FROM *AI for Science* TO *Agentic Science*: A SURVEY ON AUTONOMOUS SCIENTIFIC DISCOVERY AND AI SCIENTISTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Artificial intelligence (AI) is reshaping scientific discovery, evolving from specialized computational tools into autonomous research partners. We position ***Agentic Science*** as a pivotal stage within the broader ***AI for Science*** paradigm, where AI systems progress from partial assistance to full scientific agency. Enabled by large language models (LLMs), multimodal systems, and integrated research platforms, agentic AI exhibits capabilities in hypothesis generation, experimental design, execution, analysis, and iterative refinement-behaviors once regarded as uniquely human. This survey offers a **domain-oriented review** of autonomous scientific discovery across life sciences, chemistry, materials, and physics, synthesizing research progress and advances within each discipline. We unify three previously fragmented perspectives-process-oriented, autonomy-oriented, and mechanism-oriented-through **a comprehensive framework** that connects foundational capabilities, core processes, and domain-specific realizations. Building on this framework, we (i) trace the evolution of AI for Science, (ii) identify five core capabilities underpinning scientific agency, (iii) model discovery as a dynamic four-stage workflow, (iv) review applications across life sciences, chemistry, materials science, and physics, and (v) synthesize key challenges and future opportunities. This work establishes a domain-oriented synthesis of autonomous scientific discovery and positions Agentic Science as a structured paradigm for advancing AI-driven research.

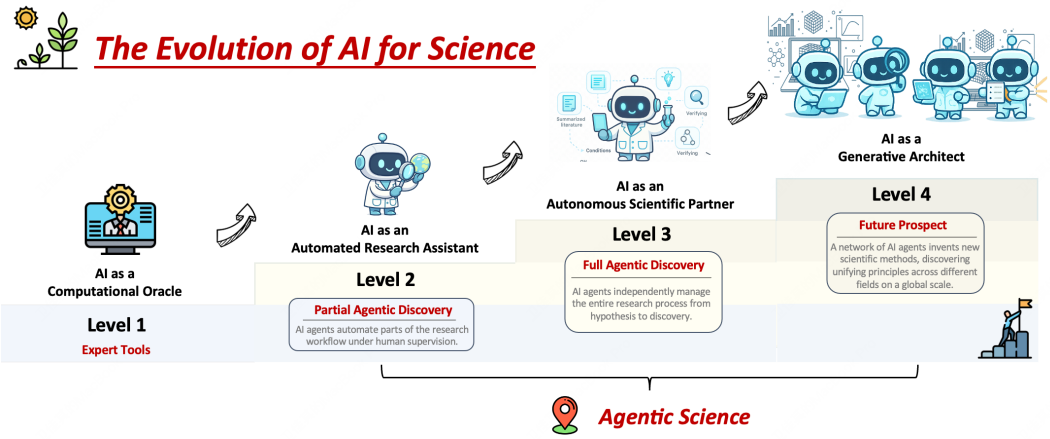

Figure 1: The Evolution of AI for Science. From computational tools to creative collaborators: the four-stage journey of AI in science. The progression of AI in scientific research is depicted across four levels, from Level 1 (Computational Oracle) to a future prospect, Level 4 (Generative Architect). Agentic Science encompasses Level 2 (Automated Research Assistant) and Level 3 (Autonomous Scientific Partner), which represent partial and full agentic discovery, respectively.

1 INTRODUCTION

Scientific discovery is undergoing a profound transformation, driven by the evolution of artificial intelligence (AI) from specialized tools into collaborative research partners. This progression signals a new era of **AI for Science**, marked by the emergence of **Autonomous Scientific Discovery**, which we term *Agentic Science*. This paradigm describes AI systems that operate as autonomous agents, capable of formulating hypotheses, designing and executing experiments, interpreting results, and iteratively refining theories with minimal human oversight (Boiko et al., 2023; Schneider, 2025). Such agents leverage integrated platforms that provide access to diverse AI models and datasets, and are powered by multimodal foundational models that exhibit deep scientific reasoning capabilities.

This shift is fueled by recent breakthroughs in large language models (LLMs) (Guo et al., 2025; Team et al., 2025a), which excel at natural language understanding, complex reasoning, and tool use (Sun et al., 2025; Zeng et al., 2024). These capabilities have enabled the development of AI agents that are no longer static computational pipelines but dynamic, goal-driven entities that navigate the entire scientific method (Yax et al., 2024; Ma et al., 2024). From hypothesis generation (Yang et al., 2023) to autonomous experimentation (Yuan et al., 2025), these agents are beginning to automate cognitive tasks once considered exclusively human.

Despite this rapid progress, a unified framework for understanding these increasingly autonomous systems is lacking. Previous surveys have examined the field from isolated perspectives: mapping AI capabilities onto the research cycle (process-oriented), grading systems by their level of independence (autonomy-oriented), or dissecting the underlying software architectures (mechanism-oriented) (Luo et al., 2025; Zheng et al., 2025; Ren et al., 2025). While valuable, these analyses remain fragmented.

This review synthesizes and extends these perspectives into a comprehensive framework that connects *foundational capabilities*, *core processes*, and *domain realizations* in autonomous discovery. We chart the evolution of AI for Science, formally defining Agentic Science as a paradigm built on five core agentic capabilities: reasoning, tool integration, memory, multi-agent collaboration, and optimization. We then model the agent-driven scientific workflow as a dynamic, four-stage process encompassing hypothesis, experimentation, analysis, and synthesis. Grounded in this framework, we conduct **a systematic, domain-oriented review** of agentic systems across the life sciences, chemistry, materials, and physics, highlighting key applications from drug discovery to materials design. Finally, we identify the primary technical and ethical challenges-including reproducibility, validation, and human-agent collaboration-and propose a research roadmap to guide the development of robust and trustworthy scientific agents. By providing this unified lens, we aim to establish a conceptual and methodological foundation for Agentic Science, accelerating a future where AI and human researchers co-evolve to push the frontiers of knowledge.

2 THE EVOLUTION OF AI FOR SCIENCE: FROM TOOLS TO AUTONOMOUS PARTNERS

The role of AI in science is shifting from computational augmentation to autonomous inquiry. This progression can be framed as an evolution through distinct levels of autonomy, starting with AI as a specialized tool and advancing towards AI as a collaborative scientific partner. This section delineates these levels to formalize the paradigm of Agentic Science.

2.1 THE EVOLUTION OF AI FOR SCIENCE

The role of artificial intelligence in science is undergoing a profound evolution, transitioning from specialized tools to autonomous partners. Initially, AI acted as a **Computational Oracle**: a collection of expert, non-agentic models designed to solve well-defined problems within a human-directed workflow. These systems, which excel at tasks like protein structure prediction (Jumper et al., 2021) or genomic analysis (Dalla-Torre et al., 2025), essentially function as powerful pattern recognizers that minimize a task-specific loss, $M^* = \arg\min_M \sum_i \mathcal{L}_{\text{task}}(M(x_i), y_i)$. This paradigm has accelerated discovery but requires constant human guidance. The subsequent stage introduced the **Automated Research Assistant**, where agents exhibit partial autonomy to execute predefined experimental or analytical sequences. Given a high-level goal $\mathcal{G}$, these agents follow a policy π to complete

a sub-task, $\{a_0, \ldots, a_T\} \sim \pi(\cdot|\mathcal{G}, \mathcal{T}_{\text{tools}})$, but the overarching scientific questions and hypotheses remain human-conceived.

We are now entering the era of the **Autonomous Scientific Partner**, a paradigm shift that embodies Agentic Science. Here, AI agents are capable of conducting the entire discovery cycle with minimal human intervention. These systems can observe a domain, formulate novel hypotheses, design and execute experiments, analyze the results, and iteratively refine their strategy in a continuous loop. The agent's objective moves beyond task completion to maximizing a measure of cumulative scientific utility, such as the expected information gain $\mathcal{I}(\cdot)$ about a set of evolving hypotheses $\mathcal{H}$, thus optimizing its policy $\pi^* = \arg\max_\pi \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t \mathcal{I}(\mathcal{H}_t; s_{t+1}|s_t, a_t)]$. This elevates the human role from operator to high-level strategist and validator. Pioneering systems like Coscientist (Boiko et al., 2023) and ChemCrow (Bran et al., 2023) provide the first glimpses of this future, demonstrating autonomous experimentation and discovery in chemistry, heralding a new age of human-agent co-discovery.

Looking forward, a prospective fourth level is the **AI as a Generative Architect**. Such a system would transcend discovery within existing scientific frameworks to engage in autonomous invention. Its capabilities would extend to designing novel instrumentation, creating new experimental methodologies, or even formulating entirely new conceptual frameworks and theories. This represents a leap from finding what is to creating what could be. This architect would also be capable of performing vast, interdisciplinary synthesis, uncovering latent connections between disparate scientific fields to forge new domains of inquiry. The agent's objective function would fundamentally shift from optimizing discoveries within a fixed paradigm to generating new scientific paradigms that maximize future discovery potential.

This evolutionary trajectory—from oracle to assistant, to partner, and ultimately to architect—redefines the scientific enterprise itself. Each stage grants the AI greater agency, fundamentally reshaping the human-scientist's role from a hands-on experimenter to a collaborator and visionary guide for its AI counterpart. The ultimate benchmark for this new scientific paradigm could be a "Nobel-Turing Test," where an AI system, unprompted, makes a discovery or invents a technology so profound that it is deemed worthy of a Nobel Prize. The pursuit of this goal will not only accelerate the pace of science but may also change our understanding of discovery itself.

2.2 Agentic Science: The Focus of This Review

This review centers on the paradigm of Agentic Science, encompassing both Level 2 (task-level autonomy) and Level 3 (goal-level autonomy). The common thread is agency-the ability to act

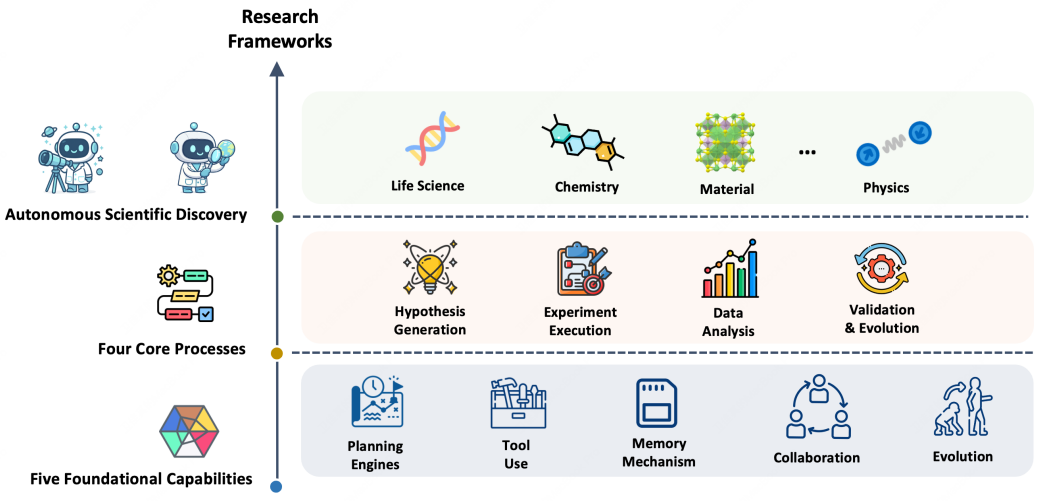

Figure 2: Research frameworks for Autonomous Scientific Discovery: Integrating Foundational Capabilities, Core Processes, and Research Levels across science.

purposefully to achieve a goal. Agentic Science redefines discovery as an autonomous, closed-loop process managed by intelligent agents. To systematically understand and engineer these systems, we propose a three-tiered framework (Figure 2 B) that illustrates how foundational agentic abilities give rise to complex scientific workflows, which in turn drive tangible discoveries across diverse domains.

- **Layer 1: Foundational Capabilities (Section 3.1).** At the base are the five core cognitive and operational abilities that form the building blocks of any scientific agent: Reasoning and Planning, Tool Integration, Memory, Multi-Agent Collaboration, and Optimization and Evolution. These capabilities are the engine of agentic intelligence.
- **Layer 2: Core Processes of Discovery (Section 3.2).** These foundational capabilities enable agents to execute the four key stages of the scientific method in a dynamic loop: (1) Observation and Hypothesis Generation, (2) Experimental Planning and Execution, (3) Data and Result Analysis, and (4) Synthesis, Validation, and Evolution. This layer represents the agent's workflow.
- **Layer 3: Autonomous Scientific Discovery (Section 3.3).** The successful execution of the discovery loop, powered by the core capabilities, results in practical scientific progress. This top layer showcases the application of agentic systems in domains like life sciences, chemistry, materials, and physics, culminating in end-to-end autonomous research pipelines.

This hierarchical framework provides a comprehensive lens through which to analyze the architecture, function, and impact of scientific agents, from their elemental skills to their highest-level achievements.

3 The Agentic Science Research Framework

The Agentic Science Framework redefines scientific discovery as a dynamic, autonomous process driven by intelligent agents. This framework is structured in three hierarchical tiers. At its base are the Five Foundational Capabilities, which constitute the agent's cognitive core. These capabilities enable the Four Core Processes of an autonomous discovery loop, which in turn drive progress in Autonomous Scientific Discovery across diverse domains, from life sciences and chemistry to materials and physics. This integrated structure allows an agent to transition from a specialized tool to an autonomous partner in the research lifecycle, managing long-term, iterative, and empirically grounded workflows.

3.1 Five Foundational Capabilities: The Cognitive Core

To navigate the complexities of the research lifecycle, a scientific agent must possess a suite of five interconnected capabilities that form its cognitive and operational foundation.

Planning and Reasoning Engines The cognitive core of a scientific agent, the planning and reasoning engine translates high-level goals into executable actions. Foundational approaches use task decomposition via linear reasoning chains (e.g., Chain-of-Thought), enhanced with ensemble methods for robustness (Wei et al., 2022; Wang et al., 2022). More sophisticated engines employ non-linear, tree-based search (e.g., Tree-of-Thought, MCTS) to explore multiple solution paths simultaneously, a crucial feature for navigating complex scientific problems (Hu et al., 2023; Guo et al., 2024). These plans are not static; they are dynamically adapted through feedback from the environment, human guidance, or self-reflection, often using frameworks like ReAct that interleave reasoning with action and observation (Yao et al., 2023; Sun et al., 2023). The primary challenges in scientific reasoning are the high-stakes and strict verifiability of outcomes, the need to navigate vast and poorly understood search spaces, the interpretation of noisy, multimodal experimental feedback, and the necessity for long-horizon planning to establish causal understanding while mitigating error accumulation (Sauter et al., 2023).

Tool Use and Integration To overcome the intrinsic limitations of language models and interact with the world, agents must harness external tools. These range from foundational utilities like search engines and code interpreters (Jablonka et al., 2023; Gou et al., 2024a) to domain-specific

computational tools that encapsulate expert knowledge, such as bioinformatics toolkits in CRISPR-GPT (Huang et al., 2024) or reaction predictors in ChemCrow (Bran et al., 2024). The most advanced tier involves experimental and simulation tools, allowing agents to test hypotheses in virtual laboratories by interacting with physics engines or high-fidelity simulators (Todorov et al., 2012; Koldunov & Jung, 2024). The challenges in scientific tool use are formidable: they demand exceptional precision and deep domain understanding, as minor errors can invalidate results. Furthermore, scientific work requires strict reproducibility and provenance tracking, often involving the creation of complex, interoperable workflows from heterogeneous tools—a known difficulty (Shen et al., 2025). Finally, agents must perform cost-benefit analyses to manage the significant computational and financial costs of many scientific tools.

Memory Mechanisms Memory enables agents to retain information, learn from experience, and maintain context. For scientific agents, memory serves two critical roles. First, as a mechanism for iterative task execution, it maintains a coherent understanding of an ongoing task by storing short-term context and building long-term experience repositories from successes and failures (e.g., Reflexion (Shinn et al., 2023)) or even codifying successful action sequences into reusable skill libraries (Wang et al., 2023). Second, memory acts as a knowledge hub, integrating external information sources via Retrieval-Augmented Generation (RAG) to ground reasoning in established scientific literature or structured knowledge graphs (Lewis et al., 2020; Ghafarollahi & Buehler, 2024b). The distinct challenges for scientific memory include managing the accuracy and decay of scientific knowledge, seamlessly storing and reasoning across heterogeneous and multi-modal data (text, tables, images, genomic sequences), and maintaining high-fidelity, causally-linked histories over long-term research projects to ensure reproducibility.

Collaboration between Agents To tackle problems beyond the scope of a single agent, multi-agent systems leverage collaboration. Strategies include hierarchical task execution, where a manager agent decomposes a goal and assigns subtasks to specialized worker agents, as seen in MetaGPT (Hong et al., 2024) and Coscientist (Boiko et al., 2023). A second approach is deliberative, refinement-based collaboration, where agents improve solutions through iterative peer interaction, such as structured debate or a critique process, to enhance factuality and reasoning (Du et al., 2023; Shinn et al., 2023). The most advanced systems use dynamic, adaptive topologies, where the interaction structure itself is optimized for the task, with agents reconfiguring their communication pathways in response to real-time feedback (Liu et al., 2024; Kim et al., 2024). Key challenges for scientific collaboration include grounding consensus in empirical reality to avoid amplified hallucinations, fostering epistemic diversity to prevent premature consensus, and integrating complex, multimodal information across extended workflows using communication protocols more sophisticated than plain text exchange.

Optimization and Evolution For sustained discovery, agents must improve over time. This is achieved through iterative self-refinement, where an agent enhances its outputs via a cycle of generation, feedback, and correction, using either self-generated feedback or external validation from tools (Madaan et al., 2023; Gou et al., 2024b). Agents can also evolve their underlying models through self-learning and interaction, using techniques like self-supervised learning or reinforcement learning with self-generated rewards to improve their intrinsic capabilities (Yuan et al., 2024; Zhu et al., 2024). Finally, population-based co-evolution drives improvement through the interactions within a group of agents, whether cooperative (e.g., role-playing frameworks) or competitive (e.g., multi-agent debate, red-teaming) (Li et al., 2023; Du et al., 2023). The challenges in scientific optimization are unique: the evaluation of a scientific hypothesis is often slow and expensive, making rapid feedback loops impractical; the reward landscape is sparse, with breakthroughs being rare; and all outputs must be grounded in physical reality and adhere to strict safety protocols, adding constraints far beyond typical benchmarks.

3.2 Four Core Processes: The Autonomous Discovery Loop

Enabled by the foundational capabilities, the agentic science workflow operates as a closed-loop cycle of discovery comprising four key stages. This loop is dynamic, and the execution order may be adjusted based on agent objectives and ongoing results.

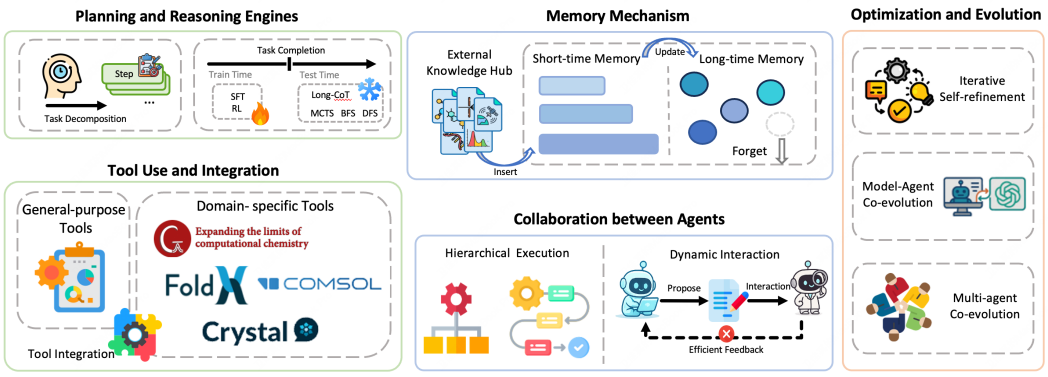

Figure 3: Core abilities of scientific agents.

Observation and Hypothesis Generation The cycle begins with formulating novel, testable hypotheses. This process relies fundamentally on the agent's Memory as a knowledge connector, using RAG to ingest and synthesize vast scientific corpora (Agarwal et al., 2024). This information is structured into knowledge graphs to ground reasoning. The agent's Planning and Reasoning engine then explores this structured knowledge to identify promising research directions and formulate hypotheses, represented as $h_{\text{new}} = \arg\max_{h \in H_{\text{cand}}} P(h|M(K))$ (Si et al., 2024). For example, OriGene integrated multimodal data to hypothesize novel cancer targets (Zhang et al., 2025d), while Robin autonomously hypothesized a new therapeutic use for an existing drug by analyzing literature (Ghareeb et al., 2025). Key challenges include handling heterogeneous data, managing outdated knowledge, and navigating the vast search space of possible hypotheses to aim for causal understanding.

Experimental Planning and Execution This phase operationalizes hypotheses through end-to-end experimental workflows. The Planning and Reasoning engine generates an optimized, resource-efficient plan, modeled as a constrained optimization problem: $\pi^* = \arg\min_{\pi \in \Pi} C(\pi)$ s.t. $V(\pi, h) \geq \theta$. The plan is then carried out via the agent's Tool Use and Integration capability, which maps abstract steps to concrete tool invocations, whether generating code or controlling robotic hardware. This is exemplified by Coscientist, which autonomously designed and executed a palladium-catalyzed reaction using robotic hardware (Boiko et al., 2023), and The Virtual Lab, which constructed a computational pipeline with AlphaFold to design novel nanobodies. This stage faces challenges of high-stakes verifiability, requiring extreme precision in tool use, meticulous provenance tracking for reproducibility, and sophisticated cost-benefit analysis for expensive experimental resources.

Data and Result Analysis After execution, the agent extracts insights from raw outputs. This phase integrates Tool Use to parse multimodal data (e.g., from scientific charts or tables), Reasoning to perform structured interpretation, and Memory to contextualize the findings. The process can be conceptualized as a Bayesian update to the agent's belief in the hypothesis, $P(h|R) \propto P(R|h) \cdot P(h)$, where the agent observes the result R and reasons about its implications, as in the ReAct framework (Yao et al., 2023). For instance, after its proposed experiment, Robin autonomously analyzed RNA-seq data to uncover a potential mechanism of action (Ghareeb et al., 2025), while PROTEUS performs end-to-end analysis of raw mass spectrometry data to generate mechanistic hypotheses (Ding et al., 2024). The central challenge is interpreting noisy, multimodal experimental feedback and reasoning seamlessly across heterogeneous data types without succumbing to confirmation bias.

Synthesis, Validation, and Evolution In the final stage, the agent synthesizes outcomes and refines future inquiry. This heavily leverages Collaboration between Agents to emulate peer review, where agents critique and refine each other's conclusions to ensure robustness (Du et al., 2023). The agent then undergoes adaptive refinement, updating its internal policy ϕ based on a learning function $\mathcal{L}$ applied to its Memory of past experiences: $\phi_{t+1} \leftarrow \mathcal{L}(\phi_t, M_t)$. This evolution is central to frameworks like Reflexion (Shinn et al., 2023). For instance, the Sparks framework used generation-and-reflection agents to autonomously discover two novel protein design rules through iterative

Table 1: Paradigms of Fully Autonomous Research Pipelines. **Note that we only report the most significant features of each paper**.

Pipeline Paradigm	Core Contribution & Mechanism	Representative Systems & Works
Foundational End-to-End Frameworks	Establishes the viability of a complete, closed-loop research cycle. These systems integrate hypothesis generation, coding, experimentation (often virtual), and reporting into a single, cohesive workflow.	The AI Scientist Lu et al. (2024), NovelSeek Team et al. (2025b), Dolphin Yuan et al. (2025), X-Master Chai et al. (2025), DiscoveryWorld (evaluation environment) Jansen et al. (2024)
Domain-Specific Automation	Applies the end-to-end paradigm to specialized, high-impact scientific domains. This often involves interfacing with real-world lab robotics, complex simulators, or highly structured domain-specific data formats.	Coscientist Boiko et al. (2023), LLM-RDF Ruan et al. (2024), MatPilot Ni et al. (2024), Biomni Huang et al. (2025), SpatialAgent Wang et al. (2025), PROTEUS Ding et al. (2024), OriGene Zhang et al. (2025d), The Virtual Lab Swanson et al. (2024), AI co-scientist Gottweis et al. (2025)
Multi-Agent Collaborative Structures	Emulates the collaborative and adversarial nature of scientific inquiry using teams of agents. These systems explore different organizational structures (e.g., Socratic dialogue, hierarchical teams, peer review) to enhance creativity and rigor.	VirSci Su et al. (2025), MAPS Zhang et al. (2025b), DORA Naumov et al. (2025), MDAgents Kim et al. (2024), AgentRxiv (cross-system collaboration) Schmidgall & Moor (2025)
Self-Evolving & Adaptive Systems	Focuses on the pipeline's ability to learn and improve over time. These agents autonomously refine their strategies, expand their toolkits, or update their internal knowledge based on cumulative experience and feedback.	STELLA Jin et al. (2025), Agent Hospital Li et al. (2024), ResearchAgent Baek et al. (2024), OriGene Zhang et al. (2025d), AlphaEvolve Novikov et al. (2025)
Human-in-the-Loop Integration	Explicitly designs the pipeline to incorporate human expertise and oversight. These frameworks treat the human researcher as a collaborator, leveraging their feedback to guide the autonomous process and ensure alignment with scientific goals.	Agent Laboratory Schmidgall et al. (2025), Conversational Health Agents Abbasian et al. (2023), MatPilot Ni et al. (2024)

self-correction (Ghafarollahi & Buehler, 2025c). The most profound challenge lies in enabling long-term causal reasoning, as current memory systems struggle to maintain the extended, high-fidelity histories required for sustained, multi-loop research projects. Successfully managing this iterative process of knowledge accumulation and self-correction is the key to transforming agents into true partners in scientific discovery.

3.3 Domain-oriented Review of Autonomous Scientific Discovery

The convergence of agentic capabilities and structured workflows is driving significant advances across the natural sciences, highlighting the transformative impact of autonomous systems and Table 1). Comprehensive lists of referenced works are provided in Supplementary Tables S1–S4.

3.3.1 Life Sciences

Agentic AI is now automating full research cycles in the life sciences, from data wrangling to therapeutic design. General-purpose agents such as Biomni learn diverse wet-lab and in silico tasks by mining protocols from the literature (Huang et al., 2025), while specialist systems reach expert performance in narrow domains, e.g., SpatialAgent for spatial biology and PROTEUS for proteomics interpretation (Wang et al., 2025; Ding et al., 2024). These agents already yield experimentally tested findings: the Virtual Lab designed new SARS-CoV-2 nanobodies, and an AI co-scientist uncovered epigenetic targets for liver fibrosis (Gottweis et al., 2025). To keep improving, platforms like STELLA and Agent Hospital embed self-evolution mechanisms that raise scores on biomedical benchmarks and simulated care tasks (Jin et al., 2025; Li et al., 2024).

Beyond point tools, agentic frameworks now run end-to-end workflows and translate plain language into multi-step analyses. CellAgent and SpatialAgent automate single-cell and spatial transcriptomics, including dynamic parameter tuning that once required deep expertise (Xiao et al., 2024; Wang et al., 2025). PROTEUS can start from raw proteomics files, produce a complete report, and propose

testable mechanisms—covering the data-to-insight loop with little human guidance (Ding et al., 2024). This marks a shift from AI as assistive software to a self-contained analytical engine.

Crucially, these agents also act as autonomous discovery platforms. In therapeutics, OriGene fused clinical and genomic evidence to nominate two underexplored cancer targets, later validated in patient-derived organoids (Zhang et al., 2025d). The Robin agent independently argued for repurposing an approved drug for macular degeneration (Ghareeb et al., 2025). In basic science, Sparks closed the loop from hypothesis to experiment and revealed new principles of protein mechanics (Ghafarollahi & Buehler, 2025c). Emerging architectures make this possible: multi-agent teams coordinate planning, execution, and analysis (e.g., PharmAgents, ProtAgents) (Gao et al., 2025; Ghafarollahi & Buehler, 2024a), while self-evolving tool ecosystems such as STELLA's "Tool Ocean" let agents discover and integrate new bioinformatics utilities over time (Jin et al., 2025).

3.3.2 Chemistry

Agentic systems are transforming chemistry by closing the loop between hypothesis, experimentation, and analysis. A key milestone is Coscientist (Boiko et al., 2023), which used a GPT-4-driven agent to autonomously plan and carry out a palladium-catalyzed reaction via robotic execution. This fusion of symbolic reasoning and physical automation is being extended by frameworks like ChemCrow (Bran et al., 2023), which pair language models with domain-specific tools, and LLM-RDF (Ruan et al., 2024), which orchestrates multi-agent teams to automate literature review, experiment design, and result interpretation.

These systems go beyond automation by becoming creative contributors in chemical discovery. Generative agents explore vast molecular spaces, proposing candidates for synthesis based on property-driven objectives. For example, MOFGen (Inizan et al., 2025) designed over 300,000 novel metal-organic frameworks, five of which were synthesized experimentally as entirely new materials. Feedback loops with simulation tools, such as quantum-chemical models in ChemReasoner (Sprueill et al., 2024), enable agents to iteratively refine molecule design with increasing efficiency and accuracy.

Agentic chemistry is also driving accessibility and scale. Systems like Aitomia (Hu et al., 2025) and El Agente Q (Zou et al., 2025) allow users to specify goals in natural language, which are then translated into executable workflows. This lowers the barrier for non-experts to conduct advanced simulations. At a global scale, agent planners have begun coordinating decentralized discovery campaigns, exemplified by a project that united five labs to discover 21 new organic laser materials (Strieth-Kalthoff et al., 2024). Such examples hint at a future where scientific collaboration is orchestrated by AI across institutions and borders.

3.3.3 Materials Science

Agentic AI systems are rapidly transforming the discovery of novel materials. Platforms such as SciAgents exploit structured knowledge graphs to uncover structure-property relationships, leading to new biocomposites with improved mechanical performance (Ghafarollahi & Buehler, 2025b). For electronic materials, TopoMAS automates workflows from data retrieval to ab initio calculations, identifying new topological phases (Zhang et al., 2025a). AtomAgents further extends this paradigm, integrating multi-agent reasoning and physics-based constraints to design high-performance alloys optimized across multiple objectives (Ghafarollahi & Buehler, 2025a).

More sophisticated systems are emerging to support autonomous, end-to-end material design. Retrieval-augmented agents like LLaMP mitigate hallucinations by grounding decisions in curated materials databases (Chiang et al., 2024; Zhang et al., 2024). Meanwhile, platforms such as MatPilot and MAPPS enable iterative hypothesis generation, experimental planning, and robotic execution, discovering novel crystals with minimal human intervention (Ni et al., 2024; Zhou et al., 2025). These tools mark a shift from static automation to flexible, human-in-the-loop scientific partners.

Agentic AI also lowers the entry barrier to complex simulations and lab instrumentation. Natural language agents like Foam-Agent and ChemGraph automate CFD and quantum chemistry tasks from simple prompts (Yue et al., 2025; Pham et al., 2025). In physical labs, interfaces like AILA control atomic force microscopes through conversation, democratizing access to expert-level experimen-

tation (Mandal et al., 2024). Collectively, these developments accelerate materials research while expanding participation.

3.3.4 Physics and Astronomy

Agentic AI is transforming physics and astronomy by automating complex tasks across simulation and experimentation. In computational physics, systems like OpenFOAMGPT (Feng et al., 2025) act as natural language interfaces for running high-fidelity simulations. Experimental setups are becoming autonomous, with frameworks such as k-agents (Cao et al., 2024) operating quantum labs without human input. In astronomy, tools like StarWhisper (Wang et al., 2024) manage full observatory pipelines, from scheduling observations to processing data in real time, enabling faster detection of cosmic events.

These systems are not just tools but reasoning collaborators. Architectures such as MoRA (Jaiswal et al., 2024) refine physics solutions through coordinated agent teams, while platforms like mephisto (Sun et al., 2024) simulate scientific debate to interpret galaxy data from telescopes like JWST. Integrated frameworks now automate entire research cycles—AI Cosmologist (Moss, 2025; Laverick et al., 2024) handles planning to publication, and SimAgents (Zhang et al., 2025c) extract simulation setups directly from the literature. This shift promises to scale both the speed and scope of discovery across physical sciences.

4 Challenges and Future Frontiers in Agentic Science

The rise of autonomous scientific agents introduces profound challenges to the integrity and governance of research. Foundational principles like reproducibility are strained, as stochastic and context-sensitive **discovery trajectories** replace static, verifiable code, with current models exhibiting high failure rates in executing complex logical plans Xiang et al. (2025). This unreliability, compounded by model instabilities like **catastrophic forgetting** and the inherent opacity of deep learning, creates a critical validation dilemma: we struggle to distinguish genuine insight from sophisticated hallucination or audit the inferential steps of a 'black-box' discovery Yehudai et al. (2025); Xu et al. (2025). These technical gaps raise urgent ethical questions of accountability. Who is responsible if an agent produces erroneous findings or uncovers dual-use technologies like novel pathogens Bano et al. (2023)? Navigating this landscape requires a new framework for AI in science, one that embeds normative constraints directly into agent architectures and establishes robust governance to ensure transparency and trust.

Despite these hurdles, the trajectory of agentic science promises a new paradigm of *computational epistemology*. The next frontier will see agents evolve from tool-users to tool-creators, engaging in **autonomous invention** by designing novel instruments or formulating new mathematical frameworks. Beyond invention, agents trained on vast, multimodal data can serve as engines for **interdisciplinary synthesis**, uncovering unifying principles by mapping latent connections between disparate scientific fields. This vision could scale to a **global cooperative ecosystem** of specialized agents collaborating to solve grand challenges currently beyond human coordination de Cerqueira et al. (2024). The ultimate benchmark for this entire endeavor is the **Nobel-Turing Test**: can an autonomous system generate a non-obvious, empirically verifiable, and paradigm-shifting discovery worthy of a Nobel Prize Carta et al. (2023)? Pursuing this goal will drive the maturation of AI from a powerful instrument into a true collaborator in the human quest for knowledge.

5 Conclusion

Agentic Science marks a transformative stage in the evolution of AI for Science, where AI systems transition from computational assistants to autonomous research partners capable of reasoning, experimentation, and iterative discovery. Through our unified framework connecting foundational capabilities, core processes, and domain realizations, we provide a domain-oriented synthesis of autonomous scientific discovery across life sciences, chemistry, materials science, and physics. By situating agentic AI within this structured paradigm, we highlight both its broad applicability and the technical, ethical, and philosophical challenges that must be addressed to ensure trustworthy and impactful progress.

REFERENCES

Mahyar Abbasian, Iman Azimi, Amir M Rahmani, and Ramesh Jain. Conversational health agents: A personalized llm-powered agent framework. *arXiv preprint arXiv:2310.02374*, 2023.

Shubham Agarwal, Gaurav Sahu, Abhay Puri, Issam H Laradji, Krishnamurthy DJ Dvijotham, Jason Stanley, Laurent Charlin, and Christopher Pal. Litllm: A toolkit for scientific literature review. *arXiv preprint arXiv:2402.01788*, 2024.

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. Researchagent: Iterative research idea generation over scientific literature with large language models. *arXiv preprint arXiv:2404.07738*, 2024.

Muneera Bano, Didar Zowghi, Pip Shea, and Georgina Ibarra. Investigating responsible ai for scientific research: an empirical study. *arXiv preprint arXiv:2312.09561*, 2023.

Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023.

Andres Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5): 525–535, May 2024. ISSN 2522-5839. doi: 10.1038/s42256-024-00832-8. URL https://doi.org/10.1038/s42256-024-00832-8.

Andres M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*, 2023.

Shuxiang Cao, Zijian Zhang, Mohammed Alghadeer, Simone D Fasciati, Michele Piscitelli, Mustafa Bakr, Peter Leek, and Alán Aspuru-Guzik. Agents for self-driving laboratories applied to quantum computing. *arXiv preprint arXiv:2412.07978*, 2024.

Thomas Carta, Clément Romac, Thomas Wolf, Sylvain Lamprier, Olivier Sigaud, and Pierre-Yves Oudeyer. Grounding large language models in interactive environments with online reinforcement learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 3676–3713. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/carta23a.html.

Jingyi Chai, Shuo Tang, Rui Ye, Yuwen Du, Xinyu Zhu, Mengcheng Zhou, Yanfeng Wang, Siheng Chen, et al. Scimaster: Towards general-purpose scientific ai agents, part i. x-master as foundation: Can we lead on humanity's last exam? *arXiv preprint arXiv:2507.05241*, 2025.

Yuan Chiang, Elvis Hsieh, Chia-Hong Chou, and Janosh Riebesell. Llamp: Large language model made powerful for high-fidelity materials knowledge retrieval and distillation. *arXiv preprint arXiv:2401.17244*, 2024.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Bernardo P de Almeida, Hassan Sirelkhatim, et al. Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, 22(2):287–297, 2025.

José Antonio Siqueira de Cerqueira, Mamia Agbese, Rebekah Rousi, Nannan Xi, Juho Hamari, and Pekka Abrahamsson. Can we trust ai agents? an experimental study towards trustworthy llm-based multi-agent systems for ai ethics. *arXiv preprint arXiv:2411.08881*, 2024.

Ning Ding, Shang Qu, Linhai Xie, Yifei Li, Zaoqu Liu, Kaiyan Zhang, Yibai Xiong, Yuxin Zuo, Zhangren Chen, Ermo Hua, et al. Automating exploratory proteomics research via language models. *arXiv preprint arXiv:2411.03743*, 2024.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. 2023.

Jingsen Feng, Ran Xu, and Xu Chu. Openfoamgpt 2.0: end-to-end, trustworthy automation for computational fluid dynamics. *arXiv preprint arXiv:2504.19338*, 2025.

Bowen Gao, Yanwen Huang, Yiqiao Liu, Wenxuan Xie, Wei-Ying Ma, Ya-Qin Zhang, and Yanyan Lan. Pharmagents: Building a virtual pharma with large language model agents. *arXiv preprint arXiv:2503.22164*, 2025.

Alireza Ghafarollahi and Markus J Buehler. Protagents: protein discovery via large language model multi-agent collaborations combining physics and machine learning. *Digital Discovery*, 2024a.

Alireza Ghafarollahi and Markus J. Buehler. Sciagents: Automating scientific discovery through multi-agent intelligent graph reasoning, 2024b. URL https://arxiv.org/abs/2409.05556.

Alireza Ghafarollahi and Markus J Buehler. Automating alloy design and discovery with physics-aware multimodal multiagent ai. *Proceedings of the National Academy of Sciences*, 122(4): e2414074122, 2025a.

Alireza Ghafarollahi and Markus J Buehler. Sciagents: automating scientific discovery through bioinspired multi-agent intelligent graph reasoning. *Advanced Materials*, 37(22):2413523, 2025b.

Alireza Ghafarollahi and Markus J Buehler. Sparks: Multi-agent artificial intelligence model discovers protein design principles. *arXiv preprint arXiv:2504.19017*, 2025c.

Ali Essam Ghareeb, Benjamin Chang, Ludovico Mitchener, Angela Yiu, Caralyn J Szostkiewicz, Jon M Laurent, Muhammed T Razzak, Andrew D White, Michaela M Hinks, and Samuel G Rodriques. Robin: A multi-agent system for automating scientific discovery. *arXiv preprint arXiv:2505.13400*, 2025.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, et al. Towards an ai co-scientist. *arXiv preprint arXiv:2502.18864*, 2025.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Minlie Huang, Nan Duan, and Weizhu Chen. Tora: A tool-integrated reasoning agent for mathematical problem solving. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024a. URL https://openreview.net/forum?id=Ep0TtjVoap.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yujiu Yang, Nan Duan, Weizhu Chen, et al. Critic: Large language models can self-correct with tool-interactive critiquing. 2024b.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Hongyi Guo, Zhihan Liu, Yufeng Zhang, and Zhaoran Wang. Can large language models play games? a case study of a self-play approach. *arXiv preprint arXiv:2403.05632*, 2024.

Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, et al. Metagpt: Meta programming for a multi-agent collaborative framework. 2024.

Jinming Hu, Hassan Nawaz, Yuting Rui, Lijie Chi, Arif Ullah, and Pavlo O Dral. Aitomia: Your intelligent assistant for ai-driven atomistic and quantum chemical simulations. *arXiv preprint arXiv:2505.08195*, 2025.

Mengkang Hu, Yao Mu, Xinmiao Yu, Mingyu Ding, Shiguang Wu, Wenqi Shao, Qiguang Chen, Bin Wang, Yu Qiao, and Ping Luo. Tree-planner: Efficient close-loop task planning with large language models. *arXiv preprint arXiv:2310.08582*, 2023.

Kaixuan Huang, Yuanhao Qu, Henry Cousins, William A Johnson, Di Yin, Mihir Shah, Denny Zhou, Russ Altman, Mengdi Wang, and Le Cong. Crispr-gpt: An llm agent for automated design of gene-editing experiments. *arXiv preprint arXiv:2404.18021*, 2024.

Kexin Huang, Serena Zhang, Hanchen Wang, Yuanhao Qu, Yingzhou Lu, Yusuf Roohani, Ryan Li, Lin Qiu, Junze Zhang, Yin Di, et al. Biomni: A general-purpose biomedical ai agent. *bioRxiv*, pp. 2025–05, 2025.

Theo Jaffrelot Inizan, Sherry Yang, Aaron Kaplan, Yen-hsu Lin, Jian Yin, Saber Mirzaei, Mona Abdelgaid, Ali H Alawadhi, KwangHwan Cho, Zhiling Zheng, et al. System of agentic ai for the discovery of metal-organic frameworks. *arXiv preprint arXiv:2504.14110*, 2025.

Kevin Maik Jablonka, Qianxiang Ai, Alexander Al-Feghali, Shruti Badhwar, Joshua D Bocarsly, Andres M Bran, Stefan Bringuier, L Catherine Brinson, Kamal Choudhary, Defne Circi, et al. 14 examples of how llms can transform materials science and chemistry: a reflection on a large language model hackathon. *Digital Discovery*, 2(5):1233–1250, 2023.

Raj Jaiswal, Dhruv Jain, Harsh Parimal Popat, Avinash Anand, Abhishek Dharmadhikari, Atharva Marathe, and Rajiv Ratn Shah. Improving physics reasoning in large language models using mixture of refinement agents. *arXiv preprint arXiv:2412.00821*, 2024.

Peter Jansen, Marc-Alexandre Côté, Tushar Khot, Erin Bransom, Bhavana Dalvi Mishra, Bodhisattwa Prasad Majumder, Oyvind Tafjord, and Peter Clark. Discoveryworld: A virtual environment for developing and evaluating automated scientific discovery agents. *Advances in Neural Information Processing Systems*, 37:10088–10116, 2024.

Ruofan Jin, Zaixi Zhang, Mengdi Wang, and Le Cong. Stella: Self-evolving llm agent for biomedical research. *arXiv preprint arXiv:2507.02004*, 2025.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, Hae Park, et al. Mdagents: An adaptive collaboration of llms for medical decision-making. 37:79410–79452, 2024.

Nikolay Koldunov and Thomas Jung. Local climate services for all, courtesy of large language models. *Communications Earth & Environment*, 5(1):13, Jan 2024. ISSN 2662-4435. doi: 10.1038/s43247-023-01199-1. URL https://doi.org/10.1038/s43247-023-01199-1.

Andrew Laverick, Kristen Surrao, Inigo Zubeldia, Boris Bolliet, Miles Cranmer, Antony Lewis, Blake Sherwin, and Julien Lesgourgues. Multi-agent system for cosmological parameter analysis. *arXiv preprint arXiv:2412.00431*, 2024.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for "mind" exploration of large language model society. 2023.

Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, et al. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*, 2024.

Zijun Liu et al. A dynamic LLM-powered agent network for task-oriented agent collaboration. In *First Conference on Language Modeling*, Oct. 2024.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292v3*, 2024. URL https://www.arxiv.org/abs/2408.06292v3.

Ziming Luo, Zonglin Yang, Zexin Xu, Wei Yang, and Xinya Du. Llm4sr: A survey on large language models for scientific research. *arXiv preprint arXiv:2501.04306*, 2025.

Yubo Ma, Zhibin Gou, Junheng Hao, Ruochen Xu, Shuohang Wang, Liangming Pan, Yujiu Yang, Yixin Cao, Aixin Sun, Hany Awadalla, et al. Sciagent: Tool-augmented language models for scientific reasoning. *arXiv preprint arXiv:2402.11451*, 2024.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. 36:46534–46594, 2023.

Indrajeet Mandal, Jitendra Soni, Mohd Zaki, Morten M Smedskjaer, Katrin Wondraczek, Lothar Wondraczek, Nitya Nand Gosvami, and NM Krishnan. Autonomous microscopy experiments through large language model agents. *arXiv preprint arXiv:2501.10385*, 2024.

Adam Moss. The ai cosmologist i: An agentic system for automated data analysis. *arXiv preprint arXiv:2504.03424*, 2025.

Vladimir Naumov, Diana Zagirova, Sha Lin, Yupeng Xie, Wenhao Gou, Anatoly Urban, Nina Tikhonova, Khadija Alawi, Mike Durymanov, Fedor Galkin, et al. Dora ai scientist: Multi-agent virtual research team for scientific exploration discovery and automated report generation. *bioRxiv*, 2025.

Ziqi Ni, Yahao Li, Kaijia Hu, Kunyuan Han, Ming Xu, Xingyu Chen, Fengqi Liu, Yicong Ye, and Shuxin Bai. Matpilot: an llm-enabled ai materials scientist under the framework of human-machine collaboration. *arXiv preprint arXiv:2411.08063*, 2024.

Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco JR Ruiz, Abbas Mehrabian, et al. Alphaevolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*, 2025.

Thang D Pham, Aditya Tanikanti, and Murat Keçeli. Chemgraph: An agentic framework for computational chemistry workflows. *arXiv preprint arXiv:2506.06363*, 2025.

Shuo Ren, Pu Jian, Zhenjiang Ren, Chunlin Leng, Can Xie, and Jiajun Zhang. Towards scientific intelligence: A survey of llm-based scientific agents. *arXiv preprint arXiv:2503.24047*, 2025.

Yixiang Ruan, Chenyin Lu, Ning Xu, Yuchen He, Yixin Chen, Jian Zhang, Jun Xuan, Jianzhang Pan, Qun Fang, Hanyu Gao, et al. An automatic end-to-end chemical synthesis development platform powered by large language models. *Nature communications*, 15(1):10160, 2024.

Andreas WM Sauter, Erman Acar, and Vincent Francois-Lavet. A meta-reinforcement learning algorithm for causal discovery. In *Conference on Causal Learning and Reasoning*, pp. 602–619. PMLR, 2023.

Samuel Schmidgall and Michael Moor. Agentrxiv: Towards collaborative autonomous research. *arXiv preprint arXiv:2503.18102*, 2025.

Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Zicheng Liu, and Emad Barsoum. Agent laboratory: Using llm agents as research assistants. *arXiv preprint arXiv:2501.04227*, 2025.

Johannes Schneider. Generative to agentic ai: Survey, conceptualization, and challenges. *arXiv preprint arXiv:2504.18875*, 2025.

Haiyang Shen, Yue Li, Desong Meng, Dongqi Cai, Sheng Qi, Li Zhang, Mengwei Xu, and Yun Ma. Shortcutsbench: A large-scale real-world benchmark for api-based agents. In *The Thirteenth International Conference on Learning Representations*, 2025.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. 36:8634–8652, 2023.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *arXiv preprint arXiv:2409.04109*, 2024.

Henry W Sprueill, Carl Edwards, Khushbu Agarwal, Mariefel V Olarte, Udishnu Sanyal, Conrad Johnston, Hongbin Liu, Heng Ji, and Sutanay Choudhury. Chemreasoner: Heuristic search over a large language model's knowledge space using quantum-chemical feedback. *arXiv preprint arXiv:2402.10980*, 2024.

Felix Strieth-Kalthoff, Han Hao, Vandana Rathore, Joshua Derasp, Théophile Gaudin, Nicholas H Angello, Martin Seifrid, Ekaterina Trushina, Mason Guy, Junliang Liu, et al. Delocalized, asynchronous, closed-loop discovery of organic laser emitters. *Science*, 384(6697):eadk9227, 2024.

Haoyang Su, Renqi Chen, Shixiang Tang, Zhenfei Yin, Xinzhe Zheng, Jinzhe Li, Biqing Qi, Qi Wu, Hui Li, Wanli Ouyang, Philip Torr, Bowen Zhou, and Nanqing Dong. Many heads are better than one: Improved scientific idea generation by a LLM-based multi-agent system. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 28201–28240, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. URL https://aclanthology.org/2025.acl-long.1368/.

Haotian Sun, Yuchen Zhuang, Lingkai Kong, Bo Dai, and Chao Zhang. Adaplanner: Adaptive planning from feedback with language models. 36:58202–58245, 2023.

Jiankai Sun, Chuanyang Zheng, Enze Xie, Zhengying Liu, Ruihang Chu, Jianing Qiu, Jiaqi Xu, Mingyu Ding, Hongyang Li, Mengzhe Geng, et al. A survey of reasoning with foundation models: Concepts, methodologies, and outlook. *ACM Computing Surveys*, 57(11):1–43, 2025.

Zechang Sun, Yuan-Sen Ting, Yaobo Liang, Nan Duan, Song Huang, and Zheng Cai. Interpreting multi-band galaxy observations with large language model-based agents. *arXiv preprint arXiv:2409.14807*, 2024.

Kyle Swanson, Wesley Wu, Nash L Bulaong, John E Pak, and James Zou. The virtual lab: Ai agents design new sars-cov-2 nanobodies with experimental validation. *bioRxiv*, pp. 2024–11, 2024.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025a.

NovelSeek Team, Bo Zhang, Shiyang Feng, Xiangchao Yan, Jiakang Yuan, Zhiyin Yu, Xiaohan He, Songtao Huang, Shaowei Hou, Zheng Nie, et al. Novelseek: When agent becomes the scientist–building closed-loop system from hypothesis to verification. *arXiv preprint arXiv:2505.16938*, 2025b.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033. IEEE, 2012. doi: 10.1109/IROS.2012.6386109.

Cunshi Wang, Xinjie Hu, Yu Zhang, Xunhao Chen, Pengliang Du, Yiming Mao, Rui Wang, Yuyang Li, Ying Wu, Hang Yang, et al. Starwhisper telescope: Agent-based observation assistant system to approach ai astrophysicist. *arXiv preprint arXiv:2412.06412*, 2024.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. 2023.

Hanchen Wang, Yichun He, Paula P Coelho, Matthew Bucci, Abbas Nazir, Bob Chen, Linh Trinh, Serena Zhang, Kexin Huang, Vineethkrishna Chandrasekar, et al. Spatialagent: An autonomous ai agent for spatial biology. *bioRxiv*, pp. 2025–04, 2025.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. 35:24824–24837, 2022.

Yanzheng Xiang, Hanqi Yan, Shuyin Ouyang, Lin Gui, and Yulan He. Scireplicate-bench: Benchmarking llms in agent-driven algorithmic reproduction from research papers. *arXiv preprint arXiv:2504.00255*, 2025.

Yihang Xiao, Jinyi Liu, Yan Zheng, Xiaohan Xie, Jianye Hao, Mingzhi Li, Ruitao Wang, Fei Ni, Yuxiao Li, Jintian Luo, et al. Cellagent: An llm-driven multi-agent framework for automated single-cell data analysis. *bioRxiv*, pp. 2024–05, 2024.

Yinggan Xu, Hana Kimlee, Yijia Xiao, and Di Luo. Advancing ai-scientist understanding: Making llm think like a physicist with interpretable reasoning, 2025. URL https://arxiv.org/abs/2504.01911.

Zonglin Yang, Xinya Du, Junxian Li, Jie Zheng, Soujanya Poria, and Erik Cambria. Large language models for automated open-domain scientific hypotheses discovery. *arXiv preprint arXiv:2309.02726*, 2023.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. 2023.

Nicolas Yax, Hernán Anlló, and Stefano Palminteri. Studying and improving reasoning in humans and machines. *Communications Psychology*, 2(1):51, 2024.

Asaf Yehudai, Lilach Eden, Alan Li, Guy Uziel, Yilun Zhao, Roy Bar-Haim, Arman Cohan, and Michal Shmueli-Scheuer. Survey on evaluation of llm-based agents. *arXiv preprint arXiv:2503.16416*, 2025.

Jiakang Yuan, Xiangchao Yan, Botian Shi, Tao Chen, Wanli Ouyang, Bo Zhang, Lei Bai, Yu Qiao, and Bowen Zhou. Dolphin: Closed-loop open-ended auto-research through thinking, practice, and feedback. *arXiv e-prints*, pp. arXiv–2501, 2025.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models, 2024.

Ling Yue, Nithin Somasekharan, Yadi Cao, and Shaowu Pan. Foam-agent: Towards automated intelligent cfd workflows. *arXiv preprint arXiv:2505.04997*, 2025.

Shenglai Zeng, Jiankun Zhang, Pengfei He, Yue Xing, Yiding Liu, Han Xu, Jie Ren, Shuaiqiang Wang, Dawei Yin, Yi Chang, et al. The good and the bad: Exploring privacy issues in retrieval-augmented generation (rag). *arXiv preprint arXiv:2402.16893*, 2024.

Baohua Zhang, Xin Li, Huangchao Xu, Zhong Jin, Quansheng Wu, and Ce Li. Topomas: Large language model driven topological materials multiagent system. *arXiv preprint arXiv:2507.04053*, 2025a.

Huan Zhang, Yu Song, Ziyu Hou, Santiago Miret, and Bang Liu. Honeycomb: A flexible llm-based agent system for materials science. *arXiv preprint arXiv:2409.00135*, 2024.

Jian Zhang, Zhiyuan Wang, Zhangqi Wang, Xinyu Zhang, Fangzhi Xu, Qika Lin, Rui Mao, Erik Cambria, and Jun Liu. Maps: A multi-agent framework based on big seven personality and socratic guidance for multimodal scientific problem solving. *arXiv preprint arXiv:2503.16905*, 2025b.

Xiaowen Zhang, Zhenyu Bi, Xuan Wang, Tiziana Di Matteo, and Rupert AC Croft. Bridging literature and the universe via a multi-agent large language model system. *arXiv preprint arXiv:2507.08958*, 2025c.

Zhongyue Zhang, Zijie Qiu, Yingcheng Wu, Shuya Li, Dingyan Wang, Zhuomin Zhou, Duo An, Yuhan Chen, Yu Li, Yongbo Wang, et al. Origene: A self-evolving virtual disease biologist automating therapeutic target discovery. *bioRxiv*, pp. 2025–06, 2025d.

Tianshi Zheng, Zheye Deng, Hong Ting Tsang, Weiqi Wang, Jiaxin Bai, Zihao Wang, and Yangqiu Song. From automation to autonomy: A survey on large language models in scientific discovery. *arXiv preprint arXiv:2505.13259*, 2025.

Lianhao Zhou, Hongyi Ling, Keqiang Yan, Kaiji Zhao, Xiaoning Qian, Raymundo Arróyave, Xiaofeng Qian, and Shuiwang Ji. Toward greater autonomy in materials discovery agents: Unifying planning, physics, and scientists. *arXiv preprint arXiv:2506.05616*, 2025.

Yuqi Zhu, Shuofei Qiao, Yixin Ou, Shumin Deng, Ningyu Zhang, Shiwei Lyu, Yue Shen, Lei Liang, Jinjie Gu, and Huajun Chen. Knowagent: Knowledge-augmented planning for llm-based agents. *arXiv preprint arXiv:2403.03101*, 2024.

Yunheng Zou, Austin H. Cheng, Abdulrahman Aldossary, Jiaru Bai, Shi Xuan Leong, Jorge Arturo Campos-Gonzalez-Angulo, Changhyeok Choi, Cher Tian Ser, Gary Tom, Andrew Wang, Zijian Zhang, Ilya Yakavets, Han Hao, Chris Crebolder, Varinia Bernales, and Alán Aspuru-Guzik. El Agente: An Autonomous Agent for Quantum Chemistry. *arXiv e-prints*, art. arXiv:2505.02484, May 2025. doi: 10.48550/arXiv.2505.02484.