Preprints.org

Article

Not peer-reviewed version

Agentic Laboratories of the Future: Towards World Models for Scientific Discovery

Kelsey Fu , Luna Lyu , Shuzhen Li , Suozhi Huang , Kaize Wu , Suheng Xu , Joy Zheng , Xuefeng Liu , Shilong Liu , Jerry Dickerson , Maxie Wu , Greg Barbone , Ying Liu , Linxi "Jim" Fan , Yuke Zhu , Matthew Davis , Kaiyi Jiang , Sherry Yang , Xu Kuang , Yuan Luo , Francesca Dominici , Christina Curtis , Xiaojie Qiu , Alberto Abadie , Caroline Uhler , Mary Tang , Joseph C. Wu , Ju Li , Jared Toettcher , Jose L. Avalos , Rebecca Rojansky , Huimin Zhao , John P. A. Ioannidis , Barry P. Rand , Emma Lundberg , Andrew S. Rosen , Zhuang Liu , Shana Kelley , Chenyan Xiong , Barbara E. Engelhardt , Thomas Montine , Christina V. Theodoris , Katherine S. Pollard , Fuxin Li , Xueyue Zhang , Tianyi Peng , Dmitri Basov , Xiang Qian , Eric Xing , Ali Yazdani , Jure Leskovec , Nigam H. Shah , Jimmy Ba , Daniel J. Cohen , Jonathan Conway , Yibin Kang , Ryan Adams , Cathy Wu , Joshua Atkinson , Xianwen Mao , Emily Carter , Chuanxia Zheng , Atlas Wang , Christina Kim , Yonggang Wen , Yu Su , Hardeep Sangha , Ellie Hong , Alexander Choi , Rajesh Raina , Hani Goodarzi , Polly M. Fordyce , Lars Steinmetz , Jonathan T.C. Liu , Eric Yang , Ansuman T. Satpathy , Howard Y. Chang , Seth Shipman , Xi Yin , Mark A. Musen , Zhenan Bao , Pietro Perona , Sanfeng Wu , Clifford Brangwynne , Le Cong * , Mengdi Wang *

Posted Date: 3 November 2026

Keywords: agentic laboratory; self-driving laboratory; AI agents for scientific discovery; world models; laboratory automation; human–AI–robot collaboration; laboratory automation; scientific reproducibility; physical AI

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC, OpenAlex.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org | NOT PEER-REVIEWED | Posted: 3 November 2026 https://doi.org/10.20944/preprints202608.0213.v2

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Agentic Laboratories of the Future: Towards World Models for Scientific Discovery

Kelsey Fu [1], Luna Lyu [2], Shuzhen Li [1], Suozhi Huang [1], Kaize Wu [28], Suheng Xu [17], Joy Zheng [1], Xuefeng Liu [2], Shilong Liu [1], Jerry Dickerson, Maxie Wu, Greg Barbone [5], Ying Liu [18], Linxi "Jim" Fan [5], Yuke Zhu [5,19], Matthew Davis [8], Kaiyi Jiang [1], Sherry Yang [10], Xu Kuang [2], Yuan Luo [7], Francesca Dominici [4], Christina Curtis [2], Xiaojie Qiu [2], Alberto Abadie [3], Caroline Uhler [3], Mary Tang [2], Joseph C. Wu [2], Ju Li [3], Jared Toettcher [1], Jose L. Avalos [1], Rebecca Rojansky [2], Huimin Zhao [12], John P. A. Ioannidis [2], Barry P. Rand [1], Emma Lundberg [2], Andrew S. Rosen [1], Zhuang Liu [1], Shana Kelley [6,7], Chenyan Xiong [1], Barbara E Engelhardt [2], Thomas Montine [2], Christina V. Theodoris [14,15], Katherine S. Pollard [6,14,15], Fuxin Li [16], Xueyue Zhang [17], Tianyi Peng [17], Dmitri Basov [17], Xiang Qian [2], Eric Xing [9,11], Ali Yazdani [1], Jure Leskovec [2], Nigam H. Shah [2], Jimmy Ba [22], Daniel J. Cohen [1], Jonathan Conway [1], Yibin Kang [1], Ryan Adams [1], Cathy Wu [23], Joshua Atkinson[1], Xianwen Mao [24], Emily Carter[1], Chuanxia Zheng [25], Atlas Wang [19], Christina Kim [1], Yonggang Wen [25], Yu Su [16], Hardeep Sangha [20], Ellie Hong [20], Alexander Choi [20], Rajesh Raina [20],Hani Goodarzi [14,26], Polly M Fordyce [2], Lars Steinmetz [2], Jonathan T.C. Liu [2], Eric Yang [2], Ansuman T. Satpathy [2], Howard Y. Chang [2,27], Seth Shipman [6,14,15], Xi Yin [4,21], Mark A Musen [2], Zhenan Bao [2], Pietro Perona [13], Sanfeng Wu [1], Clifford Brangwynne [1], Le Cong [2,*] and Mengdi Wang [1,*]

[1] Princeton University
[2] Stanford University
[3] Massachusetts Institute of Technology
[4] Harvard University
[5] NVIDIA
[6] Chan Zuckerberg Biohub
[7] Northwestern University
[8] Flagship Pioneering
[9] Carnegie Mellon University
[10] New York University
[11] Mohamed Bin Zayed University of Artificial Intelligence
[12] University of Illinois Urbana-Champaign
[13] California Institute of Technology
[14] University of California, San Francisco
[15] Gladstone Institute
[16] Oregon State University
[17] Columbia University
[18] Scale AI
[19] The University of Texas at Austin
[20] Thermo Fisher Scientific
[21] OpenAI
[22] University of Toronto
[23] University of Delaware
[24] National University of Singapore
[25] Nanyang Technological University
[26] Arc Institute
[27] Amgen
[28] University of Chicago
* Correspondence: congle@stanford.edu (L.C.); mengdiw@princeton.edu (M.W.)

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 3 November 2026 https://doi.org/10.20944/preprints202608.0213.v2

Abstract

Advances in scientific AI have accelerated hypothesis generation, prediction, and experimental design, but translating these capabilities into reliable physical experimentation remains a major bottleneck in scientific discovery. Although automation, high-throughput experimentation, foundation models, and cloud infrastructure have accelerated individual stages of the scientific workflow, they have not unified the discovery process. Scientists must still coordinate between computational models and physical laboratory systems to translate scientific designs into executable experiments and ensure their results are reliable. We argue that the next generation of laboratories will be agentic: environments in which scientists, AI systems, equipment and robotic platforms operate as collaborative discovery partners. We propose an architecture built around two complementary components: an agentic foundation that translates scientific designs into coordinated execution plans, and an evolving laboratory world model that represents physical state, predicts action outcomes, and incorporates experimental feedback. Together, they connect computational reasoning to observable and verifiable physical execution. Multimodal perception, robotics, and human–AI interfaces can make laboratory actions and evolving experimental conditions accessible to these systems, supporting adaptive execution and learning. We further propose an evaluation framework centered on physical grounding, execution fidelity, scientific validity, and adaptive coordination under human oversight, rather than autonomy alone. This Perspective outlines how agentic laboratories could connect scientific reasoning, physical experimentation, and trustworthy evidence within a human-in-the-lead discovery loop.

Keywords: agentic laboratory; self-driving laboratory; AI agents for scientific discovery; world models; laboratory automation; human–AI–robot collaboration; laboratory automation; scientific reproducibility; physical AI

1. Introduction

Scientific discovery is an iterative process of asking questions, reasoning over existing evidence, acting to generate new evidence, and updating our understanding of the world. Although computational models increasingly accelerate hypothesis generation and experimental design, scientific progress ultimately depends on testing these ideas through physical experiments. Translating computational predictions into reliable experimental evidence remains difficult, particularly when procedures involve heterogeneous instruments, variable samples, and tacit knowledge.

Automation has made experimental execution increasingly programmable, while advances in AI have expanded scientific reasoning, prediction, design and analysis. [1–4,7,8] Yet these capabilities remain fragmented: models can reason but often cannot act in the physical laboratory, automated instruments can execute but typically operate within predefined workflows, and scientists still bridge the two by translating goals into protocols, transferring information across systems, responding to failures, and deciding what to do next. [3,4,35,36] Much of this coordination depends on laboratory context that is not captured in protocols or instrument outputs, including sample conditions, procedural adjustments, and observations made during execution.[40,51]

We refer to systems that bridge this divide as agentic laboratories: human–AI–robot systems that connect scientific reasoning, evolving experimental context, and physical execution within an adaptive discovery loop (Figure 1). Their purpose is to make the transitions between scientific design, physical execution, verification, and learning increasingly machine-actionable, while keeping scientists responsible for scientific direction and consequential decisions.[17,35]

Across chemistry, materials science, biology, and physics, increasingly capable self-driving laboratories, scientific AI agents, and robotic systems demonstrate pieces of this vision, but largely

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 3 November 2026 https://doi.org/10.20944/preprints202608.0213.v2

within bounded objectives, predefined interfaces, or individual stages of discovery. [5,11,20,23,35,76–82]

In each case, progress stalls at the same boundary: models can predict and reason but often cannot act, physical instruments can act but often cannot reason or adapt, and neither can reliably integrate heterogeneous tools, data streams, and human judgment under uncertainty. [3,4,35]

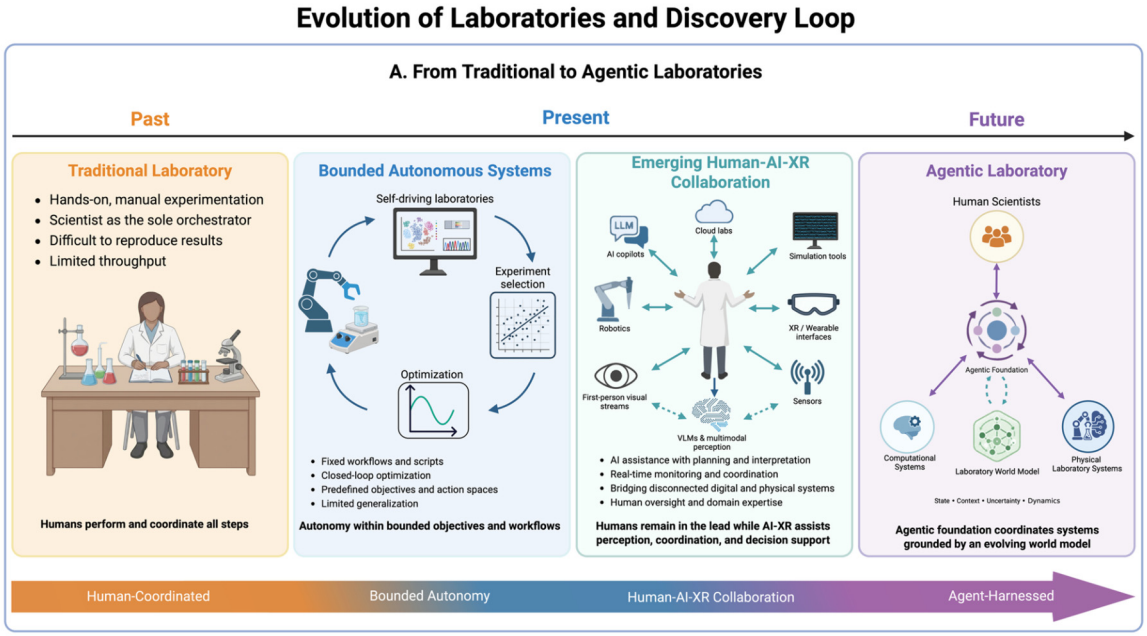


Figure 1. Evolution of laboratories toward agentic scientific discovery. Traditional laboratories rely on scientists to perform and coordinate experimental work. Present-day systems automate predefined workflows, while emerging AI–XR and multimodal technologies enable greater human–AI–robot collaboration and laboratory-state awareness. However, scientific reasoning and physical execution remain fragmented. Future agentic laboratories aim to bridge this gap through an agentic foundation that coordinates humans, computational models, and physical systems, grounded in an evolving laboratory world model that represents experimental state, context, uncertainty, and dynamics to support adaptive execution, verification, and learning.

We argue that bridging this boundary requires two foundations (Figure 1): the first is an agentic foundation that coordinates human scientists, computational systems, and physical laboratory systems across the discovery loop. The second is an evolving laboratory world model: a shared, machine-readable representation of experimental state, context, uncertainty, and dynamics that grounds reasoning and action in the physical and scientific system. The foundation translates scientific designs into executable workflows, while the world model supports state-aware planning, prediction, and verification. Together, they provide a technical basis for connecting computational reasoning to physical experimentation without removing human scientific oversight.[17,35,40]

In this Perspective, we trace the evolution from laboratory automation and self-driving laboratories toward agentic scientific discovery, develop this two-foundation architecture, propose a framework for evaluating whether agentic laboratory designs produce reliable and scientifically valid experimental evidence, and examine the technical, infrastructural, reproducibility, governance, and human challenges that must be addressed as these systems mature.

2. Evolution of Scientific Automation Toward Agentic Laboratories

Scientific automation has transformed fields ranging from high-energy physics, genome biology and neuroscience to space exploration, reflecting a broader shift toward machine-assisted scientific discovery across disciplines. Here, we focus on chemistry, biology, materials science, and quantum

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 3 November 2026 https://doi.org/10.20944/preprints202608.0213.v2

laboratories, where successive advances have made experimental execution, measurement, prediction, reasoning, and physical perception increasingly machine-actionable, while exposing recurring bottlenecks that motivate the transition toward agentic laboratories (Figure 2). Appendix A provides a detailed historical account of this progression across scientific disciplines, highlighting the methods, representative systems, achievements, and limitations underlying each transition.

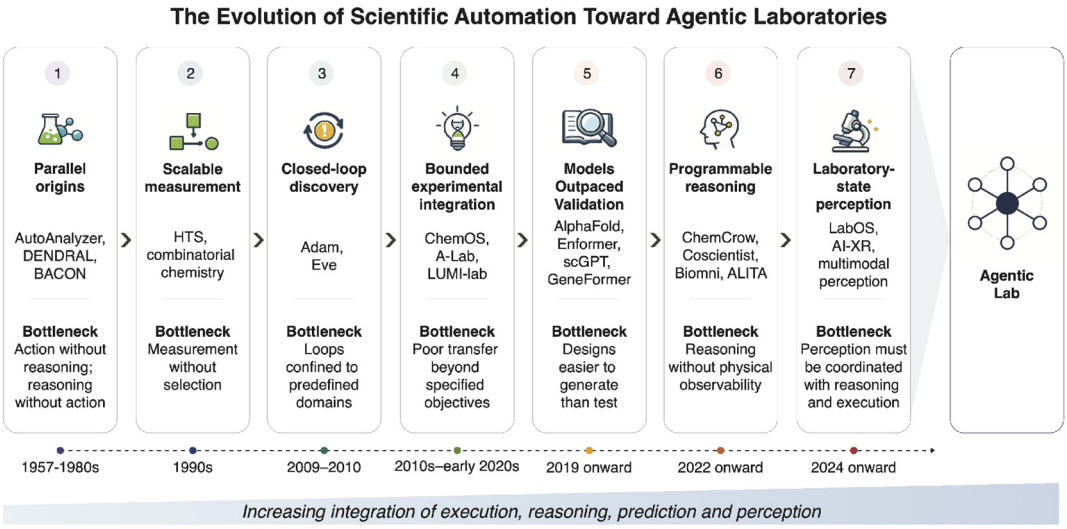


Figure 2. The evolution of scientific automation toward agentic laboratories. Scientific automation has progressed through successive advances that made individual components of laboratory work increasingly programmable while exposing new bottlenecks. Early laboratory automation and symbolic AI separated execution from reasoning; high-throughput screening scaled measurements but left experiment selection to humans; robot-scientist systems closed hypothesis–experiment loops within predefined domains; and self-driving laboratories integrated algorithms, instruments and workflows within bounded objectives. Later predictive and generative models expanded in silico design faster than validation capacity. More recently, LLM-based scientific agents made reasoning and tool use more programmable but remained weakly grounded in physical laboratory state. Emerging AI-XR and multimodal perception systems now begin to expose instruments, samples, human actions and procedural deviations as machine-readable context. Together, these stages point toward agentic laboratories, which integrate execution, reasoning, model outputs, perception, provenance and human judgment into a shared discovery loop.

2.1. From Automation to Closed-Loop Experimentation

Early laboratory automation and artificial intelligence developed complementary capabilities: automated instruments made predefined experimental procedures scalable and repeatable, while symbolic AI demonstrated computational reasoning over structured scientific representations.[63–66] For example, automated analyzers standardized sample processing and measurement, whereas DENDRAL used encoded chemical knowledge and heuristic search to infer candidate molecular structures from analytical data.[63,64] High-throughput screening and combinatorial experimentation subsequently expanded measurement capacity through robotic handling and parallel assays, but experiment selection, interpretation, and adaptation remained largely human-directed.[2,6,67] These developments exposed an early separation between execution and reasoning: machines could increasingly perform experiments or reason over representations, but could not reliably connect the two.

Robot scientists such as Adam and Eve began closing this gap by linking hypotheses, experiment selection, execution, and observations within predefined domains.[31,32] Adam used metabolic models and biological databases to propose gene–enzyme associations, then designed and executed

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 3 November 2026 https://doi.org/10.20944/preprints202608.0213.v2

yeast growth experiments to test them. Eve applied related closed-loop principles to drug screening, connecting candidate selection with robotic assays and experimental feedback.[31,32] Self-driving laboratories (SDLs) extended this paradigm by combining optimization algorithms, software orchestration, and automated instruments into iterative experimental loops.[5,11,12,23] Methods such as Bayesian optimization and active learning selected subsequent experiments based on previous measurements, enabling adaptive search within defined experimental spaces.[72,73] Such systems demonstrated increasingly autonomous experimentation across chemistry, materials science, biology, and physics, but generally remained bounded by predefined objectives, action spaces, workflows, and hardware interfaces.[5,9,35]

2.2. From Computational Models to Scientific Agents

Advances in machine learning accelerated an existing imbalance between hypothesis generation and experimental validation. Protein models, genomic and single-cell foundation models, and materials-prediction systems expanded computational prediction and design. For example, AlphaFold enabled protein-structure prediction from sequence, while genomic models learned relationships between sequence and gene regulation; materials models expanded the search for candidate structures before physical synthesis.[7,8,46,69–71,91] Meanwhile, experimental validation remained constrained by throughput, variable samples, and difficult-to-automate workflows.[19,20] As computational design accelerated, the validation bottleneck grew, making the translation of model outputs into reliable physical experiments increasingly important.

LLM-based scientific agents subsequently made reasoning and tool use increasingly programmable. Systems including ChemCrow, Coscientist, CRISPR-GPT, Google's Co-Scientist, Biomni, and related frameworks can decompose goals, invoke tools, retrieve evidence, critique outputs, and construct experimental plans.[13,24–29,47] ChemCrow connects language-model reasoning to chemistry tools, while Coscientist demonstrated the translation of natural-language research objectives into tool-mediated experimental workflows.[24,27] Yet most scientific agents remain connected to laboratories primarily through structured data, APIs, or instrument outputs.[24,27,35] They can reason about experiments without necessarily observing the physical state that determines whether these experiments are feasible, correctly executed, or scientifically interpretable.

2.3. Multimodal AI and XR Bridge Reasoning and Physical Laboratory State

Multimodal AI and extended reality (XR) provide an emerging bridge between digital reasoning and physical experimentation by making laboratory state accessible to AI systems. Experimental context is distributed not only across protocols and instrument outputs, but also across samples, objects, spatial relationships, procedural progress, human actions, environmental conditions, and unexpected deviations.[35,36,40,51] Visual, audio, sensor, and first-person data can expose this context to multimodal models, while XR interfaces can connect AI assistance to scientists as experiments unfold. In practice, fixed cameras and wearable egocentric cameras capture complementary views of laboratory activities, which can be combined with instrument telemetry and environmental sensors to contextualize physical actions and procedural changes.[40,51]

Systems such as LabOS illustrate how AI-XR can provide agents with first-person visual context, procedural information, and human corrections during physical laboratory work, supporting real-time guidance and documentation.[40] Embodied systems such as Qumus similarly combine multimodal reasoning with robotic execution and closed-loop error correction, while related systems demonstrate increasingly autonomous perception and control of scientific instruments.[54,93,94]Together, multimodal AI, XR, and embodied perception shift laboratory automation from machine-readable procedures toward machine-readable laboratory state, providing the physical grounding needed to connect reasoning with action.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 3 November 2026 https://doi.org/10.20944/preprints202608.0213.v2

2.4. Towards Agentic Laboratories

These developments reveal a recurring progression (Figure 2). Early automation made execution programmable; high-throughput systems scaled measurement; robot scientists and SDLs connected selection and execution within bounded loops; foundation models expanded prediction and design; scientific agents made reasoning and tool use programmable; and multimodal AI and XR increasingly make physical laboratory state accessible to computational systems.[1,2,5,11,24,31,32,40] Yet these capabilities remain fragmented, and physical execution still lacks the state awareness and predictive grounding needed to reliably translate scientific plans into real-world actions.

The remaining challenge is both integration and physical grounding. Agentic laboratories must connect scientific reasoning with experimental execution while maintaining awareness of laboratory state, predicting action outcomes, and adapting to physical feedback. This motivates the two complementary components developed in Section 3: an agentic foundation for coordinating scientific and physical workflows, and an evolving laboratory world model for grounding execution in physical state and dynamics.

3. The Agentic Laboratory Stack: Connecting Scientific Reasoning to Physical Discovery

The agentic laboratory stack comprises four interconnected components: cognitive intelligence, an agentic foundation, an evolving laboratory world model, and physical intelligence (Figure 3). Cognitive intelligence encompasses computational models and AI agents that support scientific reasoning, hypothesis generation, and experimental design, while physical intelligence encompasses robotic systems, instruments, and perception capabilities that enable physical experimentation and observation. The agentic foundation connects them by translating scientific designs into coordinated execution plans, supported by a world model that estimates laboratory state, predicts action outcomes, and incorporates experimental feedback. Together, these components support adaptive, verifiable experimentation under human scientific leadership.[17,35,40]

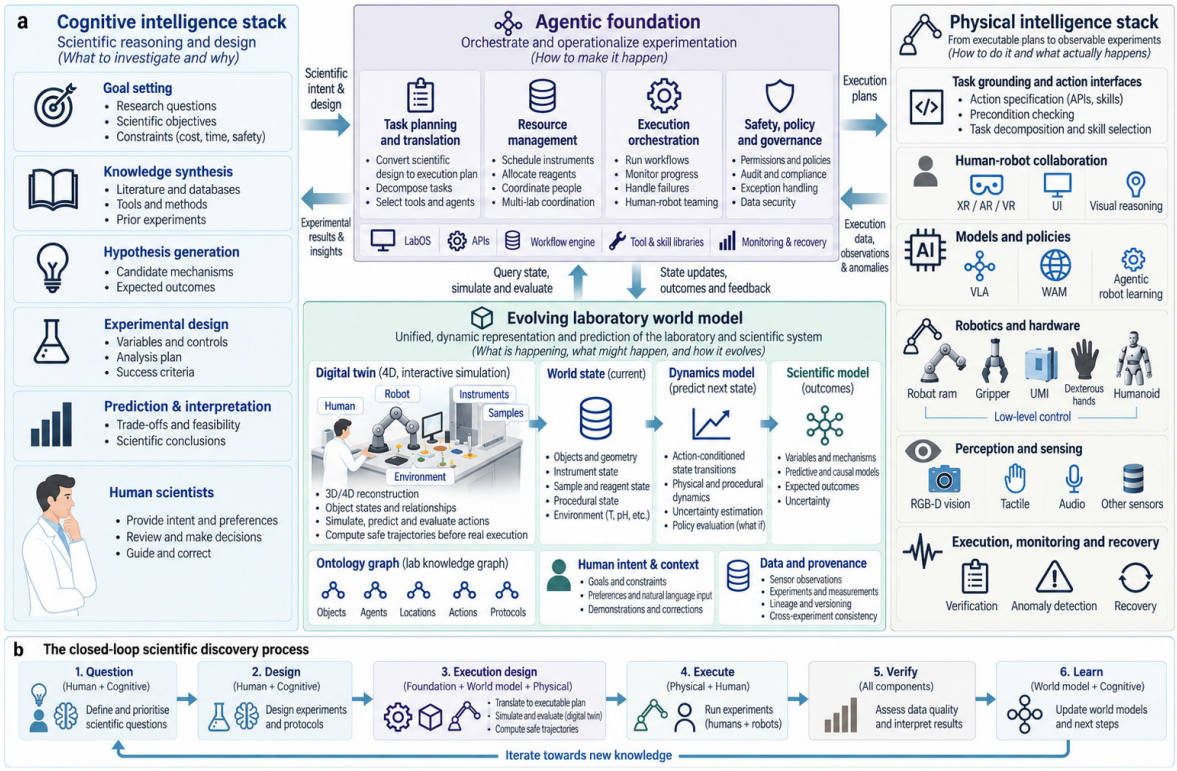


Figure 3. A technical architecture for agentic laboratories. The proposed architecture connects cognitive intelligence, which supports scientific reasoning and experimental design, with physical intelligence, which

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 3 November 2026 https://doi.org/10.20944/preprints202608.0213.v2

enables laboratory execution and observation, through two complementary components. The agentic foundation translates scientific designs into executable plans and coordinates humans, models, instruments, and robots. The evolving laboratory world model integrates digital twins as spatial representations of physical laboratory systems with state estimation, dynamics models, and scientific prediction to ground execution planning, anticipate outcomes, and incorporate experimental feedback. Together, these components support a six-stage discovery loop—question formulation, scientific design, execution design, execution, verification, and learning—in which physical observations inform subsequent decisions and model updates. Human scientists guide scientific direction and consequential decisions, enabling adaptive, verifiable experimentation.

3.1. Cognitive Intelligence: From Scientific Intent to Experimental Design

The cognitive intelligence stack works with scientists to transform research questions into testable experimental designs, specifying what should be investigated and why. Unlike execution planning, it focuses on scientific objectives, hypotheses, and the evidence needed to evaluate them.

Scientific questions are often incomplete and evolve with new evidence. Cognitive intelligence helps scientists identify candidate mechanisms, compare competing explanations, and select experiments that distinguish among them. Foundation models such as AlphaFold, Enformer, Geneformer, and scGPT support scientific prediction and representation learning, while agents including Co-Scientist, Biomni, CRISPR-GPT, and multi-agent research systems extend these capabilities to literature synthesis, hypothesis critique, and experimental planning.[7,8,13,25,26,28,47,69–71,74]Active learning and Bayesian optimization can further prioritize informative experiments under resource constraints.[5,72,73]

The resulting design specifies the hypothesis, experimental variables, controls, measurements, and criteria for interpretation. For example, a scientist investigating therapeutic resistance may use cognitive intelligence to identify competing molecular mechanisms and design perturbation experiments to distinguish them. However, specifying which perturbations to test does not determine how samples should be prepared, which instruments are available, or whether the procedure can be executed safely. These questions require execution design by the agentic foundation.

Cognitive intelligence also draws on the evolving world model, which supplies previous experimental outcomes, scientific predictions, uncertainty estimates, and records of procedural deviations. This feedback helps distinguish genuine scientific observations from execution errors before hypotheses are revised. Human scientists remain responsible for setting research priorities, evaluating scientific significance, and deciding when unexpected evidence warrants changing the research direction.[4,17,35]

3.2. Agentic Foundation: From Scientific Design to Coordinated Execution

The agentic foundation translates scientific experimental designs into executable plans and coordinates their implementation across models, instruments, robots, and human operators. Unlike fixed automation pipelines, it must adapt workflows to available resources, changing laboratory conditions, and unexpected failures.[3,4,11,36]

Its responsibilities include task planning and translation, resource management, execution orchestration, and safety governance. It decomposes scientific designs into machine-readable procedures, selects appropriate tools and agents, allocates instruments and personnel, manages task dependencies, and enforces permissions and approval requirements. Systems such as ChemOS and IvoryOS demonstrate laboratory workflow orchestration, while ChemCrow and Co-Scientist illustrate agent-based scientific tool use.[11,24,27,36] Adaptive agents may further develop and reuse task-specific capabilities across workflows.[53]

Execution design requires continuous interaction with the world model and physical intelligence. The foundation proposes candidate procedures; the world model evaluates current laboratory state, physical feasibility, predicted consequences, and uncertainty; and physical

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 3 November 2026 https://doi.org/10.20944/preprints202608.0213.v2

intelligence supplies available hardware capabilities and action constraints. The foundation then selects or revises the execution plan before dispatching authorized actions.

For example, a therapeutic-resistance experiment may require sample preparation, perturbation delivery, incubation, and imaging. The foundation coordinates these tasks while checking sample identity, instrument availability, and procedural dependencies through the world model. If a required sample is misplaced or an instrument becomes unavailable, it can request additional observations, revise the workflow, or escalate to a scientist rather than blindly following the original plan.

Verification is incorporated throughout execution. Each consequential action has expected preconditions, observable outputs, and acceptance criteria. The foundation compares reported outcomes with these requirements, coordinates recovery when appropriate, and preserves records of actions, decisions, and human interventions.[35,36,51] This allows human and robotic work to remain coordinated without assuming that every laboratory procedure can be fully automated.[17,40]

3.3. Evolving Laboratory World Model: Grounding State, Prediction, and Learning

The evolving laboratory world model provides the shared representation needed to connect scientific plans with physical reality. It integrates three complementary functions: state estimation, dynamics and simulation, and scientific prediction. These functions may use distinct models and data sources rather than a single monolithic system.[35,40,52,53]

State estimation maintains the current laboratory context, including sample identity, instrument status, object locations, procedural progress, and environmental conditions. Information from cameras, sensors, instrument APIs, and human observations can be linked through structured representations such as laboratory knowledge graphs. Provenance records connect measurements to their experimental conditions, actions, and software configurations.[35,36,51]

Digital twins provide spatially grounded representations of instruments, objects, and experimental procedures. Through 3D/4D reconstruction and multimodal observations, they can capture physical states and their evolution over time. A digital twin is a component of the broader world model, not the world model itself: it represents particular physical systems, while the world model connects these representations to dynamics, scientific predictions, and experimental context.

When equipped with appropriate dynamics and control models, interactive digital twins can support robotic simulation and policy learning. Robots can rehearse manipulations, evaluate trajectories, and identify potential failures before physical deployment, reducing disruptive trial-and-error in operational laboratories.[52,58–60] Beyond evaluating existing capabilities, digital twins can provide simulated environments for reinforcement learning, enabling robots to acquire and refine new manipulation policies through repeated interaction and feedback. However, geometric reconstruction alone does not guarantee accurate contact, fluid, or actuator dynamics. Simulation-based policies require task-specific physical modeling, calibration, and controlled real-world validation before deployment.

Dynamics models predict how laboratory states change under proposed actions, while scientific models estimate experimental outcomes, such as biological responses to perturbations. These predictions must remain distinct: successful reagent delivery does not establish that the intended biological response occurred. Scientific predictors may incorporate mechanistic models, foundation models, and probabilistic estimates, with uncertainty guiding experimental selection and interpretation.[7,8,69–74]

During execution, observations update state estimates and reveal discrepancies between predicted and actual outcomes. Validated experimental feedback may inform updates to predictive models and task-specific control policies through active learning or reinforcement learning; reliable transfer to new laboratory settings still requires task-specific validation.[5,58–60,72,73,86] The world model thereby supports both execution planning and scientific reasoning while exposing uncertainty and conditions requiring further validation.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 3 November 2026 https://doi.org/10.20944/preprints202608.0213.v2

3.4. Physical Intelligence: From Executable Plans to Observable Experiments

The physical intelligence stack executes experimental plans and captures evidence about what actually occurs. It comprises robotic systems, laboratory instruments, perception and sensing infrastructure, action models, and human–machine interfaces. Its role is to translate coordinated workflows into reliable physical operations while maintaining feedback to the world model.

Execution begins by grounding the agentic foundation's plans in available hardware capabilities. Standardized APIs and reusable skills can support pipetting, mixing, sample transfer, culturing, imaging, spectroscopy, and instrument calibration.[11,23,36] Robotic learning approaches, including vision-language-action models, can connect visual observations and task instructions to motor policies, although laboratory manipulation remains challenging because of small objects, precise handling requirements, variable instruments, and sensitive experimental conditions.[58–60]

Physical intelligence therefore operates in close coordination with the world model. The world model supplies spatial context, predicted state transitions, and simulated environments for evaluating robotic actions before deployment. Physical systems return observations that verify execution and expose discrepancies requiring model updates or replanning.

Human–robot collaboration remains essential for procedures requiring expert judgment, unfamiliar manipulations, or intervention during unexpected events. Interfaces such as XR devices, cameras, and laboratory software can support real-time guidance and capture human actions. LabOS illustrates how multimodal AI and AI-XR can connect scientific agents with researchers during physical experimentation, while egocentric observations and demonstrations may contribute to future robotic learning.[40,51]

Perception provides the feedback necessary for reliable execution. RGB-D vision, tactile sensing, audio, instrument telemetry, and environmental sensors capture complementary aspects of laboratory state. Monitoring systems compare these observations with expected outcomes to detect anomalies and determine whether an action was completed correctly.[35,36]

For example, if a sample vessel is misplaced during robotic preparation, the sensing system can identify a discrepancy, update the world model, and prompt the agentic foundation to pause or revise execution. If sample identity cannot be verified, the experiment should not continue as though the intended procedure succeeded. This distinction between command completion and verified physical execution prevents procedural errors from being mistaken for meaningful scientific outcomes.

3.5. Integrating the Stack Across the Scientific Discovery Loop

The four components form an agentic laboratory when they operate within a shared discovery loop, connecting scientific intent to executable actions, verified observations, and subsequent learning (Figure 3). The loop comprises six stages with distinct responsibilities:

1. **Question formulation (Human + Cognitive Intelligence).** Scientists define research objectives, priorities, and constraints. Cognitive intelligence supports literature synthesis, hypothesis generation, and identification of scientifically meaningful questions.[13,17]
2. **Scientific design (Human + Cognitive Intelligence).** Scientists and cognitive models develop experiments specifying hypotheses, variables, controls, measurements, and interpretation criteria. The output is a scientific design, not a robotic execution plan.[5,25,26,72,73]
3. **Execution design (Agentic Foundation + World Model + Physical Intelligence).** The foundation translates the scientific design into coordinated procedures. The world model checks current laboratory state, predicts action outcomes, and supports simulation-based evaluation, while physical intelligence defines hardware capabilities and constraints. Infeasible designs are returned for revision before execution.[11,35,36,52]
4. **Execution (Physical Intelligence + Human).** Robots, instruments, and human operators perform the experiment. The agentic foundation coordinates workflows, while sensing systems capture actions, measurements, and changes in laboratory state.[17,23,40]

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 3 November 2026 https://doi.org/10.20944/preprints202608.0213.v2

5. **Verification (All Components).** Physical intelligence checks action completion and measurement quality; the world model compares observations with expected states; the foundation checks procedural compliance and provenance; and cognitive intelligence, together with scientists, evaluates scientific validity. This distinguishes execution failures from unexpected experimental findings.[35,36,50,51]
6. **Learning (World Model + Cognitive Intelligence).** Verified observations update laboratory state, physical dynamics, scientific predictions, and uncertainty. Cognitive intelligence uses this evidence to refine hypotheses and propose subsequent experiments, while the foundation retains validated workflows and recovery experience for future execution.[5,31,32,72,73]

Across these stages, continuous verification and provenance connect scientific objectives, execution records, observations, and interpretations, supporting reproducibility within and across laboratories.[33,35,36,38,39,51]

4. Evaluating Agentic Laboratories: Can Scientific Intent Become Trustworthy Evidence?

Evaluating agentic laboratories requires more than measuring the performance or autonomy of individual models and robots. We propose assessing whether scientific designs are faithfully translated into physical experiments and whether the resulting evidence supports valid conclusions.[3–5,35,36] Evaluation must also account for human oversight, particularly when uncertainty or experimental consequences require expert judgment.[17,35]

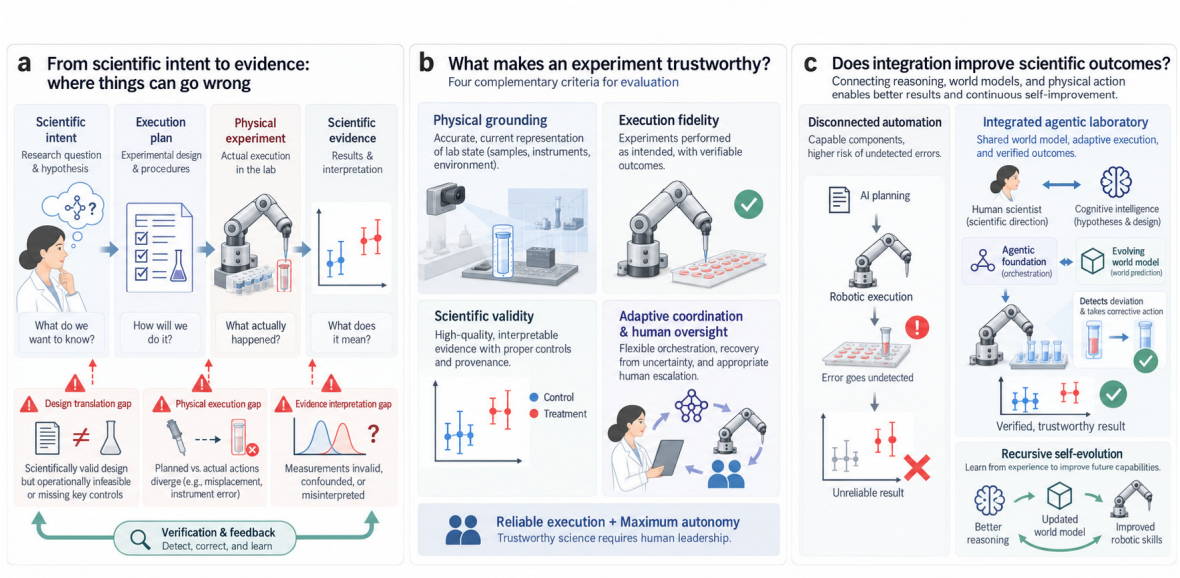


Figure 4. Evaluating agentic laboratories through trustworthy scientific experimentation. (a) Scientific intent must be translated into executable plans, physical experiments, and interpretable evidence, with potential failures arising from design translation, execution deviations, and evidence interpretation. Verification and feedback help identify and correct these gaps. (b) Four complementary criteria assess laboratory performance: physical grounding, execution fidelity, scientific validity, and adaptive coordination with human oversight. Together, they emphasize reliable experimentation rather than maximum autonomy. (c) Disconnected automation risks undetected errors and unreliable results, whereas an integrated agentic foundation and evolving laboratory world model enable state-aware execution, error detection, and corrective action. Verified experimental feedback can further support recursive self-evolution by improving scientific reasoning, world models, and robotic capabilities while preserving human scientific leadership.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 3 November 2026 https://doi.org/10.20944/preprints202608.0213.v2

4.1. Three Gaps Between Scientific Intent and Experimental Evidence

The discovery loop contains three critical transitions where errors can propagate. The design translation gap arises when scientific objectives are converted into executable procedures. An agent may correctly identify an informative perturbation but overlook instrument limitations, sample requirements, or necessary controls. The physical execution gap concerns discrepancies between planned and actual experimental actions, including instrument drift, incorrect sample handling, and unrecorded procedural modifications. The evidence interpretation gap arises when measurements are used to support conclusions without adequately establishing their validity or accounting for experimental uncertainty.[35,36,50,51]

Each gap requires distinct evidence of correctness: execution plans must preserve scientific requirements, observed actions must satisfy procedural expectations, and experimental conclusions must be supported by valid measurements. Evaluating these transitions separately identifies where errors originate and whether they propagate into subsequent scientific decisions.

4.2. Four Criteria for Trustworthy Agentic Experimentation

We propose four complementary operational criteria, evaluated within a declared experimental scope rather than combined into a single autonomy score. Further details on operationalizing these evaluation criteria and constructing representative benchmarks are provided in Appendix B.

Physical grounding measures whether the system accurately represents the laboratory state needed for decision-making. Relevant metrics include object and sample identification accuracy, instrument-state estimation, uncertainty calibration, and detection of discrepancies between expected and observed conditions.[40,51,52]

Execution fidelity measures whether experiments are performed as intended, including procedural completion, deviation detection, recovery success, and the proportion of actions supported by verifiable observations. Evaluation should distinguish successful command execution from confirmed physical outcomes.[35,36]

Scientific validity measures whether the resulting evidence supports reliable interpretation. This includes measurement quality, appropriate controls, provenance completeness, recognition of confounding factors, and the ability to distinguish unexpected scientific findings from experimental artifacts.[46,49,50,87,88]

Adaptive coordination and human oversight measure whether the system integrates tools, instruments, and personnel while responding appropriately to uncertainty. Metrics should include recovery time, resource efficiency, successful workflow handoffs, and escalation competence: whether the system recognizes when human intervention is needed, requests the appropriate expertise, and stops acting when necessary.[15–17,35,36] Human dependency should be assessed by the type and frequency of intervention, not simply minimized. A system that seeks expert judgement at consequential moments may be more reliable than one that operates independently but fails silently.

4.3. Benchmarks Must Capture Consequences Across Experiments

Static benchmarks can assess planning or error classification but cannot fully test how decisions affect subsequent laboratory states. Agentic laboratories require **state-changing evaluations** in which actions consume reagents, modify samples, occupy instruments, and introduce consequences that persist across workflow steps. Such environments expose cascading errors, resource constraints, and failures that emerge only during extended execution.[5,35,36]

For example, a benchmark could provide an agent with a perturbation protocol, reagent metadata, instrument calibration records, and partial visual observations. During execution, a sample is misplaced or an instrument drifts outside its calibrated range. The system must identify the discrepancy, determine whether the experiment remains valid, and choose whether to recover,

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 3 November 2026 https://doi.org/10.20944/preprints202608.0213.v2

repeat, revise, or escalate. Performance should be evaluated against the resulting laboratory state and scientific validity, not merely whether the agent completed the assigned workflow.

Evaluation should also test generalization across instruments, sample types, and experimental conditions, since success within a fixed workflow does not establish reliability elsewhere.[5,35,51]

4.4. Does Integration Improve Scientific Outcomes?

The proposed architecture makes a testable claim: connecting orchestration with an evolving world model should improve experimental reliability beyond what any component achieves independently. This can be evaluated through controlled comparisons between fixed automation, agentic orchestration without world-model grounding, systems with state estimation but no predictive simulation, and the complete architecture.

Comparisons should hold experimental objectives, hardware, resources, and operating conditions constant where possible. Outcomes should include verified experimental success, undetected deviations, recovery performance, human intervention requirements, and time or resources needed to obtain valid evidence.[5,11,23,35,36] Ablations can identify when world-model grounding or adaptive orchestration provides measurable benefits, rather than assuming that greater architectural complexity is always preferable.

Beyond individual experiments, agentic laboratories could progressively improve particular capabilities through accumulated, verified experience. This includes refining scientific reasoning strategies, developing reusable agentic skills, updating world models, and improving robotic policies based on verified experimental feedback.[5,53,58–60,86] Unlike ordinary experimental learning, recursive self-evolution improves the laboratory's ability to conduct future investigations. Its effectiveness should be evaluated through sustained improvements on held-out tasks, transfer across workflows, and reduced resource requirements, without compromising reliability or human oversight.[17,35]

Ultimately, the value of integration should be judged by improvements in experimental reliability, scientific validity, and resource efficiency, alongside appropriate human oversight.[17,35]

5. Challenges and Open Questions

Agentic laboratories raise challenges that extend beyond improving individual AI models or automating individual experimental procedures. Progress will depend on how reliably these systems can reason and act across changing scientific contexts, coordinate over long horizons, preserve reproducibility and human oversight, and become practical laboratory infrastructure. We highlight ten open questions below; additional examples, qualifications, and implications are discussed in Appendix C.

1. Which scientific domains are most ready for agentic laboratories?

Progress will likely vary with experimental standardization, observability, and physical complexity.[5,9,23] Chemistry, materials synthesis, quantum control, and standardized biological workflows may support earlier adoption of agentic laboratories where objectives and measurements are well defined, whereas heterogeneous biological samples and clinical settings remain more challenging because of variability, measurement noise, ethical constraints, and robotic complexity.[21,22,30,77,81]

2. Can AI discover beyond existing scientific paradigms?

Current systems are often strongest at searching or optimizing within existing representations and experimentally accessible spaces, whereas scientific discovery can require recognizing anomalies, reframing questions, or developing new measurements.[4,45,48] Agentic laboratories should therefore be evaluated not only by optimization performance, but by whether they can generate testable ideas beyond predefined search spaces and use physical experimentation to expand the data and experimental spaces available for subsequent discovery.

3. Can agentic systems remain reliable beyond predefined workflows?

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 3 November 2026 https://doi.org/10.20944/preprints202608.0213.v2

Real laboratories are exception-heavy: instruments drift, reagents vary, samples degrade, protocols change, and observations may lack clear ground truth.[35,50] Scientific agents may consequently produce outputs that appear plausible but fail physical, chemical, biological, or experimental-validity checks, particularly outside familiar conditions.

Reliability therefore needs to extend beyond successful task completion to recognizing uncertainty, detecting deviations, recovering from failures, and escalating when appropriate.[35,36,51] Evaluation should test these capabilities under changing experimental conditions rather than only within fixed workflows.

4. How can many agents collaborate over long-horizon scientific campaigns?

Scientific campaigns unfold over days to months and may require many specialized agents, tools, and physical resources.[13,28,29,47,57] Multi-agent systems can distribute planning, analysis, critique, and execution, but errors, uncertainty, provenance, and coordination failures may accumulate across long chains of dependent actions.[15,16,35]

Long-horizon coordination therefore requires persistent shared state, provenance, resource scheduling, and mechanisms for resolving disagreement and propagating uncertainty.[35,36,40] The duration and complexity of scientific campaigns may also require advances in long-context reasoning and agent architectures that maintain coherent plans, memory, and decisions across extended sequences of experiments. Multi-agent systems should ultimately be evaluated end-to-end against strong single-agent alternatives rather than assuming that additional agents improve scientific performance.

5. What standards are needed for reproducible agentic laboratories?

Reproducibility should include not only experimental protocols and data, but also the computational decisions that produced them: model and software versions, tool calls, experimental state, calibration, uncertainty, provenance, and human interventions.[35,36,51,62] Standardized machine-readable records could make reproducibility a property of the discovery process rather than only a downstream reporting requirement.

6. Who owns the scientific memory of agentic laboratories?

As researchers increasingly use foundation models and agents for scientific reasoning, planning, analysis, and experimentation, interactions with these systems may capture unpublished workflows, methodological judgment, experimental context, and other forms of institutional scientific memory.[18] Institutionally governed agentic foundations could help preserve this knowledge within research organizations while providing controlled access to external models and services. Questions of ownership, access, privacy, and portability will therefore become increasingly important as scientific knowledge accumulates not only in papers and datasets, but also in persistent agentic systems.

7. What infrastructure is needed to make agentic laboratories practical?

As AI capabilities improve, laboratory autonomy will increasingly depend on interoperable interfaces connecting models, instruments, sensors, robots, and experimental state.[35,36] Minimal machine-actionable laboratory primitives—for example pipetting, mixing, imaging, calibration reporting, and sample transfer—could provide reusable interfaces across otherwise heterogeneous workflows.

8. How can agentic laboratories become cost-effective?

Greater autonomy is useful only when scientific gains justify the costs of robotics, instrumentation, maintenance, reagents, computation, and energy.[75] Cost-effectiveness should therefore be evaluated in terms of whether agentic systems reduce the resources or time required to reach reliable scientific conclusions, rather than simply increasing the number of experiments or hypotheses generated.

9. How can agentic laboratories ensure safe physical experimentation?

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 3 November 2026 https://doi.org/10.20944/preprints202608.0213.v2

As agentic laboratories coordinate and execute increasingly complex physical experiments, failures or misuse can have consequences beyond incorrect model outputs, including unsafe experimental actions and inappropriate access to high-risk capabilities.[35] Safety should therefore be built into the agentic foundation through permissioned tool access, domain-specific validation, audit trails, human approval for high-consequence actions, and mechanisms to pause or escalate when uncertainty or risk exceeds defined limits.[35,36]

A complementary challenge is how to evaluate safety before deployment. Standardized testbeds with bounded action spaces and well-characterized hazards could provide common environments for comparing agentic laboratory designs, testing failure recovery and guardrails, and determining when agentic execution can be deployed safely and reliably. Such evaluation should consider not only accidental failures, but also whether increasingly accessible agentic infrastructure enables users with limited domain expertise or safety training to execute experiments that would otherwise require specialized oversight.

10. How should scientists lead increasingly agentic laboratories?

Near-term agentic laboratories are likely to remain collaborative systems in which AI takes on more coordination, monitoring, and execution while scientists retain responsibility for scientific direction, interpretation, and consequential decisions.[17,35] The goal is therefore not simply to minimize human involvement, but to determine where human judgment adds scientific value and how systems should recognize when that judgment is needed.

Greater automation may also change scientific training. Reducing routine laboratory work could give researchers more time for hypothesis generation, interpretation, and experimental design, but it could also remove the hands-on experiences through which scientists develop experimental judgment.[40–45] Agentic laboratories should therefore preserve opportunities for scientists and trainees to understand physical experiments even as they increasingly supervise automated systems.

Author Contributions: M.W. and L.C. conceived and initiated the project. K.F. developed the manuscript, drafted and wrote the text, coordinated and incorporated feedback from coauthors, revised and edited the manuscript, and conceptualized all figures. L.L. finalized the figures. All authors read the manuscript and provided feedback. All authors reviewed and approved the final manuscript.

Correspondence: Correspondence and requests for materials should be addressed to Mengdi Wang (mengdiw@princeton.edu) and Le Cong (congle@standford.edu).

Funding: No specific funding was received for the preparation of this manuscript.

Acknowledgements: We thank Professor Tom Griffiths and Saining Xie for helpful discussions and feedback on the manuscript. We also thank Yingcheng Wu for assistance with figure editing and for suggestions on references and content.

Competing Interests: Princeton University and Stanford University have filed patent applications related to this work.

Appendix A. Historical Evolution of Scientific Automation Toward Agentic Laboratories

The First Autonomy Gap: Execution Without Reasoning, Reasoning Without Action, 1957–1980s

In the late 1950s, automated clinical chemistry analyzers were among the first laboratory systems to make routine analytical testing automated, standardized, and repeatable at scale.[63] Clinical chemistry was an early validation domain because standardized assays could be repeated as programmable workflows.[63] The Technicon AutoAnalyzer, introduced in 1957 as one of the first widely used automated clinical chemistry analyzers, used continuous-flow analysis to automate sample handling, reagent mixing, measurement and reporting.[63] Although these systems transformed routine laboratory work by reducing manual handling and increasing throughput, their

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 3 November 2026 https://doi.org/10.20944/preprints202608.0213.v2

power was limited to repetition and standardization: they automated well-defined analytical tasks, but not scientific discovery itself.[63]

In parallel, symbolic AI developed the complementary capability: formal reasoning over scientific representations without laboratory action.[65] DENDRAL, an early expert system for chemical structure elucidation, generated candidate molecular structures from mass spectra and other analytical data using encoded chemical knowledge, constraints, and heuristic search.[64] Meta-DENDRAL, a rule-learning extension of DENDRAL, inferred fragmentation rules from known structure–spectrum pairs, showing that machines could propose, not just apply, scientific rules.[65] BACON, a symbolic discovery program, used heuristics to rediscover quantitative regularities, including Kepler's Third Law and other historical laws, from structured numerical data.[66]

Together, laboratory automation and symbolic AI exposed the first autonomy gap. Automation supplied execution without reasoning; symbolic AI supplied reasoning without action.[63,64] The next challenge became linking experimental execution with computational decision-making.[4,31,32]

Scaling Experiments Without Scaling Discovery, 1990s–2000s

By the 1990s, high-throughput screening and combinatorial chemistry had expanded the scale of drug discovery.[2,6]Robotic liquid handling, miniaturized assays, and large compound libraries allowed orders of magnitude more molecules to be tested than manual workflows could support.[2] Yet this engineering advance exposed a central limitation: more measurement did not automatically produce more discovery.[2,68]

Despite large increases in screening capacity, therapeutic productivity did not rise proportionally.[67] The bottleneck shifted from testing capacity to experimental selection. High-throughput systems encoded compound identity, assay readouts, and activity thresholds, but not the judgment needed to decide which hits were robust, artefactual, or worth pursuing.[2,68]They identified potential hits, but did not determine the next informative experiment.[2,68]

This throughput paradox motivated closed-loop systems that could connect measurement to decision-making: choosing experiments, executing them, and updating future choices from the resulting evidence.[31,32]

From Experimental Selection to Closed-Loop Discovery, 2009–2010

In 2009–2010, robot-scientist systems such as Adam and Eve began to automate hypothesis-driven discovery.[31,32]Adam integrated metabolic models, genome annotations, and biological databases to propose gene–enzyme associations, then designed and executed yeast growth experiments to test them.[32] Eve extended this closed-loop logic to early-stage drug screening by combining disease-relevant targets, compound libraries, and robotic assays.[31]

These systems showed that machines could link hypotheses, executable protocols, observations, and updates within a single experimental loop.[31,32] Automation no longer merely accelerated routine assays; it connected experimental choice to evidence.[31,32] Their limitation was boundedness: the hypothesis space, assay format, and action space were specified in advance.[31,32] They could close a loop, but only inside a carefully formalized experimental world.[31,32] This motivated self-driving laboratory architectures that treated autonomy as an integration problem across algorithms, instruments, and workflows.[5,11,12]

Self-Driving Laboratories: Closed-Loop Discovery Within Bounded Spaces, 2010s–Early 2020s

In the 2010s and early 2020s, chemistry and materials science advanced the logic of closed-loop experimentation through self-driving laboratories (SDLs).[5,9] Active learning and Bayesian optimization became central tools for SDLs because they framed experiment selection as an exploration–exploitation problem, allowing systems to choose informative experiments under limited experimental budgets.[72,73] As a software platform for autonomous experimentation, ChemOS framed autonomy as an integration problem by linking optimization algorithms, software

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 3 November 2026 https://doi.org/10.20944/preprints202608.0213.v2

control, and automated instruments into reusable workflows.[11] Thin-film SDLs and A-Lab showed that autonomous campaigns could synthesize, characterize, and update experiments when objectives, action spaces, measurements and hardware interfaces were constrained.[12,23]However, subsequent analyses clarified important limitations in materials-science autonomy: materials reported as newly discovered by autonomous experiments may not always be novel in reality, and claims of discovery can depend strongly on phase identification, validation standards, search-space design, and interpretation of prior literature.[46,87,88]

Perspectives on human–AI–robot catalysis and platforms such as LUMI-lab illustrate how SDL approaches are extending across application domains, including biomedical design.[37,57] LUMI-lab coupled a molecular foundation model, active learning, robotic synthesis and screening of lipid nanoparticles to identify ionizable lipid designs for mRNA delivery.[57] By identifying brominated lipid tails as an unexpected design feature and validating LUMI-6 for gene editing in mouse lung epithelial cells, it showed that self-driving platforms can generate new design principles, not merely optimize known variables.[57]The same closed-loop logic developed in parallel in atomic and quantum-device physics, where systems optimize within a bounded objective and a fixed control space. Online machine-learning optimization was used to tune the production of Bose–Einstein condensates over large control-parameter spaces,[83] and closed-loop machine-learning agents achieved in situ tuning of semiconductor quantum-dot qubit devices across high-dimensional gate-voltage spaces, in some cases faster than human experts;[79,84,85] reinforcement learning was likewise used to prepare non-classical states in superconducting cavities through measurement-based feedback.[86] Layered van der Waals materials also became a useful testbed for bounded autonomy. Computer vision and robotic systems automated flake identification and assembled a 29-layer graphene/hexagonal boron nitride superlattice.[89,95,96] Separately, Gaussian-process-guided scanning tunneling microscopy and scanning tunneling spectroscopy selected informative measurement locations on tungsten disulfide (WS_2).[90]

SDLs closed loops within predefined objectives, action spaces and validation criteria, but this boundedness limited transfer across goals, assays and instruments.[5,9,35] They integrated algorithms, instruments and workflows more flexibly than earlier robot-scientist systems, but remained bounded by hardware interfaces and constrained search spaces.[5,11,35,36] This exposed the need for systems that could draw on broader scientific knowledge rather than optimize only within narrow experimental spaces.[4,5,35]

Machine Learning Outpaced Experimental Validation, 2019–2022

Biology exposed a complementary bottleneck. Whereas chemistry and materials struggled to transfer closed-loop optimization across platforms, biology increasingly struggled to experimentally validate a growing number of computational hypotheses.[7,8,19,20]

From 2019 onward, machine-learning-guided protein engineering accelerated biological design, followed by AlphaFold in 2021 and later genomic and single-cell foundation models.[7,8,70,71]Machine-learning-guided protein engineering used sequence–function data to propose mutations before experimental validation, while protein language models learned statistical patterns from large sequence databases to prioritize plausible biological sequences.[7,68] AlphaFold transformed protein-structure modeling: it generated structural hypotheses from sequence far faster than traditional experimental structure determination.[8] Beyond protein structure, models such as Enformer improved sequence-based modeling of gene expression from long-range genomic context, Nucleotide Transformer extended foundation-model approaches to DNA sequence representation, Geneformer demonstrated how single-cell foundation models can prioritize experimentally validated biological mechanisms and therapeutic targets, and scGPT brought generative foundation models to single-cell and multi-omic analysis, expanding the range of biological hypotheses that could be generated in silico faster than they could be systematically validated in the wet lab.[69–71,74] Like many machine-learning systems, these models remain stronger at pattern learning than

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 3 November 2026 https://doi.org/10.20944/preprints202608.0213.v2

causal reasoning, with limits in dynamic, disordered or context-dependent biological settings.[4,7,8,69]

Biofoundries and automated evolution platforms offered a partial response by standardizing parts of the design–build–test–learn cycle, including DNA assembly, strain construction, screening and data collection.[19–22] However, they do not eliminate the validation bottleneck: many biological workflows still depend on tacit handling, variable samples, cell-state heterogeneity, reagent lots, local protocol adaptation and weakly standardized metadata.[20,35,36,51]

The key shift was that biological hypotheses became easier to generate than to test.[7,8,69–71] Biology increasingly faced the inverse problem: computational design advanced faster than automated wet-lab execution.[7,8,19,20] Structures, hypotheses and sequence designs could be generated or prioritized in silico, but still had to be selected, translated into protocols and evaluated in physical context.[7,8,20,71]By contrast, chemistry and materials science still face challenges in both physical automation and computational reasoning because of the diversity of synthesis and characterization workflows. Within bounded domains, physical execution became programmable before computational reasoning became general.

Materials science faced a related validation gap. Deep-learning platforms such as GNoME expanded the computational search space by predicting hundreds of thousands of stable inorganic structures, but such predictions are not discoveries by themselves: candidate materials still require synthesis, characterization, phase assignment, and comparison with prior experimental literature.[46,91] Subsequent commentaries and analyses have emphasized that AI-driven and autonomous materials-discovery systems must be evaluated not only by the number of candidates generated, but by whether the resulting materials satisfy the linked criteria of novelty, credibility, and utility.[46,87,88,92] This caution applies especially when models operate in large search spaces where predicted stability, apparent diversity, or automated synthesis can obscure whether a material is experimentally realizable, genuinely new, and scientifically useful.[46,91,92]

Collectively, the bottleneck shifted from generating biological and material possibilities to prioritizing, validating, and operationalizing them.[46,69,70,91,92] This machine-learning-to-validation imbalance motivated tool-using digital agents that could reason over model outputs, retrieve information, design workflows, and connect computational outputs to experimental plans.[24–27]

The Rise of AI Agents, 2022–Present

After the public emergence of ChatGPT in 2022, LLM-based scientific agents accelerated the move from static model outputs toward programmable reasoning and workflow integration.[24,27,61] ChemCrow linked LLMs to chemistry tools; Coscientist, a chemistry-focused LLM system, connected reasoning to cloud-laboratory execution; CRISPR-GPT applied agentic planning to gene-editing design; and systems such as Google's Co-Scientist, Biomni, Virtual Lab, Robin, PantheonOS and AI Scientist extended agentic workflows toward hypothesis generation, critique, biomedical tool use, multi-agent analysis and digital research automation.[13,24–29,47,56] More recent self-evolving agents such as ALITA point to a further transition: agents not limited to predefined tools or fixed workflows can construct, refine and reuse task-relevant capabilities through model-context protocols and accumulated experience.[53]In quantum device physics, agent-based and LLM-driven frameworks began translating natural-language goals into executable laboratory routines—locating readout resonators, choosing the next measurement, and reproducing measurement protocols described in the literature on superconducting-qubit hardware.[81,82] These advancements suggest a route from static tool-using agents toward adaptive scientific agents whose reasoning infrastructure can evolve across tasks.[15,16,53]

These systems make modular scientific reasoning and tool use programmable: goals can be broken down, tools invoked, evidence compared and plans critiqued across workflows.[13,24–27] Yet they face two major limitations. First, agents inherit the limits of their underlying foundation models: brittle reasoning, hallucinated claims, weak causal understanding, incomplete domain knowledge

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 3 November 2026 https://doi.org/10.20944/preprints202608.0213.v2

and poor uncertainty calibration can all propagate into tool choices and experimental plans.[4,15,16] Second, their core limitation in physical science is observability. They can reason about tools and experiments, but they often cannot see the laboratory state that determines whether a plan is feasible, correctly executed or scientifically trustworthy.[35,36,40] They depend on structured inputs, APIs and instrument responses rather than direct access to visual state, spatial layout, calibration drift, manual correction or unexpected procedural deviations.[35,36,40,51]

This combination of model limitation and physical observability gap motivates AI-XR, multimodal AI and physical-perception systems that can make laboratory state visible to computational agents.[40,52,53,59]

Collaborative Human–AI–Robot Laboratories That See, Reason and Act, 2024–Present

The next step is not simply more reasoning. Agentic laboratories require machine-readable access to physical laboratory state: instruments, samples, objects, spatial layout, procedural progress, human interventions and unexpected deviations.[35,36,40] Without this perceptual layer, even strong reasoning systems remain detached from the conditions that determine whether an experiment is feasible, correctly executed or scientifically interpretable.[35,40,53]

Current scientific agents address only part of this gap. Co-scientist and multi-agent systems can generate hypotheses, critique plans, and chain tools across literature, data analysis, and experimental design, but they remain primarily cognitive unless connected to laboratory state.[13,15,27] Adaptive digital agents may accumulate reusable reasoning patterns and procedural knowledge, but those capabilities remain trapped in software without physical perception.[40,72]

Systems such as LabOS show how AI-XR could connect digital agents to scientists during physical work by exposing first-person visual context, procedural cues and human corrections.[40] Such interfaces make tacit laboratory context available for planning, guidance and future learning, moving agents beyond structured files or instrument outputs.[40]Qumus illustrates a complementary direction in quantum materials, coupling multimodal reasoning, robotic execution and closed-loop error correction inside an embodied experimental system to accelerate two-dimensional (2D) quantum materials discovery and device fabrication.[54] Related perception-action systems have also been developed for two-dimensional materials. ATOMIC combines a vision foundation model with language-model control for autonomous optical microscopy.[93] AILA uses language-model agents to automate AFM workflows, including graphene layer identification.[94]By updating representations of objects, instruments, protocol steps, human actions, and environmental context, such systems point toward laboratory world models that connect digital reasoning to physical context.[40,52,53]

This shift matters because laboratory state is not contained only in databases, protocols or instrument logs. It is also carried by visual and spatial cues, including images, video, and egocentric observations that require visual perception and reasoning, as well as spoken corrections, instrument telemetry, environmental conditions and procedural deviations.[35,36,40,51,54] AI-XR and multimodal AI are therefore not peripheral interfaces; they are foundational infrastructure for linking digital reasoning to physical experimentation.[40,52–54,59]

Toward a Fully Agentic Laboratory Stack, Present–Future

Physical perception is necessary but not sufficient. Current systems occupy distinct layers of an emerging agentic-laboratory stack: digital agents for planning, integration systems for tool access, SDLs for bounded optimization, AI-XR for laboratory-state capture, and cloud or networked laboratories for scalable execution.[11,13,40] Each layer expands what machines can do, but none alone can support adaptive discovery across changing goals, tools, environments, and human constraints.[3–5] An agentic laboratory is not a single model, robot, or interface, but a harnessing architecture that integrates changing goals, heterogeneous tools, uncertain evidence, physical constraints, laboratory state, provenance, and human judgment; its evaluation should therefore

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 3 November 2026 https://doi.org/10.20944/preprints202608.0213.v2

extend beyond planning quality to execution feasibility, state awareness, exception handling, provenance, uncertainty calibration, and appropriate human escalation.[3–5,35,36,51]

Scientific automation has made more of the discovery loop machine-readable: laboratory automation has made procedures, measurements and reporting programmable; symbolic AI has represented rules, constraints and candidate hypotheses; high-throughput systems have scaled experimental testing; robot scientists and self-driving laboratories have closed experimental loops within bounded domains; foundation models have expanded learned representations and designs; large language model (LLM) agents have made planning, reasoning, and tool use programmable; and AI-XR and multimodal physical-AI systems are beginning to expose laboratory states by connecting human perception and action to machine-readable workflows.[1,2,4,5] Yet these capabilities remain only partially connected. Physical automation made laboratory execution programmable, while AI made scientific representation and reasoning increasingly computational.[4,64,66] The central challenge for agentic laboratories is to connect these layers into a shared harnessing architecture that represents goals, constraints, experimental state, uncertainty, provenance and human oversight across the discovery loop.[3–5] This historical progression shows how each advance solved one bottleneck while exposing the next, motivating agentic laboratories as integrated human–AI–robot discovery systems (Figure 2).[35,36,40]

Appendix B. Operationalizing the Evaluation of Agentic Laboratories

The evaluation framework introduced in Section 4 assesses how reliably agentic laboratories translate scientific designs into verified experimental evidence. Four complementary criteria measure distinct aspects of this process:

- **Physical grounding** evaluates the accuracy of laboratory state representations, including sample identity, instrument status, spatial relationships, and uncertainty estimates. Representative metrics include state-estimation accuracy, discrepancy detection, and uncertainty calibration.[40,51,52]
- **Execution fidelity** measures whether planned experimental actions occur as intended, including procedural completion, detection of deviations, recovery success, and verification of physical outcomes rather than command execution alone.[35,36]
- **Scientific validity** assesses whether experimental measurements support defensible conclusions, accounting for controls, measurement quality, provenance, confounding factors, and the distinction between unexpected discoveries and experimental artifacts.[46,49,50,87,88]
- **Adaptive coordination and human oversight** evaluate how effectively the system coordinates models, instruments, workflows, and personnel while responding to uncertainty. Relevant metrics include resource efficiency, workflow handoff success, recovery time, and escalation competence. Human dependency should be assessed by the expertise, frequency, and consequences of interventions rather than minimized indiscriminately.[15–17,35,36]

These criteria can be operationalized through benchmarks involving infeasible experiment detection, protocol-failure prediction, provenance reconstruction, and discrimination between genuine scientific signals and experimental artifacts. For example, a failed-protocol triage benchmark could provide an agent with a planned protocol, reagent-lot metadata, instrument calibration history, partial telemetry, intermediate images, and noisy experimental readouts. The agent would determine whether the outcome is valid, artifact-driven, recoverable, requires repetition or protocol revision, or warrants human escalation.

Static evaluations assess decisions against fixed records but cannot fully capture cascading failures or resource constraints. State-changing simulated laboratories should therefore allow actions to consume reagents, occupy instruments, alter samples, and affect subsequent experimental states. Performance should be measured across dependent actions, including whether the system maintains scientific validity, recognizes its limitations, and requests human intervention when appropriate.[5,35,36]

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 3 November 2026 https://doi.org/10.20944/preprints202608.0213.v2

Appendix C. Expanded Challenges and Open Questions for Agentic Laboratories

This appendix expands the ten open questions introduced in Section 5, examining the scientific, technical, infrastructural, and human challenges of developing agentic laboratories. Particular attention is given to the coordination and physical grounding needed to translate computational reasoning into reliable, verifiable experimentation.

C.1. Which Scientific Domains Are Most Ready for Agentic Laboratories?

Readiness for agentic laboratories depends on experimental standardization, physical observability, instrument interoperability, and the ability to verify outcomes.[4,5,9,23] Chemistry, catalysis, thin-film materials, and selected quantum-control tasks offer relatively structured settings with measurable objectives and programmable actions.[23,37,55,77,81]However, extending these systems beyond bounded optimization requires coordinating additional instruments, procedures, and experimental constraints.

Biology presents greater heterogeneity. Cell-free systems, engineered cells, protein engineering, and standardized perturbation screens may be more readily integrated than experiments involving primary samples, organoids, or clinical specimens, where biological variability, measurement noise, and specialized handling complicate execution.[21,22,30]Clinical laboratories may nevertheless benefit from agentic assistance in documentation, quality control, and exception monitoring, subject to strict privacy and safety requirements.[35,51]

Readiness should therefore be assessed through the ability to represent experimental state, execute intended actions, and verify resulting evidence rather than through a universal autonomy level. In partially observable biological systems, stochastic world models and uncertainty-aware planning may help evaluate candidate interventions, although their predictions still require experimental validation.

C.2. Can Agentic Laboratories Discover Beyond Existing Scientific Paradigms?

Scientific goals are rarely equivalent to fixed benchmark objectives. A stated objective—whether optimizing a material property, testing a biological mechanism, or investigating an unexpected phenomenon—exists alongside changing constraints involving feasibility, safety, reproducibility, interpretability, cost, and scientific value.[35,51] Agentic systems may therefore need representations of scientific goals that can evolve as evidence accumulates rather than treating discovery solely as optimization of a fixed objective.[4]

A deeper challenge is whether agentic systems can move beyond interpolation and local search. Transformative discoveries may require recognizing anomalies, reframing questions, or pursuing observations that are weakly represented in existing data.[48] Evidence that papers and patents have become less disruptive over time also cautions against assuming that greater automation necessarily produces greater conceptual novelty.[45] Moreover, agents restricted to already-characterized, automatable experiments may search only a biased subset of experimental space. Expanding what can physically be investigated may therefore be as important as expanding what AI systems can reason about.

C.3. Can Agentic Systems Remain Reliable Beyond Predefined Workflows?

Many laboratory failures arise outside the nominal workflow: instruments drift, reagent properties vary, samples degrade, protocols are locally adapted, and silent deviations can invalidate conclusions. Many assays also lack validated machine-readable ground truth, particularly for rare or previously unseen biological phenomena, and even human analysts may disagree when interpreting the same evidence.[49,50]

Scientific generative models introduce additional failure modes. Outputs can be formally plausible while remaining physically invalid, chemically inconsistent, biologically implausible,

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 3 November 2026 https://doi.org/10.20944/preprints202608.0213.v2

unstable, or experimentally inaccessible. A system may also arrive at an appropriate recommendation through reasoning that fails under small perturbations. Evaluation should therefore probe reasoning stability—for example, whether reordering inconsequential steps or substituting an equivalent reagent changes a recommendation without scientific justification—and test systems outside familiar workflows.

These risks become more consequential when agent decisions trigger physical actions that consume samples, incur costs, or generate misleading data.[35] Agentic laboratories may therefore require domain-specific validators and verification pipelines that test feasibility, safety, and experimental validity before consequential actions. The relevant challenge is not simply whether an individual workflow can be automated, but whether reliability persists as agents move across assays, instruments, samples, and experimental regimes.[5,35]

C.4. How Can Many Agents Collaborate over Long-Horizon Scientific Campaigns?

Scientific campaigns can extend over days or months, with results from early experiments determining later decisions. Multi-agent systems could divide this work among planners, critics, tool routers, analysts, and validity or safety checkers, enabling parallel work and complementary expertise.[13,28,29,47,57] However, long horizons also amplify failure: uncertainty accumulates, early artifacts propagate downstream, credit assignment becomes difficult, and campaign state can exceed the context available to any individual agent.[16,35] Embodied agents face an additional limitation because current robotics foundation models and vision-language-action systems are generally stronger on short-horizon tasks than open-ended workflows.[58–60]

Decomposition itself can introduce cascading errors, disagreement or artificial consensus, and diffusion of provenance and accountability.[15,35] Moreover, optimizing individual components does not guarantee better end-to-end performance; multi-agent systems should therefore be compared with strong single-agent baselines, with additional agents treated as a resource cost rather than a proxy for capability. Physical resources—including instruments, reagents, samples, and compute—are also finite, making scheduling and resource allocation part of the coordination problem.[11,35,36] Persistent shared state, provenance, uncertainty propagation, and human-in-the-lead checkpoints may consequently be important for maintaining coherence across long campaigns.[35,36,40]

C.5. What Standards Are Needed for Reproducible Agentic Laboratories?

Agentic systems could improve reproducibility through structured workflows, provenance capture, calibration tracking, and standardized metadata, but they could also introduce new failure modes through changing model versions, hidden prompts, evolving toolchains, unrecorded instrument states, or proprietary infrastructure.[50] Reproducibility therefore needs to encompass the computational and agentic decision process alongside the physical protocol, including model and software versions, prompts, tool calls, intermediate observations, uncertainty estimates, and human overrides.[35,36,51,62] Replicability is a related but distinct objective: whether findings persist across independent runs, instruments, laboratories, datasets, reagent lots, and experimental contexts.

As agents participate in planning, execution, analysis, and interpretation, the scientific record may consequently extend beyond the final paper. Machine-readable records could include raw data, protocols, analysis code, model interactions, candidate hypotheses, failed attempts, human decisions, uncertainty estimates, calibration histories, and experimental environments.[35,36] Such records could support process-level disclosure and eventually allow reviewers to inspect the provenance behind scientific claims, including missing controls or unrecorded assumptions.

Reproducibility may also need to operate across networks of laboratories rather than individual facilities. Laboratories, cloud platforms, SDLs, and scientific communities could share protocols, representations, calibration standards, provenance, and learned capabilities.[10,33,34,38,39] Early work on networked autonomous materials exploration and programmable cloud laboratories points toward this model.[14,33,38,39] Cross-laboratory transfer, however, requires more than sharing data:

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 3 November 2026 https://doi.org/10.20944/preprints202608.0213.v2

hardware configuration, reagent lots, calibration procedures, software versions, environmental conditions, and human interventions can all influence whether a workflow transfers successfully.[35,36]

C.6. Who Owns the Scientific Memory of Agentic Laboratories?

Scientific interaction data create an institutional challenge distinct from conventional data ownership. As researchers increasingly use foundation-model interfaces for coding, analysis, experimental planning, and scientific reasoning, institutions may externalize unpublished workflows, methodological judgment, reasoning traces, and accumulated scientific context alongside conventional data.[18] Institutionally governed agentic layers or model wrappers could retain prompts, experimental context, protocols, provenance, and other interaction records within local research ecosystems while still providing access to external models. In clinical and public-health settings, this may additionally require secure environments, privacy-preserving access controls, audit logs, and governance for regulated data.

Over longer timescales, institutional scientific capability may depend partly on whether experimental reasoning, domain protocols, and accumulated expertise can be represented in reusable machine-readable infrastructure. Shared agentic memory could also change how scientific knowledge circulates: knowledge currently distributed across publications, datasets, conferences, informal interactions, and individual memory could increasingly be synthesized and maintained by persistent agentic systems. This raises questions not only about ownership and access, but also about whether scientific memory should reside within individual laboratories, academic consortia, or external platforms.

C.7. What Infrastructure Is Needed to Make Agentic Laboratories Practical?

Widespread deployment may depend as much on interoperable infrastructure as on model capability. Agentic systems that plan beyond fixed workflows require interfaces linking actions to expected changes in samples, instruments, and scientific systems while representing uncertainty when predictions extend beyond available evidence.[4,35,36] Laboratory world models may support this connection by providing state and predicted dynamics for physical planning.[52,53]

One practical step is to define minimal "driverless" laboratory primitives—such as pipetting, mixing, thermal regulation, imaging, calibration reporting, and sample transfer—that can be composed across otherwise heterogeneous workflows. As scientific models improve, interfaces, standards, governance, uncertainty-aware control, and physical reliability may become increasingly important constraints; conversely, slower progress in scientific foundation models could leave model capability itself as the bottleneck.[7,8,35,36]

C.8. How Can Agentic Laboratories Become Cost-Effective?

Automation can reduce labor and accelerate iteration, but it can also produce low-value experiments while consuming reagents, compute, energy, and expensive robotic infrastructure.[75] High-mix robotic systems may additionally incur substantial maintenance costs. Cost-effectiveness should therefore be considered during system design rather than evaluated only after deployment.

Early infrastructure investments may not produce immediate returns but could make laboratories better positioned to exploit future improvements in models, robotics, sensors, and interoperability. Ultimately, the relevant measure is not simply the number of experiments or hypotheses generated, but whether an agentic system reduces the resources or time required to move from scientific inquiry to reliable validation.[5,35,75]

C.9. How Can Agentic Laboratories Ensure Safe Physical Experimentation?

Connecting AI agents to physical laboratory systems introduces risks beyond incorrect computational outputs. Agent-generated actions may involve hazardous materials, sensitive

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 3 November 2026 https://doi.org/10.20944/preprints202608.0213.v2

instruments, irreversible sample modifications, or procedures requiring specialized expertise.[35] Unlike digital errors, physical execution failures can have consequences that cannot be reversed through software correction.

Safety should therefore be enforced through the agentic foundation using permissioned tool access, operational constraints, domain-specific validators, and human approval for high-consequence actions.[35,36] The evolving laboratory world model can support pre-execution risk assessment by estimating physical state, predicting potential failures, and evaluating candidate actions in simulation. However, incomplete observations and inaccurate dynamics mean that simulation cannot replace physical safeguards or established laboratory safety procedures.[52]

A further challenge is determining when an agentic system is sufficiently reliable for deployment. Controlled testbeds should evaluate constraint violations, anomaly detection, recovery behavior, and appropriate human escalation under changing laboratory conditions. Safety assessment must also consider misuse and whether increasingly accessible automation could enable experiments beyond a user's training or authorization. Human scientists and institutions must retain authority over safety-critical decisions, with agentic systems designed to recognize and respect those boundaries.[17,35]

C.10. How Should Human Scientists Lead Agentic Laboratories?

Near-term value may arise from relatively routine but consequential activities—including sample transfer, instrument setup, workflow routing, standardized assay execution, quality control, and exception monitoring—rather than general-purpose autonomous robot scientists.[3,4,13,35,36] These tasks can provide the operational infrastructure and data needed for more capable systems while maintaining human scientific leadership.

As routine execution and coordination become increasingly automated, scientists may spend more time on scientific framing, interpretation, invention, and the development of new experimental tools and measurements. Human intuition and scientific judgment may remain particularly important for identifying consequential questions, interpreting ambiguous evidence, and deciding which unexpected observations deserve further investigation.[17,35] However, increasing distance from physical experimentation may also reduce exposure to serendipitous observations and practical details that shape scientific interpretation.

This tension is particularly important for scientific training. If agentic systems assume tasks traditionally performed by undergraduate and graduate researchers, they may remove some of the apprenticeship experiences through which experimental judgment develops. PI–trainee relationships could shift from teaching laboratory mechanics toward scientific strategy, interpretation, ethics, and judgment; AI-XR systems such as LabOS could also provide real-time guidance and protocol verification.[40,44,45] Institutions may therefore need to preserve opportunities for trainees to understand physical experiments rather than learning only to supervise automated systems.

The broader effects on scientific participation remain uncertain. Increased productivity could make more scientific questions tractable and expand demand for research, whereas fixed funding, infrastructure, or attention could instead concentrate output among fewer laboratories. The relevant question is therefore not only how much scientific work can be automated, but how increasing autonomy affects participation, training, responsibility, and the distribution of scientific capability.

References

1. Ley, S. V., Fitzpatrick, D. E., Ingham, R. J. & Myers, R. M. Organic synthesis: March of the machines. *Angew. Chem. Int. Ed.* **54**, 3449-3464 (2015).
2. Mennen, S. M. et al. The evolution of high-throughput experimentation in pharmaceutical development and perspectives on the future. *Org. Process Res. Dev.* **23**, 1213-1242 (2019).
3. Tobias, A. V. & Wahab, A. Autonomous 'self-driving' laboratories: A review of technology and policy implications. *R. Soc. Open Sci.* **12**, 250646 (2025).

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 3 November 2026 https://doi.org/10.20944/preprints202608.0213.v2

4. Musslick, S. et al. Automating the practice of science: Opportunities, challenges, and implications. *Proc. Natl Acad. Sci. USA* **122**, e2401238121 (2025).
5. Tom, G. et al. Self-driving laboratories for chemistry and materials science. *Chem. Rev.* **124**, 9633-9732 (2024).
6. Maier, W. F., Stöwe, K. & Sieg, S. Combinatorial and high-throughput materials science. *Angew. Chem. Int. Ed.***46**, 6016-6067 (2007).
7. Yang, K. K., Wu, Z. & Arnold, F. H. Machine-learning-guided directed evolution for protein engineering. *Nat. Methods* **16**, 687-694 (2019).
8. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021).
9. Seifrid, M. et al. Autonomous chemical experiments: Challenges and perspectives on establishing a self-driving lab. *Acc. Chem. Res.* **55**, 2454-2466 (2022).
10. Armer, C., Letronne, F. & DeBenedictis, E. Support academic access to automated cloud labs to improve reproducibility. *PLoS Biol.* **21**, e3001919 (2023).
11. Roch, L. M. et al. ChemOS: An orchestration software to democratize autonomous discovery. *PLoS ONE* **15**, e0229862 (2020).
12. MacLeod, B. P. et al. Self-driving laboratory for accelerated discovery of thin-film materials. *Sci. Adv.* **6**, eaaz8867 (2020).
13. Gottweis, J. et al. Accelerating scientific discovery with Co-Scientist. *Nature* **655**, 487-496 (2026).
14. Acceleration Consortium. Awesome self-driving labs. https://github.com/AccelerationConsortium/awesome-self-driving-labs (2026).
15. Tran, K.-T., Dao, D., Nguyen, M.-D., Pham, Q.-V., O'Sullivan, B. & Nguyen, H. D. Multi-agent collaboration mechanisms: A survey of LLMs. Preprint at http://arxiv.org/abs/2501.06322 (2025).
16. Shin, W. et al. The (R)evolution of scientific workflows in the agentic AI era: Towards autonomous science. Proceedings of the SC '25 Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis 2305-2316 (Association for Computing Machinery, 2025). https://doi.org/10.1145/3731599.3767580.
17. Dellermann, D., Ebel, P., Söllner, M. & Leimeister, J. M. Hybrid intelligence. *Bus. Inf. Syst. Eng.* **61**, 637-643 (2019).
18. Brynjolfsson, E. & McAfee, A. The business of artificial intelligence. https://hbr.org/2017/07/the-business-of-artificial-intelligence (2017).
19. HamediRad, M. et al. Towards a fully automated algorithm driven platform for biosystems design. *Nat. Commun.***10**, 5150 (2019).
20. Holowko, M. B., Frow, E. K., Reid, J. C., Rourke, M. & Vickers, C. E. Building a biofoundry. *Synth. Biol.* **6**, ysaa026 (2021).
21. Singh, N. et al. A generalized platform for artificial intelligence-powered autonomous enzyme engineering. *Nat. Commun.* **16**, 5648 (2025).
22. Zhang, Q. et al. Integrating protein language models and automatic biofoundry for enhanced protein evolution. *Nat. Commun.* **16**, 1553 (2025).
23. Szymanski, N. J. et al. An autonomous laboratory for the accelerated synthesis of inorganic materials. *Nature***624**, 86-91 (2023).
24. Bran, A. M., Cox, S., Schilter, O., Baldassari, C., White, A. D. & Schwaller, P. Augmenting large language models with chemistry tools. *Nat. Mach. Intell.* **6**, 525-535 (2024).
25. Qu, Y. et al. CRISPR-GPT for agentic automation of gene-editing experiments. *Nat. Biomed. Eng.* **10**, 245-258 (2026).
26. Huang, K. et al. Biomni: A general-purpose biomedical AI agent. Preprint at https://doi.org/10.1101/2025.05.30.656746 (2025).
27. Boiko, D. A., MacKnight, R., Kline, B. & Gomes, G. Autonomous chemical research with large language models. *Nature* **624**, 570-578 (2023).
28. Ghareeb, A. E. et al. A multi-agent system for automating scientific discovery. *Nature* **655**, 497-505 (2026).
29. Lu, C. et al. Towards end-to-end automation of AI research. *Nature* **651**, 914-919 (2026).
30. Fandrey, C. I. et al. NIS-Seq enables cell-type-agnostic optical perturbation screening. *Nat. Biotechnol.* **43**, 1337-1347 (2025).

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 3 November 2026 https://doi.org/10.20944/preprints202608.0213.v2

31. Sparkes, A. et al. Towards robot scientists for autonomous scientific discovery. *Autom. Exp.* **2**, 1 (2010).
32. King, R. D. et al. The automation of science. *Science* **324**, 85-89 (2009).
33. National Science Foundation. Test bed: Toward a network of programmable cloud laboratories (PCL Test Bed). https://www.nsf.gov/funding/opportunities/pcl-test-bed-test-bed-toward-network-programmable-cloud-laboratories (2025).
34. Arias, D. S. & Taylor, R. E. Scientific discovery at the press of a button: Navigating emerging cloud laboratory technology. *Adv. Mater. Technol.* **9**, 2400084 (2024).
35. Leong, S. X. et al. Steering towards safe self-driving laboratories. *Nat. Rev. Chem.* **9**, 707-722 (2025).
36. Zhang, W. et al. IvoryOS: An interoperable web interface for orchestrating Python-based self-driving laboratories. *Nat. Commun.* **16**, 5182 (2025).
37. Orouji, N. et al. Autonomous catalysis research with human-AI-robot collaboration. *Nat. Catal.* **8**, 1135-1145 (2025).
38. Canty, R. B. et al. Science acceleration and accessibility with self-driving labs. *Nat. Commun.* **16**, 3856 (2025).
39. Yoshida, N., Iwabuchi, Y., Igarashi, Y. & Iwasaki, Y. Networking autonomous material exploration systems through transfer learning. *npj Comput. Mater.* **11**, 362 (2025).
40. Cong, L. et al. LabOS: The AI-XR Co-Scientist That Sees and Works With Humans. Preprint at http://arxiv.org/abs/2510.14861 (2025).
41. Shanahan, J. O., Ackley-Holbrook, E., Hall, E., Stewart, K. & Walkington, H. Ten salient practices of undergraduate research mentors: A review of the literature. *Mentor. Tutor. Partnersh. Learn.* **23**, 359-376 (2015).
42. Limeri, L. B. et al. Development of the Mentoring in Undergraduate Research Survey. *CBE Life Sci. Educ.* **23**, ar26 (2024).
43. Hidi, S. & Renninger, K. A. The four-phase model of interest development. *Educ. Psychol.* **41**, 111-127 (2006).
44. Kestin, G. et al. AI tutoring outperforms in-class active learning: an RCT introducing a novel research-based design in an authentic educational setting. *Sci. Rep.* **15**, 17458 (2025).
45. Park, M., Leahey, E. & Funk, R. J. Papers and patents are becoming less disruptive over time. *Nature* **613**, 138-144 (2023).
46. Leeman, J. et al. Challenges in high-throughput inorganic materials prediction and autonomous synthesis. *PRX Energy* **3**, 011002 (2024).
47. Xu, W. et al. PantheonOS: An Evolvable Multi-Agent Framework for Automatic Genomics Discovery. Preprint at https://doi.org/10.64898/2026.02.26.707870 (2026).
48. Buehler, M. J. Accelerating scientific discovery with generative knowledge extraction, graph-based representation, and multimodal intelligent graph reasoning. *Mach. Learn.: Sci. Technol.* **5**, 035083 (2024).
49. Aczel, B. et al. Consensus-based guidance for conducting and reporting multi-analyst studies. *eLife* **10**, e72185 (2021).
50. Errington, T. M., Denis, A., Perfito, N., Iorns, E. & Nosek, B. A. Challenges for assessing replicability in preclinical cancer biology. *eLife* **10**, e67995 (2021).
51. Engel, J. et al. Project Aria: A New Tool for Egocentric Multi-Modal AI Research. Preprint at http://arxiv.org/abs/2308.13561 (2023).
52. Fung, P. et al. Embodied AI Agents: Modeling the World. Preprint at http://arxiv.org/abs/2506.22355 (2025).
53. Qiu, J. et al. Alita: Generalist Agent Enabling Scalable Agentic Reasoning with Minimal Predefinition and Maximal Self-Evolution. Preprint at http://arxiv.org/abs/2505.20286 (2025).
54. Shi, L. et al. Qumus: Realization of An Embodied AI Quantum Material Experimentalist. Preprint at http://arxiv.org/abs/2605.18407 (2026).
55. Inizan, T. J. et al. System of Agentic AI for the Discovery of Metal-Organic Frameworks. Preprint at http://arxiv.org/abs/2504.14110 (2025).
56. Swanson, K., Wu, W., Bulaong, N. L., Pak, J. E. & Zou, J. The Virtual Lab of AI agents designs new SARS-CoV-2 nanobodies. *Nature* **646**, 716-723 (2025).
57. Xu, Y. et al. LUMI-lab: A foundation model-driven autonomous platform enabling discovery of ionizable lipid designs for mRNA delivery. *Cell* **189**, 1620-1635.e25 (2026).

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 3 November 2026 https://doi.org/10.20944/preprints202608.0213.v2

58. Zitkovich, B. et al. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. *Proc. Mach. Learn. Res.* **229**, 2165-2183 (2023).
59. Kim, M. J. et al. OpenVLA: An Open-Source Vision-Language-Action Model. *Proc. Mach. Learn. Res.* **270**, 2679-2713 (2025).
60. Black, K. et al. π_0: A Vision-Language-Action Flow Model for General Robot Control. Preprint at https://doi.org/10.48550/arXiv.2410.24164 (2024).
61. OpenAI. Introducing ChatGPT. https://openai.com/index/chatgpt/ (2022).
62. Popova, K. Reproducibility and instruction following in the shop floor laboratory work: The case of a TMS experiment. *Sci. Technol. Hum. Values* **47**, 1294-1320 (2022).
63. Skeggs, L. T. Jr. An automatic method for colorimetric analysis. *Am. J. Clin. Pathol.* **28**, 311-322 (1957).
64. Lindsay, R. K., Buchanan, B. G., Feigenbaum, E. A. & Lederberg, J. DENDRAL: A case study of the first expert system for scientific hypothesis formation. *Artif. Intell.* **61**, 209-261 (1993).
65. Buchanan, B. G. et al. Applications of artificial intelligence for chemical inference. 22. Automatic rule formation in mass spectrometry by means of the Meta-DENDRAL program. *J. Am. Chem. Soc.* **98**, 6168-6178 (1976).
66. Bradshaw, G. F., Langley, P. W. & Simon, H. A. Studying scientific discovery by computer simulation. *Science***222**, 971-975 (1983).
67. Scannell, J. W., Blanckley, A., Boldon, H. & Warrington, B. Diagnosing the decline in pharmaceutical R&D efficiency. *Nat. Rev. Drug Discov.* **11**, 191-200 (2012).
68. Rives, A. et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl Acad. Sci. USA* **118**, e2016239118 (2021).
69. Avsec, Ž. et al. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **18**, 1196-1203 (2021).
70. Dalla-Torre, H. et al. Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nat. Methods* **22**, 287-297 (2025).
71. Cui, H. et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat. Methods* **21**, 1470-1480 (2024).
72. Shahriari, B., Swersky, K., Wang, Z., Adams, R. P. & de Freitas, N. Taking the human out of the loop: A review of Bayesian optimization. *Proc. IEEE* **104**, 148-175 (2016).
73. Kusne, A. G. et al. On-the-fly closed-loop materials discovery via Bayesian active learning. *Nat. Commun.* **11**, 5966 (2020).
74. Theodoris, C. V. et al. Transfer learning enables predictions in network biology. *Nature* **618**, 616-624 (2023).
75. Ratner, D. & Sumpter, B. G. Facilities' Current Status and Projections for Producing and Managing Large Scientific Data with Artificial Intelligence and Machine Learning. U.S. Department of Energy, Office of Science, Basic Energy Sciences https://science.osti.gov/-/media/bes/pdf/reports/2020/AI-ML_Companion_Document.pdf (2019).
76. Krenn, M., Landgraf, J., Fösel, T. & Marquardt, F. Artificial intelligence and machine learning for quantum technologies. *Phys. Rev. A* **107**, 010101 (2023).
77. Alexeev, Y. et al. Artificial intelligence for quantum computing. *Nat. Commun.* **16**, 10829 (2025).
78. Menke, T., Häse, F., Gustavsson, S., Kerman, A. J., Oliver, W. D. & Aspuru-Guzik, A. Automated design of superconducting circuits and its application to 4-local couplers. *npj Quantum Inf.* **7**, 49 (2021).
79. Moon, H. et al. Machine learning enables completely automatic tuning of a quantum device faster than human experts. *Nat. Commun.* **11**, 4161 (2020).
80. Magesan, E., Gambetta, J. M., Córcoles, A. D. & Chow, J. M. Machine learning for discriminating quantum measurement trajectories and improving readout. *Phys. Rev. Lett.* **114**, 200501 (2015).
81. Cao, S. et al. Automating quantum computing laboratory experiments with an agent-based AI framework. *Patterns* **6**, 101372 (2025).
82. Li, S. et al. Large language model-assisted superconducting qubit experiments. Preprint at https://doi.org/10.48550/arXiv.2603.08801 (2026).
83. Wigley, P. B. et al. Fast machine-learning online optimization of ultra-cold-atom experiments. *Sci. Rep.* **6**, 25890 (2016).

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 3 November 2026 https://doi.org/10.20944/preprints202608.0213.v2

84. Kalantre, S. S. et al. Machine learning techniques for state recognition and auto-tuning in quantum dots. *npj Quantum Inf.* **5**, 6 (2019).
85. Zwolak, J. P. et al. Autotuning of double-dot devices in situ with machine learning. *Phys. Rev. Appl.* **13**, 034075 (2020).
86. Sivak, V. V., Eickbusch, A., Liu, H., Royer, B., Tsioutsios, I. & Devoret, M. H. Model-free quantum control with reinforcement learning. *Phys. Rev. X* **12**, 011059 (2022).
87. Peplow, M. Robot chemist sparks row with claim it created new materials. *Nature*https://doi.org/10.1038/d41586-023-03956-w (2023).
88. Robinson, J. New analysis raises doubts over autonomous lab's materials 'discoveries'. Chemistry World https://www.chemistryworld.com/news/new-analysis-raises-doubts-over-autonomous-labs-materials-discoveries/4018791.article (2024).
89. Masubuchi, S. et al. Autonomous robotic searching and assembly of two-dimensional crystals to build van der Waals superlattices. *Nat. Commun.* **9**, 1413 (2018).
90. Thomas, J. C. et al. Autonomous scanning probe microscopy investigations over WS2 and Au{111}. *npj Comput. Mater.* **8**, 99 (2022).
91. Merchant, A. et al. Scaling deep learning for materials discovery. *Nature* **624**, 80-85 (2023).
92. Cheetham, A. K. & Seshadri, R. Artificial intelligence driving materials discovery? Perspective on the article: Scaling deep learning for materials discovery. *Chem. Mater.* **36**, 3490-3495 (2024).
93. Yang, J. et al. Zero-shot autonomous microscopy for scalable and intelligent characterization of 2D materials. *ACS Nano* **19**, 35493-35502 (2025).
94. Mandal, I. et al. Evaluating large language model agents for automation of atomic force microscopy. *Nat. Commun.* **16**, 9104 (2025).
95. Li, H. et al. Rapid and reliable thickness identification of two-dimensional nanosheets using optical microscopy. *ACS Nano* **7**, 10344–10353 (2013).
96. Han, B. et al. Deep-learning-enabled fast optical identification and characterization of 2D materials. *Adv. Mater.***32**, 2000953 (2020).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.