



AI-Guided Self-Driving Laboratories for Advanced Materials Discovery

Bernard Mallia^{✉,1,2,3,*}, Rossana Caputo^{✉,3}

¹ Equinox Group, Kalkara KKR 1184, Malta

² Institute for Research and Improvement in Social Sciences, Kalkara KKR 1184, Malta

³ Mediterranean Institute for Innovation, Communications and Technology, Kalkara KKR 1184, Malta

Article History

Received: April 01, 2026

Accepted: August 26, 2026

Abstract

The discovery of new high-performance materials and alloys is often likened to finding a needle in a haystack because the design space of chemical compositions and processing conditions is astronomically large, while the discovery process is both labor-intensive and prohibitively expensive. This perspective outlines a research agenda for autonomous materials and metallurgy discovery platforms that integrate closed-loop machine learning, robotic experimentation, and established materials engineering principles to formulate, process, and evaluate candidate materials with minimal human intervention and iteratively optimize their properties through recursive experimentation. It explores how sequential learning methods, generative modeling, and hybrid model-based strategies can analyze multimodal experimental data to propose new material compositions and processing schedules for subsequent experiments. This approach has the potential to considerably accelerate the pace of discovery, from decades to mere weeks. Representation learning for material formulations, the injection of domain knowledge into decision-making, and the mechatronic design of robotics that can safely execute harsh processing are all explored as essential components in closing the autonomous loop. Surveyed exemplars demonstrate order-of-magnitude improvements in experimental throughput and discovery efficiency. Furthermore, this paper proposes a multidimensional framework for classifying and evaluating self-driving laboratories based on their levels of autonomy, complexity, and integration. The framework is intended to provide researchers and practitioners with a structured approach for benchmarking future developments in the field. The perspective concludes with an analysis of data infrastructure, simulation coupling, robustness, and human-machine collaboration, distilled into a structured research agenda whose pursuit would broaden adoption across energy, aerospace, and other sectors.

Keywords:

autonomous experimentation; Bayesian optimization; robotics; calculation of phase diagrams; materials discovery; metallurgy; machine learning; active learning; generative models; self-driving laboratory

1. Introduction

Modern technologies require materials with unprecedented combinations of properties, ranging from turbine and nuclear fusion alloys capable of withstanding extreme temperatures to conductive polymers for flexible electronics. However, the existing catalog of discovered materials represents only a minuscule fraction of the vast space of possible materials, indicating that many more remain to be discovered or, more appropriately, designed. In this

context, naïve random exploration, although historically demonstrated as a practical mechanism for discovery, is not an optimal strategy for expanding the materials catalog because of the underlying combinatorial explosion, particularly in light of recent technological advances and their anticipated near-term trajectory. High-throughput autonomous laboratories, or self-driving labs (SDLs) as they are at times referred to, integrate robotics with artificial intelligence (AI) (and potentially, within the next

* Corresponding Author:

Bernard Mallia, Equinox Group, Kalkara KKR 1184, Malta; Institute for Research and Improvement in Social Sciences, Kalkara KKR 1184, Malta; Mediterranean Institute for Innovation, Communications and Technology, Kalkara KKR 1184, Malta, Bernard.mallia@equinoxadvisory.com



© 2026 Copyright by the Authors.

Licensed as an open access article using a [CC BY 4.0 license](https://creativecommons.org/licenses/by/4.0/).

few years, even quantum computing—if not gate-based, then quantum annealing-based) to operate a closed-loop design–synthesize–test–learn cycle that uses recursive iteration to propose and execute experiments, assimilate data, and update hypotheses with minimal human oversight [1,2]. Early examples of such SDLs demonstrate dramatic acceleration and could be to metallurgy what AlphaFold [3] was to 3D protein structures. One exemplar, the Air Force Research Laboratory’s (AFRL) Autonomous Research System (ARES), demonstrated closed-loop optimization of carbon nanotube growth in which the platform designed, executed, and analyzed its own experiments orders of magnitude faster than conventional manual practice, sustaining daily experiment counts of the order of a hundred where manual workflows typically manage only one [2,4]. Polybot and A-Lab, two other exemplars, achieved fully closed-loop operation, discovering new compounds and optimizing polymer films at scale [5–7]. In light of these results, the central thesis of this perspective is that coupling principled decision algorithms with robust automation and metallurgical knowledge can transform materials development, taking it from an incremental trial-and-error process to one underpinned by systematic exploration at machine speed, with the natural bottlenecks becoming the speed at which the robots can undertake the testing and the physics-based constraints of physically producing the alloys. To better un-

derstand the capabilities and challenges of these varied platforms, it is useful to have a classification framework, and this is exactly what the next section addresses. This perspective synthesizes recent advances in autonomous laboratory platforms, proposes a novel multidimensional classification framework for self-driving laboratories, delineates the critical technical and infrastructural bottlenecks that must be surmounted if the vision of routine autonomous materials discovery is to be realized within the present decade, and distills these considerations into a structured research agenda for the field.

2. A Multidimensional Classification Framework for Self-Driving Labs

To better contextualize the wide typological spectrum of self-driving laboratories, this perspective proposes a three-axis classification framework based on Material System State (MSS), Hardware Autonomy Level (HAL), and Software Autonomy Level (SAL) that allows for a more refined comparison of platforms beyond their stated discovery goals than is currently possible. This framework can also be explored visually via a radar chart for comparative purposes, as illustrated in Figure 1 below.

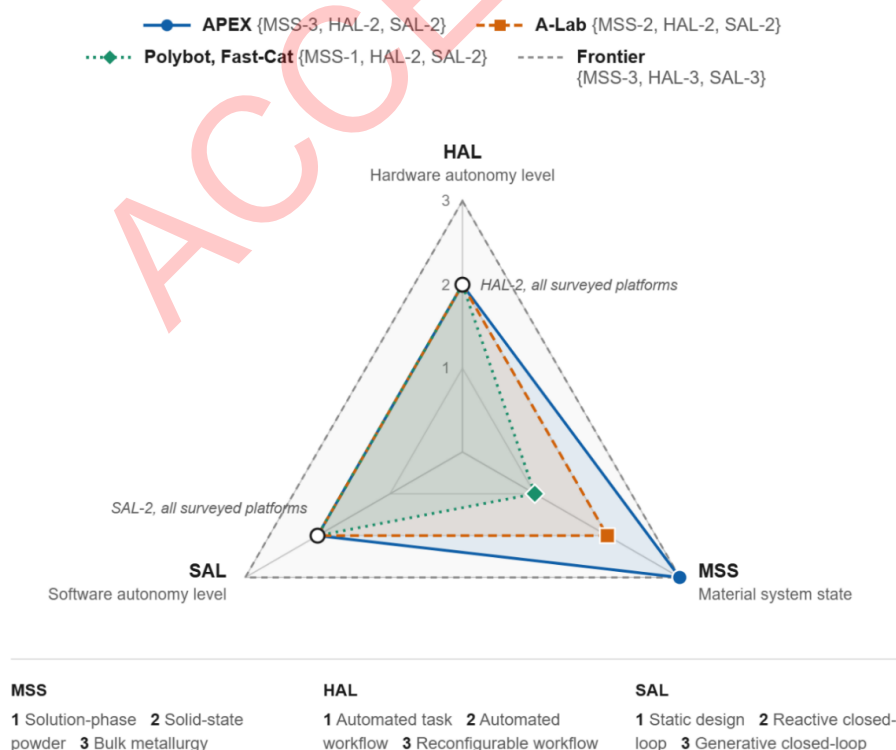


Figure 1: Cont.

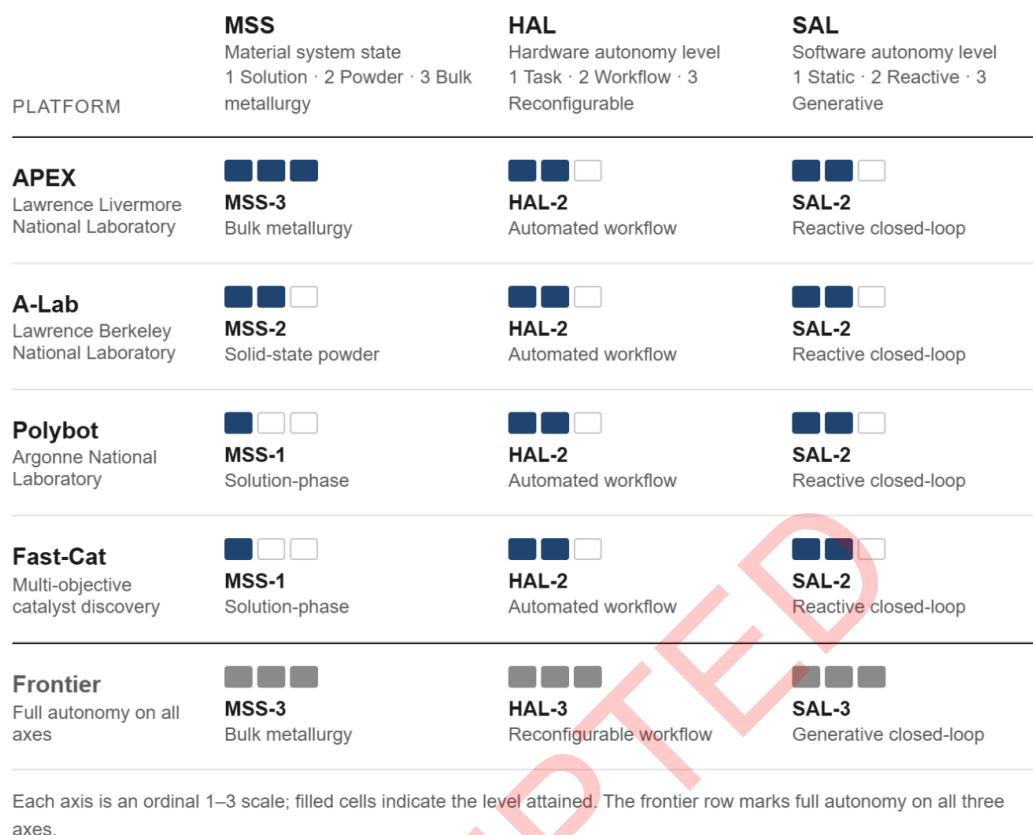


Figure 1: A multidimensional mapping of extant self-driving laboratories using the proposed taxonomy encompassing Material System State (MSS), Hardware Autonomy Level (HAL), and Software Autonomy Level (SAL). Platforms processing solid-state materials and alloys (e.g., A-Lab, APEX) face higher mechatronic complexity (MSS-2 and MSS-3), while most current platforms operate at reactive closed-loop software autonomy (SAL-2).

The proposal of a further taxonomy demands justification against those already in use. Existing classification schemes for autonomous experimentation are essentially single-axis frameworks. Coley et al. describe a continuum from automation to autonomy in chemical discovery, classifying systems according to the extent to which hypothesis generation, experimental design, and interpretation are delegated to machines [1,8]. Similarly, Hung et al. propose a six-level scale of laboratory autonomy, ranging from non-automated (L0) to fully autonomous (L5), explicitly modeled on the six-level taxonomy for driving automation established by the Society of Automotive Engineers (SAE) [9]. These schemes have real diagnostic value, but they compress into a single ordinal figure at least three properties that vary independently in practice: the physical state of the material system being handled, and hence the mechatronic difficulty of automating it; the degree of physical automation of the workflow; and the sophistication of the decision-making software. The MSS/HAL/SAL framework proposed here decouples these three dimensions. The distinction matters

because a solution-phase platform and a bulk-metallurgy platform may share identical autonomy ‘levels’ on a one-dimensional ladder while confronting utterly incommensurate engineering problems; conversely, two platforms of identical material scope may differ radically in whether their binding constraint is robotic integration or algorithmic ambition. The SAL axis proposed here is broadly commensurable with the upper rungs of the Hung et al. scale, and the HAL axis with its lower and middle rungs, but the MSS axis captures a source of difficulty absent from prior schemes altogether. The framework is therefore offered as an extension of, rather than a rival to, the existing ladders, in which their single ordinal is factored into its independent components.

2.1. Material System State (MSS)

The first axis is MSS, which describes the physical form of the material being processed and strongly correlates with the required mechatronic complexity. This axis can take on one of three values, which are subsequently explained:

1. MSS-1 (Solution-Phase) describes systems designed for liquid-phase synthesis and processing, such as those for polymer formulation or catalyst discovery (an example of this in practice is *Fast-Cat* [10]) that are able to benefit from mature liquid handling robotics and rapid in-line characterization;
2. MSS-2 (Solid-State Powder) describes systems that handle solid precursors for reactions, such as inorganic materials synthesis (with an illustrative example of this being Berkeley's *A-Lab* [5]) that require some degree of robotics for powder dispensing, mixing, and high-temperature furnace operations; and
3. MSS-3 (Bulk Metallurgy) describes systems designed for fabricating and processing bulk alloys or solid components, thus representing the highest level of complexity these systems can reach, and involve extreme temperatures, high-load mechanical testing, and advanced manufacturing techniques like additive manufacturing (with the illustrative example *par excellence* of this being Lawrence Livermore National Laboratory's (LLNL) *APEX* [11]).

2.2. Hardware Autonomy Level (HAL)

The second axis is HAL, which defines the degree of physical automation in the experimental workflow and can also take on three values, namely:

1. HAL-1 (Automated Task) systems are defined as systems that can handle a single experimental task, such as sample preparation or a specific characterization step, autonomously, but that still require manual transfer, handshaking, or handovers between the intermediate process steps in order to function as intended;
2. HAL-2 (Automated Workflow) systems are capable of handling multiple experimental modules, such as synthesis, processing and characterization, and are physically linked by robotic transport, handshaking or handovers, thereby enabling a complete synthesis-and-test sequence on a single sample without human intervention or oversight. Argonne's *Polybot* exemplifies this level on the HAL axis [7]; and
3. HAL-3 (Reconfigurable Workflow) platforms are able to autonomously execute diverse workflows by reconfiguring connections between modules or utilizing mobile robotics, thus moving beyond a fixed linear process. The mobile chemist developed by Burger et al. is an early example of this type of flexibility [12].

2.3. Software Autonomy Level (SAL)

The third axis is the SAL that the platform is capable of achieving, and it essentially classifies the sophistication of the "AI planner" that drives the decision-making loop. The three possible values on this axis are:

- SAL-1 (Static Design), where the system executes a predefined experimental array designed by a human or an offline optimization algorithm (as in experimental design) in such a way that the system loop fails to qualify for a closed-loop classification;
- SAL-2 (Reactive Closed-Loop), where the AI planner uses the results of the last experiment or batch to choose the next immediate experiment, typically using sequential learning strategies like Bayesian Optimization (BO) in order to optimize a known set of parameters, which is currently the most common strategy in extant self-driving labs; and
- SAL-3 (Generative Closed-Loop), which describes systems wherein the AI planner can modify the experimental campaign goals or expand the search space itself without needing human intervention or supervision, and entails generative models that can propose novel compositions or processing routes that were not initially conceived by human experts. This represents a shift from optimization to truly open-ended discovery, as, for instance, when an AI autonomously proposes an entirely novel crystal lattice or a hitherto unconceived alloying strategy that no domain expert would have hypothesized from the extant literature, along the same lines as the now-legendary AlphaGo move 37 [13]. A recent archetype of this capability is Microsoft's MatterGen, a diffusion-based model that can generate previously unconceived, stable inorganic crystal structures atom-by-atom based solely on target property prompts (e.g., desired magnetic density or bulk modulus) [14].

2.4. Platform Classification

Using this framework, a self-driving platform can be characterized by its three coordinates. For instance, *A-Lab* could be classified as a {MSS-2, HAL-2, SAL-2} SDL, denoting its capability for closed-loop, workflow-automated synthesis of solid-state powders. In contrast, *APEX* comes closer to a {MSS-3, HAL-2, SAL-2} SDL, highlighting the added complexity of its metallurgical focus. Extending this mapping to the remaining platforms surveyed herein: *Polybot* would qualify as a {MSS-1, HAL-2, SAL-2} SDL, reflecting its mastery of liquid-phase workflows under reactive closed-loop control, while *Fast-*

Cat sits likewise in the {MSS-1, HAL-2, SAL-2} SDL category, distinguished principally by its multi-objective Pareto-front emphasis rather than a more elevated autonomy tier. This 1–3 scale is deliberately more compressed than the SAE six-level taxonomy of driving automation, which spans Levels 0 through 5. SAL-3 is broadly analogous to SAE Level 5, representing full automation in which the system determines not only the optimal path through an existing design space but also the destination itself, without relying on a human-supplied map. This classification provides a clear snapshot of a platform’s capabilities and its position on the frontier of autonomous science.

3. AI and Machine Learning Frameworks for Autonomous Discovery

Closed-loop autonomous experimentation frames experiment selection as a sequential decision problem under uncertainty. BO, typically with Gaussian Process (GP) surrogates, has become a method of choice for sample-

efficient optimization in high-dimensional, expensive domains where GP posteriors quantify epistemic uncertainty, and acquisition functions (such as expected improvement and knowledge gradient) balance exploration and exploitation to select informative experiments [15], with extensions including heteroskedastic noise models and anisotropic kernels tailored to experimental physics. Beyond BO, evolutionary strategies and Reinforcement Learning (RL) offer alternatives for long-horizon, multi-step synthesis planning. A notable feature of recent hybrid systems is the integration of generative language models for recipe retrieval with graph search methods, such as A*, to accelerate convergence in nanochemistry, while foundation models trained on the scientific literature can support the cold start regime [5] through statistical bootstrapping strategies (Figure 2). These may also be combined with physics simulation engines acting as digital twins of sorts that formulate scientific hypotheses, narrow down the latent space of workable solutions, and finally conduct experiments to prove or disprove that hypothesis while embedding the results in the available knowledge base for incremental, evolutionary learning.

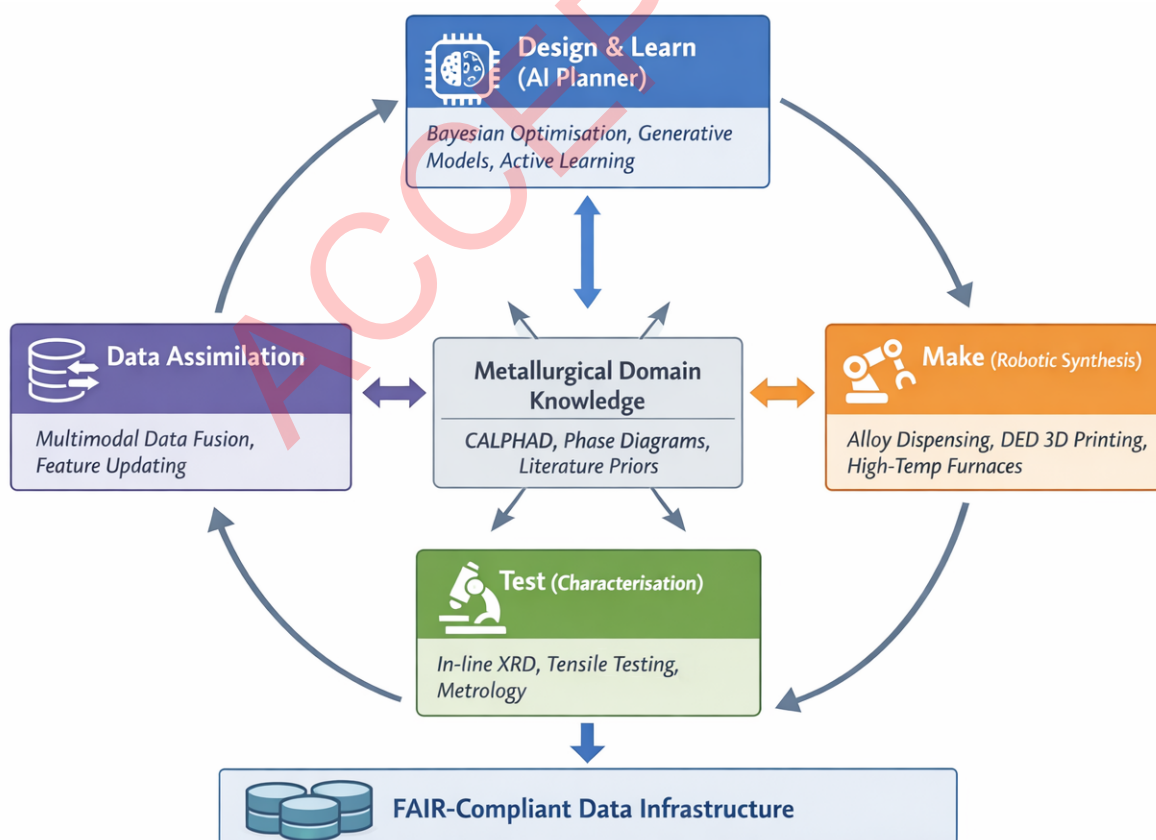


Figure 2: The core closed-loop Design–Make–Test–Learn (DMTL) architecture of an AI-guided self-driving laboratory for metallurgy. Domain knowledge is continuously injected into the AI Planner, which orchestrates robotic synthesis and automated characterization.

Representation is central to autonomous discovery. For inorganic solids and alloys, feature sets encompass elemental fractions, physically informed descriptors such as atomic size mismatch, valence electron concentration, tensile strength, conductivity, and the like, as well as outputs from physics engines such as phase fractions from Calculation of Phase Diagrams (CALPHAD), diffusion coefficients, and similar parameters, which then serve as inputs to surrogate models. For polymers and small molecules, graph-based encodings and molecular fingerprints are common, whereas for microstructures, image or diffraction embeddings via convolutional networks or scattering transformations are likely to be more useful, though this is not invariably the case. Multi-objective formulations are routine and help maximize strength subject to density and cost constraints, or co-optimize conductivity and defectivity, thus requiring Pareto-efficient acquisition strategies and constraint-aware active learning.

One of the cornerstones of this approach is injecting physics and prior knowledge into the system to avoid proposals that contravene what is possible in the realm of physics specifically, and the life sciences more generally. To this end, constraints derived from phase equilibria and the Gibbs phase rule prune infeasible regions, while physics- and life-sciences-informed engines made up of priors and simulators, such as CALPHAD, phase-field, and *ab initio*, are co-optimized with data-driven surrogates. This hybridization has underpinned recent autonomous inorganic synthesis campaigns, where computational screening over the Materials Project narrowed viable targets ahead of robotic synthesis [5], resulting in lower costs than would otherwise have been incurred.

4. Robotic Automation for Experimental Metallurgy and Materials

Self-driving labs are mechatronic systems integrating robotic manipulation, process equipment, and in-line characterization under software orchestration, where architectures commonly include multiple coordinated arms or gantries for logistics, synthesis modules for dispensing, reactors, additive manufacturing, thermal units such as furnaces with vacuum/inert atmospheres, rapid thermal processing, and sensors for tasks such as imaging, spectroscopy, diffraction, and mechanical testing. For example, Argonne's Polybot coordinates three robots for synthesis, processing, and logistics, with Python-based supervisors enabling 24/7 closed-loop operation [6, 7]. Lawrence Livermore's APEX for alloys integrates Directed-Energy Deposition (DED) metal additive manufacturing with automated post-processing and characterization, enabling rapid composition-process iteration [11].

Engineering for metallurgy imposes several non-trivial constraints, some of which include the handling of powders that melt at temperatures higher than 1500 °C, interlocked safety enclosures for lasers and hot zones, end-effectors with thermal shielding, and robust error detection and recovery [16]. Repeatability is a first-class goal to increase signal-to-noise for downstream Machine Learning (ML), whereas automated workflows minimize operator variability, enable in-line metrology, and elevate throughput by orders of magnitude relative to equivalent manual processes.

5. Incorporating Metallurgical Domain Knowledge into AI

Metallurgy contributes a dense *corpus* of priors ranging from phase diagrams, thermodynamics, and kinetics, through to process heuristics, and these need to be programmatically injected into autonomous decision loops. The lowest common denominator of priors takes the following form (see also Figure 3):

- (i) Constraint-based planning excludes immiscible or deleterious phase fields and enforces application-specific constraints (such as density ceilings, weight constraints, and cost budgets) [17];
- (ii) Physics-based model integration couples CALPHAD predictions of phase stability and fraction with surrogates for properties, guiding the search toward microstructural states consistent with target strengthening mechanisms, with solution strengthening and precipitation hardening being canonical exemplars [18];
- (iii) Literature-derived priors, extracted by natural-language models, seed initial process windows and alloying ranges, making best use of prior knowledge [19];
- (iv) Multi-objective formulations encode performance trade-offs and penalties (e.g., criticality of elements), supporting Pareto-front mapping rather than single-point optimization [10];
- (v) Continuous knowledge-based updates ensure that experimental outcomes are deposited into machine-readable repositories and linked back to the literature, such that each successive campaign inherits an ever-richer corpus of prior evidence [20]; and
- (vi) Human-in-the-loop oversight, although not strictly required and potentially detrimental to process efficiency, can enable experts to guide exploration, incorporate domain-specific heuristics, and audit the underlying rationale, thereby enhancing trust and safety [2].

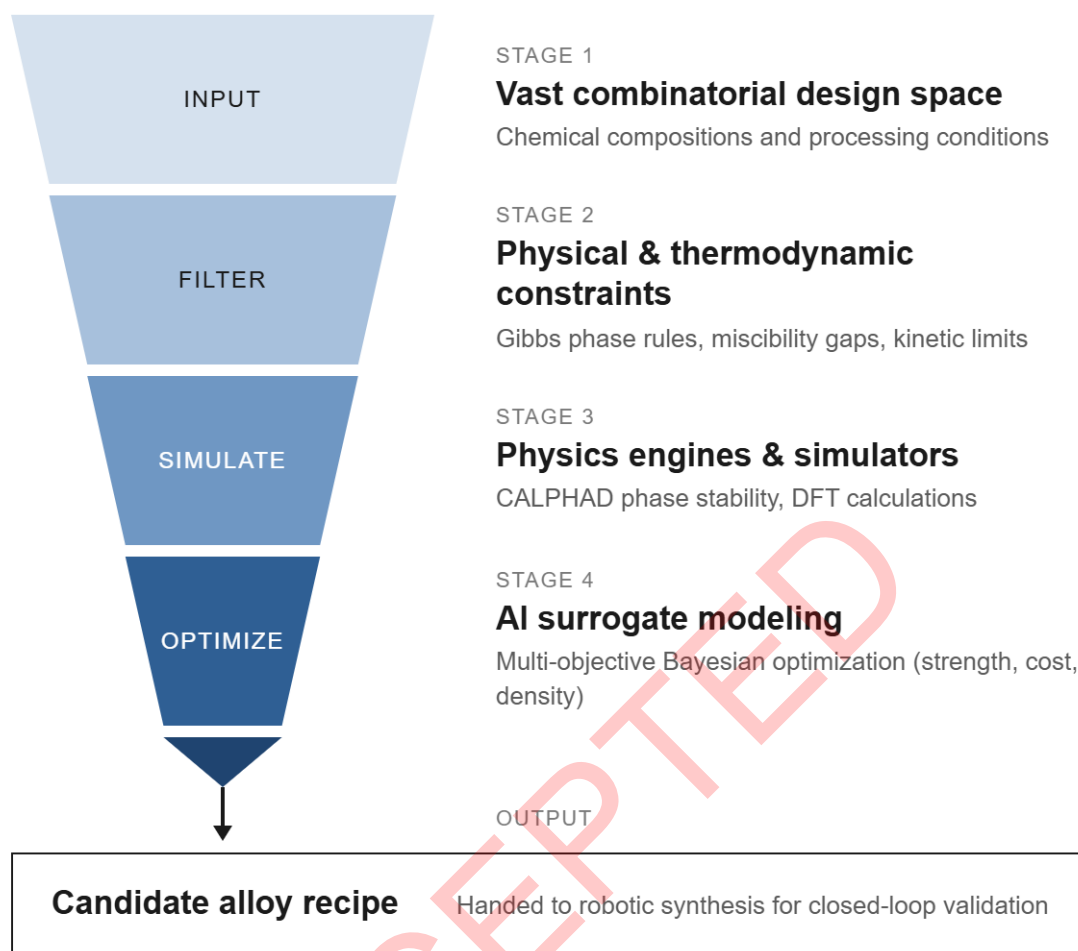


Figure 3: Diagrammatic representation of the integration of metallurgical domain knowledge into the AI decision loop, illustrating how physics-based constraints, literature-derived priors, and multi-objective formulations collectively guide the autonomous planner.

6. Self-Driving Labs Applications

6.1. Solution-Phase Platforms

Solution-phase platforms excel because of rapid feedback and the ease with which they enable automation. An illustrative example, Polybot, has demonstrated multi-parameter optimization of electronic polymer thin films (e.g., poly(3,4-ethylenedioxythiophene):poly(styrene sulfonate), or PEDOT:PSS), coordinating formulation, blade-coating, annealing, and in-line electrical and imaging metrology. Campaigns executed around 100 samples per day, each sample completing its end-to-end cycle of formulation, coating, annealing, and measurement in approximately 15 min, using multi-objective BO across seven-dimensional process spaces to elevate conductivity while suppressing defectivity [6,7]. These two figures are reconcilable because the workflow operates as a pipeline rather than as a strictly sequential process. The platform's stations, managed by coordinated robots, operate concurrently

on successive samples, such that daily throughput is determined by the residence time at the slowest station rather than by the sum of the durations of all stages. This principle is analogous to an assembly line, where takt time, rather than end-to-end lead time, determines the output rate. In catalysis, Fast-Cat integrates high-pressure flow reactors, online analytics, and autonomous planners to map Pareto fronts of yield versus selectivity, characterizing six ligands in five-day campaigns each, thereby drastically shortening months-long manual workflows while also validating scale-up fidelity [10]. Mobile robotic chemists further illustrate flexible autonomy by operating conventional lab equipment to optimize photocatalysis [12].

6.2. Autonomous Alloy Discovery

Autonomous alloy discovery introduces higher experimental complexity as it involves compositional spaces with at least five principal elements, thermal cycles

(such as solutionizing and aging), and mechanical processing. LLNL's APEX couples DED metal printing with robotic handling and characterization to iterate composition–process–property loops at scale, targeting drastic cycle-time reduction for novel alloys [11]. University initiatives focus on refractory multi-principal element alloys for ultra-high-temperature service, leveraging AI-guided exploration where human intuition is limited [21]. The demonstrated success of hybrid computational-experimental autonomy extends to identifying novel functional materials, including ternary alloys that exhibit anomalously high Curie temperatures [22], strongly underscoring the proficiency of ML-guided exploration for discovering materials with specific and advanced properties, and signaling the imminent potential for expanding full autonomy to intricately linked processing domains. The next logical frontier involves the autonomous discovery of optimal heat-treatment schedules, where an intelligent system could navigate the complex interplay of time and temperature to define ideal quench-temperature protocols and control precipitation kinetics with precision. A parallel and significant opportunity exists in autonomously optimizing manufacturing process parameters by determining, for example, the precise laser power and scanning strategies required to fabricate defect-free components through additive manufacturing, or the duration for which ionic bombardment should be applied in surface engineering.

7. Challenges and Future Directions

While the potential of self-driving laboratories is transformative, their widespread adoption and evolution from specialized platforms to routine scientific instruments hinge on overcoming several key, interconnected challenges that encompass data infrastructure, simulation integration, operational robustness, and the fundamental nature of autonomous discovery itself. Each challenge is discussed in turn.

7.1. Data Infrastructure and Interoperability

The full potential of AI in materials science can be unlocked only through large, diverse, high-quality datasets. Currently, data generated by autonomous labs often remains in bespoke, siloed formats, hindering cross-lab collaboration and the development of universal models. The primary challenge is the establishment of agreed-upon standardized schemas and ontologies for experimental materials science, following the Findable, Accessible, Inter-

operable, and Reusable (FAIR) data principles. Such standards would enable seamless data fusion from multiple labs, creating datasets of unprecedented scale and richness. This would, in turn, accelerate the development of powerful meta-learning and transfer learning models capable of generalizing across different material systems and experimental platforms [9].

7.2. Simulation Coupling and Digital Twins

Physical experiments, even when automated, remain a significant bottleneck in terms of time and cost. The tight coupling of experimental platforms with multi-fidelity simulations is a critical frontier. This involves creating “digital twins,” which may be thought of as high-fidelity virtual replicas of the self-driving lab that can run thousands of *in silico* experiments rapidly and at low total and incremental cost. The AI planner's role would evolve to arbitrate between a cheap virtual experiment and a costly but definitive physical one. This hybrid approach, where physics-based models like CALPHAD and Density Functional Theory (DFT) inform and are continuously corrected by *in situ* data, promises to maximize the information gained per unit of resource and time, thus guiding physical experimentation toward only the most promising and uncertain regions of the design space.

7.3. Transfer Learning and Platform Generality

Many current self-driving labs are highly specialized systems, representing a significant investment of time and capital to design and build for a specific task. A major challenge is reducing re-tooling costs and the “cold start” problem when applying a platform to a new domain, and we firmly believe the solution lies in developing more generalist AI agents that can leverage transfer learning. By pre-training on vast databases of literature and simulation data, these foundation models for materials science could adapt to a new experimental task with only a small number of fine-tuning experiments, thus drastically improving the agility and economic viability of autonomous experimentation. Of course, the risk of sabotage through data poisoning remains a very real concern, so transfer learning can only take place between trusted systems or in the presence of a trustless verification protocol.

7.4. Robustness, Unattended Operation, and Safety

For self-driving labs to achieve their promised throughput, they must operate reliably for extended periods with

minimal human oversight, which requires a leap in robustness and operational autonomy. Key research areas include developing sophisticated anomaly detection algorithms to identify experimental errors, sensor drift, or mechanical failures in real time. Beyond detection, platforms must also be capable of self-calibration to correct for drift and execute fail-safe recovery protocols to prevent damage or data loss. This makes achieving a true “walk-away” 24/7 operation an engineering and AI challenge that is fundamental to maximizing the return on investment for these platforms. A cognate concern, which merits explicit treatment, is that of safety and dual-use in that an autonomous system capable of navigating vast compositional spaces with minimal human oversight is, in principle, equally capable of identifying highly energetic materials, toxic precursors, or critical strategic alloys as of discovering benign structural metals. The scientific community would therefore be well-advised to develop software guardrails analogous to the content-filtering mechanisms now routinely deployed in large language models: guardrails that constrain the autonomous search to pre-approved regions of compositional and processing space, that are resistant to the adversarial removal of refusal and filtering behaviors through targeted ablation of safety-relevant model weights or activations (colloquially termed ‘ablation’ [23]), and that flag candidate materials meeting predefined hazard criteria for mandatory human review before synthesis proceeds.

7.5. Beyond Optimization to Open-Ended Discovery

Most current autonomous labs excel at optimizing a known set of parameters to maximize a target property, but a more profound challenge would be to enable open-ended discovery that can lead to the identification of truly novel materials or phenomena never conceived by human researchers. This requires moving beyond purely exploitative BO frameworks toward increasingly curiosity-driven objectives and generative models, which can be realized through system architectures that explicitly seek novelty or surprise, proposing experiments in sparsely populated regions of the feature space or generating entirely new classes of compositions to computationally mimic the process of serendipity and intuition-driven science. The transition toward this generative paradigm is already underway with the advent of property-conditioned diffusion models, such as MatterGen, which invert the traditional screening funnel by directly generating hypothetical, chemically valid materials tailored to specific constraints rather than merely sorting through known databases. The mathematical bridge between targeted op-

timization and curiosity-driven exploration is increasingly supplied by active learning frameworks, and in particular by methods such as Bayesian Active Learning by Disagreement (BALD), which directs the acquisition function toward regions of maximal epistemic uncertainty rather than maximal expected improvement, thereby institutionalizing, in algorithmic form, the very restlessness that has animated the most consequential experimental scientists in the history of science [24].

7.6. Human-AI Collaboration and Explainability

Ultimately, self-driving labs are tools that augment human scientists and that may, for narrowly circumscribed classes of experimental tasks, assume duties hitherto performed by them; the wholesale replacement of human scientific judgment, however, lies beyond the demonstrated capabilities of current platforms. The goal is not just to find new materials but to accelerate useful scientific understanding, laying the ground for better products. A significant challenge lies in developing eXplainable AI (XAI) that can articulate the rationale behind its experimental choices and help human scientists form new hypotheses and contribute their own expertise. The ideal future state, at least in the short and medium term, is a collaborative human-in-the-loop system where experts sit on the fringes of the automated pipeline but can query the AI’s reasoning, inject their own domain knowledge and intuition to steer exploration, as well as interpret the results to build deeper mechanistic insights, thereby ensuring that these powerful tools foster both discovery and comprehension [1,9].

7.7. Cost, Access, and Democratization

At present, platforms such as APEX or Polybot represent multi-million-Euro, highly bespoke investments, and are institutionally confined to national laboratories or a handful of well-capitalized research universities. This concentration of capability is antithetical to the spirit of open science and risks reproducing, at the level of experimental infrastructure, the same asymmetries that have long stratified access to computational resources. A compelling future direction is the maturation of “Laboratory-as-a-Service” (LaaS) models, whose commercial antecedents are the remotely operable cloud laboratories already run by providers such as Emerald Cloud Lab and Strateos [25], and wherein cloud-accessible self-driving labs offer metered experimental time to investigators at smaller institutions, industrial start-ups, and researchers in the Global South in much the same way that cloud-computing democratized access to high-performance computation in the preceding decade. Achieving this will require not only ad-

vances in remote tele-operation and standardized experiment description languages, but also deliberate policy choices by funding agencies and publishers to mandate that data generated on publicly funded autonomous plat-

forms be released in FAIR-compliant formats without undue embargo, as conceptualized in the architecture of **Figure 4** below.

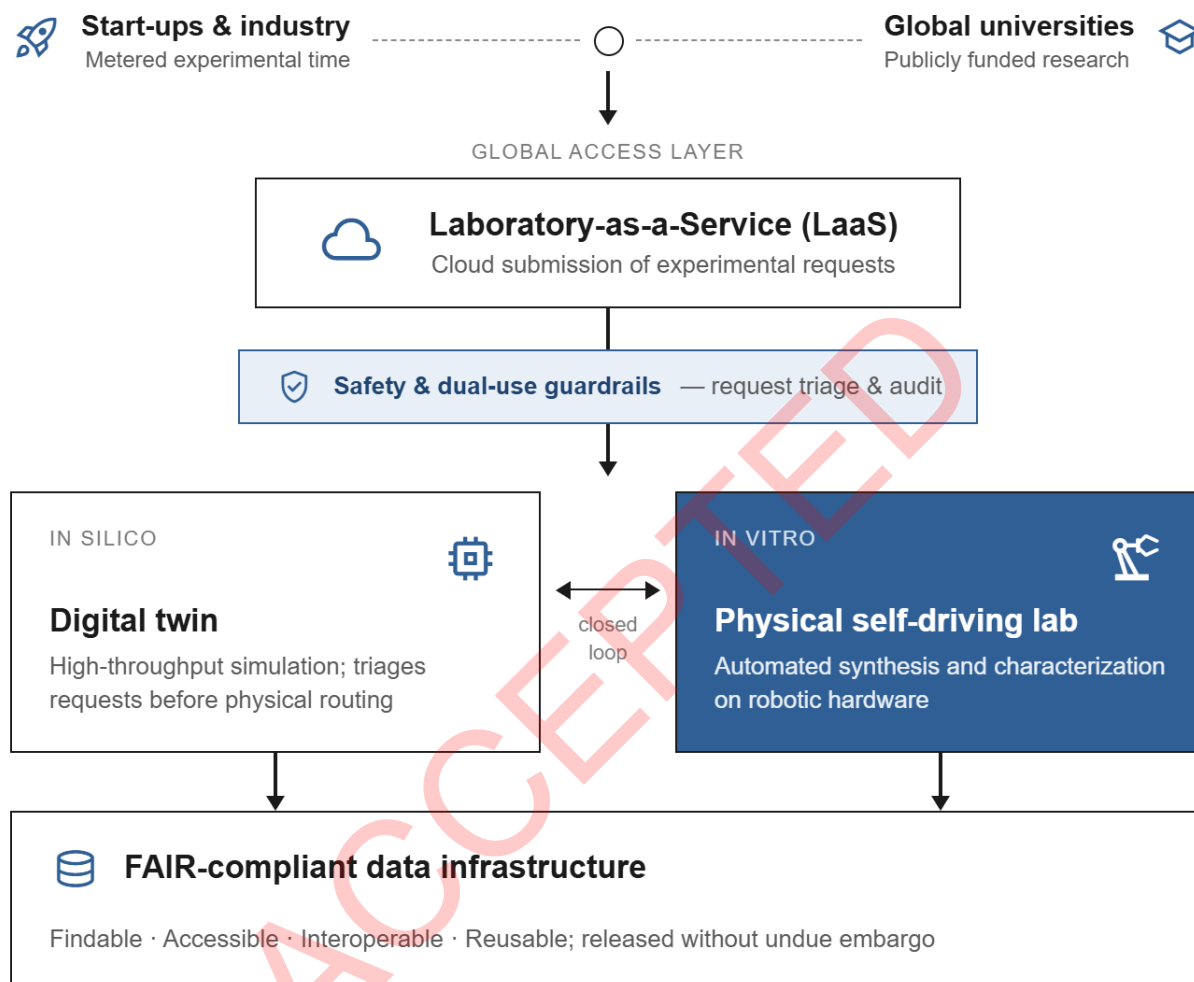


Figure 4: A proposed future architecture for democratized materials discovery. By transitioning to a ‘Laboratory-as-a-Service’ (LaaS) model, global researchers can submit experimental requests via cloud infrastructure, which are triaged by in silico digital twins before being routed to physical robotic hardware.

8. A Research Agenda for Autonomous Materials and Metallurgy Discovery

The challenges canvassed in **Section 7** are not merely obstacles to be cataloged; on inspection they resolve into a set of tractable research questions, each allowing concrete lines of attack. These are therefore distilled here into a structured agenda for the field, framed as seven open research questions (RQs), each paired with the directions of work judged most likely to answer it. The agenda is delib-

erately ordered from the infrastructural to the epistemic: from the plumbing without which nothing scales, to the questions that touch the nature of discovery itself.

RQ1 (Data infrastructure): What minimal set of shared schemas and ontologies suffices to render experimental materials data interoperable across heterogeneous platforms? Priority directions include consortium-based governance of FAIR-compliant standards for automated metallurgy, publication of reference datasets against which candidate schemas can be rigorously stress-tested, and development of machine-actionable provenance records that enable models trained on data from one

laboratory to account for and mitigate systematic biases introduced by another.

RQ2 (Simulation coupling): By what decision-theoretic criteria should an autonomous planner arbitrate between inexpensive in silico experiments and costly physical ones? The natural direction is multi-fidelity Bayesian optimization over a portfolio of simulators (CALPHAD, phase-field, DFT) and physical assets, with digital twins whose sim-to-real gap is quantified and continuously corrected by in situ measurement.

RQ3 (Transfer and generality): Can foundation models for materials science reduce the cold-start cost of retooling a platform to a new domain by an order of magnitude, and can they do so across, rather than merely within, the MSS classes of [Section 2](#)? Directions include cross-platform transfer benchmarks, pre-training corpora that fuse the literature with FAIR experimental repositories, and, given the sabotage risk noted in [Section 7.3](#), poisoning-resistant protocols for inter-laboratory model exchange.

RQ4 (Robustness and safety): Which guardrail architectures for autonomous synthesis are demonstrably resistant to adversarial removal, and what auditing standards should govern their deployment? Work is needed on hazard-aware planners that treat safety constraints as inviolable rather than merely penalized, on anomaly detection and fail-safe recovery sufficient for true walk-away operation, and on mandatory human-review triggers for candidate materials meeting predefined hazard criteria.

RQ5 (Open-endedness): Which acquisition formulations best institutionalize curiosity—Bayesian Active Learning by Disagreement and its kin—without unmooring a campaign from physical feasibility? Directions include hybrid novelty-feasibility objectives, evaluation metrics for SAL-3 systems whose goalposts move by design, and generative proposal mechanisms whose outputs are filtered through physics-informed constraint engines before any robot moves.

RQ6 (Human–AI collaboration): Which explanatory representations actually improve an expert’s capacity to steer exploration and to extract mechanistic understanding from autonomous campaigns? This requires moving beyond post hoc saliency toward explainability grounded in the domain’s own conceptual vocabulary—phase fields, strengthening mechanisms, processing windows—with comprehension gains measured empirically rather than assumed.

RQ7 (Access and democratization): What combination of technical enablers (standardized experiment description languages, secure tele-operation) and institutional arrangements (funding mandates, embargo policy, metered pricing) would render Laboratory-as-a-Service

both economically viable and equitably accessible? The prize, as [Section 7.7](#) argues, is nothing less than preventing experimental capability from stratifying along the very lines that computational capability once did.

Each of these questions is, in the authors’ judgment, answerable within the decade, and none is answerable by a single laboratory acting alone. The agenda is therefore as much institutional as it is technical, and it is offered in that spirit.

9. Conclusions

AI-guided self-driving laboratories unify computational intelligence with robotic experimentation to traverse vast materials design spaces efficiently. Early platforms have already delivered step-changes in throughput and discovery, from autonomous inorganic synthesis to polymer and catalysis optimization [[5,7,10](#)]. Embedding metallurgical knowledge that encompasses phase equilibria, kinetics, and processing heuristics within sequential decision frameworks, and engineering robust automation for harsh processing, are decisive levers for autonomous alloy discovery. The emerging convergence of data standards, simulation coupling, and trustworthy autonomy points to a near-term future where autonomous experimentation is a routine instrument of materials science, compressing development timelines for materials on demand. The path from the current state of the art to this future is neither linear nor guaranteed, and achieving it will require more than incremental advances within isolated research groups. What is urgently required is a set of concerted community commitments: the adoption of open-source hardware designs that lower barriers to entry for new participants; the establishment of an international consortium tasked with governing FAIR data-sharing standards for automated metallurgy; and proactive, transparent engagement with the dual-use and safety implications of increasingly autonomous synthesis systems before, rather than after, the occurrence of an inadvertent hazardous discovery. The materials science community possesses both the technical vocabulary and the institutional legitimacy to lead this effort; the only remaining question is whether it will summon the collective will to do so before the window of formative governance closes.

List of Abbreviations

AFRL	Air Force Research Laboratory
AI	Artificial Intelligence
APEX	Autonomous Alloy Prediction and Experimentation
ARES	Autonomous Research System

BALD	Bayesian Active Learning by Disagreement
BO	Bayesian Optimization
CALPHAD	Calculation of Phase Diagrams
DED	Directed-Energy Deposition
DFT	Density Functional Theory
DMTL	Design–Make–Test–Learn
FAIR	Findable, Accessible, Interoperable, and Reusable
GP	Gaussian Process
HAL	Hardware Autonomy Level
LaaS	Laboratory-as-a-Service
LLNL	Lawrence Livermore National Laboratory
ML	Machine Learning
MSS	Material System State
PEDOT:PSS	poly(3,4-ethylenedioxythiophene): poly(styrene sulfonate)
RL	Reinforcement Learning
RQ	Research Question
SAE	Society of Automotive Engineers
SAL	Software Autonomy Level
SDL	Self-Driving Laboratory
XAI	eXplainable AI

Author Contributions

Conceptualization, methodology, investigation, writing—original draft preparation, visualization, project administration: B.M.; validation, writing—review and editing: R.C. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

The authors declare no conflicts of interest.

Funding

The study did not receive any external funding and was conducted using only institutional resources.

AI Declaration

In accordance with COPE guidance on the use of artificial intelligence tools, the authors declare the following. Anthropic’s Claude Opus 4.8 large language model was used to evaluate the Paper and to identify shortcomings through adversarial evaluation, effectively acting as a peer reviewer of first resort before the Paper underwent a subsequent peer-review process by Scifiniti peer reviewers. All edits made after both the AI and human peer reviews were written by the authors, and all scientific content, analyses, interpretations, and conclusions were developed and elaborated by the authors, who take full responsibility for the accuracy and integrity of the work. Three of the four figures were generated using a fine-tuned QWEN Image model orchestrated through a customized fork of the open-

source PaperBanana repository, while the remaining figure was produced manually in Microsoft Excel. All scientific claims, analyses, interpretations, and conclusions are the authors’, who take full responsibility for the integrity and accuracy of the work. No AI tool is credited with authorship.

References

- [1] C. W. Coley, N. S. Eyke, and K. F. Jensen, “Autonomous discovery in the chemical sciences part I: Progress,” *Angew. Chem. Int. Ed.*, vol. 59, no. 51, pp. 22858–22893, 2020. [\[CrossRef\]](#)
- [2] B. Maruyama, “Let robots do your lab work (Fixing the Future podcast),” *IEEE Spectrum*, Feb. 21, 2024.
- [3] J. Jumper et al., “Highly accurate protein structure prediction with AlphaFold,” *Nature*, vol. 596, pp. 583–589, 2021. [\[CrossRef\]](#)
- [4] P. Nikolaev et al., “Autonomy in materials research: A case study in carbon nanotube growth,” *npj Comput. Mater.*, vol. 2, 2016, Art. no. 16031. [\[CrossRef\]](#)
- [5] N. J. Szymanski et al., “An autonomous laboratory for the accelerated synthesis of novel materials,” *Nature*, vol. 624, pp. 86–91, 2023. [\[CrossRef\]](#)
- [6] C. Wang et al., “Autonomous platform for solution processing of electronic polymers,” *Nat. Commun.*, vol. 16, 2025, Art. no. 1498. [\[CrossRef\]](#)
- [7] T. Xu, “Argonne’s Polybot joins the self-driving lab club,” *IEEE Spectrum*, May 10, 2023.
- [8] C. W. Coley, N. S. Eyke, and K. F. Jensen, “Autonomous discovery in the chemical sciences part II: Outlook,” *Angew. Chem. Int. Ed.*, vol. 59, no. 52, pp. 23414–23436, 2020. [\[CrossRef\]](#)
- [9] L. Hung, J. A. Yager, D. Monteverde, D. Baiocchi, H.-K. Kwon, S. Sun, and S. Suram, “Autonomous laboratories for accelerated materials discovery: A community survey and practical insights,” *Digit. Discov.*, vol. 4, no. 7, pp. 1273–1279, 2024. [\[CrossRef\]](#)
- [10] J. A. Bennett, N. Orouji, M. Khan, S. Sadeghi, J. Rodgers, and M. Abolhasani, “Autonomous reaction Pareto-front mapping with a self-driving catalysis laboratory,” *Nat. Chem. Eng.*, vol. 1, pp. 240–250, 2024. [\[CrossRef\]](#)
- [11] Lawrence Livermore National Laboratory (LLNL), “Self-driving lab to automate the discovery of novel alloys,” *LLNL News*, Jul. 21, 2025.
- [12] B. Burger et al., “A mobile robotic chemist,” *Nature*, vol. 583, pp. 237–241, 2020. [\[CrossRef\]](#)
- [13] D. Silver et al., “Mastering the game of Go with deep neural networks and tree search,” *Nature*, vol. 529, pp. 484–489, 2016. [\[CrossRef\]](#)
- [14] C. Zeni et al., “A generative model for inorganic materials design,” *Nature*, vol. 639, pp. 624–632, 2025. [\[CrossRef\]](#)
- [15] M. M. Noack et al., “Autonomous materials discovery driven by Gaussian process regression with inho-

- homogeneous noise and anisotropic kernels,” *Sci. Rep.*, vol. 10, 2020, Art. no. 17663. [CrossRef]
- [16] E. Stach et al., “Autonomous experimentation systems for materials development: A community perspective,” *Matter*, vol. 4, no. 9, pp. 2702–2726, 2021. [CrossRef]
- [17] A. Debnath et al., “Design and validation of refractory alloys using machine learning, CALPHAD, and experiments,” *Int. J. Refract. Met. Hard Mater.*, vol. 118, 2024, Art. no. 106673. [CrossRef]
- [18] L. Kaufman and J. Agren, “CALPHAD, first and second generation—birth of the materials genome,” *Scr. Mater.*, vol. 70, pp. 3–6, 2014. [CrossRef]
- [19] E. A. Olivetti et al., “Data-driven materials research enabled by natural language processing and information extraction,” *Appl. Phys. Rev.*, vol. 7, no. 4, 2020, Art. no. 041317. [CrossRef]
- [20] M. D. Wilkinson et al., “The FAIR Guiding Principles for scientific data management and stewardship,” *Sci. Data*, vol. 3, no. 1, 2016, Art. no. 160018.
- [21] S. A. Kube, “Autonomous Alloy Discovery Lab,” University of Wisconsin-Madison. Accessed: Oct. 5, 2025. [Online]. Available: <https://grainger.wisc.edu/centers/a2dl/>
- [22] Y. Iwasaki, “Autonomous search for materials with high Curie temperature using ab initio calculations and machine learning,” *Sci. Technol. Adv. Mater. Methods*, vol. 4, no. 1, 2024, Art. no. 2399494. [CrossRef]
- [23] A. Arditi, O. Obeso, A. Syed, D. Paleka, N. Panickssery, W. Gurnee, and N. Nanda, “Refusal in language models is mediated by a single direction,” in *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*. San Diego, CA, USA: NeurIPS, 2024.
- [24] N. Houlsby, F. Huszár, Z. Ghahramani, and M. Lengyel, “Bayesian Active Learning for Classification and Preference Learning,” 2011, arXiv:1112.5745.
- [25] C. Arnold, “Cloud labs: where robots do the research,” *Nature*, vol. 606, pp. 612–613, 2022. [CrossRef]

Disclaimer/Publisher’s Note: The views expressed in this article are those of the author(s) and do not necessarily reflect the views of the publisher or editors. The publisher and editors assume no responsibility for any injury or damage resulting from the use of information contained herein.