AI-Guided Self-Driving Laboratories for Advanced Materials Discovery

Bernard Mallia [1,2,3,] Rossana Caputo[3]

[1] *Equinox Group, Kalkara, 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*

Abstract

Discovering 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, and the discovery process is laborious and prohibitively costly. This perspective articulates a research agenda for autonomous materials and metallurgy discovery platforms that combine closed-loop machine learning with robotic experimentation and traditional materials engineering precepts to formulate, process, and test candidate materials with reduced human intervention, and to iterate recursively for stepwise optimization. It explores how sequential learning methods, generative modeling, and hybrid model-based strategies can analyze multimodal experimental data and propose new compositions and processing schedules for subsequent experiments. This approach has the potential to considerably accelerate the pace of discoveries, 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 show order-of-magnitude improvements in experimental throughput and discovery efficiency. Furthermore, the paper proposes a multidimensional framework for

classifying and evaluating the autonomy, complexity, and integration levels of self-driving laboratories, intended to give researchers and practitioners a structured approach to 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 optimisation; robotics; calculation of phase diagrams; materials discovery; metallurgy; machine learning; active learning; generative models; self-driving laboratory.

1. Introduction

Modern technologies require materials, from turbine and nuclear fusion alloys tolerating extreme temperatures to conductive polymers for flexible electronics, with unprecedented combinations of *ad hoc* properties, but the extant catalogue of discovered materials represents only a minuscule subset of all possible materials, meaning that many more are still to be discovered or, as would be more apt to call the underlying process, designed. In this milieu, naïve, random exploration, while proven over time as a mechanism of discovery that works in practice, is not the optimal approach to augmenting this catalogue due to underlying combinatorial explosion issues, particularly given recent developments in technology and its likely near-future

trajectory. High-throughput autonomous laboratories, or self-driving labs (SDLs) as they are at times referred to, integrate robotics with Artificial Intelligence (AI) (and probably within the next few years even quantum computing — if not gate-based, quantum annealing-based) to operate a closed-loop design–synthesise–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 optimisation of carbon nanotube growth in which the platform designed, executed and analysed 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 one [2, 4]. Polybot and A-Lab, two other exemplars, achieved fully closed-loop operation, discovering new compounds and optimising polymer films at scale [5, 6, 7]. Along the lines 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 understand the capabilities and challenges of these varied platforms, it is useful to have a classification framework, and this is exactly what we turn our attention to in the next section. This perspective synthesises 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 realised within the present decade, and distils these considerations into a structured research agenda for the field.

2. A Multidimensional Classification Framework for Self-Driving Labs

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

The proposal of a further taxonomy demands justification against those already in circulation. Existing classification schemes for autonomous experimentation are, in essence, single-axis ladders: Coley et al. chart a continuum from automation to autonomy in chemical discovery, grading systems by the extent to which hypothesis generation, experimental design, and interpretation are ceded to the machine [1, 8], while Hung et al. propose a six-level scale of laboratory autonomy, from non-automated (L0) to fully autonomous (L5), explicitly modelled on the six-level taxonomy for driving automation promulgated 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. Our SAL axis is broadly commensurable with the upper rungs of the Hung et al. scale, and our 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 that for MSS, which describes the physical form of the material being processed, and strongly correlates with the mechatronic complexity required. This axis can take on 1 of 3 values, which are subsequently explained:

- **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 characterisation;
- **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
- **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, which 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 that for HAL, which defines the degree of physical automation in the experimental workflow, and which can also take on 3 values, namely:

- **HAL-1 (Automated Task)** systems are defined as systems that can handle a single experimental task, like sample preparation or a specific characterisation step autonomously, but that still requires manual transfer, handshaking or handovers between the intermediate process steps in order to function as intended;

- **HAL-2 (Automated Workflow)** systems are capable of handling multiple experimental modules, like, say, synthesis, processing and characterisation, 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

- **HAL-3 (Reconfigurable Workflow)** platforms are able to autonomously execute diverse workflows by reconfiguring connections between modules or utilising 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 and last 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 3 values that are possible on this axis are:

- **SAL-1 (Static Design)**, where the system executes a pre-defined experimental array designed by a human or an offline optimisation 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 Optimisation (BO) in order to optimise a known set of parameters, this currently being 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 that entails generative models that can propose novel compositions or processing routes that were not initially conceived by human experts. This represents a shift from optimisation to truly open-ended discovery, as, for instance, when an AI autonomously proposes an entirely novel crystal structure lattice or a hitherto unconceived alloying strategy that no domain expert would have hypothesised from the extant literature, on the same lines of 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 characterised 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 at 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. It is worth noting that this 1–3 scale is deliberately more compressed than the SAE six-level taxonomy of driving automation (Levels 0 through 5): SAL-3 is broadly analogous to SAE Level 5 full automation, wherein the system determines not merely the optimal path through an extant design space but the very coordinates of the destination itself, without recourse to a human-supplied map. This classification provides a clear snapshot of a platform's capabilities and its position on the frontier of autonomous science.

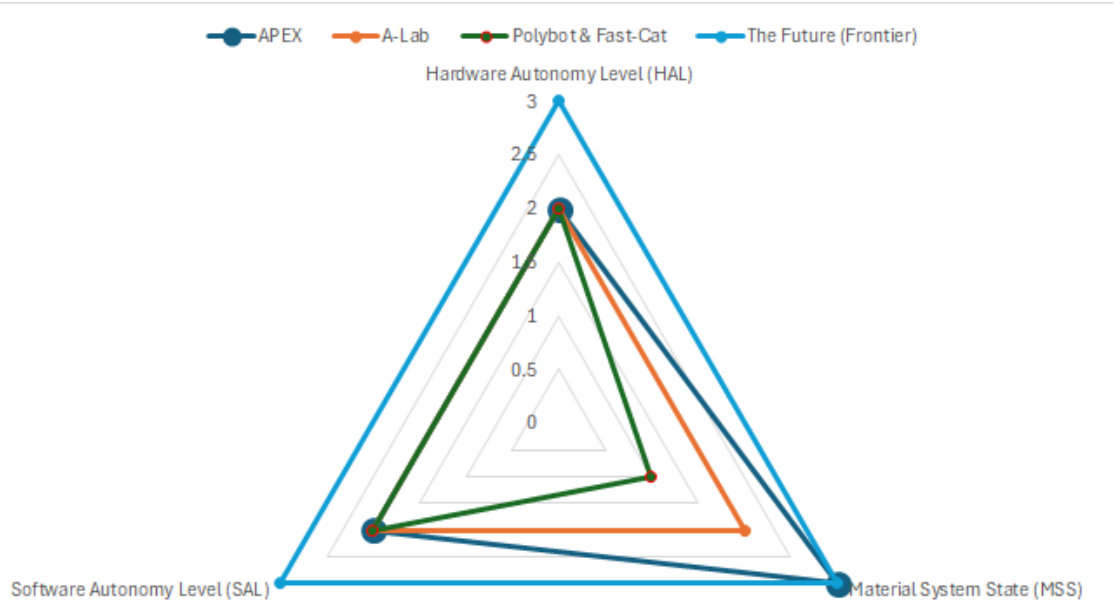

Figure 1. A multidimensional mapping of extant self-driving laboratories using the proposed taxonomy encompassing Material System State (MSS), Hardware Autonomy Level (HAL), and Software Autonomy Level (SAL). Platforms processing solid-state materials and alloys (e.g., A-Lab, APEX) face higher mechatronic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

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 optimisation in high-dimensional, expensive domains where GP posteriors quantify epistemic uncertainty, and acquisition functions (like, for example, 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, with a salient feature of recent hybrid systems being their combination of generative language models for recipe retrieval with graph-search (e.g., A*) to accelerate convergence in nanochemistry, while foundation models trained on literature fuel the cold-start regime [5] through statistical bootstrapping strategies (see **Figure 2** below). These may also be combined with physics simulation engines acting as digital twins of sorts that formulate scientific hypotheses, narrow down a workable solutions latent space, and finally conduct the experimentation to prove or disprove that hypothesis while embedding the results in the available knowledge base for incremental, evolutionary learning.

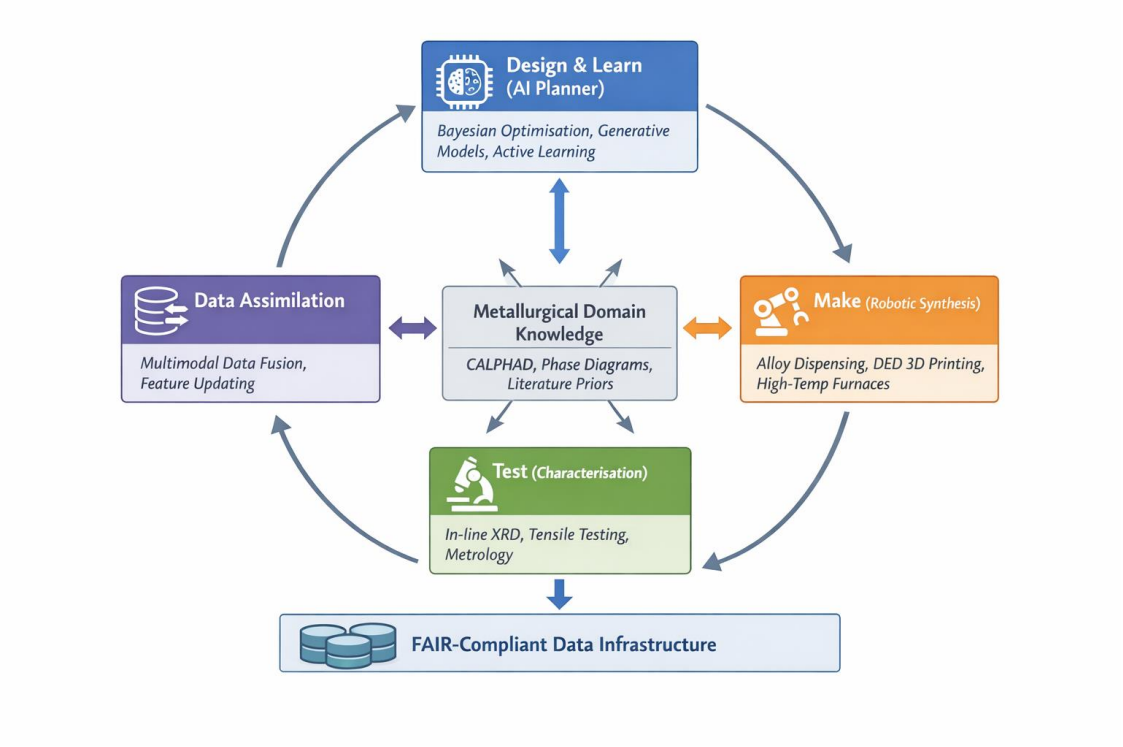

Figure 2. The core closed-loop Design–Make–Test–Learn (DMTL) architecture of an AI-guided self-driving laboratory for metallurgy. Domain knowledge is continuously injected into the AI Planner, which orchestrates robotic synthesis and automated characterisation.

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 maximise strength subject to density and cost constraints, or co-optimise conductivity and defectivity, thus requiring Pareto-efficient acquisition strategies and constraint-aware active learning.

One of the cornerstones of this approach entails physics and prior knowledge to be injected into the system with a view to avoid proposals that go against what is possible in the realm of Physics specifically, and 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, like CALPHAD, phase-field, and ab initio, are co-optimised with data-driven surrogates. This hybridisation has underpinned recent autonomous inorganic synthesis campaigns, where computational screening over the Materials Project narrowed viable targets ahead of robotic synthesis [5], resulting in costs that were lower than they would otherwise have been.

4. Robotic Automation for Experimental Metallurgy and Materials

Self-driving labs are mechatronic systems integrating robotic manipulation, process equipment, and in-line characterisation under software orchestration, where architectures commonly include multiple coordinated arms or gantries for logistics, synthesis modules for dispensing, reactors, additive manufacturing, thermal units like furnaces with vacuum/inert atmospheres, rapid thermal processing, and sensors for jobs like imaging, spectroscopy, diffraction, and mechanical testing. By way of illustration, 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 characterisation, enabling rapid composition-process iteration [11].

Engineering for metallurgy imposes several non-trivial constraints, some of which include handling powders that melt at temperatures higher than 1500 °C, interlocked safety enclosures for lasers and hot zones, end-effectors with thermal shielding, as well as 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 minimise operator variability, enable in-line metrology, and elevate throughput by orders of magnitude relative to equivalent manual processes and practices.

5. Incorporating Metallurgical Domain Knowledge into AI

Metallurgy contributes a dense *corpus* of priors ranging from phase diagrams, thermodynamics, and kinetics, all the way 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 (like, for example, 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 optimisation [10];

(v) Continuous knowledge-base 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 (which is not strictly required and may slow the process down) sometimes enables experts to steer exploration, inject heuristics, and audit rationale, enhancing trust and safety [2].

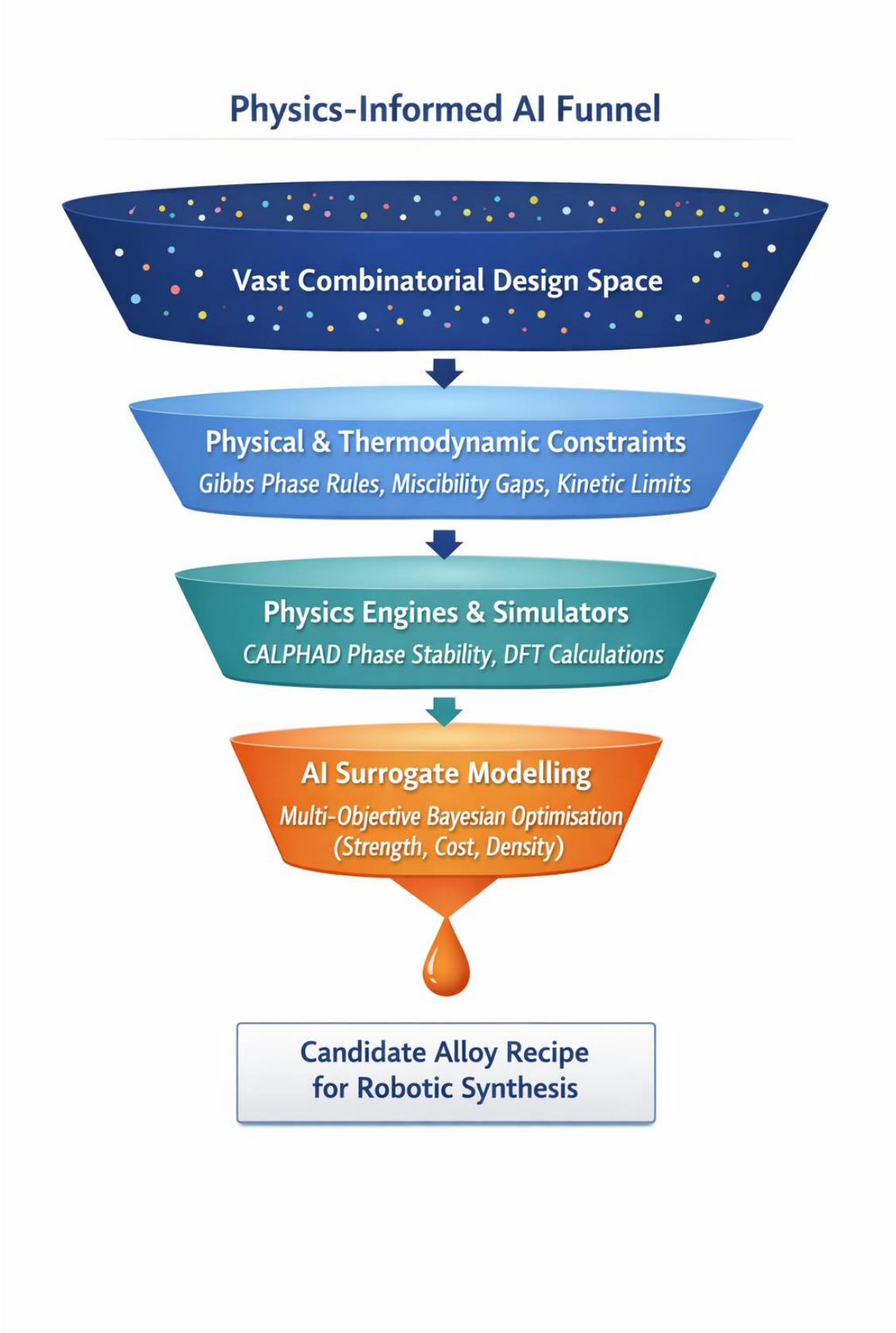

Figure 3. Diagrammatic representation of the integration of metallurgical domain knowledge into the AI decision loop, illustrating how physics-based constraints,

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 due to rapid feedback and the ease with which they enable automation. An illustrative example, Polybot, has demonstrated multi-parameter optimisation 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 minutes, 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 is pipelined rather than strictly sequential: the platform's stations, tended by coordinated robots, operate concurrently on successive samples, so that daily throughput is governed by the residence time at the slowest station rather than by the sum of all stage durations — precisely as the takt time of an assembly line, and not its end-to-end lead time, dictates the line's output. In catalysis, Fast-Cat integrates high-pressure flow reactors, online analytics, and autonomous planners to map Pareto fronts of yield versus selectivity, characterising six ligands in five-day campaigns each, thereby drastically cutting down on months-long manual workflows while at the same time validating scale-up fidelity [10]. Mobile robotic chemists further

illustrate flexible autonomy by operating conventional lab equipment to optimise photocatalysis [12].

6.2. Autonomous Alloy Discovery

Autonomous alloy discovery introduces higher experimental complexity as it requires compositional spaces with at least 5 principal elements, thermal cycles (like solutionising and ageing), and mechanical processing. LLNL's APEX couples DED metal printing with robotic handling and characterisation 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 thin [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 signalling 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-temper protocols and control precipitation kinetics with precision. A parallel and significant opportunity exists in autonomously optimising 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 amount of time ionic bombardment should take place in surface engineering.

7. Challenges and Future Directions

While the potential of self-driving laboratories is transformative, their widespread adoption and evolution from specialised 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 of these challenges is discussed and elaborated on next.

7.1. Data Infrastructure and Interoperability

The full potential of AI in materials science can only be unlocked through large, diverse, and 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 standardised schemas and ontologies for experimental materials science, following the Findable, Accessible, Interoperable, 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 generalising 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”, that 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 a low total and incremental cost level. The AI planner's role would evolve to arbitrate between executing a cheap virtual experiment or 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 maximise 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 specialised systems, representing a significant investment of time and capital to design and build for a specific task. A major challenge is reducing the re-tooling cost and the "cold start" problem when applying a platform to a new domain, and we firmly believe that 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 the development of

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 the achievement of a true "walk-away" 24/7 operation an engineering and AI challenge that is fundamental to maximising 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 it is 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, 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 behaviours through targeted ablation of safety-relevant model weights or activations (colloquially termed 'abliteration' [23]), and that flag candidate materials meeting predefined hazard criteria for mandatory human review before synthesis proceeds.

7.5. Beyond Optimisation to Open-Ended Discovery

Most current autonomous labs excel at optimising a known set of parameters to maximise a target property, but a more profound challenge would be to enable open-ended discovery that can be conducive 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 planned for by systems architecture designs 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 optimisation 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 institutionalising, 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 judgement, 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 helping human scientists form new hypotheses and contribute their own expertise. The ideal future state, at least in the short- and medium-run, 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 Democratisation

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-capitalised 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 pretty much in the same way as cloud-computing democratised access to high-performance computation in the preceding decade. Achieving this will require not only advances in remote tele-operation and standardised experiment description languages, but also deliberate policy choices by funding agencies and publishers to mandate that data generated on publicly funded autonomous platforms be released in FAIR-compliant formats without undue embargo, as conceptualised in the architecture of **Figure 4** below.

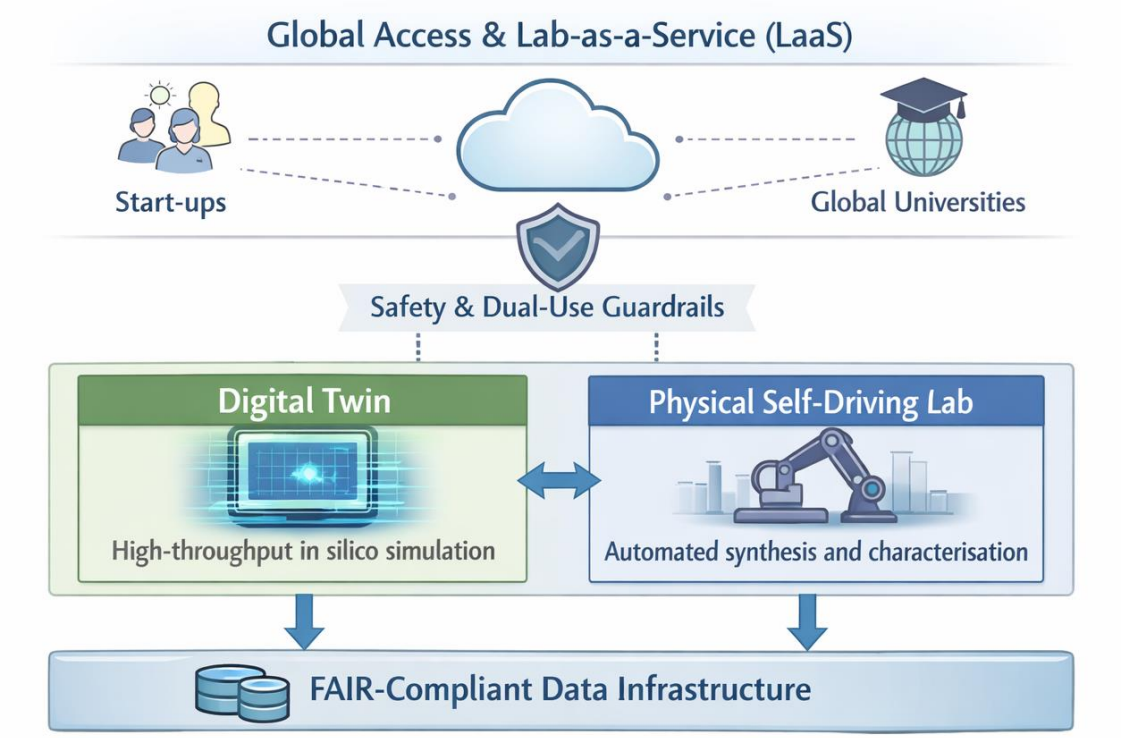

Figure 4. A proposed future architecture for democratised materials discovery. By transitioning to a 'Laboratory-as-a-Service' (LaaS) model, global researchers can submit experimental requests via cloud infrastructure, which are triaged by in silico digital twins before being routed to physical robotic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

8. A Research Agenda for Autonomous Materials and Metallurgy Discovery

The challenges canvassed in Section 7 are not merely obstacles to be catalogued; on inspection they resolve into a set of tractable research questions, each admitting of concrete lines of attack. We therefore distil them here into a structured agenda for the field, framed as seven open research questions (RQs), each paired with the directions of work we judge most likely to answer it. The agenda is deliberately 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, the publication of reference datasets against which candidate schemas can be stress-tested, and machine-actionable provenance records that allow a model trained on one laboratory's output to discount the systematic biases of another's.

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 optimisation over a portfolio of simulators (CALPHAD, phase-field, DFT) and physical assets, with digital twins whose sim-to-real gap is itself quantified and continuously corrected by in situ measurement.

RQ3 (Transfer and generality): Can foundation models for materials science reduce the cold-start cost of re-tooling 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 penalised, 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 institutionalise 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 democratisation): What combination of technical enablers (standardised 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 our judgement, 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 optimisation [10, 5, 7]. Embedding metallurgical knowledge that encompasses phase equilibria, kinetics, and processing heuristics within sequential decision frameworks, while 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 present state of the art to that future is neither linear nor guaranteed, and it will not be traversed by incremental tinkering within isolated research groups. What is required, urgently, is a set of concerted communal commitments: the adoption of open-source hardware designs that lower the barrier of entry for new entrants; the establishment of an international consortium mandated to govern FAIR data-sharing standards for automated metallurgy; and a frank, proactive engagement with the dual-use and safety implications of increasingly autonomous synthesis systems before, rather than after, the first 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; BALD, Bayesian Active

APEX, Autonomous Alloy Prediction and Experimentation

ARES, Autonomous Research System

Learning by Disagreement; BO, Bayesian Optimisation; 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.

Funding

This research received no external funding.

Conflict of Interest

The authors declare no conflict of interest.

Declaration on the Use of Artificial Intelligence

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 adversarial identify shortcomings, effectively acting as a peer-reviewer of first resort before the Paper went through a successive peer review process by Scifiniti peer reviewers. All edits made subsequent to both the AI and the 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 customised 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] Coley CW, Eyke NS, Jensen KF. Autonomous discovery in the chemical sciences part I: Progress. Angew Chem Int Ed 2020;59(51):22858-22893.

[2] Maruyama B. Let robots do your lab work (Fixing the Future podcast). IEEE Spectrum; 2024 Feb 21.

[3] Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596:583-589.

[4] Nikolaev P, Hooper D, Webber F, et al. Autonomy in materials research: a case study in carbon nanotube growth. npj Comput Mater 2016;2:16031.

[5] Szymanski NJ, Rendy B, Fei Y, et al. An autonomous laboratory for the accelerated synthesis of novel materials. Nature 2023;624:86-91.

[6] Wang C, Kim Y-J, Vriza A, Batra R, et al. Autonomous platform for solution processing of electronic polymers. Nat Commun 2025;16:1498.

[7] Xu T. Argonne's Polybot joins the self-driving lab club. IEEE Spectrum; 2023 May 10.

[8] Coley CW, Eyke NS, Jensen KF. Autonomous discovery in the chemical sciences part II: Outlook. Angew Chem Int Ed 2020;59(52):23414-23436.

[9] Hung L, Yager JA, Monteverde D, Baiocchi D, Kwon H-K, Sun S, Suram S. Autonomous laboratories for accelerated materials discovery: a community survey and practical insights. Digital Discovery 2024;3(7):1273-1279.

[10] Bennett JA, Orouji N, Khan M, Sadeghi S, Rodgers J, Abolhasani M. Autonomous reaction Pareto-front mapping with a self-driving catalysis laboratory. Nat Chem Eng 2024;1:240-250.

[11] Lawrence Livermore National Laboratory (LLNL). Self-driving lab to automate the discovery of novel alloys. LLNL News; 2025 Jul 21.

[12] Burger B, Maffettone PM, Gusev VV, et al. A mobile robotic chemist. Nature 2020;583:237-241.

[13] Silver D, Huang A, Maddison CJ, et al. Mastering the game of Go with deep neural networks and tree search. Nature 2016;529:484-489.

[14] Zeni C, Pinsler R, Zügner D, et al. A generative model for inorganic materials design. Nature 2025;639:624-632.

[15] Noack MM, Lewis A, Yu T, et al. Autonomous materials discovery driven by Gaussian process regression with inhomogeneous noise and anisotropic kernels. Sci Rep 2020;10:17663.

[16] Stach E, De Yoreo J, Sadowski P, et al. Autonomous experimentation systems for materials development: A community perspective. Matter 2021;4(9):2702-2726.

[17] Debnath A, Li W, Raman L, et al. Design and validation of refractory alloys using machine learning, CALPHAD, and experiments. International Journal of Refractory Metals and Hard Materials 2024;118:106673.

[18] Kaufman L, Agren J. CALPHAD, first and second generation—birth of the materials genome. Scripta Materialia 2014;70:3-6.

[19] Olivetti EA, Cole JM, Kim E, et al. Data-driven materials research enabled by natural language processing and information extraction. Applied Physics Reviews 2020;7(4):041317.

[20] Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 2016;3(1):1-9.

[21] Kube SA. Autonomous Alloy Discovery Lab. University of Wisconsin-Madison. Retrieved October 5, 2025, from https://grainger.wisc.edu/centers/a2dl/.

[22] Iwasaki Y. Autonomous search for materials with high Curie temperature using ab initio calculations and machine learning. Sci Technol Adv Mater Methods 2024;4(1):2399494.

[23] Arditi A, Obeso O, Syed A, Paleka D, Panickssery N, Gurnee W, Nanda N. Refusal in language models is mediated by a single direction. In: Advances in Neural Information Processing Systems 37 (NeurIPS 2024); 2024.

[24] Houlsby N, Huszár F, Ghahramani Z, Lengyel M. Bayesian Active Learning for Classification and Preference Learning. arXiv preprint arXiv:1112.5745 (2011).

[25] Arnold C. Cloud labs: where robots do the research. Nature 2022;606:612-613.