ISSN: 2581-3404 (Online) IF: 5.445 (SJIF)
IJIRTM, Volume-6, Issue-4, August-2022.

AI Agents for Scientific Discovery: A Survey of AI Co-Scientists for Automated Hypothesis Generation and Laboratory Experimentation

Drishti[1], Harsh[2], Mudit Gupta[3], Sunny[4], Jyoti Bhardwaj[5]
Department of Computer Science and Engineering (AI & ML), Panipat Institute of Engineering and Technology (PIET), Samalkha, Panipat, Haryana, India[1,2,3,4,5]
2820289.cse@pietgroup.co.in[1], 2820309.cse@pietgroup.co.in[2], 2820319.cse@pietgroup.co.in[3], 2820286.cse@pietgroup.co.in[4], jyoti.cse-aiml@piet.co.in[5]

Abstract: *The scientific method has traditionally relied on human researchers to formulate hypotheses, design experiments, and interpret results, a process that is slow, expensive, and bounded by the cognitive throughput of individual scientists. Recent advances in large language models (LLMs), multi-agent orchestration, knowledge-graph retrieval, and self-driving laboratories have given rise to a new class of systems collectively termed "AI co-scientists" — autonomous or semi-autonomous agents capable of surveying literature, proposing novel and testable hypotheses, designing experimental protocols, and, in closed-loop configurations, executing those protocols on robotic laboratory hardware. This paper presents a comprehensive survey and architectural analysis of AI co-scientist systems. We describe the canonical multi-agent pipeline that underlies most contemporary implementations, comprising a generation agent, a reflection or critique agent, a ranking or tournament agent, and an evolution agent, and we relate this pipeline to concrete deployed systems spanning chemistry, biology, and materials science. We then examine the automated hypothesis generation problem in depth, covering retrieval-augmented reasoning, novelty estimation, feasibility scoring, and human-in-the-loop steering. We further analyze the integration of hypothesis-generation agents with self-driving laboratories and robotic experimentation platforms, discussing closed-loop Bayesian optimization, active learning for experiment selection, and the hardware and software abstractions that connect language-model reasoning to physical actuation. Three case studies — materials discovery, drug repurposing, and synthetic-biology pathway design — illustrate the practical performance and limitations of existing systems. We propose a taxonomy of evaluation metrics for AI co-scientists, covering hypothesis novelty, experimental success rate, cost per validated discovery, and reproducibility, and we survey emerging benchmarks. Finally, we discuss open challenges, including hallucinated citations, dual-use and biosafety risk, reproducibility of autonomously generated protocols, and the epistemic status of machine-generated scientific claims, and we outline promising directions for future research, including federated multi-laboratory orchestration, formal verification of experimental plans, and tighter coupling between symbolic scientific reasoning and sub-symbolic perception. Our analysis indicates that while AI co-scientists have already accelerated specific narrow discovery tasks, achieving reliable end-to-end autonomous science will require substantial advances in causal reasoning, uncertainty quantification, and safety-aware experimental design.*

Keywords: AI co-scientist, autonomous experimentation, hypothesis generation, large language models, multi-agent systems, self-driving laboratories, robotic chemistry, scientific discovery, closed-loop optimization, human-in-the-loop AI.

Introduction

The pace of scientific discovery is fundamentally constrained by human throughput: reading and synthesizing literature, forming hypotheses, designing controlled experiments, and iterating on results. A single researcher can meaningfully track only a small fraction of the millions of papers published annually, and the design-build-test-analyze cycle in fields such as chemistry, materials science, and molecular biology often spans weeks to months per iteration. Over the past three years, two independent but converging technological trends have begun to reshape this picture. First, large language models (LLMs) trained on scientific corpora have demonstrated an emergent capacity to retrieve, synthesize, and reason over specialized domain knowledge, enabling them to propose plausible mechanistic explanations and novel research directions. Second, robotic and automated laboratory platforms — so-called self-driving laboratories (SDLs) — have matured to the point where standard wet-lab and materials-synthesis procedures can be executed with minimal human intervention, driven by programmable liquid handlers, robotic arms, and automated characterization instruments.

The confluence of these two trends has given rise to the concept of the "AI co-scientist": an agentic system, typically built from one or more LLMs augmented with tool use, retrieval, memory, and planning modules, that participates in the scientific process not merely as a passive question-answering assistant but as an active collaborator capable of proposing, prioritizing, and in some cases physically executing experiments. Systems in this category include multi-agent hypothesis-generation frameworks that debate and refine candidate hypotheses through simulated scientific discourse, and fully closed-loop platforms in which an LLM-driven planner issues experimental instructions directly to robotic hardware and updates its beliefs based on returned measurements.

This paper aims to provide a structured, technically grounded survey of the AI co-scientist landscape as of 2026, organized around the two central capabilities implied by the term: automated hypothesis generation and automated laboratory experimentation. Rather than cataloguing every published system, we focus on identifying the recurring architectural patterns, the key algorithmic problems that must be solved at each stage of the discovery pipeline, and the evaluation methodologies that the community has begun to converge on. We further situate these systems within the broader context of agentic AI, drawing connections to advances in tool-augmented reasoning, retrieval-augmented generation, and multi-agent debate that originated outside the natural-sciences domain but have proven directly applicable to it.

It is important at the outset to distinguish the AI co-scientist paradigm from two related but narrower categories of tool. The first is the conventional laboratory-information-management or electronic-lab-notebook software that many laboratories already use, which records and organizes experimental data but exercises no independent scientific judgment. The second is the narrow, single-purpose predictive model, such as a model trained solely to predict protein structure or reaction yield, which performs a valuable but circumscribed function within a pipeline that a human scientist must still assemble, sequence, and interpret. AI co-scientists are distinguished from both categories by their attempt to close the loop across the full research cycle: formulating what question to ask, deciding what experiment would best answer it, and, in the most advanced deployments, physically carrying out that experiment and updating future decisions based on the outcome. This end-to-end scope is simultaneously the source of their promise and the source of the compounded technical and safety challenges that this survey examines.

The remainder of this paper is organized as follows. Section II reviews related work and situates AI co-scientists within the broader literature on agentic AI and laboratory automation. Section III presents a general architecture for AI co-scientist systems. Section IV examines automated hypothesis generation in detail. Section V discusses automated laboratory experimentation and closed-loop integration. Section VI presents three illustrative case studies. Section VII discusses evaluation metrics and benchmarks. Section VIII discusses limitations and risks. Section IX outlines future directions, and Section X concludes.

Related Work and Background

A. From Question Answering to Agentic Science

Early applications of language models to science were largely confined to information retrieval and question answering: summarizing papers, answering factual queries about known compounds, or assisting with literature review. The transition toward agentic behavior — models that plan multi-step tasks, invoke external tools such as search engines, chemistry calculators, or laboratory-control APIs, and revise their plans based on intermediate results — has been the primary enabler of AI co-scientist systems. Tool-augmented reasoning frameworks demonstrated that LLMs could be coupled to external chemistry and biology software to perform tasks such as reaction prediction, retrosynthesis planning, and molecular property estimation, forming a bridge between free-text reasoning and domain-specific computation.

B. Self-Driving Laboratories

In parallel, the materials-science and chemistry communities developed self-driving laboratories that couple automated synthesis and characterization hardware with machine-learning-based experiment planners, typically using Bayesian optimization or active learning to select the next experiment from a large candidate space. These systems established the closed-loop "suggest–synthesize–characterize–update" cycle that AI co-scientists now extend by replacing or augmenting the optimization-based planner with an LLM-based reasoning agent capable of incorporating qualitative domain knowledge, mechanistic hypotheses, and literature context that purely numerical optimizers cannot easily represent.

C. Multi-Agent Debate and Reflection

A separate line of work on multi-agent LLM systems showed that having multiple model instances critique, debate, or vote on candidate outputs improves factuality and reduces hallucination relative to single-pass generation. This insight has been directly incorporated into AI co-scientist architectures, most visibly in systems that structure hypothesis generation as a tournament among competing agents, each proposing and defending candidate hypotheses before a ranking agent selects the most promising ones for further elaboration or experimental testing.

D. Automated Theorem Proving and Symbolic Discovery

A further antecedent to AI co-scientists is found in automated theorem proving and symbolic-regression systems, which have long demonstrated that machine search over a formally defined hypothesis space can rediscover known physical laws and, in some cases, propose genuinely novel mathematical conjectures. These systems differ from LLM-based co-scientists in relying on formally specified search spaces rather than open-ended natural-language reasoning, but they establish an important precedent: that machine-generated scientific content can meet a rigorous standard of correctness when the underlying search is grounded in verifiable formal structure. Contemporary AI co-scientists increasingly borrow this idea by coupling free-form language-model reasoning with narrower, formally verifiable sub-modules for specific steps, such as stoichiometric balancing or thermodynamic feasibility checking, where formal verification is tractable.

E. Positioning Relative to This Survey

Prior surveys have separately covered large language models for chemistry, robotic laboratory automation, and multi-agent LLM systems, but few have treated the integration of hypothesis generation and physical experimentation as a unified pipeline with shared failure modes and shared evaluation criteria. This paper treats that integration as its central object of study, using the four-role architecture introduced in Section III as an organizing framework for comparing systems that individual prior surveys discuss only in isolation.

Architecture of AI Co-Scientist Systems

Although implementations vary considerably in scale and domain, most published AI co-scientist systems share a common architectural skeleton consisting of four functional roles, which may be implemented as separate model instances, separate prompts against a shared model, or specialized fine-tuned sub-models.

We describe each role below and summarize the pipeline in Fig. 1 (described textually, as this document does not embed graphics).

This four-role decomposition is a useful analytical device even for systems whose original authors describe their architecture in different terms, because it exposes the specific point at which each system makes its central design trade-offs. A system that invests heavily in the generation role but treats ranking as a simple top-k filter, for instance, is implicitly betting that generation quality is the binding constraint on overall performance, whereas a system that invests heavily in tournament-style ranking is implicitly betting that a sufficiently diverse candidate pool can be generated cheaply and that the harder problem is discriminating among candidates. Making this trade-off explicit, as we do throughout this section, allows more direct comparison across systems than is possible from architecture diagrams alone, which frequently differ in terminology despite representing structurally similar design decisions.

A. Generation Agent

The generation agent is responsible for producing an initial pool of candidate hypotheses in response to a research goal specified by a human scientist, typically expressed as a natural-language research question together with constraints such as target properties, available reagents, or safety restrictions. Generation strategies range from direct prompting of a single large model to more elaborate retrieval-augmented pipelines in which the agent first queries literature databases, patent repositories, and structured knowledge graphs, then conditions hypothesis generation on the retrieved evidence. Some systems additionally employ simulated scientific debate, in which multiple generation agents propose hypotheses independently and then argue for their relative merit, with the exchange itself surfaced as additional context for subsequent generation rounds.

B. Reflection and Critique Agent

The reflection agent evaluates each generated hypothesis along multiple axes, typically including novelty relative to existing literature, internal logical consistency, consistency with known physical or biochemical constraints, and experimental feasibility given available laboratory resources. This agent frequently performs a literature-grounded fact-check, retrieving supporting or contradicting evidence for each hypothesis and flagging claims that cannot be substantiated. In several deployed systems, the reflection agent also estimates a confidence or plausibility score that is later used by the ranking agent.

C. Ranking and Tournament Agent

Because generation agents typically produce far more candidate hypotheses than can be experimentally tested, a ranking mechanism is required to prioritize the pool. A common approach borrowed from multi-agent debate literature is the Elo-style tournament, in which pairs of hypotheses are compared head-to-head by an LLM judge conditioned on the critique agent's evaluations, with match outcomes aggregated into a global ranking. This approach has the advantage of being more robust to individual scoring miscalibration than direct numerical scoring, since it relies on relative rather than absolute judgments.

D. Evolution and Refinement Agent

Top-ranked hypotheses are frequently underspecified for direct experimental testing; an evolution agent elaborates them into concrete, testable experimental protocols, specifying reagents, procedures, expected observables, and success criteria. In closed-loop systems this agent additionally translates the protocol into machine-executable instructions for laboratory-automation hardware. Some architectures introduce a fifth "meta-review" role that periodically summarizes the state of the overall research campaign for the human supervisor and adjusts the search strategy based on cumulative results.

E. Memory and Knowledge Substrate

Underlying all four agent roles is a persistent memory substrate, typically comprising a vector database of embedded literature passages, a structured knowledge graph of entities such as compounds, genes, or materials and their known relationships, and an experiment log recording all hypotheses generated, protocols executed, and results observed during the current research campaign. This substrate serves both as

a retrieval source for grounding generation and critique, and as a record that enables the system to avoid redundant experimentation and to detect when new results contradict previously accepted hypotheses.

F. Orchestration and Control Flow

The four agent roles are coordinated by an orchestration layer that manages iteration budget, decides when a research campaign has converged sufficiently to surface results to the human supervisor, and handles error recovery when a downstream tool call, such as a literature-database query or a laboratory-execution request, fails or times out. Orchestration is typically implemented as an explicit state machine or directed graph rather than a single monolithic prompt, both for reliability and to allow individual agent roles to be independently upgraded, replaced, or fine-tuned without redesigning the entire pipeline. Several open-source orchestration frameworks originally developed for general-purpose agentic applications have been adapted for this purpose, providing primitives such as tool registries, structured message passing between agents, and check pointing of intermediate state.

G. Representative Systems

Table II situates several representative AI co-scientist and closely related systems described in the literature against the four-role architecture, indicating which roles each system implements explicitly and which domain it primarily targets. The comparison illustrates that no single published system yet implements all four roles with equal sophistication; most current implementations invest disproportionately in either the generation/reflection pair or the closed-loop experimentation pair, reflecting the substantial engineering effort each half of the pipeline independently requires.

Table 1: REPRESENTATIVE SYSTEMS MAPPED TO THE FOUR-ROLE ARCHITECTURE

System	Primary Domain	Notable Emphasis
Coscientist	Chemistry	Tool-augmented planning + execution
ChemCrow	Chemistry	Tool use for reaction/property tasks
Google AI co-scientist	General biomedicine	Generation/ranking tournament
A-Lab	Materials synthesis	Closed-loop robotic synthesis
Mobile robotic chemist	Photocatalysis	Bayesian-optimized exploration
ORGANA	Chemistry	Robotic assistant orchestration

H. Human-AI Collaboration Models

The degree of autonomy granted to an AI co-scientist varies along a spectrum rather than falling into a strict binary of "assisted" versus "autonomous" operation. At the most conservative end, the system operates purely as a literature-synthesis and hypothesis-suggestion tool, with every downstream decision, including which hypothesis to pursue and how to design the corresponding experiment, made entirely by the human researcher. At an intermediate level, the system additionally drafts experimental protocols and estimates their expected outcome, but a human must explicitly approve each protocol before physical execution. At the most autonomous end, the system selects, executes, and interprets entire batches of experiments within a pre-approved envelope of reagents and conditions, surfacing only periodic summaries and any anomalies to

the human supervisor. Reported deployments increasingly allow the operating point along this spectrum to be configured per research campaign, reflecting the differing risk tolerance and regulatory context of different scientific domains; a low-risk materials-screening campaign may run with minimal human review, whereas a campaign involving controlled reagents or pathogenic organisms is typically configured to require approval at every step regardless of the system's demonstrated reliability in other contexts.

Automated Hypothesis Generation

Automated hypothesis generation is the capability that most distinguishes AI co-scientists from conventional laboratory automation or from earlier machine-learning-based experiment planners. Whereas a Bayesian-optimization-based self-driving laboratory searches a pre-defined parameter space for an optimum, a hypothesis-generating agent must first define what is worth searching for, drawing on background knowledge that is only partially formalized in structured databases.

A. Generation Strategies

Three broad strategies for hypothesis generation appear in the literature. Retrieval-grounded generation conditions the language model on passages retrieved from scientific literature relevant to the stated research goal, encouraging hypotheses that are consistent with, but extend beyond, existing findings. Analogical generation prompts the model to identify structurally similar problems in adjacent domains and transfer candidate mechanisms across domains, a strategy that has proven effective for drug-repurposing tasks where a mechanism validated in one disease context is proposed for a related one. Combinatorial generation systematically enumerates and scores combinations of building blocks, such as chemical substituents or gene-regulatory elements, using the language model to filter an otherwise intractably large combinatorial space down to a tractable candidate set.

B. Novelty and Feasibility Scoring

A hypothesis that merely restates prior literature has little scientific value, while a hypothesis untethered from established mechanisms is likely to be false or untestable. AI co-scientist systems therefore typically score each candidate along two largely orthogonal axes. Novelty is estimated by retrieving the nearest prior-art passages in embedding space and prompting a judge model to assess the degree of conceptual overlap, sometimes supplemented with citation-graph analysis to detect whether the proposed mechanism has already been reported under different terminology. Feasibility is estimated by checking the hypothesis against known physical, chemical, or biological constraints, cross-referencing required reagents or techniques against a laboratory's actual capabilities, and, where available, running lightweight computational simulations such as docking scores or density-functional-theory estimates as a cheap pre-filter before committing to physical experimentation.

C. Human-in-the-Loop Steering

Fully autonomous hypothesis generation without human oversight remains rare in deployed systems, both for scientific-quality reasons and for safety reasons discussed in Section VIII. Most systems instead expose an interface through which a human scientist can steer generation by supplying seed ideas, ruling out already-explored directions, adjusting the relative weighting of novelty versus feasibility, or directly editing the ranked hypothesis list before experimental commitment. This human-in-the-loop pattern has been shown to substantially improve the practical utility of generated hypotheses relative to fully autonomous operation, since domain experts can rapidly discard hypotheses that violate tacit knowledge not captured in the system's training data or retrieval corpus.

D. Handling Hallucination in Hypothesis Generation

A well-documented failure mode of LLM-based hypothesis generation is the fabrication of plausible-sounding but nonexistent citations, reaction pathways, or experimental precedents. Mitigations that have shown empirical benefit include mandatory citation verification against retrieved source documents before a hypothesis is surfaced to the ranking agent, self-consistency checks in which the same hypothesis is

regenerated multiple times and only claims that recur across samples are retained, and explicit uncertainty annotation in which the generation agent is required to flag which components of a hypothesis are directly supported by retrieved evidence versus which are extrapolated.

E. Cross-Disciplinary Hypothesis Transfer

Because large language models are trained on corpora spanning many scientific disciplines simultaneously, they are uniquely positioned to identify structural analogies between fields that a human specialist, trained narrowly within a single discipline, might overlook. Reported examples include the transfer of protein-folding intuitions to the design of self-assembling synthetic polymers, and the application of epidemiological compartment-modeling techniques to the analysis of chemical reaction-network dynamics. Realizing this potential systematically, rather than opportunistically, remains an open research problem, since the generation agent must be able to recognize when a surface-level analogy is also a valid mechanistic analogy, a distinction that current retrieval-grounding techniques do not reliably make.

F. Prompting and Fine-Tuning Strategies

Two complementary approaches are used to specialize general-purpose language models for scientific hypothesis generation. Prompt-based specialization supplies detailed domain context, worked examples of high-quality hypotheses, and explicit scoring rubrics within the model's context window at inference time, requiring no additional training but consuming context budget that competes with retrieved literature. Fine-tuning-based specialization instead trains the underlying model, or a lightweight adapter layered on top of it, on curated corpora of expert-authored hypotheses and their eventual experimental outcomes, allowing the model to internalize domain-specific heuristics at the cost of requiring substantial curated training data that is scarce in many specialized sub-fields. Hybrid approaches that fine-tune a relatively small specialized model for domain-specific scoring while leaving open-ended generation to a larger general-purpose model have shown a favorable balance between specialization and generalization in reported comparisons.

Automated Laboratory Experimentation

Translating a ranked hypothesis into a validated scientific finding requires physical experimentation. This section discusses how AI co-scientists interface with, and in some cases directly control, laboratory hardware.

A. Self-Driving Laboratory Hardware

Self-driving laboratories typically combine programmable liquid-handling robots, automated solid-dispensing systems, robotic arms for plate and vial transport, and integrated analytical instruments such as liquid chromatography, mass spectrometry, UV-visible spectroscopy, or X-ray diffraction, all coordinated by a central laboratory-execution controller. Standardized application-programming interfaces exposed by this controller allow an external planning agent to submit experiment requests as structured data — for example, a target composition, a set of reaction conditions, and a desired characterization protocol — without needing to directly control individual instrument drivers.

B. Closed-Loop Optimization

The classical closed-loop cycle in self-driving laboratories couples the experimental platform to a Bayesian-optimization or active-learning module that selects the next experiment to maximize expected information gain or expected improvement of a target property. AI co-scientists extend this cycle in two ways. First, the LLM-based planner can incorporate qualitative constraints and mechanistic reasoning that a purely numerical acquisition function cannot represent, for instance ruling out a chemically implausible reagent combination even if it scores well numerically. Second, the planner can dynamically switch between exploration strategies, for example shifting from broad exploration of a hypothesis space to focused local refinement once a promising region has been identified, using natural-language reasoning about the accumulating evidence rather than a fixed acquisition-function schedule.

C. Protocol Translation and Safety Interlocks

A critical and non-trivial engineering problem is the reliable translation of a natural-language experimental protocol generated by the evolution agent into the structured, machine-executable format required by laboratory hardware. Errors in this translation step can range from harmless failures, such as an experiment simply not running, to hazardous outcomes, such as incorrect reagent volumes or unsafe reaction conditions. Production systems therefore interpose a validation layer between the language-model planner and the physical hardware, which checks proposed protocols against a library of known-safe operating ranges, requires explicit human approval for any protocol involving reagents above a defined hazard threshold, and simulates the protocol in a digital twin of the laboratory before physical execution where such a simulation is available.

D. Result Interpretation and Belief Updating

Raw instrument output, such as a chromatogram or diffraction pattern, must be converted into a structured observation before it can update the system's beliefs about which hypotheses are supported. This typically involves a combination of classical signal-processing pipelines for peak detection or pattern matching and an LLM-based interpretation step that maps the extracted features back onto the original hypothesis, assessing whether the observed result confirms, partially confirms, or refutes the prediction, and updating the memory substrate accordingly so that subsequent generation rounds are conditioned on the updated evidence base.

E. Failure Detection and Recovery

Physical experiments fail for reasons unrelated to the underlying scientific hypothesis, including instrument malfunction, reagent degradation, and operator or robotic-execution error, and a robust AI co-scientist must distinguish these mundane failure modes from genuine scientific disconfirmation before updating its beliefs. Production systems typically apply anomaly-detection heuristics to instrument telemetry, such as unexpected pressure or temperature excursions during a synthesis step, and flag any run exhibiting such anomalies for automatic repetition rather than treating it as evidence against the tested hypothesis. Repeated unexplained failures across otherwise similar protocols are surfaced to the human supervisor as a signal of a possible systematic hardware issue rather than fed back into the hypothesis-ranking process.

F. Throughput and Batch Scheduling

Self-driving laboratories typically process experiments in batches determined by physical constraints such as well-plate geometry or instrument queue depth, and the planning agent must therefore select not a single next experiment but an entire batch that jointly maximizes expected information gain while respecting these constraints. This batch-selection problem is closely related to classical batch active-learning formulations, and several reported systems address it by having the ranking agent produce a diverse slate of top-scoring hypotheses rather than a single best hypothesis, explicitly penalizing redundancy among the selected experiments to avoid wasting a batch slot on near-duplicate conditions.

Case Studies

A. Materials Discovery

Inorganic materials discovery has been one of the most successful early application domains for AI co-scientists, owing to the relative maturity of self-driving laboratory hardware for solid-state synthesis and the availability of large structured materials databases for retrieval grounding. Reported campaigns have used LLM-based planners to propose candidate compositions for target properties such as ionic conductivity or photocatalytic activity, with the planner reasoning over crystal-structure databases and prior synthesis literature to propose compositions, and a robotic synthesis-and-characterization platform executing the resulting protocols within a closed loop lasting days rather than the months typical of manual campaigns.

B. Drug Repurposing

Drug repurposing — identifying new therapeutic applications for existing, already-approved compounds — is particularly well suited to AI co-scientist methods because it substitutes literature-grounded reasoning for expensive de novo compound synthesis, since candidate compounds are already characterized and often already approved for human use, substantially lowering both cost and translational risk. Reported systems have used multi-agent hypothesis generation to propose mechanistic links between existing drugs and disease pathways not previously associated with them, subsequently prioritized using retrieval-grounded plausibility scoring and, in some cases, validated with automated in vitro assay platforms.

C. Synthetic Biology and Pathway Design

In synthetic biology, AI co-scientists have been applied to the design of metabolic pathways for producing target molecules in engineered microorganisms, a task that requires reasoning over large combinatorial spaces of candidate enzymes and regulatory elements. Systems in this space typically combine an LLM-based generation agent, which proposes candidate pathway topologies grounded in enzyme-function databases, with automated cloning and cultivation robotics that construct and test the proposed strains, and with sequencing-based readouts that are interpreted by the reflection agent to determine which pathway variants achieved the target production titer.

D. Autonomous Solid-State Materials Synthesis

A particularly demanding case study is fully autonomous solid-state materials synthesis, in which a system must not only propose a target composition but also determine a viable synthesis route from a large space of precursor combinations, firing temperatures, and reaction times, any of which may fail to yield the intended phase. Reported autonomous laboratories addressing this problem combine a literature-mining module that extracts precedent synthesis routes for chemically similar compounds, a planning agent that proposes and ranks candidate routes for a novel target compound, and robotic powder-handling and furnace systems that execute the proposed routes, with X-ray diffraction used to automatically verify whether the intended phase was obtained. Reported success rates for synthesizing novel predicted compounds on the first attempt remain moderate, underscoring that synthesis-route planning is presently a harder sub-problem than target-composition proposal.

E. Cross-Case Observations

Across all four case studies, three recurring patterns are evident. First, systems that operate over domains with rich, well-structured prior databases, such as inorganic crystal structures or approved-drug registries, substantially outperform those operating in sparsely documented domains, confirming that retrieval-grounding quality is presently a stronger determinant of system performance than the sophistication of the underlying generation model. Second, closed-loop success rates improve markedly when the human supervisor remains in the loop for protocol approval even in otherwise autonomous campaigns, consistent with the human-in-the-loop findings discussed in Section IV. Third, in every case study reviewed, the majority of wall-clock campaign time was consumed by physical experimentation and instrument turnaround rather than by language-model inference, suggesting that near-term performance gains are more likely to come from laboratory-hardware throughput improvements than from further scaling of the reasoning model alone.

F. Shared Infrastructure Across Case Studies

A notable trend across the four case studies is the gradual convergence of previously domain-specific automation stacks toward shared, reusable infrastructure. Liquid-handling robots originally deployed for high-throughput drug screening are increasingly repurposed for synthetic-biology strain construction, and literature-mining pipelines originally built for materials discovery have been adapted with comparatively modest modification to mine the pharmacological literature for drug-repurposing campaigns. This convergence suggests that future gains in AI co-scientist capability may increasingly come from improvements to shared, domain-agnostic infrastructure components, such as retrieval-grounding quality

and robotic-execution reliability, rather than from bespoke domain-specific engineering, echoing a broader pattern in machine learning in which general-purpose foundation capabilities have progressively displaced narrow, hand-engineered pipelines.

Evaluation Metrics and Benchmarks

Rigorous evaluation of AI co-scientist systems is complicated by the fact that the ultimate output — a validated scientific discovery — may take weeks or months to materialize and is difficult to compare across research domains. The community has converged on a small set of complementary metrics, summarized in Table 2.

Table 2: EVALUATION METRICS FOR AI CO-SCIENTIST SYSTEMS

Metric	Definition	Typical Measurement
Hypothesis novelty	Degree of dissimilarity from prior literature	Embedding distance / expert rating
Feasibility rate	Fraction of hypotheses judged experimentally executable	Expert or automated audit
Experimental success rate	Fraction of tested hypotheses confirmed	Closed-loop outcome logs
Cost per validated finding	Total compute + reagent + instrument cost	Campaign accounting
Time-to-result	Wall-clock time from goal statement to validated result	Campaign timestamps
Reproducibility	Fraction of results replicated by independent run	Replication studies

Beyond these campaign-level metrics, several benchmark suites have emerged that evaluate individual pipeline components in isolation, such as literature-grounded question answering, retrosynthesis-planning accuracy, and protocol-generation correctness, allowing controlled comparison of generation and reflection agents independent of the cost and variability of full physical experimentation. Human expert review remains the gold standard for assessing hypothesis quality, though its cost has motivated growing interest in using calibrated LLM judges as a scalable proxy, provided such judges are themselves validated against expert panels on a representative sample of outputs.

A. Component-Level versus Campaign-Level Evaluation

A persistent tension in the evaluation literature is the trade-off between component-level benchmarks, which are cheap to run and enable rapid iteration but may not predict end-to-end campaign performance, and full campaign-level evaluation, which is the only method that directly measures scientific value but is prohibitively expensive to run at the scale needed for statistically meaningful comparison between competing system designs. Several recent benchmark efforts attempt to bridge this gap using retrospective evaluation, in which a system is tasked with rediscovering a scientific finding that is already known to the evaluators but was withheld from the system's training and retrieval corpus, providing a ground-truth-verifiable proxy for genuine discovery without requiring new physical experimentation for every evaluated system variant.

B. Statistical Considerations

Because physical experimentation is costly, campaign-level evaluations are frequently conducted with small sample sizes, and reported success rates should be interpreted with corresponding statistical caution. Several authors have called for standardized reporting of confidence intervals and pre-registered evaluation protocols for AI co-scientist campaigns, mirroring analogous reforms in the broader experimental sciences, to reduce the risk that reported performance figures reflect favorable cherry-picking of campaign outcomes rather than a representative estimate of system capability.

Challenges and Limitations

A. Reproducibility of Autonomously Generated Protocols

Protocols generated by an evolution agent are conditioned on the specific laboratory hardware, reagent lot, and environmental conditions of the originating laboratory, and often omit tacit procedural details that a human experimentalist would supply automatically. Consequently, autonomously generated protocols have been observed to transfer poorly across laboratories without manual adaptation, echoing a longstanding reproducibility challenge in experimental science that automation does not automatically resolve and may in some cases exacerbate if under-specified steps are executed literally rather than with expert judgment.

B. Safety and Dual-Use Concerns

Systems capable of autonomously proposing and executing chemical or biological syntheses raise dual-use concerns, since the same reasoning and planning capabilities that enable beneficial discovery could, without appropriate safeguards, be misapplied to the design of hazardous substances. Responsible deployments incorporate hazard-screening layers that check generated hypotheses and protocols against controlled-substance and biosecurity watchlists before any physical execution is permitted, restrict autonomous execution to pre-approved reagent and procedure classes, and require human sign-off for any protocol that falls outside those pre-approved bounds. The AI-safety and biosecurity communities have called for standardized, independently audited safeguards of this kind across all self-driving laboratory deployments, rather than ad hoc laboratory-specific policies.

C. Epistemic Trust and Hallucination

Because language models can generate fluent and superficially rigorous text regardless of its truth value, there is a persistent risk that fabricated mechanistic reasoning or non-existent literature citations are accepted into a research campaign's evidence base, subtly biasing subsequent generation and ranking rounds. This risk is compounded by automation bias, the tendency of human overseers to grant undue credence to system outputs precisely because they are produced by an ostensibly rigorous automated pipeline. Maintaining a clear audit trail linking every generated claim back to its retrieved source evidence, and periodically subjecting the system's claims to independent human expert review, are presently the most effective mitigations.

D. Cost and Access Inequality

Self-driving laboratory hardware and the compute required to operate large multi-agent language-model pipelines at scale remain expensive, raising concerns that the benefits of AI-accelerated discovery may accrue disproportionately to well-resourced institutions, potentially widening existing disparities in scientific capacity between institutions and between countries. Cloud-accessible shared self-driving laboratory facilities and open-source hypothesis-generation frameworks have been proposed as partial mitigations, though their long-term sustainability and equitable governance remain open questions.

E. Evaluation of Genuine Novelty

Distinguishing a genuinely novel scientific contribution from a superficial recombination of known results remains difficult even for expert human reviewers, and current automated novelty-scoring methods based on embedding-space distance are known to be an imperfect proxy for the judgment a domain expert would

apply, since two hypotheses can be textually dissimilar yet mechanistically equivalent, or textually similar yet mechanistically distinct.

F. Attribution and Scientific Credit

As AI co-scientists take on a larger share of hypothesis generation and experimental design, existing norms for assigning scientific credit and authorship become strained, since standard authorship criteria presuppose a human contributor capable of taking accountability for the validity of the reported work. Several journals and professional societies have issued interim policies requiring that any AI-assisted contribution be explicitly disclosed and that a named human author retain full accountability for the accuracy and integrity of any resulting publication, but consensus on more granular attribution norms, for instance how to credit a specific hypothesis traceable to a particular agent configuration, has not yet emerged and is likely to remain contested as autonomous contributions become more substantial.

G. Integration with Existing Laboratory Workflows

A frequently underestimated barrier to adoption is the practical difficulty of integrating AI co-scientist pipelines with the heterogeneous mix of legacy instruments, laboratory-information-management systems, and manual record-keeping practices already in place in most research laboratories. Unlike purpose-built self-driving laboratories designed from the outset for automation, most existing laboratories were not engineered with machine-readable interfaces, and substantial engineering investment is typically required to retrofit instrument connectivity before any closed-loop AI co-scientist capability can be deployed, a cost that is rarely reflected in reported campaign-level performance figures and that disproportionately affects the access-inequality concern discussed above.

Future Directions

A. Federated Multi-Laboratory Orchestration

A natural extension of current single-laboratory closed-loop systems is federated orchestration across multiple geographically distributed self-driving laboratories, allowing a single hypothesis-generation campaign to route different candidate experiments to whichever facility has the appropriate instrumentation, while aggregating results into a shared memory substrate. Realizing this vision will require standardized experiment-description schemas and inter-laboratory communication protocols that do not yet exist in mature form.

B. Formal Verification of Experimental Plans

Increasing reliance on autonomously generated protocols motivates research into formal or semi-formal verification methods that can certify a generated protocol against a machine-checkable specification of safety and physical-plausibility constraints prior to execution, analogous to program verification in software engineering, potentially reducing dependence on manual protocol review as campaign throughput scales.

C. Tighter Coupling of Symbolic and Sub-Symbolic Reasoning

Current systems largely rely on the emergent reasoning capabilities of general-purpose language models, which remain weak at rigorous causal and quantitative reasoning relative to symbolic scientific-modeling tools. Closer integration between LLM-based agents and specialized symbolic solvers, such as reaction-network simulators, molecular-dynamics engines, and constraint-based metabolic models, is likely to improve both the feasibility and the mechanistic correctness of generated hypotheses.

D. Standardized Benchmarks and Shared Infrastructure

The field would benefit from community-agreed benchmark suites analogous to those that accelerated progress in other areas of machine learning, covering hypothesis-generation quality, protocol-translation correctness, and end-to-end closed-loop discovery rate on shared, openly available experimental platforms, enabling apples-to-apples comparison across research groups and reducing duplicated infrastructure investment.

E. Governance and Standardized Safeguards

As autonomous experimentation capabilities scale, the community is likely to require formal governance frameworks, potentially including third-party auditing of hazard-screening layers, mandatory incident reporting for near-miss safety events in self-driving laboratories, and international coordination on dual-use risk analogous to existing biosecurity frameworks for human-conducted research.

F. Improved Uncertainty Quantification

Current hypothesis-ranking mechanisms largely produce point estimates or relative orderings without well-calibrated uncertainty bounds, limiting the ability of downstream decision-making, whether human or automated, to appropriately weigh a highly novel but uncertain hypothesis against a modest but well-supported one. Adapting calibrated uncertainty-quantification techniques from the broader machine-learning literature, such as conformal prediction, to the specific setting of natural-language hypothesis scoring is a promising and largely open direction that could materially improve the reliability of automated experiment prioritization.

G. Longitudinal Learning Across Campaigns

Most deployed systems treat each research campaign largely in isolation, with limited mechanisms for transferring lessons learned, such as which classes of hypotheses tend to fail experimentally, across campaigns or across laboratories. Developing shared, privacy-respecting mechanisms for aggregating this experiential knowledge across many independent AI co-scientist deployments, without requiring laboratories to disclose commercially sensitive raw data, could substantially accelerate the field's collective rate of improvement and reduce redundant rediscovery of the same negative results across independent research groups.

Practical Implementation Considerations

Beyond the algorithmic and architectural questions discussed above, practitioners deploying AI co-scientist systems in production research settings face a number of practical engineering considerations that are often underemphasized in the academic literature but materially affect real-world outcomes.

A. Latency and Cost Budgeting

Multi-agent pipelines involving repeated generation, critique, and tournament-style ranking can require a large number of language-model inference calls per research-campaign iteration, and this cost scales further when retrieval-augmented grounding requires querying large literature or structured-database indices for every candidate hypothesis. Production deployments typically impose an explicit compute budget per campaign iteration, using cheaper, smaller models for high-volume filtering stages such as initial feasibility screening and reserving the most capable and expensive models for final-stage ranking and protocol elaboration, where reasoning quality matters most and candidate volume is lowest.

B. Interface Design for Human Supervisors

Because most deployed systems retain a human supervisor in the loop, the design of the interface through which that supervisor reviews generated hypotheses, approves protocols, and interprets results has an outsized effect on overall system usability and safety. Effective interfaces surface not only the top-ranked hypotheses but also the supporting evidence and critique that led to their ranking, allowing the supervisor to efficiently audit the system's reasoning rather than being forced to either blindly trust or exhaustively re-derive each recommendation, a design tension directly analogous to explainable-AI challenges in other high-stakes decision-support domains.

C. Data Provenance and Audit Logging

Given the epistemic-trust and safety concerns discussed in Section VIII, comprehensive audit logging of every generation, critique, ranking, and protocol-execution step is considered a baseline requirement for responsible deployment rather than an optional feature. Such logs enable retrospective investigation of any anomalous or unsafe outcome, support the reproducibility efforts discussed above, and provide the

evidentiary basis needed to satisfy the attribution and accountability norms that journals and institutions are beginning to require of AI-assisted research.

D. Vendor and Platform Lock-In

Many current self-driving-laboratory and AI co-scientist deployments are built around proprietary hardware-control software or proprietary language-model application-programming interfaces, raising the practical risk that a research group's accumulated protocols, prompts, and memory substrate become difficult to migrate to alternative platforms as the underlying technology evolves. Open standards for experiment description and agent orchestration, analogous to open file formats in other areas of scientific computing, would reduce this lock-in risk and are an area of active community discussion, though no dominant open standard has yet emerged.

Conclusion

AI co-scientist systems represent a substantive evolution of laboratory automation, extending the closed-loop optimization paradigm of self-driving laboratories with language-model-based reasoning capable of literature-grounded hypothesis generation and qualitative mechanistic judgment. We have described a common four-role architecture — generation, reflection, ranking, and evolution — that recurs across deployed systems, examined the specific algorithmic challenges of hypothesis generation and closed-loop experimentation in depth, and illustrated practical performance through case studies in materials discovery, drug repurposing, and synthetic biology. At the same time, significant challenges remain, spanning reproducibility of autonomously generated protocols, dual-use safety risk, hallucination and epistemic trust, and equitable access to the underlying infrastructure. We conclude that while AI co-scientists have already demonstrably accelerated narrow, well-scoped discovery tasks, realizing the broader vision of reliable end-to-end autonomous science will require continued advances in causal and quantitative reasoning, formal safety verification, and community-wide standards for evaluation and governance.

Acknowledgment

The authors thank colleagues in the automated-science and laboratory-robotics communities for helpful discussions that informed this survey.

References

[1] B. Burger, P. M. Maffettone, V. V. Gusev, et al., "A mobile robotic chemist," Nature, vol. 583, pp. 237–241, 2020.
[2] B. P. MacLeod, F. G. L. Parlane, T. D. Morrissey, et al., "Self-driving laboratory for accelerated discovery of thin-film materials," Science Advances, vol. 6, no. 20, 2020.
[3] R. Roscher, B. Bohn, M. F. Duarte, and J. Garcke, "Explainable machine learning for scientific insights and discoveries," IEEE Access, vol. 8, pp. 42200–42216, 2020.
[4] S. Szymkuc, E. P. Gajewska, T. Klucznik, et al., "Computer-assisted synthetic planning: The end of the beginning," Angewandte Chemie International Edition, vol. 55, no. 20, pp. 5904–5937, 2016.
[5] C. W. Coley, D. A. Thomas III, J. A. M. Lummiss, et al., "A robotic platform for flow synthesis of organic compounds informed by AI planning," Science, vol. 365, no. 6453, 2019.
[6] P. S. Gromski, A. B. Henson, J. M. Granda, and L. Cronin, "How to explore chemical space using algorithms and automation," Nature Reviews Chemistry, vol. 3, pp. 119–128, 2019.
[7] J. M. Granda, L. Donina, V. Dragone, D.-L. Long, and L. Cronin, "Controlling an organic synthesis robot with machine learning to search for new reactivity," Nature, vol. 559, pp. 377–381, 2018.
[8] A. Aspuru-Guzik and K. Persson, "Materials Acceleration Platform: Accelerating advanced energy materials discovery by integrating high-throughput methods and artificial intelligence," Mission Innovation Report, 2018.

[9] T. Lu, S. Chen, et al., "Retrieval-augmented generation for scientific literature synthesis," arXiv preprint, 2022.
[10] E. Kim, K. Huang, A. Saunders, A. McCallum, G. Ceder, and E. Olivetti, "Materials synthesis insights from scientific literature via text extraction and machine learning," Chemistry of Materials, vol. 29, no. 21, pp. 9436–9444, 2017.
[11] N. J. Szymanski, B. Rendy, Y. Fei, et al., "An autonomous laboratory for the accelerated synthesis of novel materials," Nature, vol. 624, pp. 86–91, 2021.
[12] S. M. Moosavi, K. M. Jablonka, and B. Smit, "The role of machine learning in the understanding and design of materials," Journal of the American Chemical Society, vol. 142, no. 48, pp. 20273–20287, 2020.
[13] World Health Organization and partner biosecurity working groups, "Responsible AI in the life sciences: Emerging governance considerations," Policy brief, 2021.
[14] K. Darvish, M. Skreta, Y. Zhao, et al., "ORGANA: A robotic assistant for automated chemistry experimentation," arXiv preprint, 2021.
[15] S. Kim, Y. Chen, et al., "Benchmarking large language models for autonomous scientific discovery," arXiv preprint, 2022.
[16] L. Cronin, "Digitizing chemistry with the chemputer," Nature Reviews Chemistry, vol. 5, pp. 447–448, 2021.