceptual genomics tasks (specifically, outlining workflows for quality control, RNA sequencing alignment and viral genome assembly). Although they continue to struggle with complex code generation, these tools can substantially reduce the burden of manual pipeline design.

Experimental orchestration and discovery

The most compelling promise of agentic AI lies in the transformation of the laboratory from a sequence of disconnected operations into a continuous reasoning system (Table 3). In line with the recent surge of interest in autonomous platforms[7], agentic AI has begun to exhibit capabilities that resemble aspects of human reasoning, turning the laboratory practice into a navigable decision landscape in which prior knowledge, experimental outcomes and robotic execution continuously inform one another.

A relatively early demonstration of this approach was Coscientist[36], a GPT-4-powered planner that combined web search, code execution



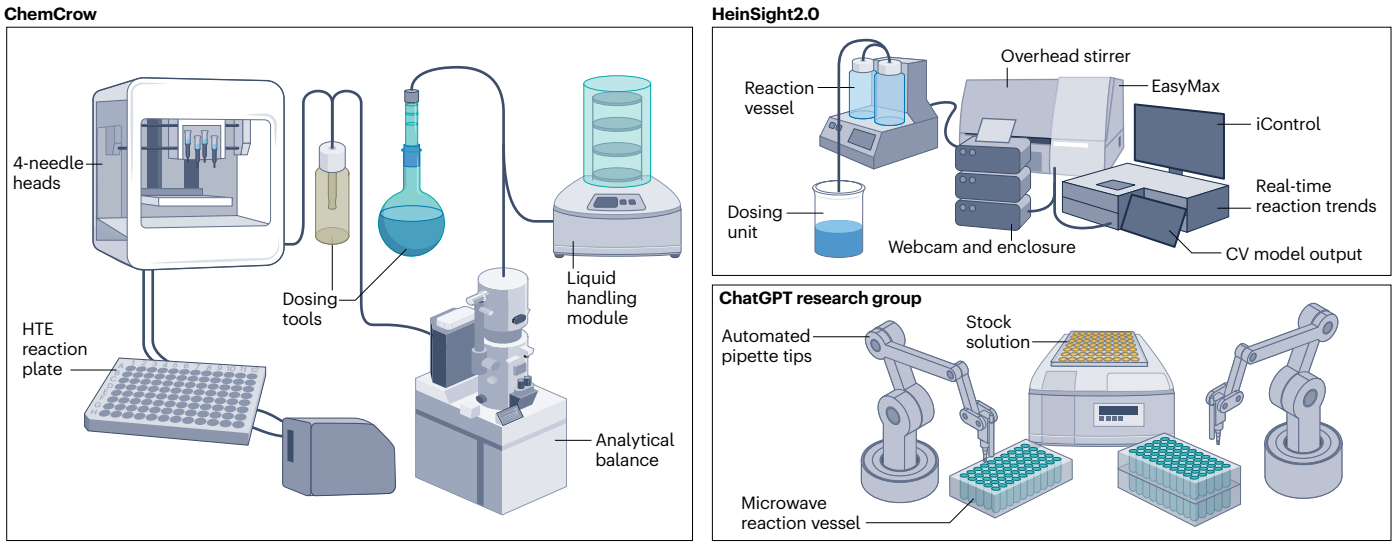


Fig. 5 | Selected agentic AI applications and future closed-loop integration. **a**, Workflow of the ChatGPT Chemistry Assistant, in which classification and summarization agents mine chemistry and materials science literature and convert synthesis conditions for around 800 metal–organic frameworks (MOFs) into tabulated data with greater than 90% precision, recall and F1 score[26]. The structured dataset assembled by the agents from heterogeneous literature sources was then used to predict crystallization outcomes with more than 87% accuracy, providing an early example of agentic literature mining for synthesis prediction. **b**, Overview of the CoDHy system architecture. The workflow proceeds from a user-defined scientific question to task-specific knowledge graph construction, followed by multi-agent hypothesis generation, validation and ranking to produce prioritized, evidence-grounded hypotheses. **c**, Overview of El Agente Q, a large language model (LLM)-based multi-agent workflow that uses a hierarchical design to orchestrate computational chemistry tasks[126]. **d**, The Virtual Lab emulates a research group through iterative collaboration among a coordinating agent acting as a principal investigator, role-specialized agents assigned to immunology, machine learning and computational biology, and a scientific critic until a final output is reached[134]. **e**, Examples of agentic laboratories, highlighting recent progress in wet-lab agentic automation and experimental closed-loop integration. ChemCrow interfaces multi-needle liquid handling, automated solid dosing and weighing, a high-throughput experimentation (HTE) reaction plate and optical characterization through 18 integrated tools. The agent used these tools together with IBM RoboRXN and local hardware to autonomously plan, execute and correct the syntheses of *N*,*N*-diethyl-*meta*-toluamide (DEET), three thiourea organocatalysts and a novel chromophore[41]. HeinSight2.0 integrates an EasyMax reactor with dosing and stirring hardware, a webcam, computer vision and iControl process data for real-time monitoring and feedback control of operations including distillation, crystallization, mixing and extraction[255]. The ChatGPT Research Group is a multi-agentic system that couples automated liquid handling with microwave synthesis for iterative optimization of MOF and covalent organic framework (COF) synthesis[26]. AI, artificial intelligence; CV, computer vision.

Table 3 | Representative applications of agentic AI systems with relevance for experimental discovery

Model name	Scientific domain	Task type	LLM	Agent type(s)	Reported outcome
SAGA[85, a]	Biology and chemistry	Experimental discovery	GPT-5 and Claude Sonnet 4.5	Cognitive; collaborative	Design of antibiotics, antibody protein binders, functional DNA sequences, magnet materials and chemical separation processes
GPT-5-driven autonomous laboratory[151, a]	Biochemistry	Experimental discovery	GPT-5	Cognitive; embodied	27% increase in protein production yield
ReactionSeek[131]	Organic chemistry	Literature mining	GLM-4V	Cognitive	Generation of a dataset from 102 volumes of the organic syntheses collection
CREst[163]	Materials science	Experimental discovery	GPT-4V	Cognitive; embodied; collaborative	Identification of metal catalysts for electrochemical formate oxidation
Virtual Lab[134]	Drug discovery	Experimental discovery	GPT-4o	Cognitive; collaborative	Design of nanobody binders for SARS-CoV-2 variants
ChemAgents[95]	Materials science	Experimental discovery	Llama-3.1	Cognitive; embodied; collaborative	Identification of heterogeneous catalyst for oxygen evolution reactions
CRISPR-GPT[35]	Biology	Experimental discovery	GPT-4o	Cognitive; collaborative	Achieved four CRISPR–Cas12a gene knockouts in lung adenocarcinoma cells and two dCas9 activations in melanoma cells
ORGANA[244]	General chemistry	Experimental orchestration	GPT-3.5	Cognitive; embodied; collaborative	Automation of various general laboratory tasks leveraging visual detector feedback
MERMaid[247]	General chemistry	Literature mining	Florence-2 and GPT-4o	Cognitive	Demonstration of automated reaction mining from scientific literature using both text and image data
El Agente Q[126] and Quntur[138]	Theoretical chemistry	Computational orchestration	GPT-4.1 and Claude Sonnet 3.7	Cognitive; collaborative	Automation of various quantum chemical calculation tasks
BioAgents[149]	Biology	Computational orchestration	Phi-3-mini	Cognitive; collaborative	Multistep bioinformatics workflow construction and evaluation on conceptual genomics tasks
CASCADE[248, a]	General science	Computational and experimental discovery	GPT-5, O3, Claude Sonnet 4.5	Cognitive; embodied; collaborative	Self-evolving skill acquisition across computational analysis and autonomous laboratory synthesis
LLM-RDF[249]	Catalysis	Experimental discovery	GPT-4	Cognitive; embodied; collaborative	Identification of adapted reaction conditions with improved performance; mechanistic analysis of the adapted method
ChemCrow[41]	Organic chemistry	Experimental orchestration	GPT-4	Cognitive; embodied	Planning and execution of several synthesis tasks; synthesis of a novel chromophore
Literature mining model[111]	Materials science	Literature mining	GPT-3 and Llama-2	Cognitive	Demonstration in various data extraction tasks from scientific text
Coscientist[36]	Organic chemistry	Experimental orchestration	GPT-4	Cognitive; embodied; collaborative	Planning and execution of several synthesis tasks
ChatGPT Chemistry Assistant[26]	Materials science	Literature mining	GPT-3.5 and GPT-4	Cognitive	Text mining of MOF synthesis conditions from diverse formats and styles
ChatGPT Research Group[33]	Materials science	Experimental orchestration	GPT-4	Cognitive; embodied; collaborative	Planning and execution of crystalline materials synthesis tasks

AI, artificial intelligence; LLM, large language model; MOF, metal–organic framework. [a]Available as a preprint as of 26 September 2026.

and documentation retrieval to operate a liquid handler. In optimization tasks, Coscientist was evaluated on Suzuki–Miyaura and Buchwald–Hartwig cross-coupling reaction datasets, where it subsequently achieved a normalized maximum advantage comparable with a standard Bayesian optimization baseline within a strictly bounded 20-iteration budget[36]. Despite these capabilities, the platform remained semi-autonomous, necessitating human intervention such as plate transport. At the same time, a materials discovery research team based on a multi-agentic system leveraged automated liquid handlers, programmable microwave reactors and automated powder X-ray diffraction to achieve the fast optimization of crystallinity and porosity of covalent organic framework materials[33]. More recently, a similar workflow was deployed to discover a stable perovskite absorber[150].

Subsequent systems expanded this paradigm into biological settings[151]. In a preliminary study, a GPT-5-guided cloud laboratory for cell-free protein synthesis reduced the specific cost of a benchmark protein by 40% relative to the best previously published formulation, with a 27% higher titre[151]. The largest single jump in performance coincided with human intervention, when the model was given internet and computing tools together with the state-of-the-art report, and the DNA template and cell lysate were independently improved. In parallel, CRISPR-GPT automates the design of CRISPR-based gene-editing

experiments, providing guidance in system selection, RNA design, choice in delivery method and protocol drafting, and has been validated in real gene-knockout and epigenetic-activation experiments[35].

Recent AI-XR laboratory systems complement instrument-centred automation with a more human-centred form of agentic embodiment at the laboratory bench. Through egocentric perception, these systems can observe experimental operations from the researcher's perspective and couple this information with agentic reasoning to provide real-time protocol guidance, detect errors and coordinate hand-offs between human and robotic actions[152]. Similarly, the Virtual Lab coordinated specialized agents through structured research meetings to design nanobody binders for SARS-CoV-2 variants[134] (Fig. 5d). Although human researchers still supplied essential high-level guidance and tool selection, the experimental screening of 92 agent-designed nanobodies, which successfully yielded 2 functional nanobodies with improved binding to target variants, highlights the efficacy of agentic teams in navigating complex, interdisciplinary biological spaces.

These studies suggest that although fully autonomous agentic laboratories for experimental discovery have not yet been realized, in the near term agentic systems will function more as adaptive co-pilots, augmenting human expertise within structured experimental workflows while progressively expanding their scope as integration with physical systems and validation frameworks matures (Fig. 5e). This trajectory may be particularly valuable for experimental domains that require simultaneous coordination of heterogeneous variables, ranging from food formulation[153] and shelf-life optimization[154] to biomaterials, synthetic biology, environmental monitoring and translational medicine.

Reproducibility and data deposition

Given the increasing role of learning-powered autonomy and path-dependent reasoning of agentic AI systems, ensuring the reproducibility of experimental discoveries generated by these complex workflows is of paramount importance. Extensive efforts are being directed towards constructing databases and benchmark environments to preserve the executable scientific cognition underlying agentic workflows.

Reproducibility

In current research on agentic AI systems, reproducibility is typically achieved through sharing open-source code and data[6]. Yet this standard is insufficient for agentic workflows, where subtle perturbations in prompting, retrieval context or tool availability may redirect downstream reasoning. Reproducibility should cover prompts, tool interfaces, memory and retrieval strategies, intermediate steps and any human involvement. Beyond details about the origin of retrieved content, reproducibility records should trace the origin and transformation history of agent outputs and decisions, including contributing tool results, agent critiques or consensus steps and human interventions. Furthermore, where code or models cannot be released because of confidentiality constraints[155], operational documentation becomes important for reproduction[156]. For instance, a study advocates the use of Model Cards[157] to outline model details, intended uses and relevant contextual factors. This concept refers to a short, standardized document that accompanies a released model and records the information needed for its appropriate use (for example, model characteristics, intended applications and pertinent limitations). Similar to an instrument specification sheet, it enables subsequent users to identify the experimental tasks for which the model is suited and recognize the caveats that accompany its use.

Furthermore, prompts deserve special attention as they have a central role in determining an agent's behaviour. A change in wording, context window, retrieval instruction or output structure can redirect planning and alter experimental outcomes, which explains the emergence of prompt engineering as a practical discipline[158]. Even when the same prompt is used, agentic systems can generate different outputs because of variability in model generation and retrieval, differences in tool outputs and updates to external databases[159,160]. As such, reproducibility reporting should distinguish between two classes of repetition: technical replicates, which repeat output generation from an already configured agentic system (for example, submitting the same prompt multiple times); and independent replicates, which recreate the agentic system from the documented experimental specification before repeating the complete workflow (for example, initializing the specified model, tools and information sources anew). Essentially, the former reveals how much the output can vary when the same system is repeatedly queried, whereas the latter establishes whether the same result can be obtained when the entire workflow is independently reconstructed and executed. Both classes should document whether final outputs and intermediate trajectories remain stable or diverge through quantitative measures, such as agreement rates, variance across runs and success rates or qualitative inspection of representative divergent trajectories.

Nevertheless, when stochastic agents sample diverse candidates from large search spaces, exact-output reproducibility may be neither attainable nor the most informative criterion. In such cases, combining results across multiple models or consensus across independent trajectories can identify findings that recur across runs. Because repeated-run and consensus analyses can be computationally expensive, reports should also state the number of runs and the associated computational budget. In this regard, the tools available to an agent can be viewed as part of its experimental apparatus and should hence be documented with the same care as instruments in a wet-lab. Beyond the identity and version of each tool, this includes the function it performs, the inputs it accepts, the outputs it returns and the conditions that govern its operation (for example, rate limits, authentication requirements, expected units or known failure modes). Such specifications matter as a change in how a tool receives or returns information can alter the evidence available to the agent and, in turn, its subsequent decisions. For example, ChemCrow[41] is supplied with a suite of tool names, functional descriptions and precise input and output specifications, thereby allowing the agent to expand its capabilities through the integration of new tools.

Another critical factor affecting the reproducibility of agentic AI systems is human intervention. Human intervention can support reproducibility when it is predefined and recorded as part of the workflow, such as human approval at safety-critical checkpoints[125,161], validation of intermediate outputs[57,162,163] or correction of failed tool calls according to explicit criteria[33,41,164]. Conversely, it can detract from reproducibility when researchers manually repair failed steps, change the search space, discard outputs, select candidates post hoc or modify experimental conditions without documenting the rationale. Reporting this clarifies for researchers when and how humans should step into the experimental workflow, which is especially crucial in partially automated systems. For example, the PARTNR benchmark[129] explicitly provides a HITL tool to evaluate how agents collaborate with real humans, specifically reporting the steps, task offloading, and performance with human intervention.

Data deposition and benchmarks

Regarding the data generated by agentic AI systems, deposition practices should strictly adhere to the FAIR principles[165]. Although data

Table 4 | Representative benchmarks for evaluating the capabilities of agentic AI systems across scientific and research-oriented tasks

Resource	Scale[a]	Task type[b]
AAAR-1.0 (ref. 250)	4 types of scientific tasks	Scientific discovery
AgentBench[128]	8 task environments	Daily interaction
ScienceAgentBench[127]	102 tasks	Scientific discovery
SUPER[251]	801 tasks	Scientific code repository reproduction
CORE-Bench[231]	270 tasks	Computational experiment reproduction
SWIF^{2}T (ref. 252)	300 tasks	Scientific writing feedback
MLGym[253]	13 tasks	Scientific discovery
DiscoveryWorld[254]	24 tasks	Scientific discovery
LAB-Bench[227]	2,400+ tasks (questions)	Biological research capabilities

[a]The size of each benchmark using the unit defined in the original study (for example, individual tasks or questions, complete task environments or broader task categories); these values are not directly comparable across resources. [b]The principal scientific capability or activity evaluated by each benchmark. The resources shown provide a representative cross-section of current evaluation practice across distinct capabilities relevant to scientific artificial intelligence (AI) agents.

derived from different experimental tasks are highly diverse and possess unique characteristics, it is strongly recommended to encapsulate the workflow, working environment and execution context together during deposition[166–168]. Submissions to data repositories should consistently include version numbers, comprehensive documentation, appropriate licensing and, where possible, containerized execution environments that package the software and dependencies needed to reproduce the workflow, together with environment files, such as Conda environment files, Docker or Singularity images, as well as files specifying exact software dependency versions or workflow records[165,167]. These materials help to preserve the software stack used in the original workflow and reduce the risk that changes in package versions, external tools or hosted services prevent future reproduction. For proprietary or remotely hosted components that cannot be fully containerized, authors should document the non-containerizable components, access dates, tool interfaces and execution logs to support approximate reproduction and auditing[165,167]. Several popular benchmark datasets for evaluating agentic AI systems across various domains are listed in Table 4, encompassing diverse task types that correspond to experimental discoveries across different scientific disciplines and complexity levels.

Limitations and optimizations

Although AI systems are compelling for experimental discovery, they often introduce vulnerabilities that constrain their reliability[45,169,170]. As autonomy increases and workflows lengthen, small imperfections in reasoning, knowledge or execution are likely to propagate and compound across successive steps, occasionally cascading into failures that range from subtly flawed conclusions to compromised laboratory safety[36,171–173]. Anticipating these failure modes is therefore imperative for deploying agents responsibly within real discovery pipelines, and we thus examine interrelated challenges that recur across current systems, including error accumulation over extended reasoning chains, knowledge defects such as hallucination and inherited bias, tool misuse, context saturation, coordination bottlenecks in multi-agent settings and the broader safety and ethical risks of autonomous experimentation. For each, we illustrate the mitigation and optimization strategies that have been and continue to be developed, underscoring that these limitations represent not fundamental barriers but active frontiers of engineering and governance.

The relative salience of these limitations also varies across the three agent categories. Cognitive agents, which often rely heavily on language-based reasoning, are particularly susceptible to error accumulation, knowledge defects and context saturation. Embodied agents, through their direct interaction with laboratory tools and physical environments, warrant particular attention to tool misuse, execution errors and experimental safety. Collaborative agents face more pronounced challenges involving communication distortion, task coordination and state synchronization. These risks are not independent, however; failures in one component may propagate through interconnected workflows and amplify downstream errors[90,174].

Error accumulation

The very feature that renders agentic systems compelling for scientific discovery — their capacity for sustained multistep reasoning — also creates an inherent vulnerability to error accumulation[175,176]. Although agents possess certain planning capabilities, their concrete decision-making is path-dependent, with each step strongly conditioning the next. Even correctly executed early steps do not immunize later decisions against drift; prolonged trajectories can privilege local gains over global scientific intent, so a small displacement can lead to the wrong destination[177]. To mitigate error accumulation, explicit planning strategies that separate the planning phase from reasoning and execution have proven highly effective[178,179]. Additionally, iterative self-refinement[180–182] and dynamic replanning approaches[183,184] have been proposed.

Knowledge defects

Hallucinations and unsupported inference arise from multiple, often co-occurring causes, including retrieval failure or the absence of relevant evidence[185], the stochastic nature of decoding[186], gaps and biases in prompts or training data[187] and an incomplete grasp of the causal relationships between actions[188,189]. Current approaches seek to constrain these tendencies through retrieval anchoring[190,191], implementing self-verification mechanisms[192] or quantitatively assessing the hallucinations[193]. Moreover, because scientific experimental data used in model training are often sparse, with underlying patterns or structures that reflect the systems being studied, agents may inherit biases from these training data, thereby skewing the final outcomes[32,194]. One way to limit this influence is to separate experimental decisions from the expectations already encoded by the agent through randomization and blinding[195]. Specifically, randomization determines which conditions are tested without regard to which outcome the agent considers most promising. In turn, blinding can extend this separation to experimental execution and analysis, as the identity of experimental groups is withheld from those who might influence the outcome. Remarkably, neither measure corrects the model itself, removes biases from its training data or guarantees an unbiased result; instead, they limit the extent to which inherited biases can pass from model predictions to experimental decisions and, ultimately, scientific conclusions. Finally, agents may

arrive at conclusions that appear well reasoned despite lacking the evidence required to support them[196], a failure referred to as mirage reasoning[197]. Put simply, the agent appears to reason from scientific evidence even when that evidence has not actually been provided. The extent of this problem was illustrated in a recent study, in which multimodal models generated detailed descriptions and reasoning traces for images that were never supplied; remarkably, a text-only model with no access to images ranked first on a standard chest X-ray benchmark[197]. Such results reveal that apparently strong benchmark performance can reflect patterns in the surrounding text rather than genuine interpretation of the underlying data. A direct way to detect this failure is to remove or alter the evidence and determine whether the agent changes its conclusion accordingly; persistence of the same interpretation despite the loss of its supporting evidence provides a warning that the reasoning may be spurious. Although such failures can be exposed through controlled tests, their prevention remains less straightforward. The same study found that performance declined when models were explicitly informed that the expected input was absent, which suggests that clearly specifying what information is and is not available can reduce opportunities for unsupported inference. More generally, controlled tests involving withheld or perturbed inputs can reveal the extent to which an agent's conclusions actually depend on the evidence provided. Resolving these reasoning deficits remains an important step towards reliable scientific agents.

Tool misuse

Tool misuse is a unique vulnerability in agentic AI systems. Tool use requires an agent to determine which tool to invoke, what inputs and operating parameters to provide, and when and in what sequence the tool should be used. Misuse can entail selection of an inappropriate tool, operation outside its intended conditions or execution at the wrong stage of a workflow. In a physical laboratory, these decisions can translate an error in reasoning into an experimental action with real-world consequences. Such errors not only degrade the reliability of results but also pose severe safety risks, including potential damage to physical laboratory equipment[198]. The most prevalent issue is incorrect tool selection[172]. Research indicates that these errors frequently stem from low-quality retrieval processes, suggesting that model fine-tuning via specialized training datasets can offer a viable solution[199]. Furthermore, agents may execute unauthorized operations or experience communication breakdowns. Current optimization strategies to counter these issues include strict privilege separation[172], self-verification protocols[41,200] and standardized orchestration layers[201].

Context saturation

During extended experimental workflows, context length limitations prevent agents from fully processing comprehensive experimental protocols and upstream results. As the workflow lengthens, response latency increases and the accumulated record eventually exceeds the context window. The earliest material is then truncated or summarized away, so the agent can no longer consult the original protocol, the constraints it was given or the results of early runs. This resembles forgetting, but nothing signals that it has occurred[133,172]. The agent continues to respond fluently and may quietly violate a limit set at the outset, such as a temperature ceiling or an excluded reagent. To overcome this bottleneck, retrieval-augmented generation has been widely adopted[202], whose core principles involve segmenting and indexing experimental documents, followed by retrieving only the most task-relevant fragments to serve as context for agents. Additional optimization methods include context compaction[78], iterative summaries generated by LLMs[76,79] and hard clearance techniques to systematically manage memory load.

Coordination bottleneck

Multi-agent systems leverage collaboration among agents to execute tasks that surpass the capabilities of any single entity. However, overall performance does not necessarily scale linearly with the number of agents; beyond a certain point, or without adequate coordination, adding more agents can degrade system performance[203]. This deterioration is commonly attributed to uneven task distribution and communication barriers among agents[204,205]. Establishing a robust task allocation system is therefore essential, and requires careful optimization based on the specific capabilities and resource constraints of distinct execution units[206]. Furthermore, managing complex and hierarchical context information is vital for eliminating communication bottlenecks and ensuring that individual agent objectives align with overarching goals. Recent studies have successfully implemented graph-structured memory management frameworks to achieve such alignment[84,207].

Safety and ethical issues

Similar to human scientists, agentic AI systems may encounter risks during experimentation, including chemical explosions, pathogen leaks and unsafe autonomous operation[208]. They also introduce information risks, as AI-generated data can re-enter training corpora and contaminate future systems with unverified claims[209,210], whereas transparent retrieval and connected laboratory records may expose private or proprietary information[211].

At a fundamental level, the concern is twofold: an agent may generate scientific content that bears the appearance of evidence without having a verifiable scientific basis, or it may accept such content as evidence when introduced by another agent, a human or an external information source. In materials science, for example, AI-generated microscopy images can be sufficiently realistic that even experts may struggle to distinguish them from authentic experimental data[173]. One can consider this problem as being especially consequential in agentic scientific discovery, where generated information may not remain a final output; it can be passed to other agents, used to select subsequent experiments or retained as knowledge for later decisions. Indeed, redundant publications derived from a single public dataset increased 17-fold between 2022 and 2024 (ref. 212), and increasingly realistic synthetic scientific images continue to challenge available methods for identifying fabricated content[213]. At issue, then, is not only the generation of unsupported information; once admitted into an agentic workflow, it may become the basis for subsequent reasoning and experimental action.

This risk becomes still more difficult to control when the systems producing scientific outputs are insufficiently transparent. Many frontier models are proprietary, which limits access to information needed to establish how an output arose (that is, training data, model specifications and inference settings)[214,215]. Moreover, hosted models can change over time, such that the system available when a result is revisited may no longer be identical to that used in the original study. Some journals have consequently raised concerns about scientific claims that depend on closed models that cannot be independently inspected[216]. For agentic scientific discovery, one can view provenance as being part of the evidential record itself; the model version, prompt, retrieved sources and intermediate steps establish how information entered the workflow and how it contributed to subsequent decisions.

Glossary

Abductive inference
A form of reasoning that proposes the most plausible explanation for an observation, as distinct from deducing a conclusion that follows necessarily from its premises.

Agent harness
Runtime scaffolding that coordinates model and tool interactions, manages context and state, and supports iterative execution in an agentic system.

Agentic tree search
A multi-path planning strategy in which an agent explores alternative action sequences as branches of a tree, expands promising branches and abandons those that fail at intermediate steps.

Chain-of-thought
The step-by-step thinking mode that an artificial intelligence (AI) model uses to work from a question towards an answer or decision.

Cloud laboratories
Remotely accessible experimental infrastructures in which users submit experimental tasks online, after which automated laboratory platforms or service providers perform the experiments and return data or results.

Context compaction
The reduction of accumulated information in a model's context by summarizing, compressing or removing earlier material.

Context window
The maximum amount of text a model can take into account at one time, which covers instructions, retrieved material, tool outputs and the record of preceding steps.

Dry-run primitives
A preliminary pass that tests inputs, parameters and physical configurations before a job or experiment is formally executed, in order to identify invalid settings surface in advance.

Egocentric perception
A first-person mode of sensing in which head-mounted or wearable cameras record the scene from the researcher's own viewpoint.

Ethical drift
The gradual departure of a field's practices or a system's behaviour from agreed ethical standards, which happens incrementally rather than through a single clear breach.

Foundation models
Artificial intelligence (AI) models trained on broad and diverse datasets that can be adapted or used directly for many downstream tasks, such as language processing, reasoning, prediction and multimodal analysis.

Graph-based retrieval
A retrieval strategy that traverses a graph of entities and their relationships, so that connected information can be gathered rather than isolated passages.

Hallucination
A response generated by an artificial intelligence (AI) model that includes plausible-sounding but unsupported or fabricated information.

High-throughput experimentation
Experimental strategies in which large numbers of experiments are performed in parallel or in rapid sequence on automated platforms.

Human-in-the-loop
(HITL). A design principle in which human experts remain involved in artificial intelligence (AI) or automated workflows through supervision, judgement, approval or intervention at selected stages.

Internet of Things
Networked systems in which sensors, software and communication technologies connect laboratory equipment, instruments, robots, computers and databases to enable real-time data acquisition, transfer and exchange.

Lab-as-a-Service
A service model in which laboratory capabilities are provided on demand, often supported by cloud laboratories or automated experimental platforms that perform experimental tasks for remote users.

Large language models
(LLMs). Artificial intelligence (AI) models trained on large-scale text corpora, capable of processing and generating human-like natural language.

Machine learning
A computational approach in which algorithms identify patterns and relationships within data and use them to make predictions or improve performance on subsequent tasks.

Mirage reasoning
Output in which a model produces detailed descriptions and reasoning about an input it was never given, such as an image that was not, in fact, supplied.

Model Context Protocol
(MCP). A protocol for connecting agents with external environments, tools and data sources, allowing models to access contextual information and perform actions through more standardized interfaces.

Multi-hop retrieval
Retrieval performed in successive steps, in which the results of one query are used to formulate the next; this makes it possible to answer questions that require several pieces of evidence to be chained.

Prompt engineering
The systematic design and refinement of instructions given to an artificial intelligence (AI) model to guide its behaviour and outputs.

Prompts
The input that goes into artificial intelligence (AI).

Reasoning models
Artificial intelligence (AI) models, often large language models (LLMs), designed to perform multistep analysis before producing an answer or selecting an action.

Retrieval
The process of selecting relevant items from an external store of documents or data and supplying them to a model as context.

Retrieval-augmented generation
A generation strategy in which an agent retrieves relevant information from an external knowledge base before producing an answer.

Reusable capability modules
Packaged units of agent functionality that are described in machine-readable form so that other agents can discover and invoke them, and that can be reused and composed across different tasks.

Self-driving laboratories
Experimental platforms that integrate automated instrumentation, artificial intelligence (AI) systems and closed-loop decision strategies to allow experimental outcomes to inform the selection of subsequent experiments.

Skills
Structured collections of actions or capabilities that are packaged in formats accessible to agents, enabling task execution through reusable and composable functional modules.

Surrogate models
Computational models trained or constructed to approximate the output of a more expensive calculation, simulation or experiment, allowing candidate conditions or designs to be screened more rapidly.

Symbiotic autonomy
An operating model in which automated systems and human researchers each retain control over the decisions they handle best.

Box 1 | Safeguards for agentic scientific discovery

The autonomy that makes agentic artificial intelligence (AI) systems powerful also creates safety, security and societal risks that must be carefully managed.

- Human oversight and limits on autonomy. Agents should be given greater autonomy only after demonstrating safe performance[171]. High-risk actions should remain subject to expert review, audit trails and institutional oversight (for example, with Institutional Review Board-style approval where appropriate), supported by governance mechanisms that combine human supervision, agent alignment and environmental feedback[174].
- Dual-use screening. Agentic systems should screen for hazardous materials, dual-use applications and unsafe procedures. Safeguards such as input and output filtering, restricted access to sensitive capabilities and protection against jailbreak and prompt-injection attacks can help to reduce misuse[57,208]. More specifically, the point at which the malicious instruction enters determines the type of attack. In a jailbreak, a request devised by the user leads the agent to answer something it should have refused; in a prompt injection, the instruction is concealed within material the agent reads while it works, such as a web page, a supplementary file or a database field, which also makes it easier to reach the agent.
- Managing physical risks. In autonomous laboratories, robots handling hazardous materials can pose substantial safety risks. Protective measures such as hardware safety controls, computer-vision monitoring and formal risk assessments should be built into both laboratory equipment and control software[171].
- Verifying information quality. Grounding outputs in reliable external sources can improve accuracy and traceability[202]. Because information sources can be manipulated or corrupted, provenance checks, source verification and cross-checking against trusted references are important safeguards.
- Protecting scientific data. Standardized workflows, provenance tracking and execution records support reproducibility[201]. These measures should be complemented by clear labelling of AI-generated data and preservation of human-generated records to maintain data quality and prevent over-reliance on synthetic content.
- Preserving privacy and confidentiality. Shared laboratory records and retrieval systems may expose proprietary or sensitive information. Privacy-preserving approaches such as federated learning and differential privacy can help to protect data while still enabling collaboration and model development[217]. In federated learning, the model is sent to each site and trained on data that never leave it, so only parameter updates are exchanged. Differential privacy takes a different route and adds calibrated noise, so that the output cannot reveal whether any single record was included.

Only with this chain intact can researchers determine which conclusions are grounded in evidence and which arose from unsupported model generation. In response, scientists have proposed risk-stratification models and regulatory frameworks that enforce rigorous human oversight to mitigate automation biases and risks[171,174,217] (Box 1).

A further concern is that an agent drawing indiscriminately on the literature of a field may also produce hazardous knowledge, which in chemistry and biology means information whose application endangers human health or life. Where a conventional language model supplies such information as text, an agent plans routes and drives instruments, and can therefore carry the same information towards a material product; this capacity for execution is what raises the stakes of hazardous chemical and biological knowledge. The planning capability creates a difficulty of its own, as an agent can decompose a prohibited objective into subtasks that each appear innocuous in isolation, so that safeguards screening one request at a time may be satisfied at every step whereas the trajectory as a whole is not. Restricting hazardous knowledge at the source is the obvious alternative, yet dual use blocks that route as well because knowledge developed for beneficial ends can be turned towards harm, in some cases with little more than a change of objective. One example of an agentic system designed to address these concerns is El Agente Seguro[57], a framework for chemical safety that coordinates specialized AI agents to detect dual-use risks and select appropriate protective equipment using structured information exchange. Prioritizing early dual-use screening and prohibited-chemical constraints in the reasoning process reduces the risk of generating actionable synthesis guidance for sensitive substances, even under adversarial jailbreaking attempts. At the same time, experts are calling for new governance frameworks to control the threat of hazardous knowledge posed by agents[208], and such safeguards should focus more on the intent of the user[218]. On this basis, human oversight and dual-use screening have already been applied to address such risks (Box 1).

In addition to laboratory hazards, the computational intensity of agentic pipelines carries a substantial environmental cost, where the model training and repeated inference that underpin sustained multistep reasoning consume considerable energy and water and generate carbon emissions that scale with the length and frequency of autonomous workflows[219–221]. Model capacity and reasoning budget should therefore be allocated according to the complexity, uncertainty and risk of each subtask, using smaller or specialized models for routine operations and reserving more capable or computationally intensive models for steps that require them. Equally pressing are equity concerns, as the computational resources, proprietary foundation models and instrumented laboratories required for agentic discovery remain concentrated within a small number of well-resourced institutions and nations, risking a widening disparity that could exclude researchers in less-resourced settings from both contributing to and benefiting from these advances[222].

Beyond technical hurdles, the field must also guard against ethical drift. Agent capabilities can be overstated and misinterpreted; hence, reports should distinguish autonomous action from curated search spaces, protected toolboxes and expert correction. Their findings should also be examined carefully to preserve scientific rigour and mandate objective and factual reporting. Furthermore, although endowing agents with autonomous discovery capabilities remains a primary focus, foundational tasks such as database curation and infrastructure maintenance must not be marginalized despite their perceived lack of

novelty. Overstated autonomy also obscures a harder question, namely who answers for an error or a harmful outcome. Responsibility cannot rest with the system itself, as an agent has no legal standing and cannot be held to account[223], which is why scientific publishing holds human authors fully answerable for every part of a manuscript, including the parts an AI tool produced[224]. In autonomous laboratories the same principle points to the approval step, where named researchers who understand the experiment sign off on the plan and the executable code before it runs[106]. These problems are compounded by persistent barriers in interdisciplinary communication. Ultimately, fostering cross-disciplinary participation and formulating equitable regulations will therefore be central to the ethical evolution of agentic AI systems[225].

Outlook

As AI continues to advance, scientific experimentation is increasingly shifting from isolated trial and error towards coordinated systems. Although agentic AI is increasingly envisioned as a general methodology for research, the central challenge ahead lies in cultivating scientific agency itself rather than continually pursuing bespoke models. The field must therefore move towards systems judged by the discoveries they enable and by the traceable paths through which those discoveries are reached, thereby advancing a more calibrated, reproducible and accountable form of science.

Trustworthy scientific agency

Current discussions surrounding error accumulation and hallucination suggest that agentic AI systems remain incompletely controllable. Crucially, agents must be capable of accurately quantifying and communicating uncertainty in their own conclusions, and of triggering human oversight or validation processes when necessary. A particularly important capability for agentic AI systems is accurate calibration[226]. The confidence expressed by a system should closely reflect its true likelihood of successful task completion. This correspondence should be maintained throughout the entire course of action, not merely in the final output. Consequently, future scientific agents may be assessed not only by their success rates but also by how reliably they identify and communicate their failures.

Standardized evaluation

At present, the field of agentic AI still lacks a unified and comparable evaluation framework. Although emerging benchmarks increasingly probe scientific capabilities[127,227,228], different systems frequently demonstrate their proficiency through isolated case studies or highly specific workflows[39]. The field is in need of a shared discovery test bed with the organizing force that MNIST[229] and CIFAR-10 (ref. 230) once provided for image classification, albeit with scientific breadth beyond static labels. Based on most currently proposed benchmarks, agentic AI systems achieve relatively low accuracy rates in task completion[231,232], largely because they must navigate dynamic and interactive environments rather than static data. Consequently, evaluation must move beyond final-answer correctness towards metrics at the process and system levels, including trajectory quality, tool-use robustness, human-intervention burden, cross-run stability and downstream experimental utility.

Interdisciplinary expansion

The modular structure of agentic systems renders them intrinsically transferable across disciplines and suitable for expansion into real experimental infrastructure[39]. To forge a deeper coupling, agents must be seamlessly connected with cloud laboratories, the Internet of Things, robotic platforms and Lab-as-a-Service architectures[15,36,233]. The ability to couple digital reasoning with real-time spatial perception, procedural guidance and embodied assistance in physical environments is equally important[42]. The laboratory can then become a navigable landscape in which the literature, computation, measurement and actuation inform one another in real time instead of remaining confined to reasoning within a digital space. Unlike traditional, highly specialized predictive models, agentic AI systems can use a common underlying architecture across different scientific domains. This transferability is particularly relevant for domains with heterogeneous experimental variables, from food formulation[153] and shelf-life optimization[154] to biomaterials[234] and synthetic biology[12]. Looking ahead, this structural independence will facilitate the breaking down of disciplinary silos, paving the way for innovative scientific discoveries across highly intersecting fields[39,41,235].

Collective intelligence

By its very nature, experimental discovery has historically depended on networks of researchers, instruments, institutions and accumulated knowledge. Given that the concept of collective intelligence has been extensively studied in the context of human societies[236], it is envisioned that agentic AI will extend this idea into coordinated human–machine ecosystems composed of researchers, agents, databases, robotic platforms, simulation engines and continuous feedback mechanisms[123,134,237]. In this paradigm, discovery no longer depends on a single super-model, but on distributed systems capable of shared memory, iterative refinement and collaborative specialization.

In summary, agentic AI systems are extending beyond the role of task assistants to research methodologies that coordinate memory, tools, simulations and experiments within a shared scientific trajectory. Their ultimate significance lies in augmenting the efficiency of researchers while sustaining reproducibility and accountable decision-making across experimental workflows. In this regard, agentic experimentation has the capacity to establish a new collaborative mode for scientific discovery and thereby evolve into a continuously learning research tool.

Published online: 22 October 2026

References

1. Lyu, H., Ji, Z., Wuttke, S. & Yaghi, O. M. Digital reticular chemistry. *Chem* **6**, 2219–2241 (2020).
2. Mennen, S. M. et al. The evolution of high-throughput experimentation in pharmaceutical development and perspectives on the future. *Org. Process. Res. Dev.* **23**, 1213–1242 (2019).
3. Mounet, N. et al. Two-dimensional materials from high-throughput computational exfoliation of experimentally known compounds. *Nat. Nanotechnol.* **13**, 246–252 (2018).
4. Bräse, S. Digital chemistry: navigating the confluence of computation and experimentation—definition, status quo, and future perspective. *Digit. Discov.* **3**, 1923–1932 (2024).
5. Himanen, L., Geurts, A., Foster, A. S. & Rinke, P. Data-driven materials science: status, challenges, and perspectives. *Adv. Sci.* **6**, 1900808 (2019).
6. Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
7. Tom, G. et al. Self-driving laboratories for chemistry and materials science. *Chem. Rev.* **124**, 9633–9732 (2024).
8. King, R. D. et al. The automation of science. *Science* **324**, 85–89 (2009).
9. Sparkes, A. et al. Towards Robot Scientists for autonomous scientific discovery. *Autom. Exp.* **2**, 1 (2010).
10. Granda, J. M., Donina, L., Dragone, V., Long, D.-L. & Cronin, L. Controlling an organic synthesis robot with machine learning to search for new reactivity. *Nature* **559**, 377–381 (2018).
11. Burger, B. et al. A mobile robotic chemist. *Nature* **583**, 237–241 (2020).
12. Martin, H. G. et al. Perspectives for self-driving labs in synthetic biology. *Curr. Opin. Biotechnol.* **79**, 102881 (2023).
13. Dai, T. et al. Autonomous mobile robots for exploratory synthetic chemistry. *Nature* **635**, 890–897 (2024).

14. Rapp, J. T., Bremer, B. J. & Romero, P. A. Self-driving laboratories to autonomously navigate the protein fitness landscape. *Nat. Chem. Eng.* **1**, 97–107 (2024).
15. Szymanski, N. J. et al. An autonomous laboratory for the accelerated synthesis of inorganic materials. *Nature* **624**, 86–91 (2023).
16. Xu, Y. et al. LUMI-lab: a foundation model-driven autonomous platform enabling discovery of ionizable lipid designs for mRNA delivery. *Cell* **189**, 1620–1635.e25 (2026).
17. Bayley, O., Savino, E., Slattery, A. & Noël, T. Autonomous chemistry: navigating self-driving labs in chemical and material sciences. *Matter* **7**, 2382–2398 (2024).
18. Beal, J. & Rogers, M. Levels of autonomy in synthetic biology engineering. *Mol. Syst. Biol.* **16**, MSB202010019 (2020).
19. Senior, A. W. et al. Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710 (2020).
20. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
21. Feng, B. et al. Interpreting X-ray diffraction patterns of metal–organic frameworks via generative artificial intelligence. *J. Am. Chem. Soc.* **148**, 869–878 (2026).
22. Xie, Y., Sattari, K., Zhang, C. & Lin, J. Toward autonomous laboratories: convergence of artificial intelligence and experimental automation. *Prog. Mater. Sci.* **132**, 101043 (2023).
23. Sanchez-Lengeling, B. & Aspuru-Guzik, A. Inverse molecular design using machine learning: generative models for matter engineering. *Science* **361**, 360–365 (2018).
24. Schwaller, P. et al. Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chem. Sci.* **11**, 3316–3325 (2020).
25. Gao, W., Luo, S. & Coley, C. W. Generative AI for navigating synthesizable chemical space. *Proc. Natl Acad. Sci. USA* **122**, e2415665122 (2025).
26. Zheng, Z., Zhang, O., Borgs, C., Chayes, J. T. & Yaghi, O. M. ChatGPT Chemistry Assistant for text mining and the prediction of MOF synthesis. *J. Am. Chem. Soc.* **145**, 18048–18062 (2023).
27. Antunes, L. M., Butler, K. T. & Grau-Crespo, R. Crystal structure generation with autoregressive large language modeling. *Nat. Commun.* **15**, 10570 (2024).
28. Ma, Z. et al. ProtTeX: structure-in-context reasoning and editing of proteins with large language models. *J. Chem. Inf. Model.* **65**, 6599–6612 (2025).
29. Pan, H. et al. Quantum many-body physics calculations with large language models. *Commun. Phys.* **8**, 49 (2025).
30. Xie, Z. et al. Autonomous chemistry and materials innovation driven by scientific agents. *JACS Au* https://doi.org/10.1021/jacsau.6c00213 (2026).
31. MacKnight, R. et al. Rethinking chemical research in the age of large language models. *Nat. Comput. Sci.* **5**, 715–726 (2025).
32. Zhang, Y. et al. Exploring the role of large language models in the scientific method: from hypothesis to discovery. *npj Artif. Intell.* **1**, 14 (2025).
33. Zheng, Z. et al. ChatGPT Research Group for optimizing the crystallinity of MOFs and COFs. *ACS Cent. Sci.* **9**, 2161–2170 (2023).
34. Jablonka, K. M., Schwaller, P., Ortega-Guerrero, A. & Smit, B. Leveraging large language models for predictive chemistry. *Nat. Mach. Intell.* **6**, 161–169 (2024).
35. Qu, Y. et al. CRISPR-GPT for agentic automation of gene-editing experiments. *Nat. Biomed. Eng.* **10**, 245–258 (2026).
36. Boiko, D. A., MacKnight, R., Kline, B. & Gomes, G. Autonomous chemical research with large language models. *Nature* **624**, 570–578 (2023).
37. Gao, S. et al. Empowering biomedical discovery with AI agents. *Cell* **187**, 6125–6151 (2024).
38. Schmidgall, S. et al. Agent Laboratory: using LLM agents as research assistants. In *Findings of the Association for Computational Linguistics: EMNLP 2025* (eds Christodoulopoulos, C. et al.) 5977–6043 (Association for Computational Linguistics, 2025).
39. Wang, H. et al. Scientific discovery in the age of artificial intelligence. *Nature* **620**, 47–60 (2023).
40. Gridach, M., Nanavati, J., Mack, C., El Abidine, K. Z. & Mendes, L. Agentic AI for scientific discovery: a survey of progress, challenges, and future directions. In *ICLR 2025 Workshop on Agentic AI for Science* https://iclr.cc/virtual/2025/33138 (ICLR, 2025).
41. M Bran, A. et al. Augmenting large language models with chemistry tools. *Nat. Mach. Intell.* **6**, 525–535 (2024).
42. Wu, Y. C. et al. MedOS: AI-XR-Cobot world model for clinical perception and action. Preprint at *medRxiv* https://doi.org/10.64898/2026.02.18.26345936 (2026).
43. Lu, C. et al. Towards end-to-end automation of AI research. *Nature* **651**, 914–919 (2026).
44. Coley, C. W. et al. A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science* **365**, eaax1566 (2019).
45. Ramos, M. C., Collison, C. J. & White, A. D. A review of large language models and autonomous agents in chemistry. *Chem. Sci.* **16**, 2514–2572 (2025).
46. Liang, H. et al. Real-time experiment–theory closed-loop interaction for autonomous materials science. *Sci. Adv.* **11**, eadu7426 (2025).
47. Koscher, B. A. et al. Autonomous, multiproperty-driven molecular discovery: from predictions to measurements and back. *Science* **382**, eadi1407 (2023).
48. Nisa, U., Shirazi, M., Saip, M. A. & Pozi, M. S. M. Agentic AI: the age of reasoning—a review. *J. Autom. Intell.* **5**, 69–89 (2026).
49. Sinha, A., Arun, A., Goel, S., Staab, S. & Geiping, J. The illusion of diminishing returns: measuring long horizon execution in LLMs. In *Proc. 14th Int. Conf. on Learning Representations* (eds Vondrick, C. et al.) 35184–35211 (ICLR, 2026).
50. Zhao, Q. et al. Uncertainty propagation on LLM agent. In *Proc. 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (eds Che, W. et al.) 6064–6073 (Association for Computational Linguistics, 2025).
51. Hou, X., Zhao, Y., Wang, S. & Wang, H. Model Context Protocol (MCP): landscape, security threats, and future research directions. *ACM Trans. Softw. Eng. Methodol.* **35**, 302 (2026).
52. Wu, Y. et al. VILA-U: a unified foundation model integrating visual understanding and generation. In *Proc. 13th Int. Conf. on Learning Representations* (eds Yue, Y. et al.) 93620–93638 (ICLR, 2025).
53. NVIDIA et al. Nemotron 3 Nano Omni: efficient and open multimodal intelligence. Preprint at https://doi.org/10.48550/arXiv.2604.24954 (2026).
54. Cao, S. et al. Automating quantum computing laboratory experiments with an agent-based AI framework. *Patterns* **6**, 101372 (2025).
55. Cao, S. et al. QCalEval: benchmarking vision-language models for quantum calibration plot understanding. Preprint at https://doi.org/10.48550/arXiv.2604.25884 (2026).
56. Munguia-Galeano, F. et al. Chemist Eye: a visual language model-powered system for safety monitoring and robot decision-making in self-driving laboratories. *Digit. Discov.* **5**, 2209–2220 (2026).
57. Kang, Y. et al. El Agente Seguro: an agent for chemical safety. Preprint at https://doi.org/10.26434/chemrxiv.15004378/v1 (2026).
58. Choi, C. et al. El Agente Estructural: an artificially intelligent molecular editor. Preprint at https://doi.org/10.48550/arXiv.2602.04849 (2026).
59. Jamali, V., Aghazadeh, A. & Kacher, J. Thinking microscopes: agentic AI and the future of electron microscopy. *npj Comput. Mater.* **12**, 149 (2026).
60. OpenAI API Platform Documentation. *OpenAI Developers* https://developers.openai.com/api/docs (2026).
61. McNaughton, A. D. et al. CACTUS: Chemistry Agent Connecting Tool Usage to Science. *ACS Omega* **9**, 46563–46573 (2024).
62. Wu, M. et al. CheMatAgent: enhancing LLMs for chemistry and materials science through tree-search based tool learning. Preprint at https://doi.org/10.48550/arXiv.2506.07551 (2025).
63. Chandrasekhar, V., Sharma, N., Schaub, J., Steinbeck, C. & Rajan, K. Cheminformatics Microservice: unifying access to open cheminformatics toolkits. *J. Cheminformatics* **15**, 98 (2023).
64. Hicks, C. B. & Martinez, T. J. Massively scalable workflows for quantum chemistry: BigChem and ChemCloud. *J. Chem. Phys.* **160**, 142501 (2024).
65. Li, S.-C. & Burns, J. The Chemprop Model Context Protocol. Preprint at https://doi.org/10.26434/chemrxiv-2025-tsx5s (2025).
66. Yang, K. et al. Analyzing learned molecular representations for property prediction. *J. Chem. Inf. Model.* **59**, 3370–3388 (2019).
67. Zhang, Z. et al. El Agente Forjador: task-driven agent generation for quantum simulation. Preprint at https://doi.org/10.48550/arXiv.2604.14609 (2026).
68. Qiu, J. et al. Alita: generalist agent enabling scalable agentic reasoning with minimal predefinition and maximal self-evolution. Preprint at https://doi.org/10.48550/arXiv.2505.20286 (2025).
69. Sumers, T. R., Yao, S., Narasimhan, K. & Griffiths, T. L. Cognitive architectures for language agents. In *Trans. Mach. Learn. Res.* (2024).
70. Guo, D. et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature* **645**, 633–638 (2025).
71. Li, S. et al. GraphReader: building graph-based agent to enhance long-context abilities of large language models. In *Findings Assoc. Comput. Linguist.: EMNLP 2024* (eds Al-Onaizan, Y., Bansal, M. & Chen, Y.-N.) 12758–12786 (Association for Computational Linguistics, 2024).
72. Ma, Q., Zhou, Y. & Li, J. Automated retrosynthesis planning of macromolecules using large language models and knowledge graphs. *Macromol. Rapid Commun.* **47**, 2500065 (2026).
73. K-Dense. Scientific Agent Skills. *GitHub* https://github.com/K-Dense-AI/scientific-agent-skills (2026).
74. Edge, D. et al. From local to global: a graph RAG approach to query-focused summarization. Preprint at https://doi.org/10.48550/arXiv.2404.16130 (2025).
75. Bai, J. et al. El Agente Gráfico: structured execution graphs for scientific agents. Preprint at https://doi.org/10.48550/arXiv.2602.17902 (2026).
76. Chhikara, P., Khant, D., Aryan, S., Singh, T. & Yadav, D. Mem0: building production-ready AI agents with scalable long-term memory. In *ECAI 2025* (eds Lynce, I. et al.) 2993–3000 (SAGE Publications, 2026).
77. Gao, Y. et al. Retrieval-augmented generation for large language models: a survey. Preprint at https://doi.org/10.48550/arXiv.2312.10997 (2024).
78. Xu, W. et al. A-MEM: agentic memory for LLM agents. *Adv. Neural Inf. Process. Syst.* **38**, 17577–17604 (2026).
79. Kang, J., Ji, M., Zhao, Z. & Bai, T. Memory OS of AI agent. In *Proc. 2025 Conf. Empir. Methods Nat. Lang. Process.* (eds Christodoulopoulos, C. et al.) 25961–25970 (Association for Computational Linguistics, 2025).
80. Shen, Y. et al. SciAgentGym: benchmarking multi-step scientific tool-use in LLM agents. Preprint at https://doi.org/10.48550/arXiv.2602.12984 (2026).
81. Muhoberac, M. et al. State and memory is all you need for robust and reliable AI agents. Preprint at https://doi.org/10.48550/arXiv.2507.00081 (2025).
82. Liu, Y. et al. Benchmarking LLM-based agents for single-cell omics analysis. *Genome Biol.* **27**, 123 (2026).
83. Alam, K. & Roy, B. From prompt to pipeline: large language models for scientific workflow development in bioinformatics. Preprint at https://doi.org/10.48550/arXiv.2507.20122 (2025).
84. Bai, J. et al. A dynamic knowledge graph approach to distributed self-driving laboratories. *Nat. Commun.* **15**, 462 (2024).

85. Du, Y. et al. Accelerating scientific discovery with autonomous goal-evolving agents. Preprint at https://doi.org/10.48550/arXiv.2512.21782 (2026).
86. Asche, S., Cooper, G. J. T., Keenan, G., Mathis, C. & Cronin, L. A robotic prebiotic chemist probes long term reactions of complexifying mixtures. *Nat. Commun.* **12**, 3547 (2021).
87. Bran, A. M., Neukomm, T. A., Armstrong, D., Jončev, Z. & Schwaller, P. Chemical reasoning in LLMs unlocks strategy-aware synthesis planning and reaction mechanism elucidation. *Matter* **9**, 102812 (2026).
88. Fushimi, K. et al. Development of the autonomous lab system to support biotechnology research. *Sci. Rep.* **15**, 6648 (2025).
89. Wang, W. et al. Agentic Lab: an agentic-physical AI system for cell and organoid experimentation and manufacturing. Preprint at *bioRxiv* https://doi.org/10.1101/2025.11.11.686354 (2025).
90. Kim, Y. et al. Capable language models can outgrow the benefits of collaboration. *Nat. Mach. Intell.* **8**, 1157–1172 (2026).
91. Zhang, Z. et al. A survey on the memory mechanism of large language model-based agents. *ACM Trans. Inf. Syst.* **43**, 1–47 (2025).
92. Vriza, A., Prince, M. H., Zhou, T., Chan, H. & Cherukara, M. J. Operating advanced scientific instruments with AI agents that learn on the job. *npj Comput. Mater.* **12**, 160 (2026).
93. Besta, M. et al. Graph of thoughts: solving elaborate problems with large language models. *Proc. AAAI Conf. Artif. Intell.* **38**, 17682–17690 (2024).
94. Yamada, Y. et al. The AI Scientist-v2: workshop-level automated scientific discovery via agentic tree search. Preprint at https://doi.org/10.48550/arXiv.2504.08066 (2025).
95. Song, T. et al. A multiagent-driven robotic AI chemist enabling autonomous chemical research on demand. *J. Am. Chem. Soc.* **147**, 12534–12545 (2025).
96. Yao, S. et al. ReAct: synergizing reasoning and acting in language models. In *Proc. 11th Int. Conf. on Learning Representations* (ICLR, 2023).
97. Campbell, Q., Cox, S., Medina, J., Watterson, B. & White, A. D. MDCrow: automating molecular dynamics workflows with large language models. *Mach. Learn. Sci. Technol.* **7**, 025037 (2026).
98. Ghafarollahi, A. & Buehler, M. J. SciAgents: automating scientific discovery through bioinspired multi-agent intelligent graph reasoning. *Adv. Mater.* **37**, 2413523 (2025).
99. Lyu, Y. et al. EvoScientist: towards multi-agent evolving AI scientists for end-to-end scientific discovery. Preprint at https://doi.org/10.48550/arXiv.2603.08127 (2026).
100. Su, H. et al. Many heads are better than one: improved scientific idea generation by a LLM-based multi-agent system. In *Proc. 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (eds Che, W. et al.) 28201–28240 (Association for Computational Linguistics, 2025).
101. Fehlis, Y. et al. Accelerating drug discovery through agentic AI: a multi-agent approach to laboratory automation in the DMTA cycle. Preprint at https://doi.org/10.48550/arXiv.2507.09023 (2025).
102. Ghareeb, A. E. et al. A multi-agent system for automating scientific discovery. *Nature* **655**, 497–505 (2026).
103. Kusne, A. G. & McDannald, A. Scalable multi-agent lab framework for lab optimization. *Matter* **6**, 1880–1893 (2023).
104. Holzinger, A., Zatloukal, K. & Müller, H. Is human oversight to AI systems still possible? *N. Biotechnol.* **85**, 59–62 (2025).
105. Dadfar, B., Alemdag, B. & Kabay, G. The alchemist, the scientist, and the robot: exploring the potential of human–AI symbiosis in self-driving polymer laboratories. *Macromol. Rapid Commun.* **46**, e00380 (2025).
106. Tobias, A. V. & Wahab, A. Autonomous 'self-driving' laboratories: a review of technology and policy implications. *R. Soc. Open. Sci.* **12**, 250646 (2025).
107. Haque, A. et al. Human–AI collaborative autonomous synthesis with pulsed laser deposition for remote epitaxy. Preprint at https://doi.org/10.48550/arXiv.2511.11558 (2025).
108. Thiyagalingam, J., Shankar, M., Fox, G. & Hey, T. Scientific machine learning benchmarks. *Nat. Rev. Phys.* **4**, 413–420 (2022).
109. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).
110. King, R. D. et al. The Robot Scientist Adam. *Computer* **42**, 46–54 (2009).
111. Dagdelen, J. et al. Structured information extraction from scientific text with large language models. *Nat. Commun.* **15**, 1418 (2024).
112. Lookman, T., Balachandran, P. V., Xue, D. & Yuan, R. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. *npj Comput. Mater.* **5**, 21 (2019).
113. Saichandran, K. S., Elzokm, K., Guney, O. B. & Kolachalama, V. B. Agentic AI for automated hypothesis testing in Alzheimer's disease and related dementias. Preprint at *medRxiv* https://doi.org/10.64898/2025.12.02.25341517 (2025).
114. Wang, Q., Downey, D., Ji, H. & Hope, T. SciMON: scientific inspiration machines optimized for novelty. In *Proc. 62nd Annu. Meet. Assoc. Comput. Linguist. (Volume 1 Long Papers)* (eds Ku, L.-W., Martins, A. & Srikumar, V.) 279–299 (Association for Computational Linguistics, 2024).
115. Younis, R., Basak, S., Chavez, L. & Ahmadi, Z. From literature to hypotheses: an AI Co-Scientist system for biomarker-guided drug combination hypothesis generation. Preprint at https://doi.org/10.48550/arXiv.2603.00612 (2026).
116. Jiang, H., Shi, S., Yao, Y., Jiang, C. & Li, Q. HypoChainer: a collaborative system combining LLMs and knowledge graphs for hypothesis-driven scientific discovery. *IEEE Trans. Vis. Comput. Graph.* **32**, 298–308 (2026).
117. Gu, X. & Krenn, M. Interesting scientific idea generation using knowledge graphs and LLMs: evaluations with 100 research group leaders. Preprint at https://doi.org/10.48550/arXiv.2405.17044 (2026).
118. Saeedi, D., Buckner, D. K., Aponte, J. C. & Aghazadeh, A. AstroAgents: a multi-agent AI for hypothesis generation from mass spectrometry data. In *ICLR 2025 Workshop on Agentic AI for Science* https://iclr.cc/virtual/2025/33150 (ICLR, 2025).
119. Huynh, D. L. et al. Human-supervised agentic AI for hypothesis generation and experimental assistance in drug repurposing. Preprint at *bioRxiv* https://doi.org/10.64898/2026.04.20.719538 (2026).
120. Martinek, V. et al. Agentomics: an agentic system that autonomously develops novel state-of-the-art solutions for biomedical machine learning tasks. *Bioinformatics* **42**, btag250 (2026).
121. Dawson, W. et al. LARA-HPC: validation-driven agentic supercomputer workflows for atomistic modeling. *PRX Intell.* https://doi.org/10.1103/cp6v-t93q (2026).
122. Feng, T. et al. Towards autonomous mathematics research. Preprint at https://doi.org/10.48550/arXiv.2602.10177 (2026).
123. Gottweis, J. et al. Accelerating scientific discovery with Co-Scientist. *Nature* **655**, 487–496 (2026).
124. Guan, Y. et al. AI-assisted drug re-purposing for human liver fibrosis. *Adv. Sci.* **12**, e08751 (2025).
125. Zheng, Z. et al. A GPT-4 reticular chemist for guiding MOF discovery. *Angew. Chem. Int. Ed.* **62**, e202311983 (2023).
126. Zou, Y. et al. El Agente: an autonomous agent for quantum chemistry. *Matter* **8**, 102263 (2025).
127. Chen, Z. et al. ScienceAgentBench: toward rigorous assessment of language agents for data-driven scientific discovery. *Int. Conf. Learn. Representations* **2025**, 96934–96990 (2025).
128. Liu, X. et al. AgentBench: evaluating LLMs as agents. In *Proc. 12th Int. Conf. on Learning Representations* (eds Kim, B. et al.) 52989–53046 (ICLR, 2024).
129. Chang, M. et al. Partnr: a benchmark for planning and reasoning in embodied multi-agent tasks. *Int. Conf. Learn. Representations* **2025**, 65205–65268 (2025).
130. Polak, M. P. & Morgan, D. Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nat. Commun.* **15**, 1569 (2024).
131. Li, J., Li, M., Yang, Q. & Luo, S. ReactionSeek: LLM-powered literature data mining and knowledge discovery in organic synthesis. *Nat. Commun.* **17**, 3356 (2026).
132. Jeyakumar, S. K., Ahmad, A. & Gabriel, A. Advancing agentic systems: dynamic task decomposition, tool integration and evaluation using novel metrics and dataset. In *NeurIPS 2024 Workshop on Open-World Agents* https://neurips.cc/virtual/2024/100915 (2024).
133. Prince, M. H. et al. Opportunities for retrieval and tool augmented large language models in scientific facilities. *npj Comput. Mater.* **10**, 251 (2024).
134. Swanson, K., Wu, W., Bulaong, N. L., Pak, J. E. & Zou, J. The Virtual Lab of AI agents designs new SARS-CoV-2 nanobodies. *Nature* **646**, 716–723 (2025).
135. Brachman, M. et al. Building appropriate mental models: what users know and want to know about an agentic AI chatbot. In *Proc. 30th Int. Conf. Intelligent User Interfaces* 247–264 (Association for Computing Machinery, 2025).
136. Kim, S. et al. Materealize: a multi-agent deliberation system for end-to-end material design and synthesis. Preprint at https://doi.org/10.48550/arXiv.2601.15743 (2026).
137. Kumar, S. G. H. et al. El Agente Sólido: a new age(nt) for solid state simulations. Preprint at https://doi.org/10.48550/arXiv.2602.17886 (2026).
138. Pérez-Sánchez, J. B. et al. El Agente Quntur: a research collaborator agent for quantum chemistry. *Chem* https://doi.org/10.1016/j.chempr.2026.103259 (2026).
139. Gustin, I. et al. El agente cuántico: automating quantum simulations. *Rep. Prog. Phys.* **89**, 077602 (2026).
140. Zhang, W. et al. Fine-tuning large language models for chemical text mining. *Chem. Sci.* **15**, 10600–10611 (2024).
141. Pruyn, T. M. et al. MOF-ChemUnity: literature-informed large language models for metal–organic framework research. *J. Am. Chem. Soc.* **147**, 43474–43486 (2025).
142. Zheng, Z. et al. Image and data mining in reticular chemistry powered by GPT-4V. *Digit. Discov.* **3**, 491–501 (2024).
143. Papidocha, S. M., Burger, A., Bernales, V. & Aspuru-Guzik, A. The elephant in the lab: synthesizability in generative small-molecule design. *Curr. Opin. Chem. Eng.* **51**, 101217 (2026).
144. Schilling-Wilhelmi, M. et al. From text to insight: large language models for chemical data extraction. *Chem. Soc. Rev.* **54**, 1125–1150 (2025).
145. Hu, J. et al. Aitomia: an agentic framework for AI-driven atomistic and quantum chemical simulations. *J. Chem. Theory Comput.* **22**, 6557–6571 (2026).
146. Liu, G., Yang, S. & Zhong, Y. Masgent: an AI-assisted materials simulation agent. *Digit. Discov.* **5**, 2151–2171 (2026).
147. Chaudhari, A., Ock, J. & Barati Farimani, A. Modular large language model agents for multi-task computational materials science. *Commun. Mater.* **7**, 131 (2026).
148. Liu, G. et al. CatGo: bridging CLI coding agents with interactive structure and workflow management for computational chemistry. *Digit. Discov.* https://doi.org/10.1039/d6dd00273k (2026).
149. Mehandru, N. et al. BioAgents: bridging the gap in bioinformatics analysis with multi-agent systems. *Sci. Rep.* **15**, 39036 (2025).
150. Guo, J. et al. AI-guided design of efficient perovskite solar cells operationally stable at 100 °C. *Science* **392**, 724–728 (2026).

151. Smith, A. A. et al. Using a GPT-5-driven autonomous lab to optimize the cost and titer of cell-free protein synthesis. Preprint at *bioRxiv* https://doi.org/10.64898/2026.02.05.703998 (2026).
152. Cong, L. et al. LabOS: The AI-XR Co-Scientist that sees and works with humans. Preprint at https://doi.org/10.48550/arXiv.2510.14861 (2025).
153. Datta, B. et al. Artificial intelligence for food innovation. *Nat. Food* **7**, 644–654 (2026).
154. Rashvand, M. et al. Artificial intelligence for prediction of shelf-life of various food products: recent advances and ongoing challenges. *Trends Food Sci. Technol.* **159**, 104989 (2025).
155. Comeau, D. S., Bitterman, D. S. & Celi, L. A. Preventing unrestricted and unmonitored AI experimentation in healthcare through transparency and accountability. *npj Digit. Med.* **8**, 42 (2025).
156. van Genderen, M. E. et al. Charting a new course in healthcare: early-stage AI algorithm registration to enhance trust and transparency. *npj Digit. Med.* **7**, 119 (2024).
157. Mitchell, M. et al. Model Cards for model reporting. In *Proc. Conf. Fairness, Accountability, and Transparency* 220–229 (Association for Computing Machinery, 2019).
158. Schulhoff, S. et al. The prompt report: a systematic survey of prompt engineering techniques. Preprint at https://doi.org/10.48550/arXiv.2406.06608 (2025).
159. Alampara, N. et al. General-purpose models for the chemical sciences: LLMs and beyond. *Chem. Rev.* **126**, 2484–2549 (2026).
160. Zheng, Z. et al. Large language models for reticular chemistry. *Nat. Rev. Mater.* **10**, 369–381 (2025).
161. Hellert, T., Montenegro, J. & Sulc, A. Osprey: production-ready agentic AI for safety-critical control systems. *APL Mach. Learn.* **4**, 016103 (2026).
162. Lupoiu, R. et al. A multi-agentic framework for real-time, autonomous freeform metasurface design. *Sci. Adv.* **11**, eadx8006 (2025).
163. Zhang, Z. et al. A multimodal robotic platform for multi-element electrocatalyst discovery. *Nature* **647**, 390–396 (2025).
164. Yu, B. et al. Tooling or not tooling? The impact of tools on language agents for chemistry problem solving. In *Findings of the Association for Computational Linguistics: NAACL 2025* (eds Chiruzzo, L. et al.) 7635–7655 (Association for Computational Linguistics, 2025).
165. Wilkinson, S. R. et al. Applying the FAIR principles to computational workflows. *Sci. Data* **12**, 328 (2025).
166. Sicilia, M.-A., García-Barriocanal, E. & Sánchez-Alonso, S. Community curation in open dataset repositories: insights from Zenodo. *Procedia Comput. Sci.* **106**, 54–60 (2017).
167. Gustafsson, O. J. R. et al. WorkflowHub: a registry for computational workflows. *Sci. Data* **12**, 837 (2025).
168. Jain, S. M. *Introduction to Transformers for NLP: With the Hugging Face Library and Models to Solve Problems* (Apress, 2022).
169. Stach, E. et al. Autonomous experimentation systems for materials development: a community perspective. *Matter* **4**, 2702–2726 (2021).
170. Pyzer-Knapp, E. O. et al. Foundation models for materials discovery—current state and future directions. *npj Comput. Mater.* **11**, 61 (2025).
171. Leong, S. X. et al. Steering towards safe self-driving laboratories. *Nat. Rev. Chem.* **9**, 707–722 (2025).
172. Mandal, I. et al. Evaluating large language model agents for automation of atomic force microscopy. *Nat. Commun.* **16**, 9104 (2025).
173. Reeves-McLaren, N. & Moth-Lund Christensen, S. Data integrity in materials science in the era of AI: balancing accelerated discovery with responsible science and innovation. *J. Mater. Chem. A* **14**, 276–283 (2026).
174. Tang, X. et al. Risks of AI scientists: prioritizing safeguarding over autonomy. *Nat. Commun.* **16**, 8317 (2025).
175. Duan, J. et al. Gtbench: uncovering the strategic reasoning capabilities of llms via game-theoretic evaluations. *Adv. Neural Inf. Process. Syst.* **37**, 28219–28253 (2024).
176. Jin, M. et al. The impact of reasoning step length on large language models. In *Findings Assoc. Comput. Linguist.: ACL 2024* (eds Ku, L.-W., Martins, A. & Srikumar, V.) 1830–1842 (Association for Computational Linguistics, 2024).
177. Wang, Z. et al. Why reasoning fails to plan: a planning-centric analysis of long-horizon decision making in LLM agents. Preprint at https://doi.org/10.48550/arXiv.2601.22311 (2026).
178. Erdogan, L. E. et al. Plan-and-Act: improving planning of agents for long-horizon tasks. In *Proc. 42nd Int. Conf. on Machine Learning* (eds Singh, A. et al.) 15419–15462 (PMLR, 2025).
179. Sel, B., Jia, R. & Jin, M. LLMs can plan only if we tell them. In *Proc. 13th Int. Conf. on Learning Representations* (eds Yue, Y. et al.) 78024–78064 (2025).
180. Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K. & Yao, S. Reflexion: language agents with verbal reinforcement learning. *Adv. Neural Inf. Process. Syst.* **36**, 8634–8652 (2023).
181. Madaan, A. et al. Self-refine: iterative refinement with self-feedback. *Adv. Neural Inf. Process. Syst.* **36**, 46534–46594 (2023).
182. Wang, G. et al. Voyager: an open-ended embodied agent with large language models. *ML Anthology* https://mlanthology.org/tmlr/2024/wang2024tmlr-voyager/ (2024).
183. Xu, T., Zhang, D., Mitra, K. & Hruschka, E. Verification-aware planning for multi-agent systems. In *Proc. 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* (eds Demberg, V., Inui, K. & Marquez, L.) 7528–7546 (Association for Computational Linguistics, 2026).
184. Wen, B., Zhang, R., Chen, Y., Xie, H. & Guo, L.-Z. Aligning progress and feasibility: a neuro-symbolic dual memory framework for long-horizon LLM agents. Preprint at https://doi.org/10.48550/arXiv.2604.02734 (2026).
185. Shi, F. et al. Large language models can be easily distracted by irrelevant context. In *Proc. 40th Int. Conf. on Machine Learning* (eds Krause, A. et al.) 31210–31227 (PMLR, 2023).
186. Holtzman, A., Buys, J., Du, L., Forbes, M. & Choi, Y. The curious case of neural text degeneration. In *Proc. 8th Int. Conf. on Learning Representations* (ICLR, 2020).
187. Huang, L. et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.* **43**, 42:1–42:55 (2025).
188. Ortega, P. A. et al. Shaking the foundations: delusions in sequence models for interaction and control. Preprint at https://doi.org/10.48550/arXiv.2110.10819 (2021).
189. Zečević, M., Willig, M., Dhami, D. S. & Kersting, K. Causal parrots: large language models may talk causality but are not causal. *ML Anthology* https://mlanthology.org/tmlr/2023/zecevic2023tmlr-causal/ (2023).
190. Luo, L., Li, Y.-F., Haffari, R. & Pan, S. Reasoning on graphs: faithful and interpretable large language model reasoning. *Int. Conf. Learn. Representations* **2024**, 14400–14423 (2024).
191. Ji, Z. et al. Towards mitigating LLM hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (eds Bouamor, H., Pino, J. & Bali, K.) 1827–1843 (Association for Computational Linguistics, 2023).
192. Gao, L. et al. RARR: researching and revising what language models say, using language models. In *Proc. 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (eds Rogers, A. et al.) 16477–16508 (Association for Computational Linguistics, 2023).
193. Farquhar, S., Kossen, J., Kuhn, L. & Gal, Y. Detecting hallucinations in large language models using semantic entropy. *Nature* **630**, 625–630 (2024).
194. Wagner, M. R. & Kleiner, M. How thoughtful experimental design can empower biologists in the omics era. *Nat. Commun.* **16**, 7263 (2025).
195. Townsend, H. G. et al. A call to action to address critical flaws and bias in laboratory animal experiments and preclinical research. *Sci. Rep.* **15**, 30745 (2025).
196. Zhou, S. et al. Automating expert-level medical reasoning evaluation of large language models. *npj Digit. Med.* **9**, 34 (2025).
197. Asadi, M. et al. Mirage the illusion of visual understanding. Preprint at https://doi.org/10.48550/arXiv.2603.21687 (2026).
198. Qin, Y. et al. Toolllm: facilitating large language models to master 16000+ real-world APIs. *Int. Conf. Learn. Representations* **2024**, 9695–9717 (2024).
199. Shi, Z. et al. Retrieval models aren't tool-savvy: benchmarking tool retrieval for large language models. In *Findings of the Association for Computational Linguistics: ACL 2025* (eds Che, W. et al.) 24497–24524 (Association for Computational Linguistics, 2025).
200. Pagel, S., Jirasek, M. & Cronin, L. Verification and execution of the scientific literature via chemputation augmented by large language models. *Commun. Chem.* **9**, 191 (2026).
201. Sim, M. et al. ChemOS 2.0: an orchestration architecture for chemical self-driving laboratories. *Matter* **7**, 2959–2977 (2024).
202. Lewis, P. et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33* (eds Larochelle, H. et al.) 9459–9474 (Curran Associates, 2020).
203. Gu, Z. et al. Agentgroupchat-v2: divide-and-conquer is what LLM-based multi-agent system need. Preprint at https://doi.org/10.48550/arXiv.2506.15451 (2025).
204. Han, S., Zhang, Q., Jin, W. & Xu, Z. LLM multi-agent systems: challenges and open problems. Preprint at https://doi.org/10.48550/arXiv.2402.03578 (2026).
205. Zhang, Y. et al. SILO-BENCH: a scalable environment for evaluating distributed coordination in multi-agent LLM systems. In *Proc. 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (eds Liakata, M. et al.) 29379–29398 (Association for Computational Linguistics, 2026).
206. Zhou, J. et al. A multi-robot–multi-task scheduling system for autonomous chemistry laboratories. *Digit. Discov.* **4**, 636–652 (2025).
207. Zhang, G. et al. G-memory: tracing hierarchical memory for multi-agent systems. *Adv. Neural Inf. Process. Syst.* **38**, 12988–13018 (2026).
208. Zhang, Z. et al. Generative AI for biosciences: emerging threats and roadmap to biosecurity. Preprint at https://doi.org/10.48550/arXiv.2510.15975 (2025).
209. Shumailov, I. et al. AI models collapse when trained on recursively generated data. *Nature* **631**, 755–759 (2024).
210. Yang, J. et al. Poisoning medical knowledge using large language models. *Nat. Mach. Intell.* **6**, 1156–1168 (2024).
211. Badani, A., de Moraes, F. Y., Vollmuth, P., Chung, C. & Mansouri, A. AI and innovation in clinical trials. *npj Digit. Med.* **8**, 683 (2025).
212. Maupin, D., Suchak, T., Barnett, A. & Spick, M. Dramatic increases in redundant publications in the generative AI era. *BMC Med.* **24**, 29 (2025).
213. Kwon, D. AI-generated images threaten science—here's how researchers hope to spot them. *Nature* **635**, 268–269 (2024).
214. McLoughlin, K. L., Palmer, A., Matias, J. N. & Crockett, M. J. Proprietary LLMs impede scientific transparency and reproducibility. *Nat. Rev. Psychol.* **5**, 513–514 (2026).
215. Spirling, A. Why open-source generative AI models are an ethical way forward for science. *Nature* **616**, 413–413 (2023).
216. Wang, W., Wang, G., Marivate, V. & Hufton, A. L. On the transparency of large AI models. *Patterns* **4**, 100797 (2023).

217. Bhanbhro, J., Nisticò, S. & Palopoli, L. Issues in federated learning: some experiments and preliminary results. *Sci. Rep.* **14**, 29881 (2024).
218. Urbina, F., Lentzos, F., Invernizzi, C. & Ekins, S. Dual use of artificial-intelligence-powered drug discovery. *Nat. Mach. Intell.* **4**, 189–191 (2022).
219. Strubell, E., Ganesh, A. & McCallum, A. Energy and policy considerations for modern deep learning research. *Proc. AAAI Conf. Artif. Intell.* **34**, 13693–13696 (2020).
220. Schwartz, R., Dodge, J., Smith, N. A., Etzioni, O. & Green, A. I. *Commun. ACM* **63**, 54–63 (2020).
221. Ren, S., Tomlinson, B., Black, R. W. & Torrance, A. W. Reconciling the contrasting narratives on the environmental impact of large language models. *Sci. Rep.* **14**, 26310 (2024).
222. Lehdonvirta, V., Wú, B. & Hawkins, Z. Compute north vs. compute south: the uneven possibilities of compute-based AI governance around the globe. *Proc. AAAIACM Conf. AI Ethics Soc.* **7**, 828–838 (2024).
223. Matthias, A. The responsibility gap: ascribing responsibility for the actions of learning automata. *Ethics Inf. Technol.* **6**, 175–183 (2004).
224. COPE Council. Authorship and AI tools. *COPE: Committee on Publication Ethics* https://publicationethics.org/guidance/cope-position/authorship-and-ai-tools (2023).
225. Channing, G. & Ghosh, A. AI for scientific discovery is a social problem. *Patterns* **7**, 101497 (2026).
226. Gawlikowski, J. et al. A survey of uncertainty in deep neural networks. *Artif. Intell. Rev.* **56**, 1513–1589 (2023).
227. Laurent, J. M. et al. LAB-Bench: measuring capabilities of language models for biology research. Preprint at https://doi.org/10.48550/arXiv.2407.10362 (2024).
228. Mirza, A. et al. A framework for evaluating the chemical knowledge and reasoning abilities of large language models against the expertise of chemists. *Nat. Chem.* **17**, 1027–1034 (2025).
229. LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2324 (1998).
230. Krizhevsky, A. *Learning Multiple Layers of Features from Tiny Images*. Technical Report, University of Toronto (2009).
231. Siegel, Z. S., Kapoor, S., Nadgir, N., Stroebl, B. & Narayanan, A. CORE-Bench: fostering the credibility of published research through a computational reproducibility agent benchmark. In *Trans. Mach. Learn. Res.* (2025).
232. Starace, G. et al. PaperBench: evaluating AI's ability to replicate AI research. In *Proc. 42nd Int. Conf. on Machine Learning* (eds Singh, A. et al.) 56843–56873 (PMLR, 2025).
233. Abolhasani, M. & Kumacheva, E. The rise of self-driving labs in chemical and materials sciences. *Nat. Synth.* **2**, 483–492 (2023).
234. Hickman, R. J., Bannigan, P., Bao, Z., Aspuru-Guzik, A. & Allen, C. Self-driving laboratories: a paradigm shift in nanomedicine development. *Matter* **6**, 1071–1081 (2023).
235. Moor, M. et al. Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).
236. Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N. & Malone, T. W. Evidence for a collective intelligence factor in the performance of human groups. *Science* **330**, 686–688 (2010).
237. Vaccaro, M., Almaatouq, A. & Malone, T. When combinations of humans and AI are useful: a systematic review and meta-analysis. *Nat. Hum. Behav.* **8**, 2293–2303 (2024).
238. Floris Jacobs, P. & Pollice, R. Developing large language models for quantum chemistry simulation input generation. *Digit. Discov.* **4**, 762–775 (2025).
239. Wang, Z. et al. DREAMS: density functional theory based research engine for agentic materials simulation. Preprint at https://doi.org/10.48550/arXiv.2507.14267 (2026).
240. Li, J. et al. Can LLM already serve as a database interface? A big bench for large-scale database grounded text-to-SQLs. in *Proc. 37th Int. Conf. Neural Inf. Process. Syst.* 42330–42357 (ACM, 2023).
241. Gaudin, T. et al. 'Just ask': agentic pilot to empower chemists. Preprint at https://doi.org/10.26434/chemrxiv.15001010/v2 (2026).
242. Yuen, N. H.-Y., Li, C.-H., Kwok, K. C.-M. & Leung, R. Laboratory Information Management System (LIMS) 4.0: a model to solve the current pain points for commercial electrical and electronic laboratories. *Int. J. Eng. Bus. Manag.* **17**, 18479790251385743 (2025).
243. Opentrons. Python Protocol API. *Opentrons Documentation* https://docs.opentrons.com/python-api/ (2026).
244. Darvish, K. et al. ORGANA: a robotic assistant for automated chemistry experimentation and characterization. *Matter* **8**, 101897 (2025).
245. Yoshikawa, N. et al. Large language models for chemistry robotics. *Auton. Robot.* **47**, 1057–1086 (2023).
246. Fitzpatrick, D. E., Battilocchio, C. & Ley, S. V. A novel internet-based reaction monitoring, control and autonomous self-optimization platform for chemical synthesis. *Org. Process. Res. Dev.* **20**, 386–394 (2016).
247. Leong, S. X., Pablo-García, S., Wong, B. & Aspuru-Guzik, A. MERMaid: universal multimodal mining of chemical reactions from PDFs using vision-language models. *Matter* **8**, 102331 (2025).
248. Huang, X. et al. CASCADE: Cumulative Agentic Skill Creation through Autonomous Development and Evolution. Preprint at https://doi.org/10.48550/arXiv.2512.23880 (2026).
249. Ruan, Y. et al. An automatic end-to-end chemical synthesis development platform powered by large language models. *Nat. Commun.* **15**, 10160 (2024).
250. Lou, R. et al. AAAR-1.0: assessing AI's potential to assist research. In *Proc. 42nd Int. Conf. on Machine Learning* (eds Singh, A. et al.) 40361–40383 (PMLR, 2025).
251. Bogin, B. et al. SUPER: evaluating agents on setting up and executing tasks from research repositories. In *Proc. 2024 Conf. Empirical Methods in Natural Language Processing* (eds Al-Onaizan, Y. et al.) 12622–12645 (Association for Computational Linguistics, 2024).
252. Chamoun, E., Schlichtkrull, M. & Vlachos, A. Automated focused feedback generation for scientific writing assistance. In *Findings of the Association for Computational Linguistics: ACL 2024* (eds Ku, L.-W., Martins, A. & Srikumar, V.) 9742–9763 (Association for Computational Linguistics, 2024).
253. Nathani, D. et al. MLGym: a new framework and benchmark for advancing AI research agents. Preprint at https://doi.org/10.48550/arXiv.2502.14499 (2025).
254. Jansen, P. et al. DiscoveryWorld: a virtual environment for developing and evaluating automated scientific discovery agents. *Adv. Neural Inf. Process. Syst.* **37**, 10088–10116 (2024).
255. El-khawaldeh, R. et al. Keeping an "eye" on the experiment: computer vision for real-time monitoring and control. *Chem. Sci.* **15**, 1271–1282 (2024).

Acknowledgements

Z. Zheng acknowledges support provided by the EQT Foundation Breakthrough Science Grants (no. 059994) and thanks support from the Royal Society of Chemistry through a Researcher Collaborations Grant (GR0042769). The Matter Laboratory (A.M., S.M.P., Z. Zhang, S.K., V.B. and A.A.-G.) work was supported by the Defense Advanced Research Projects Agency (DARPA) under Agreement No. HR0011262E022 and the AI2050 programme of Schmidt Sciences. S.M.P. thanks the Swiss National Science Foundation (SNSF) for a Postdoc. Mobility (P500-2_235461) fellowship. A.A.-G. thanks A. G. Frøseth for generous support and acknowledges the generous support of Natural Resources Canada and the Canada 150 Research Chairs programme. A.A.-G. and V.B. are supported by the University of Toronto's Acceleration Consortium, which receives funding from the CFREF-2022-00042 Canada First Research Excellence Fund. D.M.S., S.T. and P.S. acknowledge support from the NCCR Catalysis (grant number 225147), a National Centre of Competence in Research funded by the SNSF. A.K., D.M.S. and S.T. acknowledge funding by the SNSF (grant number 226509). A.M. and S.K. acknowledge funding from Department of the Navy, Office of Naval Research (grant number N00014-19-1-2134). The authors thank A. M. Bran for providing the engineering schematic associated with the ChemCrow project.

Author contributions

Introduction (B.F., P.W., R.A.C., Z. Zhang, S.C.P., V.B., W.G. and Z. Zheng); Experimentation (B.F., L.Z., R.A.C., S.K., Z. Zhang, S.M.P., S.T., V.B. and Z. Zheng); Results (B.F., S.K., A.M., Z. Zhang, S.M.P., D.M.S., V.B. and Z. Zheng); Applications (B.F., L.Z., A.M., Z. Zhang, S.M.P., A.K., D.M.S., V.B., W.G., X.Y., L.C. and Z. Zheng); Reproducibility and data deposition (B.F., S.K., S.M.P., A.K., D.M.S., V.B. and Z. Zheng); Limitations and optimizations (B.F., A.M., S.M.P., D.M.S., V.B. and Z. Zheng); Outlook (B.F., P.W., L.Z., S.C.P., W.G. and Z. Zheng); overview of the Primer (B.F., P.W., R.A.C., S.C.P., V.B., W.G., X.Y., L.C., P.S., A.A.-G. and Z. Zheng).

Competing interests

The authors declare no competing interests.

Additional information

Peer review information *Nature Reviews Methods Primers* thanks Nubi Achebo; James Evans; Li Shen, who co-reviewed with Shu Yang; and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Related links

Agent-to-Agent protocol: https://a2a-protocol.org/latest
Claude Code: https://code.claude.com/docs/en/overview
Gemini command line interface: https://developers.google.com/gemini-code-assist/docs/gemini-cli
Hermes Agent: https://hermes-agent.nousresearch.com/
OpenAI Agents SDK: https://openai.github.io/openai-agents-python/
Open AI Codex: https://openai.com/codex
OpenClaw: https://openclaw.ai/
OpenCode: https://opencode.ai/
Open Knowledge Format: https://cloud.google.com/blog/products/data-analytics/how-the-open-knowledge-format-can-improve-data-sharing
LangChain: https://docs.langchain.com/oss/python/langchain/agents
LangGraph: https://www.langchain.com/langgraph
Perplexity: https://www.perplexity.ai
Pydantic AI: https://pydantic.dev/docs/ai/overview/
Python Protocol API Documentation: https://docs.opentrons.com/python-api/
Tavily: https://tavily.com

© Springer Nature Limited 2026

[1]Department of Chemistry, Washington University, St. Louis, MO, USA. [2]Institute of Materials Science & Engineering, Washington University, St. Louis, MO, USA. [3]IBM Research, IBM Thomas J. Watson Research Center, New York, NY, USA. [4]Department of Chemistry, University of Toronto, Toronto, Ontario, Canada. [5]Vector Institute for Artificial Intelligence, Toronto, Ontario, Canada. [6]Department of Computer Science, University of Toronto, Toronto, Ontario, Canada. [7]NVIDIA, Toronto, Ontario, Canada. [8]Laboratory of Artificial Chemical Intelligence, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland. [9]National Centre of Competence in Research (NCCR) Catalysis, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland. [10]Acceleration Consortium, Toronto, Ontario, Canada. [11]Department of Chemical and Biomolecular Engineering, University of Pennsylvania, Philadelphia, PA, USA. [12]Department of Grain Science and Food Science, Kansas State University, Manhattan, KS, USA. [13]Department of Pathology, Stanford University School of Medicine, Stanford, CA, USA. [14]Department of Chemical Engineering and Applied Chemistry, University of Toronto, Toronto, Ontario, Canada. [15]Department of Materials Science & Engineering, University of Toronto, Toronto, Ontario, Canada. [16]Canadian Institute for Advanced Research (CIFAR), Toronto, Ontario, Canada.