Article

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

Kelsey Fu [1], Luna Lyu [2], Shuzhen Li [1], Suozhi Huang [1], Suheng Xu [3], Joy Zheng [1], Xuefeng Liu [2], Shilong Liu [1], Greg Barbone [4], Ying Liu [5], Linxi Jim Fan [4], Yuke Zhu [4,6], Matthew Davis [7], Kaiyi Jiang [1], Sherry Yang [8], Xu Kuang [2], Yuan Luo [9], Francesca Dominici [10], Christina Curtis [2], Xiaojie Qiu [2], Alberto Abadie [11], Caroline Uhler [11], Mary Tang [2], Joseph C. Wu [2], Ju Li [11], Jared Toettcher [1], Jose L. Avalos [1], Rebecca Rojansky [2], Huimin Zhao [12], John P. A. Ioannidis [2], Barry Rand [1], Emma Lundberg [2], Andrew S. Rosen [1], Zhuang Liu [1], Shana Kelley [9,13], Chenyan Xiong [1], Barbara E Engelhardt [2], Thomas Montine [2], Christina V. Theodoris [14,15], Katherine S. Pollard [13,14,15], Fuxin Li [16], Xueyue Zhang [3], Tianyi Peng [11], Dmitri Basov [11], Xiang Qian [2], Eric Xing [17,18], Ali Yazdani [1], Jure Leskovec [2], Nigam Shah [2], Zhenan Bao [2], Pietro Perona [19], Sanfeng Wu [1], Clifford Brangwynne [1], Le Cong [2] and Mengdi Wang [1,*]

[1] Princeton University
[2] Stanford University
[3] Columbia University
[4] NVIDIA
[5] Scale AI
[6] The University of Texas at Austin
[7] Flagship Pioneering
[8] New York University
[9] Northwestern University
[10] Harvard University
[11] Massachusetts Institute of Technology
[12] University of Illinois Urbana-Champaign
[13] Chan Zuckerberg Biohub
[14] University of California, San Francisco
[15] Gladstone Institute
[16] Oregon State University
[17] Carnegie Mellon University
[18] Mohamed Bin Zayed University of Artificial Intelligence
[19] California Institute of Technology
* Correspondence: mengdiw@princeton.edu

Abstract

Scientific discovery is fundamentally a problem-solving process involving distributed intelligence. Human intuition, computational reasoning, and experimental execution are distributed across people, instruments, and software systems, limiting the speed and scale of 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. We hypothesize that the next generation of laboratories will be agentic: environments in which scientists, AI systems, and robotic platforms operate as collaborative discovery partners, with humans contributing the parts of discovery that remain hardest to make explicit: asking the right questions and holding provisional mechanistic models of how a system works. The key missing layer is an agentic harnessing layer that continuously integrates hypothesis, literature-derived evidence, experimental data, uncertainty, and experimental state into a shared "laboratory world model"—a dynamic representation of the scientific system and its evolving context. By maintaining and updating this lab world model, the agentic harnessing layer enables coordinated decision-making,

adaptive planning, and increasingly autonomous scientific workflows across humans and machines. A central challenge is that much of the scientific research process remains inaccessible to machines, including tacit knowledge, human observations, adaptive decision-making, and evolving experimental context. Advances in multimodal AI and immersive interfaces may help bridge this gap, allowing humans, agents, and robotic systems to collaborate seamlessly in scientific discovery rather than simply automating isolated tasks. Agentic laboratories could provide a new architecture for science, integrating human, artificial, and physical intelligence into a unified discovery system.

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 research advances by asking questions, collecting and interpreting data, and deciding what evidence should be generated next. In modern laboratories, this depends on two intertwined functions: experimental execution, which measures and manipulates physical systems, and scientific coordination, which links protocols, models, instruments, interpretation, and decisions about what to test next. Some tasks once performed manually are now possible to be automated, from liquid handling and high-throughput screening to robotic synthesis, cloud experimentation, and digital data capture.[1,2] These advances have improved throughput, reproducibility, and access, while AI has accelerated reasoning, prediction, and design; however, broadly autonomous discovery remains limited because AI's software capabilities are still weakly connected to scalable physical execution, especially in wet-lab settings.[3,4,7,8]

Despite these advances, scientific discovery remains fragmented. Humans still translate goals into protocols, move information between models and instruments, assess data quality, respond to failures, and decide what to test next.[35,36] The central opportunity is to build an agentic harnessing layer that connects models, physical instruments, and human judgment into a unified discovery process (Figure 1).[3,4] Rather than replacing scientists, this layer shifts them from manually coordinating fragmented workflows toward defining scientific goals, interpreting evidence, and governing risk, while agentic infrastructure manages routine execution and coordination across models, instruments, protocols, and laboratory states.[17,35]

This opportunity extends across scientific disciplines. In chemistry and materials science, self-driving laboratories (SDLs), robotic platforms, and integration systems have accelerated synthesis, optimization, and characterization of molecular, material, structural, device and functional properties under well-defined objectives and bounded action spaces[5,11,23] In biology, foundation models, protein language models, genome-editing design systems, optical perturbation screens, and biofoundries have made parts of experimental design and measurement increasingly computational, but many wet-lab workflows remain difficult to standardize, physically automate, or transfer across contexts.[20,35,36] In quantum science and device engineering, optimization algorithms, neural networks, reinforcement learning, and foundation models have, foundation models have accelerated device design, calibration, and data analysis, although these advances remain concentrated in individual subtasks rather than integrated across entire experimental campaigns.[82–88] Across these trajectories, 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]

We define the agentic laboratory as a human–AI–robot discovery system organized around an agentic harnessing layer: it interprets evolving human goals, develops plans to achieve those goals, proposes theoretical, computational, and physical tools, integrates observations and human judgment, revises strategies, and surfaces its reasoning for human review, escalating uncertainty or failure to human oversight.[3,4] This layer functions as an operating system for scientific discovery:

it makes human judgment, model reasoning, laboratory observations, literature evidence, and physical execution composable within a closed loop.[11,36] Together, these capabilities move laboratories from isolated automation toward adaptive human–AI–robot discovery.[3,40] This transition is also emerging as a US national science-infrastructure priority: the U.S. Department of Energy's Genesis Mission lists "Achieving AI-Driven Autonomous Laboratories" among its national science and technology challenges, emphasizing robotics, edge AI, real-time analysis, intelligent feedback, hypothesis generation, and data curation as core components of future experimental workflows.[74]

Two conclusions emerge. First, the near-term objective is reliable human–AI–robot co-discovery rather than full laboratory autonomy because scientific goals evolve, experiments are costly and exception-heavy, and physical AI remains far from robust laboratory deployment. Agentic systems should therefore manage integration, monitoring, and experimental execution while scientists retain responsibility for framing scientific questions, interpreting evidence, and governing risk.[3,4,35] Second, such systems depend on a shared, machine-readable representation of laboratory state—a laboratory world model—because experimental conditions are dynamic, partial, and context-dependent. Without this representation, agents may produce plans that appear plausible but are physically infeasible or irreproducible. Accordingly, agentic laboratories should be judged by their ability to coordinate scientific discovery, generate new knowledge, reduce human operational burden, and execute experiments reliably, rather than by automation level alone.

In this Perspective, we first trace the historical evolution from laboratory automation to agentic laboratories, showing how each advance solved one bottleneck while exposing the next (Section 2). We then define the agentic harnessing layer and laboratory world models as the architectural foundation for scientific discovery (Section 3), introduce a framework for measuring laboratory autonomy and operational performance that pairs an L0-L5 ladder with reliability criteria (Section 4), and finally discuss the major technical, human, governance, and infrastructural challenges that will shape future agentic laboratories (Section 5).

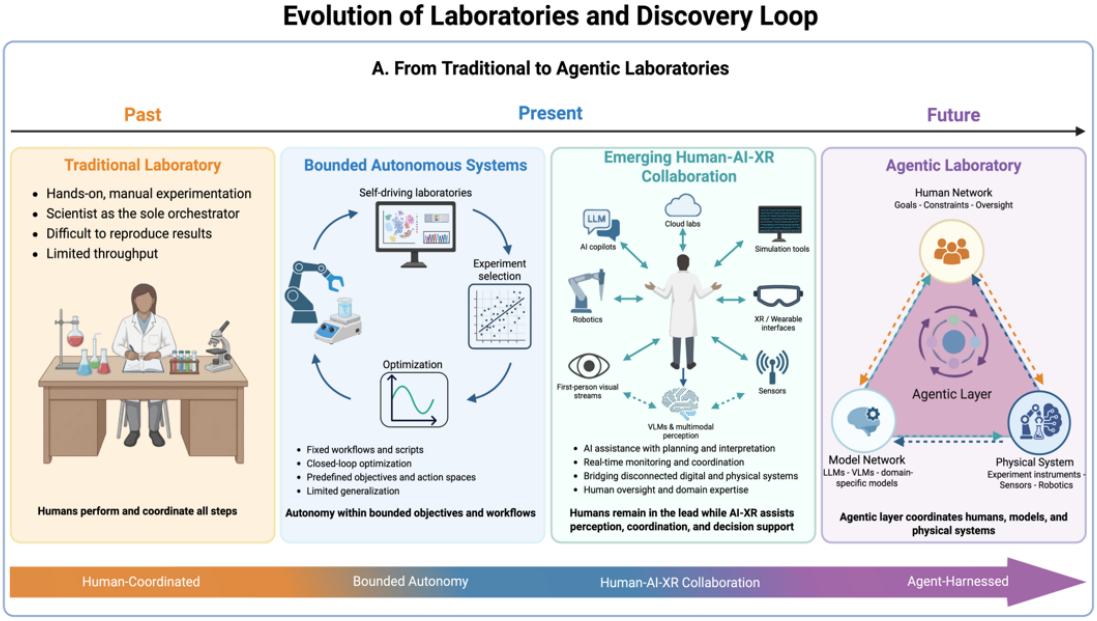

Figure 1. Evolution of laboratory autonomy toward agentic laboratories. Traditional laboratories rely on humans to perform and coordinate experimental work. Present-day systems increasingly automate pre-defined workflows and begin to support human–AI–robot collaboration through XR devices and multimodal AI, but humans still bridge fragmented models, instruments, sensors, cloud laboratories, and physical procedures. Future agentic laboratories integrate human, model, and physical networks through an agentic layer centered

on a shared world model that maintains situational awareness, contextual understanding, and state estimation across digital and physical environments. This world model enables coordinated planning, perception, execution, and oversight among humans, AI agents, robots, and laboratory infrastructure.

We envision that the most plausible next stage of the agentic lab will be human-in-the-lead systems in which agents increasingly handle integration, monitoring, workflow execution and iterative planning, while scientists remain responsible for framing objectives, interpreting anomalies, setting constraints, and governing risk.[17,35] Over time, these systems are likely to become networked rather than monolithic, with multiple laboratories, cloud platforms, models, and scientific communities sharing protocols, provenance, representations, and experimental capabilities.[33,38,39] This transition will reshape not only how experiments are performed, but also how scientific institutions manage data ownership, train researchers, evaluate reproducibility, allocate credit, publish results and teach science.[41,46]

2. Evolution of Scientific Automation Toward Agentic Laboratories

Scientific automation has transformed many fields, from high-energy physics and neuroscience to space exploration. Here, we focus on chemistry, biology, materials science, and quantum laboratories, where successive advances in automation exposed recurring bottlenecks in execution, reasoning, feedback, and system integration that ultimately motivated the concept of agentic laboratories (Figure 2).

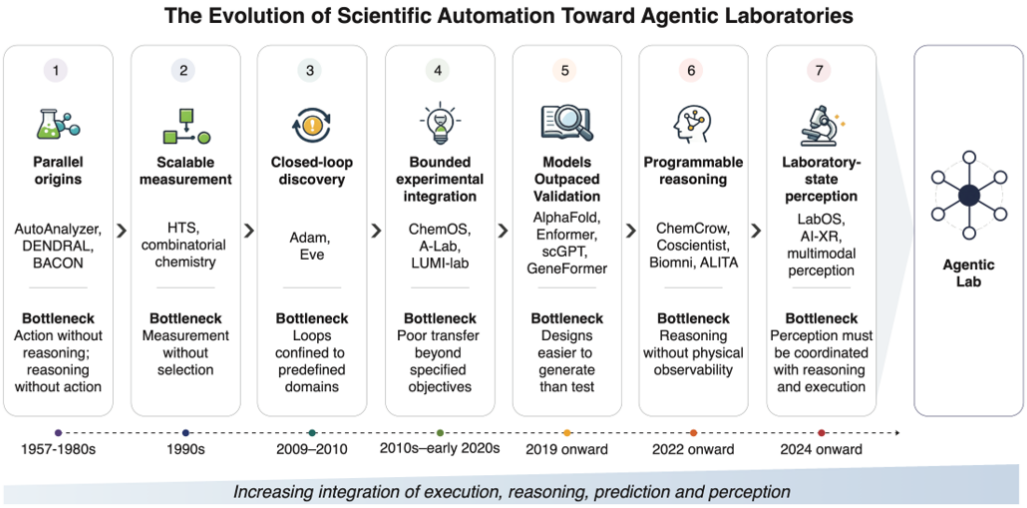

Figure 2. The evolution of scientific automation toward agentic laboratories. Scientific automation has advanced through successive partial representations of the discovery process, each making one repetitive component of laboratory work more streamlined while exposing the next bottleneck. Early laboratory 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.

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.[65] Clinical chemistry was an early validation domain because standardized assays could be repeated as programmable workflows.[65] 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.[65] Although these systems transformed routine laboratory work by reducing manual handling and increasing throughput, their power was limited to repetition and standardization: they automated well-defined analytical tasks, but not scientific discovery itself.[65]

In parallel, symbolic AI developed the complementary capability: formal reasoning over scientific representations without laboratory action.[67] 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.[66] 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.[67] BACON, a symbolic discovery program, used heuristics to rediscover quantitative regularities, including Kepler's Third Law and other historical laws, from structured numerical data.[68]

Together, laboratory automation and symbolic AI exposed the first autonomy gap. Automation supplied execution without reasoning; symbolic AI supplied reasoning without action.[65,66] 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.[2,70]

Despite large increases in screening capacity, therapeutic productivity did not rise proportionally.[69] 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.[2,70] They identified potential hits, but did not determine the next informative experiment.[2,70]

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.[75,76] As a software platform for autonomous experimentation, ChemOS framed autonomy as an integration problem by linking optimization algorithms, software 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.[48,96,97]

Autonomous catalysis platforms and LUMI-lab showed how SDL logic expanded from chemistry and materials into biomedical design.[37,59] LUMI-lab coupled a molecular foundation model, active learning, robotic synthesis, lipid nanoparticle formulation, and mRNA-delivery screening to discover ionizable lipid designs.[59] By identifying brominated lipid tails as an unexpected design feature and validating LUMI-6 for in vivo lung epithelial gene editing, it showed that self-driving platforms can generate new design principles, not merely optimize known variables.[60] 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,[89] 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;[85,90,91] reinforcement learning was likewise used to prepare non-classical states in superconducting cavities through measurement-based feedback.[92] 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.[98,104,105] Separately, Gaussian-process-guided scanning tunneling microscopy and scanning tunneling spectroscopy selected informative measurement locations on tungsten disulfide (WS_2).[99]

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.[7,8,72,73] 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.[7,70] 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.[71–73,78] Like many machine-learning systems, these models remain stronger at pattern learning than causal reasoning, with limits in dynamic, disordered or context-dependent biological settings.[4,7,8,71]

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.[20,35,36,53]

The key shift was that biological hypotheses became easier to generate than to test.[7,8,71–73] 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.[7,8,20,73] 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.48,100 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.48,96,97,101 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.[48,100,101]

Collectively, the bottleneck shifted from generating biological and material possibilities to prioritizing, validating, and operationalizing them.[48,71,72,100,101] 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.[24,27,63] 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.[12,24–29,49,58,59] 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.[55] 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.[87,88] These

advancements suggest a route from static tool-using agents toward adaptive scientific agents whose reasoning infrastructure can evolve across tasks.[15,16,75]

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 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.[35,36,40,53]

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.[40,54,55,61

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.[35,40,55]

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.[40,75]

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.[56] 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.[102] AILA uses language-model agents to automate AFM workflows, including graphene layer identification.[103] 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.[40,54,55]

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.[35,36,40,53,56] AI-XR and multimodal AI are therefore not peripheral interfaces; they are foundational infrastructure for linking digital reasoning to physical experimentation.[40,54–56,61]

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.[3–5,35,36] The field therefore needs end-to-end evaluations that measure not only planning quality, but also execution feasibility, instrument-state awareness, exception handling, provenance completeness, uncertainty calibration and appropriate human escalation.[5,35,36]

Scientific automation made more of the discovery loop machine-readable: laboratory automation made procedures, measurements and reporting programmable; symbolic AI represented rules, constraints and candidate hypotheses; high-throughput systems scaled experimental testing; robot scientists and self-driving laboratories closed experimental loops within bounded domains; foundation models expanded learned representation and design; large language model (LLM) agents made planning, reasoning, and tool use programmable; and AI-XR and multimodal physical-AI systems are beginning to expose laboratory state 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.[4,66,68] 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]

3. The Agentic Lab Layer: An Operating System for Scientific Discovery

As physical AI, multimodal models, robotics and scientific reasoning systems converge, agentic laboratories are becoming technically feasible .[4,40,55] An agentic laboratory is not simply an automated laboratory, a cloud laboratory, or a laboratory with an LLM interface.[10,11,24] We define an agentic laboratory as a human-AI-robot discovery system in which an agentic harnessing layer coordinates human judgment, computational reasoning, and physical execution toward evolving scientific goals under uncertainty.

This definition distinguishes agentic laboratories from two important precursors. A pipeline executes a fixed sequence of computational or physical steps; it may be reliable, but it does not replan at runtime.[11,36] An SDL adds feedback, often through active learning or Bayesian optimization, but usually operates within a predefined objective, workflow, and action space.[5,9,23] These methods can make experiment choice adaptive, but within a fixed design space; they do not revise the goal, update the available tools, or incorporate broader contextual evidence.[4,5]

Continuous learning in an agentic laboratory is broader than updating only a surrogate model or acquisition function; it also updates plans, assumptions, protocols, tool choices, memory, uncertainty estimates, explanations, and criteria for human escalation as new evidence accumulates.[35,36] An agentic laboratory therefore requires at least eight capabilities : (i) interpret human goals, which may be underspecified, conflicting and evolving ; (ii) read the literature and place those goals in context; (iii) develop actionable plans to achieve those goals ; (iv) execute plans by flexibly combining models, instruments, and databases through shared interfaces ; (v) use new observations to revise plans and initiate new actions ; (vi) identify when predictions, actions, or decisions are unreliable or unexpected and surprising and escalate to human oversight; and (vii) preserve provenance, uncertainty, and decision histories to support reproducibility and explain experimental choices, failures, and recommendations; and (viii) engage with human scientists to interpret results, identify productive next steps, disseminate knowledge, and enable effective human supervision.[3,4,35]

The Agentic Harnessing Layer

We call the system that enables these capabilities the agentic harnessing layer (Figure 1).[3,4] This layer sits between three interacting domains: the human domain, which defines goals, constraints, values and interpretations; the model domain, which includes LLMs, vision language models (VLMs), domain-specific models, predictors and simulators, retrieval systems, scientific databases and the literature; and the physical domain, which includes instruments, robotics, laboratory information systems, cloud laboratories, sensors and experimental environments.[7,8,24] These domains may exchange information directly, but reliable interoperation requires shared representations of the goal, laboratory state, possible actions, uncertainty, and provenance. [35,36] For example, a model may recommend increasing a reagent concentration, but a robot cannot safely act unless it also knows the available reagent lot, sample volume, instrument calibration state, protocol constraints, and rationale for the recommendation.

The analogy to an operating system is useful: it does not replace applications or hardware, but makes them composable.[17,36] Similarly, an agentic harnessing layer does not replace humans, models or instruments; it provides the abstractions needed to link them, including task decomposition, tool routing, state tracking, memory, exception handling, safety constraints, uncertainty escalation and audit trails.[3,4,35] The central gap lies not within any single component, but between them: a model cannot trigger an experiment, a robot cannot judge what is worth testing and an LLM cannot know whether an instrument is calibrated or an observation is trustworthy unless those states are exposed in machine-readable form.[35,36]

Representing Scientific Goals and Constraints

A crucial role of the agentic harnessing layer is to represent scientific goals.[3,4] Unlike industrial optimization targets, scientific goals are often incomplete, evolving and partially implicit.[4,35] They may involve hypothesis testing, model discrimination, mechanism discovery, product development or exploration of unexpected phenomena.[4,52] They also include constraints on cost, safety, reproducibility and ethical concerns .[35] Scientific value may also emerge beyond the original objective.

Agentic systems must therefore avoid reducing science to static reward maximization.[4,35] This does not mean that all objectives should be fluid: stable meta-level objectives may be essential, such as sharpening hypotheses, generating reasonable alternatives, designing experiments that discriminate among them, and identifying which hypotheses should be abandoned. Instead, agentic systems require goal representations that can evolve with evidence and allow human scientists to revise the question being asked.[17] The reward should operate less as a fixed target for a desired result and more as a guide for good scientific practice: clarifying hypotheses, maximizing information gain, exposing uncertainty, and supporting decisions about when to continue, stop, or reformulate the inquiry. In this sense, the goal representation is not only an input to the system, but also a mechanism of human oversight: it determines what the agent is allowed to optimize, when it should stop, and when the scientific question itself needs to change.

Building Laboratory World Models

A second function of the agentic harnessing layer is to build and update laboratory world models. A laboratory world model is a shared, machine-readable representation of the evolving laboratory state and the scientific system being studied. It should represent not only static metadata, but also samples, reagents, instruments, protocols, observations, human interventions, model outputs, provenance, uncertainty, and environmental context as states that change over time. Its core function is to separate state estimation from prediction under intervention: first, the system must maintain an up-to-date description of what is currently true in the laboratory; second, it must learn how that state is expected to change under candidate experimental actions.

Beyond maintaining state, laboratory world models should support counterfactual reasoning. Scientists rarely ask only "What is happening now?" but also "What would happen if we changed this variable instead?" A useful world model should therefore evaluate alternative experimental trajectories before committing physical resources, estimate which assumptions drive different predictions, and identify interventions that would most efficiently distinguish competing

hypotheses. This capability transforms the world model from a passive record of laboratory state into an active substrate for scientific reasoning and experimental design.

In agentic laboratories, this representation should separate two related but distinct layers: physical-action world models and scientific-system world models. Physical-action world models capture the immediate consequences of embodied laboratory actions, such as liquid transfer, transferring samples, positioning plates, opening instruments, or detecting protocol deviations from video, egocentric images, telemetry, and sensor data. Scientific-system world models address the harder task of predicting how a biological system, material, reaction, instrument, or workflow will evolve under experimental intervention. The former is likely to be more immediately useful for physical AI and automated laboratories, whereas the latter is a longer-term goal for discovery. This distinction prevents the term "world model" from conflating two different problems: knowing what is happening in the laboratory and predicting what the scientific system will do next. Digital twins—dynamically updated virtual representations of specific instruments, samples, workflows, or laboratory environments—can serve as physically grounded components of physical-action world models. A laboratory world model is broader than any individual twin, integrating multiple asset- and process-level twins with scientific hypotheses, provenance, uncertainty, and models of intervention effects to support planning across an experimental campaign. Digital twins therefore capture localized states, whereas laboratory world models integrate them into a unified planning framework.

In practice, laboratory world models are likely to be hierarchical. Low-level representations capture immediate instrument state, robot actions, sample identity, and procedural execution. Mid-level representations summarize experiments, workflows, and causal relationships among variables. High-level representations encode scientific objectives, hypotheses, mechanistic models, and research strategy. Maintaining consistency across these levels allows local experimental observations to influence global scientific reasoning while enabling high-level goals to constrain low-level planning.

One useful implementation is a typed relational graph in which samples, reagents, instruments, protocols, observations, human interventions, model outputs, uncertainty estimates, and provenance records are represented as typed nodes and edges. Such a graph would give humans and AI agents a common substrate for planning, monitoring, error detection, causal learning, and reproducibility. It also turns the laboratory world model into a concrete learning problem: estimating latent state from heterogeneous measurements and predicting the consequences of proposed experimental interventions.

World models allow agents to reason about experiments before execution, anticipate possible outcomes, identify uncertainties, and adapt plans as new evidence becomes available. They may be implemented through multimodal foundation models, simulators, code-generating agents, executable analysis pipelines, structured provenance records, instrument telemetry, and learned predictive representations. In future agentic laboratories, they could bridge abstract reasoning and physical action by grounding decisions in the dynamics of the system being studied rather than only in text, prior literature, or tool calls.[4,55] Unlike conventional laboratory information management systems that primarily archive completed experiments, laboratory world models should be continuously updated as evidence accumulates. New observations should revise latent state estimates, modify uncertainty, invalidate outdated assumptions, and propagate changes throughout downstream plans. This continual state updating allows the laboratory to behave as a persistent learning system rather than a sequence of disconnected experiments.

This representation must integrate formal records with tacit and multimodal context, including visual cues, instrument telemetry, protocol deviations, human corrections and uncertainty estimates.[53] It should not merely store what happened, but help infer whether an observation is trustworthy, whether a protocol remains on track and whether the next planned action is still appropriate. The same applies to model state: the layer should flag when a predictor lacks sufficient training-data support or biological evidence, when a generative model is out of distribution, or when

a vision-language model is interpreting unfamiliar imagery, rather than passing such outputs downstream as uniformly reliable evidence.[7,55]

World models should not be treated as omniscient predictors. Uncertainty can accumulate across multi-step interventions, underspecified plans can create too many possible action branches to evaluate exhaustively, and some outcomes reflect genuine gaps in scientific knowledge rather than model failure. Their role is not to predict every experiment with confidence, but to clarify what is known, what is uncertain, which assumptions support a recommendation, and when new evidence or human oversight is needed. The same logic should apply to model updating: before fine-tuning or revising predictors, the agentic layer should check data coverage, data quality, validation performance, uncertainty calibration, and distributional similarity to decide whether model updating is justified or whether more data collection is needed.

A major obstacle to laboratory world models is data scarcity. Unlike text or natural images, high-quality experimental data are slow and expensive to acquire because each measurement may consume scarce samples, fabricated devices, cryostat or beamline time, and reagents. Quantum device engineering illustrates this challenge: a single processor may take weeks to fabricate and characterize and where the accessible design and control space vastly exceeds what can be exhaustively measured. Several complementary strategies could let world models be trained, and make decisions, under this regime. First, embedding known physics—Hamiltonians, master equations, or reduced device models—as structural priors can sharply lower the data requirement, as shown by physics-informed and gray-box models that jointly characterize and control quantum devices from limited experimental records.[94] Second, models can be pretrained in simulation and transferred to hardware with limited real-world fine-tuning, as in reinforcement-learning policies learned in simulation and deployed on cold-atom and superconducting-qubit experiments.[92,93] Third, active learning, Bayesian optimization, and reinforcement learning from demonstration can maximize the information gained from each costly experiment.

A more speculative question is whether a model trained on one system can then act competently in a new, equally data-scarce setting—a different device, qubit modality, or material. Early evidence that models trained on small or low-cost systems transfer to larger or previously unseen ones—scaling from small circuits to many-qubit chips, or across qubit types through transfer learning—suggests that a new experiment could inherit prior structure rather than starting from scratch, reaching useful performance far faster than exhaustive manual characterization.[95] If such transfer holds, world models that also search regions of parameter space that human intuition tends to avoid could occasionally surface non-obvious designs, protocols, or operating points that exceed established practice. However, transfer across devices, batch effects, and distribution shift can fail silently, so models trained under data scarcity will require calibrated uncertainty and human oversight to separate genuine extrapolation from confident error.

Making the Discovery Loop Machine-Actionable

A third function of the agentic harnessing layer is to make the discovery loop machine-actionable. Scientific discovery can be described as a loop: read and synthesize the literature, hypothesize, design, simulate, experiment, observe, update, and disseminate results.[31,32] Literature synthesis is already one of the more mature uses of agentic systems, because models can retrieve, compare, and summarize large bodies of prior work before proposing hypotheses or experiments. Earlier robot scientists such as Adam and Eve closed parts of this loop, and later integration systems extended the logic to chemistry and materials discovery.[11,23,31] What changes in agentic laboratories is not the existence of the loop, but the degree to which the stages and transitions between stages become machine-mediated.[3,4]

Beyond automating individual stages, agentic laboratories must automate the transitions between them: outputs from one tool must become inputs to another, exceptions must be resolved, and each result must inform the next analysis or experiment.[35,36] Scientists currently bridge these gaps manually, coordinating fragmented models, instruments, data and workflows through many routine but error-prone translation tasks, and this limits scale.[35,36] The goal is not to remove human

review from these transitions. Human inspection can detect errors, reinterpret unexpected outcomes and generate new ideas. Rather, agentic infrastructure should initially handle predictable, low-judgment handoffs, such as formatting data, routing results, checking protocol constraints, tracking instrument state and preparing inputs for the next analysis, while escalating ambiguous failures, surprising results and high-consequence decisions to scientists. The agentic layer moves this coordination into infrastructure by making handoffs machine-readable, logging decisions and state changes, enforcing escalation rules, and requiring human approval for ambiguous, surprising, or high-consequence steps, while preserving scientific judgment, interpretation and responsibility as human roles.[17,35]

In future agentic laboratories, the harnessing layer will increasingly manage transitions across computational and physical actions while humans remain responsible for scientific framing, oversight and judgment.[17] For this loop to become reliable scientific infrastructure, actions and transitions must remain reproducible across model versions, software environments, instrument states, calibration histories and experimental conditions.[35,53]

Selecting Experiments Under Uncertainty

Selecting the next experiment is a decision-theoretic challenge.[4,35] It requires balancing expected information gain against cost, time, feasibility, model uncertainty, safety, data latency, resource use and the possibility that the current objective is itself poorly specified.[5,35] In wet-lab settings, this trade-off is especially consequential: an agent should not simply choose the experiment predicted to be most informative, but also take into account overall turnaround time, reagent cost and availability, instrument time, and failure risk.[35,53] Active learning, Bayesian optimization, reinforcement learning and uncertainty-aware planning offer useful foundations to select the best next experiment.

Scientific decision-making extends beyond selecting the next experiment within a fixed search space.[9,75,76] It often involves deciding whether to change the search space, revise the hypothesis, repeat an experiment and collect more reliable data, escalate to a human, or stop optimizing a misleading proxy.[4] Agentic systems appear promising in this respect because they can explore large experimental spaces in silico, through modeling and simulation, before committing resources to costly physical validation, analogous to how robotics uses simulation to evaluate policies before deployment in the physical world.[7,8] They can also preserve and revisit "promising routes not taken," helping prioritize future experiments and open research questions.

Preserving Provenance and Human Responsibility

The agentic harnessing layer is not merely a convenience layer for automating workflows (supplementary note 5).[3,4] It is the infrastructure through which scientific intent becomes machine-actionable.[17] Its value depends not only on faster execution, but also on preserving provenance, exposing uncertainty, documenting decisions, supporting reproducibility, and keeping humans responsible for scientific framing, interpretation, and risk governance.[17,53]

Human control is a design goal, not an automatic consequence of the technology: increasingly autonomous systems may automate more of discovery, but safe deployment requires auditable logs, escalation rules, and autonomy levels (section 4) that preserve accountability. In practice, human coordination should define the boundaries within which agents operate: preventing hallucinated claims from becoming actions, approving access to sensitive data or high-risk tools, enforcing no-harm and bias-mitigation principles, weighing social and environmental impact, and deciding when efficiency gains, including computational cost and energy use, are not worth the scientific or societal risk. These records could also support publication and peer review by allowing authors, peer reviewers, and publishers to inspect how automated systems generated, revised, and validated scientific claims.

4. Measuring Levels of Autonomy of Science Laboratories

A definition of agentic labs must be paired with an operational framework for measuring autonomy, because systems can appear similarly “autonomous” while relying on different forms of

reasoning, execution, and human involvement.[3,5] Two systems may both be called autonomous while differing sharply in what they can reason about, what they can physically execute, how often they require human rescue, and whether their actions remain reliable outside narrow workflows.[5,35,36] Two aims are easily conflated: showing that a system augments an accountable scientist, and showing that it may act without case-by-case review. Augmentation is judged by the quality of assistance and whether errors are caught; delegated action is judged by behaviour when no one is checking — including the failure distribution, calibration under distribution shift, and awareness of its own limits. The ladder below addresses this second, stricter case, so each level represents delegated authority, not capability alone.[3,4,35]

Levels of Laboratory Autonomy L0-L5

We propose a six-level autonomy ladder for scientific laboratories organized around cognitive autonomy and physical autonomy (Figure 3a-c).[3,35] This framing is inspired by autonomy scales in automated driving, where levels of automation helped industry and regulators distinguish driver assistance from full automation.[77] However, laboratory autonomy is harder to classify than vehicle autonomy. Driving operates in environments where relevant state is often directly observable, whereas laboratory state may be latent, distributed across samples, instruments, protocols, reagents, data histories and tacit assumptions, and may take human scientists days to understand. Moreover, scientific agents introduce a dimension largely absent from vehicle-autonomy ladders: the possibility that a system may improve not only its actions, but also its own scientific capabilities. Accordingly, the highest level adds recursive self-improvement, the ability to improve cognitive and physical capabilities over time.

Cognitive autonomy is a system's ability to interpret goals, reason over scientific context, design experiments, revise plans, critique outputs, and decide when evidence changes the direction of a campaign.[13,24] Cognitive autonomy depends on perception, learning, and reasoning: converting laboratory context into machine-readable state, updating models and uncertainty estimates from new data, and linking goals, evidence, and constraints to plans and decisions.[4,35,55] Physical autonomy is a system's ability to execute actions through robotics, instruments, data pipelines and laboratory APIs with minimal manual intervention.[11,23,36] Current LLM agents can support parts of cognitive autonomy but remain largely physically disembodied unless connected to instruments, sensors, robots, or laboratory APIs.[24,27,40]

The L0–L5 ladder provides a descriptive taxonomy inspired by autonomy-level frameworks used for self-driving vehicles, but adapted to scientific discovery rather than transportation (Figure 3d).[77] Because cognitive and physical autonomy are partly orthogonal, the L0–L5 ladder is best viewed as a coarse projection of a two-dimensional space. At the highest level, L5 combines high cognitive and physical autonomy with recursive self-improvement, the ability to use accumulated evidence and experience to increase future scientific capability. Human personae are included only as illustrative analogies; the levels themselves are defined by system capabilities:

- **L0: Manual laboratories.** Humans design, coordinate and carry out the work.[1,2]
- **L1: Fixed automation.** Automated systems use fixed workflows or scripts to carry out well-specified tasks, improving efficiency without meaningful adaptability.[11,36] This is roughly analogous to an undergraduate intern following a predefined protocol under close supervision.
- **L2: Copilot systems.** Systems assist with planning, coding, analysis or interpretation, while humans retain control over decisions and execution.[17,24] This resembles the role of an early graduate student who can contribute to analysis and planning but still requires frequent guidance on experimental direction.
- **L3: Bounded closed-loop systems.** Systems propose, execute and update experiments within a pre-specified objective, measurable success criterion, available instruments, and allowed set of experimental actions.[5,12,23] L3 marks the first major inflection point because cognitive and physical autonomy become coupled within a bounded experimental loop. This is analogous to a senior graduate student who can independently run a constrained project but still works within a clearly defined research question and oversight structure.

- **L4: Generalist campaign-level agentic systems.** Systems integrate multiple tools, workflows and experimental contexts under structured human oversight.[3,35,40] At this level, the relevant paradigm is not merely human-in-the-loop but human-in-the-lead: scientists shift from managing individual loops to supervising research campaigns, evaluating AI-generated hypotheses, resolving cross-context anomalies and governing the boundaries of autonomous discovery.[17] We frame this as lab-in-the-loop (LITL): humans, AI and instruments share one loop that any party can enter, which places the burden on the system to remain transparent — able to state where it is in the hypothesis-generation process at each step — rather than only to escalate, since an agent cannot reliably escalate a thread it does not recognize as off track. This is roughly analogous to a postdoctoral fellow coordinating a research campaign across methods, collaborators, and unexpected results, while still operating within human-defined scientific and ethical boundaries.
- **L5: Self-improving agentic systems.** Systems operate across broad scientific task classes with minimal human intervention, can redefine goals and strategies over time, and improve their own capabilities by accumulating experience across experiments, workflows, models, and scientific domains. This level is not simply "more autonomous" than L4; it marks a qualitative shift from executing and coordinating research campaigns to recursively strengthening the system's own capacity for future discovery. In other words, L5 combines high cognitive and physical autonomy with recursive self-improvement, rather than replacing autonomy with self-improvement. This goes beyond the analogy to self-driving vehicles: an L5 agentic laboratory would not merely drive itself along a defined route, but would resemble a research group that becomes increasingly capable through its own discoveries, failures, and accumulated memory. In this sense, self-improvement is a capability beyond autonomy: research progress would not only accumulate linearly, but could accelerate as the system becomes better at generating hypotheses, selecting experiments, building models, and reusing institutional knowledge. Fully open-ended autonomy may eventually be technically achievable, but in domains where human judgment, ethics, and societal priorities shape what should be studied, human responsibility must be built into the design of the system rather than treated as an optional safeguard (supplementary note 2).[4,35]

Most current systems occupy L1–L3, even when described as autonomous. A-Lab and related SDLs demonstrate strong physical autonomy and closed-loop optimization in bounded domains.[23] Coscientist shows how LLM-based systems can orchestrate chemistry workflows with partial experimental grounding.[27] CRISPR-GPT, Biomni, AI co-scientist, PantheonOS, and AI Scientist demonstrate increasing cognitive autonomy, but most remain physically disconnected or domain-constrained.[13,29,49] These systems are important precursors, but not yet general agentic laboratories.

Higher autonomy is not inherently better, safer, or more scientifically valuable. A robust L3 system operating within a well-defined experimental loop may be more useful than a fragile L4 system that generalizes poorly, requires frequent rescue, or acts beyond reliable oversight.[5,35]

Operational Criteria for Agentic Labs

Autonomy levels describe the scope of what a system is allowed to do, whereas operational criteria evaluate how well it performs within that scope. To evaluate whether a laboratory system reliably translates reasoning into action, we propose four operational criteria (Figure 3e):

- **Coordination capacity** measures the diversity of models, tools, instruments, data sources, and workflows that a system can dynamically integrate within a campaign.
- **Scientific generativity** measures a system's ability to propose original, useful, and testable questions, hypotheses, plans, or alternative explanations — including whether it can frame the right question and articulate a candidate mechanism, the largely tacit mental model of how a system works that scientists rarely write down — rather than only optimizing within a predefined search space. It captures a form of scientific creativity, but emphasizes ideas that can be evaluated experimentally.

- **Execution reliability** measures how consistently planned actions are carried out as intended, including failure detection, exception recovery, calibration-drift detection, and avoidance of silent deviations. It should also capture whether agent-generated outputs remain scientifically valid under downstream verification.
- **Human dependency** measures how much performance degrades as human intervention is reduced, including intervention for decision-making, error recovery, tacit knowledge, and physical execution, as well as the system's escalation competence: its ability to recognize when human input is needed, what kind of expertise is required, and when to stop acting independently. This criterion should specify which human expertise is being replaced or requested, because dependence on an undergraduate assistant, senior graduate student, postdoctoral fellow, PI, or external expert reflects different forms of autonomy. It should also capture whether the system can diagnose its own limits and ask the appropriate human for help at the right time.

These criteria can be estimated from dynamically invoked tools and knowledge sources, human corrections and escalations, and discrepancies between intended and executed actions. The ladder defines capability but not the evidence required to justify each level. A system reaches L4 only by meeting the Figure 3e criteria within a declared scope—including the assays, instruments, samples, and failure modes evaluated. L5 sets the highest bar, requiring evidence that capability improves with experience without degrading reliability or escalation competence, a standard not yet supported by current evaluation frameworks. This limitation extends beyond laboratories. Across health AI, competencies are often declared without specifying the evidence required to support them, while evaluations remain focused on narrow, non-clinical tasks that rarely reflect deployment. Claims of autonomy should therefore be tied to explicit, scope-defined evaluation criteria from the outset.

These criteria could also be operationalized as benchmark tasks, such as detecting infeasible experiments, predicting protocol failure from partial context, reconstructing provenance, distinguishing biological signal from batch artefacts, and recommending when to repeat, revise, stop, or escalate. For example, a failed-protocol triage benchmark could provide an agent with a planned protocol, reagent-lot metadata, instrument calibration history, partial telemetry, images of intermediate states, and noisy readouts, then score whether the agent correctly classifies the outcome as valid, artifact-driven, failed but recoverable, requiring repetition, requiring protocol revision, or requiring escalation to a human. Scored against a fixed record, this only tests classification on a frozen snapshot; cascading errors, resource depletion and long-horizon credit assignment appear only when each action changes the state the next one meets. Evaluation should move into simulated laboratories where actions consume a reagent lot, occupy instrument time or commit a sample, so an early misjudgement propagates as it would in a real campaign. Comparable environments already exist in clinical evaluation, where agents act inside simulated hospitals and administrative interfaces whose state evolves over many dependent steps.

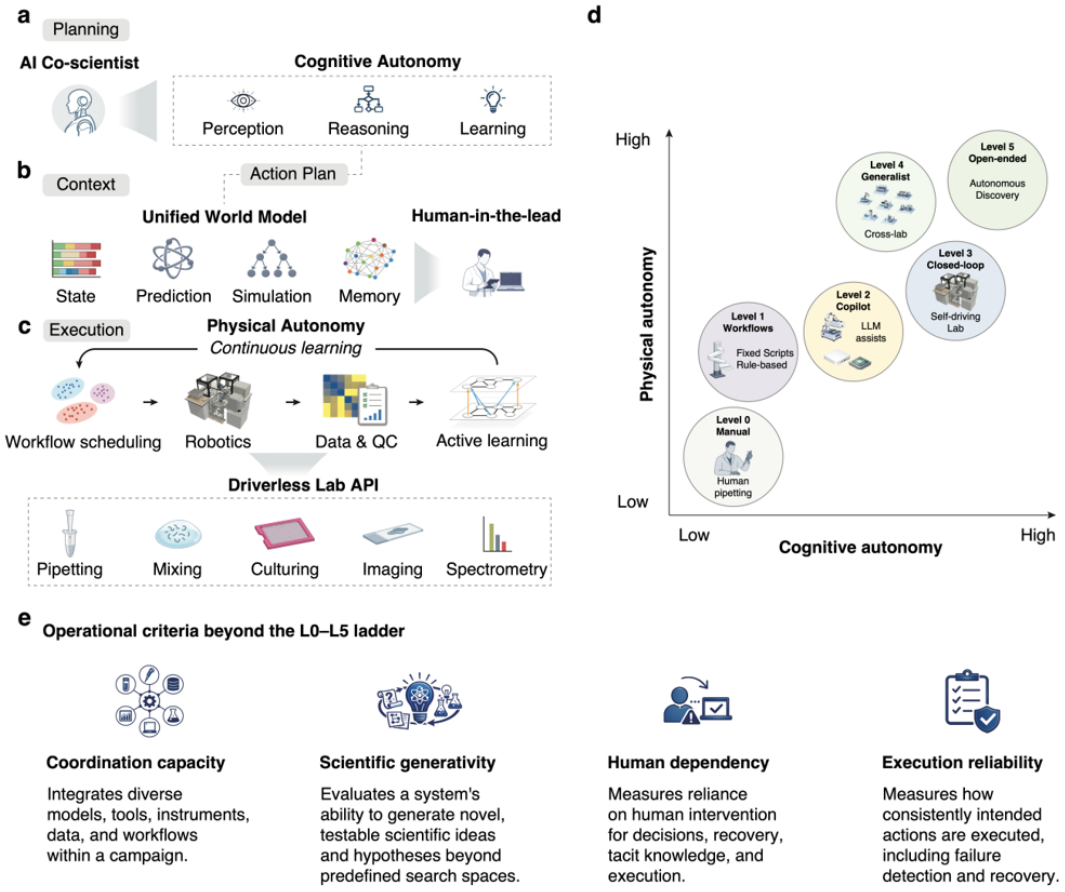

Figure 3. Architecture and evaluation criteria for agentic laboratory autonomy. (a) Cognitive autonomy describes the planning layer of an agentic laboratory, in which an AI co-scientist supports perception, reasoning, learning, and action-plan generation. **(b)** A unified world model provides context for action by integrating laboratory state, prediction, simulation, memory, and human-in-the-lead oversight. **(c)** Physical autonomy links plans to execution through workflow scheduling, robotics, data capture and quality control (QC), continuous learning, active learning, and driverless laboratory APIs for tasks such as pipetting, mixing, culturing, imaging, and spectrometry. **(d)** The L0–L5 autonomy landscape positions laboratory systems by increasing cognitive and physical autonomy, ranging from manual human pipetting and fixed rule-based workflows to AI copilots, closed-loop self-driving laboratories, cross-lab generalist systems, and recursively self-improving agentic laboratories capable of open-ended scientific discovery. **(e)** Operational criteria beyond the L0–L5 ladder include coordination capacity, scientific generativity, human dependency, and execution reliability, which assess whether a system can integrate diverse tools and workflows, generate novel testable ideas, reduce reliance on human intervention, and execute intended actions consistently. Together, these panels frame agentic laboratory autonomy as the integration of cognitive reasoning, world-model context, physical execution, recursive self-improvement, continuous learning, reliability, and human-in-the-lead control.

5. Challenges and Open Questions

While recent advances in AI, robotics, and laboratory automation have brought the vision of agentic laboratories closer to reality, fundamental challenges remain. Addressing these challenges will shape the next generation of autonomous discovery systems. These challenges span four interconnected themes: scientific and technical capabilities, human factors, governance and trust, and the practical deployment of agentic laboratory infrastructure.

Scientific and Technical Capabilities

The first set of questions concerns how agentic laboratories can move beyond today's bounded systems. Progress depends not only on stronger AI models, but also on reliable reasoning, physical

execution, long-horizon planning, and the ability to generate scientifically meaningful discoveries across diverse experimental domains.

(1) **Which Sciences Will Reach Autonomy First?**

The path toward agentic laboratories will be uneven because autonomy depends as much on experimental standardization, infrastructure and governance as on model capability. Purely computational fields such as mathematics are likely to reach autonomy sooner, because mathematical proofs can be formally verified computationally. However, agentic laboratory autonomy requires coupling reasoning to reliable physical experimentation. Agentic laboratories must integrate intent parsing, plan decomposition, tool invocation, continuous learning, hypothesis generation, and failure recovery to be closed in the same loop (Figure 4a).[4,13] Yet these capabilities will mature at different rates across scientific domains.

Autonomy will likely scale first where experiments are standardized, objectives are well defined, action spaces are bounded, and measurements are reproducible.[5,9,23] Synthesis of small-molecules, catalysis, thin-film materials, metal-organic frameworks[57], and some crystal-growth workflows could be domains for early adoption.[23,37] Extending autonomy from synthesis to functional properties and performance optimization is substantially more challenging.

More broadly, autonomous beamline operation, microscopy, quantum control, and materials-characterization workflows suggest that physics and materials laboratories may reach agentic autonomy first when the experimental state can be measured, modeled, and adjusted in real time.[83,87]

Biology will progress more heterogeneously: cell-free systems, cellular engineering workflows, protein engineering, and standardized perturbation screens may advance earlier, while primary samples, organoids, animal models, and clinical settings will remain harder because of biological variability, missingness, measurement noise, ethical constraints, and robotic complexity (Figure 4b).[21,22,30] Clinical laboratories may be a partial exception: although deployment requires safety, ethical and regulatory layers, many workflows already involve defined processes, quality-control standards and repeatable decision points, making them plausible early settings for agentic assistance in validation, documentation and exception monitoring.[35,53] Beyond discovery, agentic systems may also support translational and public-health workflows, including regulatory evidence generation for drug approval, pharmacovigilance, pandemic preparedness, and climate-adaptation monitoring, where the key challenge is not only designing new interventions but coordinating evidence, models, protocols, and decisions across institutions.

(2) **Can AI Discover Beyond Existing Scientific Paradigms?**

Can autonomous systems generate genuinely disruptive scientific discoveries, or only optimize within existing paradigms?[4,52] Scientific goals are rarely as clean as benchmark rewards. Whether optimizing a material property, testing a mechanism, or exploring a poorly understood phenomenon, the stated goal sits within broader constraints: cost, interpretability, safety, feasibility, reproducibility, and scientific value, which may shift with evidence.[35,53] Agentic systems must therefore represent goals as evolving scientific intentions rather than fixed objective functions.[4] Many current systems are well suited to interpolation or local search, but transformative discovery often requires recognizing anomalies, reframing questions, or pursuing observations weakly represented in prior data.[50] Evidence that papers and patents have become less disruptive over time raises a caution: more automation does not necessarily imply more conceptual novelty.[46] More fundamentally, agents confined to automatable, already-characterized experiments search a small and biased slice of experimental space, reproducing well-trodden biology rather than probing the larger space of experiments no platform yet performs; expanding what can be executed may matter as much as improving what can be reasoned.

(3) **Can Agentic Systems Remain Reliable Beyond Predefined Workflows?**

Another open question is how agentic systems can remain reliable outside narrow, well-constrained workflows (supplementary note 1).[5] Real laboratories are exception-heavy:

instruments drift, reagents vary, samples degrade, protocols are adapted locally, and silent deviations can invalidate conclusions. Many assays lack validated, machine-readable ground truth, especially for rare or previously unobserved biological behaviors.[52] Even experienced researchers may interpret the same data differently, as shown by multi-analyst studies.[51]

A related failure mode is that scientific generative models may produce outputs that are formally plausible but physically invalid, unstable, chemically inconsistent, biologically implausible or synthetically inaccessible. This differs from ordinary LLM hallucination: the output may look scientifically structured while failing downstream physics, chemistry, biological or experimental-validity checks. A subtler variant gives the right answer for the wrong reasons. In clinical evaluation, models with high scores have relied on reasoning that collapses when the question is perturbed, revealing pattern matching rather than inference. The laboratory analogue is an agent that proposes a sound experiment for unsound reasons, failing once conditions leave the training distribution. Validation should therefore probe reasoning stability under perturbation: reordering steps or substituting an equivalent reagent should not change a recommendation, only its justification where warranted.

These challenges become more severe when agents act on physical systems, because errors can consume expensive samples, create safety risks, or generate misleading data at scale.[35] Agentic laboratories will therefore need domain-specific validators, verification pipelines, and reward or alignment mechanisms that penalize outputs failing feasibility, safety or experimental-validation constraints. Even when one workflow succeeds, the next informative experiment may require a different assay, instrument, sample type, or customized protocol. The central challenge is therefore not whether individual workflows can be automated, but whether reliability can be preserved as agents move across workflows, contexts, and experimental regimes (Figure 4a).[5,35]

(4) **How Can Many Agents Collaborate Over Long-Horizon Campaigns?**

Reaching L4 and especially L5 autonomy will require agents that can sustain exploration and foundational building over long horizons—days to months—and, plausibly, not a handful but hundreds of agents working together. Genuine breakthroughs are rarely achieved in one shot or a few shots; they are built cumulatively, with each result standing on those before it, much as the members of a large human research group divide and conquer while building on one another's work. Current agentic systems already demonstrate elements of this collaborative, cumulative potential [13,28,29,49,59] but remain limited by hallucination and by coordination failures that compound over long horizons: uncertainty accumulates across multi-step interventions, early errors and artifacts propagate downstream, credit assignment is hard because a late failure may originate in a much earlier choice, and campaign state exceeds any single context window and must be externalized and continuously updated.[16,35] This is also the regime in which embodied agents are weakest, since robotics foundation models and vision-language-action systems perform best on short-horizon tasks rather than open-ended workflows.[60–62]

A leading response is to decompose the work across a multi-agent system of specialized agents—planners, critics, tool routers, analysts, and safety or validity checkers—enabling parallelism and adversarial critique, though this introduces its own failure modes, including cascading errors in which one agent's fabricated claim becomes another's trusted input, consensus or deadlock under disagreement, and diffusion of provenance and accountability.[15,35]

Whether decomposition actually improves end-to-end performance also deserves scrutiny.Optimizing each component need not optimize the pipeline: in clinical settings, systems built from individually best components underperform end-to-end, because stage-level metrics do not compose. Whether this holds in laboratories is untested, but its causes — sequential dependence, error propagation, per-stage metrics — are all present. Multi-agent designs should be judged end-to-end against a strong single-agent baseline, with the number of agents treated as a cost, not a proxy for capability. At this scale, physical resources—instruments, reagents, samples, and compute—are finite and shared, so coordinating many agents across a long campaign becomes an explicit scheduling and throughput-optimization problem; ultimately the objective is to convert bounded

resources and time into the most scientific discovery, motivating client-side and system-level agent optimization such as model allocation and speculative execution.[80,81] Progress therefore depends on multi-agent coordination mechanisms that are robust to idiosyncratic, per-agent errors, anchored by a persistent shared world model with provenance and human-in-the-lead checkpoints, while jointly optimizing scientific throughput under resource and time constraints.[35,36,40]

Governance, Reproducibility, and Trust

Increasing laboratory autonomy also raises broader questions of safety, transparency, and scientific trust. Agentic laboratories must therefore be designed with governance mechanisms that ensure reproducibility, responsible deployment, and appropriate oversight as autonomy increases. Because an agentic harnessing layer connects model reasoning to physical execution, misuse guardrails belong inside that layer rather than bolted on after deployment: permissioned tool access, synthesis and pathogen screening at the point of order, audit logs, human approval for high-risk actions, and policies requiring agents to pause or escalate. These are the same provenance and escalation mechanisms reproducibility already requires.

(1) **What Standards Are Needed for Reproducible Agentic Labs?**

Reproducibility should be treated not as a downstream reporting requirement, but as a design criterion for agentic laboratories (Figure 4d).[52] Agentic systems could improve reproducibility by enforcing structured workflows, recording provenance, tracking calibration, and standardizing metadata. However, they could also create new reproducibility failures if decisions depend on opaque model versions, hidden prompts, changing toolchains, unlogged instrument states, or proprietary cloud infrastructure. Reproducibility must include not only the physical protocol, but also the computational and agentic decision process: model versions, software environments, prompts, tool calls, intermediate observations, uncertainty estimates, and human overrides.[35,36,53,64] Replicability should also be treated as a separate goal: agentic systems should help determine whether a result persists across independent runs, instruments, laboratories, datasets, reagent lots, and experimental contexts, rather than merely reproducing the same workflow in the same setting.

As agents increasingly participate in literature synthesis, planning, execution, analysis, and interpretation, the scientific record may extend beyond the final paper. Future publications may need to include machine-readable logs of raw data, protocols, model interactions, candidate hypotheses, human decisions, failed attempts, and uncertainty estimates alongside the narrative account of goals, results, and conclusions. Supporting this vision will require interoperable workflow representations, standardized metadata, end-to-end provenance tracking, and machine-readable descriptions of calibration state, maintenance history, resource availability, software versions, and experimental environments (Figure 4b).[35,36]

Automation could make data and discovery sharing part of the workflow itself by exporting standardized, machine-readable records of protocols, raw data, analysis code, model interactions, failed attempts, human decisions, uncertainty estimates, and decision histories. Scientific publishing may evolve from polished narratives alone toward process-level disclosure, with journals asking authors to share agentic logs supporting scientific claims and reviewers using agentic tools to inspect them for missing controls, hidden assumptions, or reproducibility risks.

(2) **How Can Reproducibility Scale Across Networked Agentic Laboratories?**

Over the medium term, agentic laboratories are likely to become networked rather than monolithic, with reproducibility emerging as one of the clearest beneficiaries.[33,34,38,39] A single laboratory is unlikely to contain every model, instrument, dataset and expertise needed for general scientific autonomy.[3] Instead, multiple laboratories, cloud platforms, self-driving systems and human communities may share representations, protocols, calibration standards, provenance and learned skills, allowing reproducibility to scale across networks rather than being re-established in each laboratory.[33,34,38,39] Early work on networking autonomous materials exploration systems through transfer learning, NSF's programmable cloud laboratory efforts and SDL consortia point

toward autonomy and reproducibility developing as shared infrastructure rather than isolated facilities.[14,33,38,39]

In academia, this trajectory is likely to emerge bottom-up, as small, useful and easily adoptable autonomous modules gradually accumulate shared standards and interoperability across heterogeneous laboratories. Reproducibility would then become a measurable system property, evaluated through cross-laboratory workflow transfer, standardized API execution, reduced human intervention per experimental cycle and publication of machine-readable workflows with end-to-end provenance. However, cross-lab transfer will require more than data sharing: hardware configurations, reagent lots, calibration protocols, software versions, environmental metadata and human interventions must also be documented, since batch effects and fragile inter-instrument handoffs can limit transfer even when data are shared.[35,36]

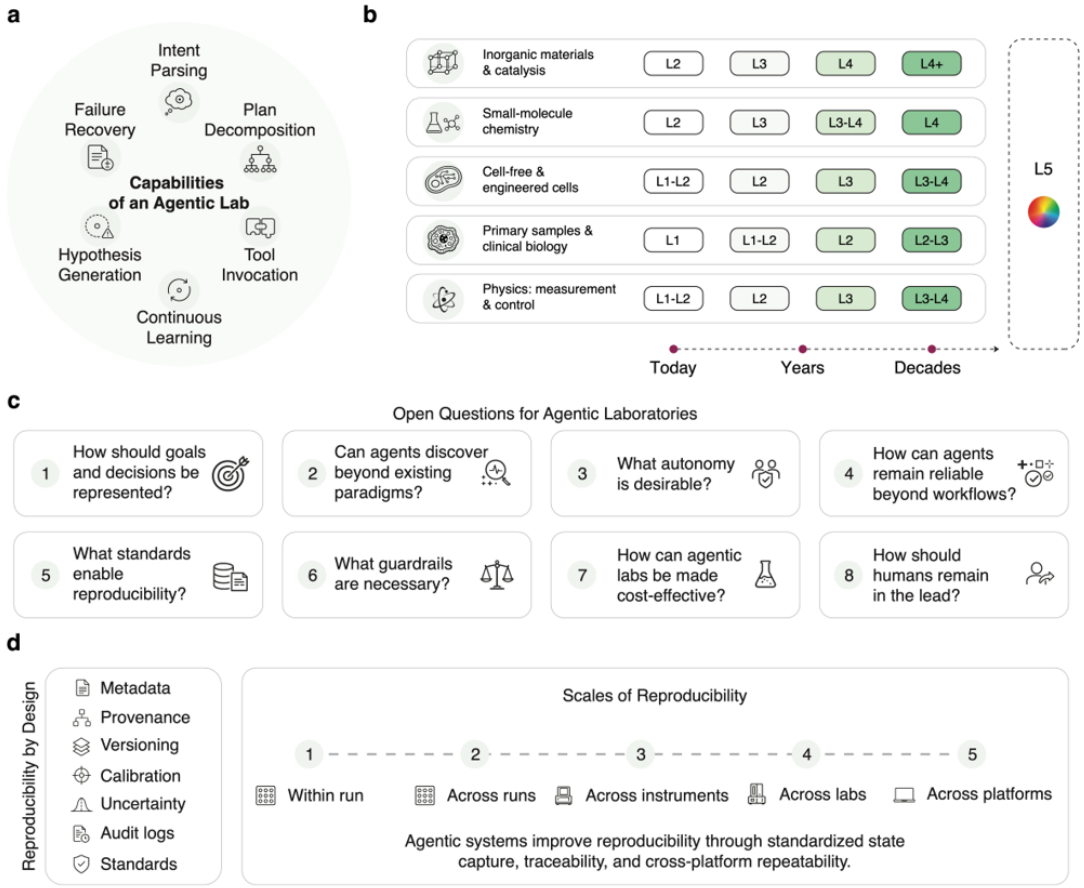

Figure 4. Functional capabilities and future directions of agentic laboratories. (a) Core capabilities of agentic laboratories, including intent interpretation, planning, tool use, learning, hypothesis generation, and recovery from failures. **(b)** Conceptual trajectory of laboratory autonomy across scientific domains. Most current systems operate at L1–L3 autonomy. Near-term advances are expected in areas such as inorganic materials, catalysis, and small-molecule chemistry, where workflows are relatively standardized. Progress in biology is likely to be more uneven, with cell-free and engineered-cell systems advancing earlier than primary-sample and clinical applications. Over time, some domains may approach cooperative L4 or L4+ autonomy, whereas L5 open-ended autonomy remains speculative. **(c)** Key open questions for agentic laboratories. **(d)** Reproducibility by design. Metadata, provenance, versioning, calibration, uncertainty estimates, audit logs, and standards enable reproducibility to be assessed within and across runs, instruments, laboratories, and platforms.

Infrastructure and Deployment

Even if agentic laboratories become scientifically capable and responsibly governed, widespread adoption will ultimately depend on practical infrastructure and economic viability. Interoperable

laboratory systems, scalable interfaces, and sustainable deployment models may become as important as advances in AI models themselves.

(1) **Will Infrastructure Become the Main Bottleneck for Agentic Laboratories?**

The bottleneck may therefore shift from model capability alone to the combined challenge of trustworthy models, interoperable infrastructure, and reliable physical execution.[35,36] If agentic systems are to plan beyond fixed workflows, they will need representations that connect actions to expected changes in the underlying scientific system, while also exposing uncertainty when those predictions leave the regime supported by data.4 This makes world models especially important for physical autonomy, where the agent must reason not only about what experiment is scientifically interesting, but also about how the system, instrument, or sample is likely to evolve once an action is taken.[54,55] If foundation models and scientific predictors continue to improve, the limiting factors will increasingly be interfaces, standards, governance, uncertainty-aware control, and physical reliability.[35,36] Conversely, if progress in scientific foundation models slows, model capability may again become the limiting factor..[7,8] One actionable first 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 biology, chemistry, and materials workflows (supplementary note 4).

(2) **How To Make Agentic Laboratories Cost-Effective?**

The economics of this transition will also matter. Autonomous experimentation is valuable only if the scientific gains justify the resources consumed. High-throughput systems can reduce labor and accelerate iteration, but they can also generate low-value data, consume costly reagents, tokens and increase infrastructure and energy demands.[79] Maintenance of high-mix robotic systems may add to the cost substantially. Cost-effectiveness should therefore be treated as a design criterion rather than an afterthought. However, early investment may not pay off immediately; its value may lie in making laboratories more ready to exploit future advances in models, robotics, sensors, and interoperable infrastructure. The ultimate test is not whether a system can generate more hypotheses, but whether it can shorten the path from discovery to validated intervention, a process that has historically taken decades.[47]

Human Factors and Scientific Practice

Agentic laboratories will reshape not only how experiments are performed, but also how scientists work, collaborate, and are trained. The central challenge is to determine which responsibilities should remain human-led while preserving scientific creativity, judgment, institutional knowledge, and accountability.

(1) **What Form of Human–AI–Robot Collaboration Is Realistic in the Near Term?**

The most defensible near-term forecast is not L5 fully autonomous laboratories, but L4 cooperative laboratories.[3,35] General-purpose robot scientists may not come to existence soon.[4] Early industrial value may come from mundane but economically meaningful tasks: sample transfer, instrument setup, adaptive workflow routing, standardized assay execution, surgical preparation, automated quality control, and exception monitoring.[13,35,36] These tasks may appear limited individually, but together they create the operational scaffolding and data needed for broader autonomy.[35,36]

(2) **Who Owns the Scientific Memory of Agentic Laboratories?**

A related governance challenge is ownership of scientific interaction data. As researchers increasingly use foundation-model interfaces for coding, analysis, experimental planning and scientific reasoning, universities may externalize not only data, but also unpublished workflows, methodological judgment and institutional scientific memory.[18] Institutionally governed agentic layers or model wrappers could keep prompts, experimental context, reasoning traces, protocols and provenance within local research ecosystems while still allowing researchers to use advanced models. For clinical and public-health applications, these systems will also need secure deployment environments that can handle regulated data such as electronic medical records, including privacy-

preserving access controls, audit logs, data-use governance, and compliance with institutional and government security standards such as the Federal Information Security Modernization Act (FISMA) where applicable.

In the long term, academic advantage may depend not only on access to models or robots, but on whether institutions can formalize faculty expertise, experimental reasoning and domain protocols into machine-readable infrastructure that compounds across disciplines. One possible response is a consortium of academic institutions that builds institutionally governed foundation models for science rather than relying entirely on commercial AI systems.

Agentic systems may also reshape scientific communities. Today, knowledge is distributed across papers, conferences, informal conversations, and individual memory; no scientist can read or retain more than a fraction of a field. Agentic systems may change this balance by rapidly synthesizing published literature, methods, datasets, and experimental records, raising new questions about whether scientific exchange will remain centered on human meetings and papers or increasingly occur through shared agentic memory systems.

(3) **What Roles Will Humans Play as Science Becomes Increasingly Autonomous?**

Finally, agentic laboratories must remain human-led. Current AI discourse often uses "human-in-the-loop," implying that humans intervene when machines fail.[17] In scientific settings, we argue that a better principle is human-in-the-lead.[17,35] Humans should remain responsible for framing objectives, setting constraints, interpreting surprising results, deciding when optimization has become misleading, and governing ethical or societal risk.[17,35] These roles depend not only on prediction, but on intuition, creativity, scientific taste, accountability, and judgment about meaning, values, and acceptable risk. Agentic systems may handle more routine integration, execution, monitoring, and documentation, but science still requires humans to decide which questions matter, when ambiguous observations deserve attention, and when a technically efficient path is scientifically or ethically inappropriate.[4,17]

The scientist's job may therefore shift rather than disappear. Freed from some routine equipment operation, repetitive execution, and workflow translation, scientists may spend more time reading, imagining, inventing, testing a broader range of ideas and creating new experimental tools or new types of measurements. This could be especially valuable because human intuition and scientific taste remain central to recognizing important questions, interpreting ambiguous results, and deciding which unexpected observations are worth pursuing. By reducing time spent on routine execution, agentic laboratories could give scientists and trainees more opportunity to develop these higher-level forms of judgment. Lowering the cost of trying an idea could make scientists more exploratory and willing to pursue unconventional hypotheses. However, greater distance from the bench may also reduce exposure to serendipitous observations and practical details that shape interpretation, a problem already familiar in large laboratories where principal investigators can become separated from day-to-day experimental practice. A related risk is that, as autonomous laboratories become more capable, they may weaken the human training pipeline needed to use them well. If early agentic systems take over tasks once performed by undergraduate or graduate research assistants, they may also displace the apprenticeship stages through which scientists learn experimental judgment. Laboratories could become more autonomous while fewer humans acquire the practical knowledge needed to interpret their outputs, set goals, or intervene when systems fail.

PI–trainee relationships may also change. If agentic systems reduce the need for trainees as manual labor, and if trainees can use agents to execute and analyze more independently, mentorship may shift from the mechanics of laboratory work toward scientific taste, strategy, interpretation, ethics, and judgment. This could improve training if institutions deliberately preserve opportunities for trainees to understand experimental details rather than only supervise automated systems. Agentic laboratories may also reshape scientific training: systems such as LabOS could provide protocol verification, real-time feedback, and context-aware assistance, but institutions must ensure that trainees still develop judgment, independence, and tolerance for uncertainty (supplementary note 3).[40,45,46]

Whether this transition creates more or fewer scientists will depend on institutional incentives and resource constraints. If scientific productivity becomes cheaper, Jevons-like effects may increase demand for scientific work because more questions become tractable. If funding, infrastructure, or attention remain fixed, automation could instead concentrate output among fewer laboratories. The central policy question is therefore not only how much work agentic systems can automate, but whether they broaden participation in discovery or narrow it.

If designed around reproducibility, safety, and human leadership, agentic laboratories could transform automation from a tool for faster execution into infrastructure for asking more ambitious questions and driving discoveries that would otherwise remain experimentally out of reach.

Acknowledgments: We thank 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.

Conflicts of Interest: Princeton University and Stanford University have filed patent applications related to this work.

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).
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).
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. Ryan, R. M. & Deci, E. L. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *Am. Psychol.* **55**, 68-78 (2000).
42. 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).
43. Limeri, L. B. et al. Development of the Mentoring in Undergraduate Research Survey. *CBE Life Sci. Educ.* **23**, ar26 (2024).
44. Hidi, S. & Renninger, K. A. The four-phase model of interest development. *Educ. Psychol.* **41**, 111-127 (2006).
45. 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).

46. Park, M., Leahey, E. & Funk, R. J. Papers and patents are becoming less disruptive over time. *Nature* **613**, 138-144 (2023).
47. Drucker, D. J. Mechanisms of action and therapeutic application of glucagon-like peptide-1. *Cell Metab.* **27**, 740-756 (2018).
48. Leeman, J. et al. Challenges in high-throughput inorganic materials prediction and autonomous synthesis. *PRX Energy* **3**, 011002 (2024).
49. 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).
50. 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).
51. Aczel, B. et al. Consensus-based guidance for conducting and reporting multi-analyst studies. *eLife* **10**, e72185 (2021).
52. Errington, T. M., Denis, A., Perfito, N., Iorns, E. & Nosek, B. A. Challenges for assessing replicability in preclinical cancer biology. *eLife* **10**, e67995 (2021).
53. Engel, J. et al. Project Aria: A New Tool for Egocentric Multi-Modal AI Research. Preprint at http://arxiv.org/abs/2308.13561 (2023).
54. Fung, P. et al. Embodied AI Agents: Modeling the World. Preprint at http://arxiv.org/abs/2506.22355 (2025).
55. 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).
56. Shi, L. et al. Qumus: Realization of An Embodied AI Quantum Material Experimentalist. Preprint at http://arxiv.org/abs/2605.18407 (2026).
57. 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).
58. 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).
59. 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).
60. Zitkovich, B. et al. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. *Proc. Mach. Learn. Res.* **229**, 2165-2183 (2023).
61. Kim, M. J. et al. OpenVLA: An Open-Source Vision-Language-Action Model. *Proc. Mach. Learn. Res.* **270**, 2679-2713 (2025).
62. 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).
63. OpenAI. Introducing ChatGPT. https://openai.com/index/chatgpt/ (2022).
64. 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).
65. Skeggs, L. T. Jr. An automatic method for colorimetric analysis. *Am. J. Clin. Pathol.* **28**, 311-322 (1957).
66. 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).
67. 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).
68. Bradshaw, G. F., Langley, P. W. & Simon, H. A. Studying scientific discovery by computer simulation. *Science* **222**, 971-975 (1983).
69. 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).
70. 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).
71. Avsec, Ž. et al. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **18**, 1196-1203 (2021).

72. Dalla-Torre, H. et al. Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nat. Methods* **22**, 287-297 (2025).
73. Cui, H. et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat. Methods* **21**, 1470-1480 (2024).
74. U.S. Department of Energy. Genesis Mission National Science and Technology Challenges. https://www.energy.gov/documents/genesis-mission-science-and-technology-challenges (2026).
75. 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).
76. Kusne, A. G. et al. On-the-fly closed-loop materials discovery via Bayesian active learning. *Nat. Commun.* **11**, 5966 (2020).
77. National Highway Traffic Safety Administration. Levels of Automation. https://www.nhtsa.gov/document/levels-automation (2022).
78. Theodoris, C. V. et al. Transfer learning enables predictions in network biology. *Nature* **618**, 616-624 (2023).
79. 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).
80. Ye, N., Ahuja, A., Liargkovas, G., Lu, Y., Kaffes, K. & Peng, T. Speculative actions: A lossless framework for faster AI agents. International Conference on Learning Representations (ICLR, 2026).
81. Hua, W. et al. AgentOpt v0.1 technical report: Client-side optimization for LLM-based agent. Preprint at http://arxiv.org/abs/2604.06296 (2026).
82. Krenn, M., Landgraf, J., Fösel, T. & Marquardt, F. Artificial intelligence and machine learning for quantum technologies. *Phys. Rev. A* **107**, 010101 (2023).
83. Alexeev, Y. et al. Artificial intelligence for quantum computing. *Nat. Commun.* **16**, 10829 (2025).
84. 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).
85. Moon, H. et al. Machine learning enables completely automatic tuning of a quantum device faster than human experts. *Nat. Commun.* **11**, 4161 (2020).
86. 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).
87. Cao, S. et al. Automating quantum computing laboratory experiments with an agent-based AI framework. *Patterns* **6**, 101372 (2025).
88. Li, S. et al. Large language model-assisted superconducting qubit experiments. Preprint at https://doi.org/10.48550/arXiv.2603.08801 (2026).
89. Wigley, P. B. et al. Fast machine-learning online optimization of ultra-cold-atom experiments. *Sci. Rep.* **6**, 25890 (2016).
90. Kalantre, S. S. et al. Machine learning techniques for state recognition and auto-tuning in quantum dots. *npj Quantum Inf.* **5**, 6 (2019).
91. Zwolak, J. P. et al. Autotuning of double-dot devices in situ with machine learning. *Phys. Rev. Appl.* **13**, 034075 (2020).
92. 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).
93. Reinschmidt, M., Fortágh, J., Günther, A. & Volchkov, V. V. Reinforcement learning in cold atom experiments. *Nat. Commun.* **15**, 8532 (2024).
94. Ni, X., Zhao, H.-H., Wang, L., Wu, F. & Chen, J. Integrating quantum processor device and control optimization in a gradient-based framework. *npj Quantum Inf.* **8**, 106 (2022).
95. Ai, H. & Liu, Y.-X. Scalable parameter design for superconducting quantum circuits with graph neural networks. *Phys. Rev. Lett.* **135**, 040601 (2025).
96. Peplow, M. Robot chemist sparks row with claim it created new materials. *Nature* https://doi.org/10.1038/d41586-023-03956-w (2023).

97. 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).
98. Masubuchi, S. et al. Autonomous robotic searching and assembly of two-dimensional crystals to build van der Waals superlattices. *Nat. Commun.* **9**, 1413 (2018).
99. Thomas, J. C. et al. Autonomous scanning probe microscopy investigations over WS2 and Au{111}. *npj Comput. Mater.* **8**, 99 (2022).
100. Merchant, A. et al. Scaling deep learning for materials discovery. *Nature* **624**, 80-85 (2023).
101. 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).
102. Yang, J. et al. Zero-shot autonomous microscopy for scalable and intelligent characterization of 2D materials. *ACS Nano* **19**, 35493-35502 (2025).
103. Mandal, I. et al. Evaluating large language model agents for automation of atomic force microscopy. *Nat. Commun.* **16**, 9104 (2025).
104. Li, H. et al. Rapid and reliable thickness identification of two-dimensional nanosheets using optical microscopy. *ACS Nano* **7**, 10344–10353 (2013).
105. Han, B. et al. Deep-learning-enabled fast optical identification and characterization of 2D materials. *Adv. Mater.* **32**, 2000953 (2020).