REVIEW Open access

Institutional Oversight Models for AI-Directed Materials Innovation

Carlos Vega[1*], Maria Hernandez[1]

Abstract

The convergence of machine learning, high-throughput computation, and autonomous experimentation has transformed materials discovery into an AI-directed process capable of closed-loop, data-driven innovation at unprecedented speed. This narrative review examines the computational and data-driven materials engineering ecosystem, with a specific focus on the governance, regulatory, and institutional oversight frameworks required to steward these capabilities responsibly. We synthesize developments in materials informatics, representation learning, graph neural networks, active learning, uncertainty quantification, and simulation–experiment integration, showing how these tools have enabled autonomous laboratories and inverse design. Particular attention is given to community-driven calls for standards, explainability, and scientific responsibility that have emerged alongside the technology. By integrating technical literature with explicit discussions of data governance, reproducibility, and ethical deployment, we articulate the need for structured institutional oversight models that span standards bodies, regulatory readiness, and multi-stakeholder governance regimes. These models must operate at the infrastructure level—embedding accountability into discovery pipelines rather than retrofitting them. The review positions institutional oversight not as a constraint on innovation but as an essential enabler that ensures AI-directed materials engineering delivers safe, equitable, and societally beneficial outcomes.

Keywords Materials informatics, Autonomous laboratories, AI-directed materials innovation, Institutional oversight, Governance regimes, Regulatory readiness

*Correspondence:
Carlos Vega
carlos.vega@gmail.com
[1] Department of Computational Materials Science, Faculty of Engineering, Polytechnic University of Valencia, Valencia, Spain

Introduction

Materials science has entered an era in which artificial intelligence no longer merely assists discovery but actively directs it. The Materials Genome Initiative and subsequent computational infrastructure laid the groundwork for large-scale data generation and sharing [1]. Within a few years, machine learning methods moved from proof-of-concept demonstrations to production-scale tools that now underpin entire discovery pipelines [2-4]. Today, graph neural networks, multimodal datasets, representation learning, and active learning systems routinely propose, evaluate, and refine candidate materials with minimal human intervention [5, 6].

This shift is most visible in the rise of autonomous laboratories and closed-loop platforms. On-the-fly Bayesian active learning has enabled real-time materials discovery without exhaustive screening [7]. Self-driving laboratories now execute synthesis, characterization, and iteration cycles with little or no human oversight [8-10]. Foundation models and large language models are being integrated into workflow orchestration, further accelerating the transition from data to hypothesis to validated material [11, 12]. These capabilities promise solutions to pressing societal challenges—energy storage, catalysis, sustainable manufacturing—but they also introduce systemic risks that cannot be addressed by technical excellence alone.

The scale and autonomy of these systems amplify well-known concerns in data-driven science: bias propagation from training data, lack of explainability in black-box models, irreproducibility of high-throughput results, and the potential for unintended material properties with safety or environmental consequences [13, 14]. Moreover, the concentration of advanced computational infrastructure in a small number of well-resourced institutions risks widening global disparities in materials innovation capacity.

Institutional oversight models are therefore required at three interconnected levels. First, governance regimes must define accountability for AI-generated hypotheses and the materials they produce. Second, regulatory readiness frameworks must anticipate safety, intellectual property, and export-control issues specific to AI-directed discovery. Third, standards bodies must establish interoperable data formats, metadata schemas, and validation protocols that make autonomous results auditable and reproducible [15].

Scientific responsibility further demands that researchers, institutions, and funders explicitly address the societal implications of accelerated discovery. This includes equitable access to data and models, transparent documentation of uncertainty, and mechanisms for community oversight of high-risk applications.

Existing literature already contains the seeds of these oversight models. Community perspectives on autonomous experimentation explicitly call for shared infrastructure, standardized reporting, and ethical guidelines [10, 13]. Reviews of explainable machine learning emphasize the necessity of interpretability for regulatory acceptance [14]. Papers on data standards highlight the urgent need for FAIR (findable, accessible, interoperable, reusable) principles tailored to AI-driven workflows [15].

The governance vulnerabilities associated with each stage of AI-directed discovery and the corresponding oversight mechanisms are summarized in **Table 1**.

Table 1. Governance infrastructure requirements across AI-directed materials discovery pipelines.

Discovery Stage	Technical Capability	Governance Risk Vector	Ove Mech
Data Ingestion	Multimodal dataset fusion	Data bias, undocumented provenance	Met stan audi
Model Training	Representation learning, GNNs	Opaque decision pathways	Expla frame
Hypothesis Generation	Inverse design, generative models	Unrealistic or hazardous candidates	Sa scre prot
Autonomous Experimentation	Robotic synthesis & testing	Experimental hazard escalation	Labo autho sys
Validation	High-throughput characterization	Reproducibility gaps	Bench & rep au
Model Updating	Active learning loops	Error propagation	Iterati checl

This review synthesizes the technical landscape of computational and data-driven materials engineering and the parallel emergence of autonomous and closed-loop systems, framing both within the explicit context of institutional oversight. By providing an original integrative analysis rather than a chronological summary, we identify leverage points where governance can be embedded into existing discovery infrastructure. The goal is to move the conversation from “what is technically possible” to “what oversight structures are required to ensure responsible deployment.”

Landscape of Computational & Data-Driven Materials Engineering

Foundations of materials informatics and machine learning

The emergence of computational and data-driven materials engineering as a coherent scientific paradigm can be traced to the early recognition that descriptor-based machine learning frameworks could dramatically accelerate property prediction relative to conventional physics-based simulation workflows. Traditional first-principles methods—while highly accurate—remained computationally intensive and thus impractical for exhaustive exploration of vast compositional spaces. The introduction of statistical

learning architectures trained on curated materials datasets reframed prediction as a data-efficient inference task rather than a purely mechanistic calculation problem. Early foundational reviews documented successful deployments across diverse material classes, including metallic alloys, polymer systems, and inorganic crystalline compounds, demonstrating that carefully engineered descriptors—capturing composition, structure, and electronic attributes—could enable predictive accuracies sufficient for high-throughput screening and prioritization [2- 4].

These formative studies did not merely validate machine learning as a surrogate modeling tool; they also articulated the methodological constraints that would shape the field's maturation. Chief among these was the recognition that predictive performance is inseparable from data quality. Noise, sampling bias, and representational sparsity were shown to propagate directly into model uncertainty, thereby affecting downstream design decisions. As a result, uncertainty quantification emerged early as a methodological imperative rather than a supplementary analytical layer. Probabilistic modeling, ensemble prediction, and calibration diagnostics were advocated as safeguards against overconfident extrapolation, particularly when models were deployed beyond the statistical support of their training distributions [16].

By the early 2020s, these methodological foundations had coalesced into a mature informatics ecosystem characterized by integrated databases, interoperable platforms, and scalable predictive infrastructures. Public web environments such as MaterialsAtlas.org exemplified this transition, offering curated repositories coupled with user-friendly analytical interfaces that democratized access to machine-learning predictions for both computational specialists and experimental practitioners [17]. Concurrently, large-scale deep learning initiatives demonstrated that models trained on millions of density-functional theory–derived structures could approach chemical accuracy across expansive chemical domains [6]. Such achievements marked a turning point: the principal bottleneck in discovery pipelines shifted from algorithmic capability to data infrastructure robustness, dataset lineage transparency, and interpretability of increasingly complex model architectures.

Representation learning and graph neural networks

As materials informatics matured, attention progressively shifted from handcrafted descriptor engineering toward automated representation learning. Graph-based encodings of atomic structure emerged as the dominant paradigm, motivated by their capacity to natively capture relational, topological, and geometric information at the atomic scale. In these architectures, atoms are represented as nodes and interatomic interactions as edges, enabling message-passing neural networks to iteratively propagate structural information across bonding environments. Such frameworks have consistently outperformed traditional descriptor sets in predicting formation energies, electronic band gaps, elastic tensors, and thermodynamic stability metrics [5].

The performance gains associated with graph neural networks are not solely attributable to architectural novelty but also to their scalability. Recent investigations demonstrate that predictive accuracy continues to improve as model depth, parameter count, and training dataset volume increase, mirroring scaling behaviors observed in natural language and vision foundation models [6]. These findings have catalyzed discussion around the feasibility of universal materials foundation models capable of learning transferable atomic representations spanning inorganic crystals, molecular systems, and hybrid interfaces.

Beyond forward prediction, representation learning has enabled inverse design paradigms in which target properties serve as generative constraints. In these workflows, learned latent spaces encode structure–property relationships in continuous manifolds, allowing optimization algorithms to navigate toward candidate structures exhibiting desired functional attributes [18]. When integrated with generative modeling techniques—such as variational autoencoders or diffusion-based samplers—these systems can explore chemical spaces that would be computationally intractable through enumeration alone [18]. Nevertheless, the literature consistently cautions that generative capacity must be accompanied by rigorous epistemic safeguards. Without calibrated uncertainty estimates and interpretable decision pathways, inverse-designed candidates may lack physical plausibility or experimental realizability, particularly in safety-critical or regulated materials sectors [14].

High-throughput computation and active learning strategies

The predictive capabilities of modern machine-learning systems are deeply rooted in the expansion of high-throughput computational infrastructures. Large density-functional theory repositories provided the foundational training corpora necessary to parameterize early surrogate models, enabling statistical learning across chemically diverse datasets. Yet static databases alone proved insufficient for mapping the combinatorial vastness of materials phase space. Active learning strategies emerged as a dynamic extension, embedding machine intelligence directly into simulation acquisition processes.

Within these frameworks, models iteratively identify regions of maximal epistemic uncertainty or predicted performance gain and selectively trigger new first-principles calculations. This closed computational loop dramatically reduces the number of simulations required to achieve comprehensive phase mapping. Adaptive sampling strategies guided by uncertainty metrics have demonstrated particular efficacy in identifying metastable polymorphs, optimizing magnetic compositions, and discovering rare-earth-free functional materials [16].

Over time, these iterative workflows have evolved into fully formalized discovery pipelines that algorithmically balance exploration and exploitation. Exploration routines map underrepresented compositional regions, while exploitation strategies refine high-performance candidate clusters. However, the reliability of these pipelines remains contingent upon the fidelity of their uncertainty quantification layers. Poorly calibrated uncertainty estimates can misdirect sampling priorities, leading to premature convergence or overlooked discovery zones. Consequently, standardized validation protocols and benchmarking frameworks have been proposed to ensure robustness and reproducibility across active learning deployments [15].

Multimodal datasets and simulation–experiment integration

A defining feature of the contemporary materials informatics landscape is the rapid expansion of multimodal data fusion. Whereas early machine-learning models relied predominantly on computational outputs, modern discovery ecosystems increasingly integrate heterogeneous data streams encompassing experimental measurements, spectroscopic signatures, microscopy imaging, and in situ process diagnostics. Multimodal learning architectures capable of jointly ingesting density-functional theory outputs and electron microscopy imagery have begun to demonstrate the capacity to infer structure–property relationships without reliance on intermediate descriptor engineering [19].

This convergence of simulation and experiment introduces both opportunity and complexity. On one hand, multimodal integration enhances model realism by grounding computational predictions in experimentally observed phenomena. On the other, it exposes the fragility of models trained exclusively on idealized simulation conditions. Discrepancies arising from synthesis defects, environmental variability, or measurement noise can produce distributional shifts that degrade predictive reliability.

Uncertainty quantification therefore assumes heightened importance within multimodal pipelines. Bayesian inference frameworks, deep ensemble modeling, and evidential learning approaches are increasingly deployed to identify epistemically uncertain regions requiring targeted experimental validation [14, 16]. Beyond predictive calibration, these uncertainty signals function as resource allocation guides—prioritizing laboratory effort where informational gain is maximal. The resulting simulation–experiment feedback architectures exemplify a hybrid epistemology in which computational foresight and empirical verification co-evolve within unified discovery loops.

Emerging standards and community infrastructure

Parallel to advances in modeling, computation, and data fusion, the materials community has recognized that infrastructural standardization is indispensable for sustaining scalable innovation. Autonomous discovery systems, distributed databases, and collaborative experimentation networks cannot function effectively without interoperable data schemas and reporting conventions. As a result, coordinated initiatives have emerged to define shared metadata standards, ontology alignment protocols, and provenance tracking mechanisms capable of capturing the full lineage of materials data—from synthesis conditions to model inference pathways [15].

These infrastructural efforts are not merely technical formalities; they constitute the epistemic backbone of reproducible science. Community surveys examining autonomous experimentation platforms consistently identify

the absence of standardized reporting frameworks as a primary barrier to cross-institutional collaboration, regulatory evaluation, and independent verification [10, 13]. Without harmonized data structures, the outputs of self-driving laboratories remain siloed, limiting cumulative knowledge integration.

Consequently, calls for infrastructure-level harmonization increasingly intersect with governance discourse. Standardized metadata, uncertainty annotation, and audit trails are being positioned as prerequisites for institutional oversight, safety certification, and policy compliance. In this sense, community infrastructure does not simply support discovery—it enables accountability. The maturation of computational and data-driven materials engineering therefore rests not only on algorithmic sophistication but also on the development of shared epistemic architectures capable of sustaining transparent, trustworthy, and globally interoperable innovation ecosystems.

Autonomous & closed-loop discovery systems

The transition from offline prediction regimes toward real-time, adaptive discovery infrastructures marks a foundational shift in computational and data-driven materials engineering. Early machine-learning deployments largely operated in retrospective analytical modes—training predictive models on static datasets and validating performance against held-out benchmarks. In contrast, contemporary on-the-fly closed-loop discovery platforms embed machine intelligence directly within experimental execution environments. Bayesian active learning systems exemplify this transition, functioning as autonomous decision architectures capable of selecting candidate materials, initiating synthesis, triggering characterization workflows, and assimilating the resulting data streams without human intervention. Within such platforms, the discovery process is formalized as a sequential optimization problem: each experimental measurement updates both the surrogate model representing structure–property relationships and the acquisition function governing subsequent exploration decisions. This dual updating mechanism allows the system to continuously recalibrate its epistemic landscape, dynamically balancing exploitation of high-confidence regions with exploration of uncertain compositional spaces. Empirical studies demonstrate that, when uncertainty is rigorously quantified and propagated, these closed loops can identify novel compositions and functional materials orders of magnitude faster than traditional Edisonian experimentation paradigms, fundamentally redefining the tempo of laboratory innovation [7].

Building upon these algorithmically steered feedback systems, fully autonomous or “self-driving” laboratories have emerged as integrated cyber-physical infrastructures. These environments fuse robotic synthesis platforms, in-line characterization instrumentation, automated sample handling, and machine-learning decision engines into unified operational ecosystems [8-10]. Initial implementations concentrated on thin-film deposition and organic synthesis, where experimental parameter spaces were relatively constrained and automation barriers lower. More recent advances, however, have expanded autonomy into solid-state inorganic synthesis, combinatorial alloy production, and high-throughput electrochemical testing for energy materials [8]. The scaling of these platforms has catalyzed parallel discourse on infrastructural standardization. Community analyses emphasize that the transformative potential of autonomous laboratories is maximized when robotic hardware, data schemas, and experimental ontologies are harmonized across institutions, enabling interoperability, reproducibility, and distributed learning [10, 13, 20].

An important development within this trajectory is the emergence of low-cost “frugal twin” laboratory architectures. These systems demonstrate that modular robotics, open-source control software, and affordable sensing technologies can replicate many functional capabilities of high-capital autonomous facilities [20]. By lowering financial and technical entry barriers, frugal twins expand global participation in AI-driven discovery and mitigate infrastructural inequities between well-resourced and emerging research ecosystems. Yet the democratization of autonomy simultaneously intensifies governance demands. As experimental decision authority shifts from human researchers to algorithmic agents, new questions arise concerning validation, accountability, and safety certification. Who bears responsibility for autonomously discovered compounds that exhibit hazardous properties? How are synthesis protocols documented, audited, and regulated when generated dynamically by machine reasoning systems? These concerns position governance not as an external constraint but as a co-evolving design requirement of autonomous experimentation infrastructures.

Parallel to the rise of robotic autonomy, recent research has begun integrating large language models and broader foundation model architectures into discovery workflows [9, 12]. Unlike task-specific predictive systems, these models operate as cognitive orchestration layers capable of synthesizing multimodal knowledge streams. They mine scientific literature, extract latent hypotheses, propose experimental pathways, and translate high-level research objectives into machine-readable laboratory instructions. In this capacity, foundation models function as epistemic intermediaries between human scientific intent and robotic execution. Early demonstrations show promise in automating experiment planning, reagent selection, and parameter optimization. However, the literature consistently underscores that such deployments remain contingent on advances in explainability, traceability, and verification. Without transparent reasoning pathways and human-in-the-loop override mechanisms, the delegation of experimental design authority to generative models poses significant scientific and regulatory risks, particularly within safety-critical or highly regulated materials domains [14].

As autonomy deepens across both physical and cognitive layers of discovery systems, scholarly attention has increasingly turned toward institutional oversight architectures. Technical capability is no longer the sole frontier; governance readiness has emerged as an equally critical determinant of responsible innovation. Community surveys and policy analyses identify the absence of standardized ethical review mechanisms, cross-border data-sharing agreements, and audit frameworks as systemic vulnerabilities [10, 15]. In response, explainability infrastructures are being reframed not merely as interpretability tools for scientists but as compliance instruments necessary for regulatory acceptance and public trust [14]. Standards organizations and research consortia are advancing protocols that embed provenance metadata, uncertainty quantification records, and reproducibility markers directly into the outputs of autonomous experimental runs [15]. Such measures ensure that every machine-directed discovery carries an auditable lineage—from data ingestion through hypothesis generation to validation outcomes. The multi-layer integration of technical, institutional, and regulatory governance within autonomous discovery infrastructures is conceptualized in **Figure 1**.

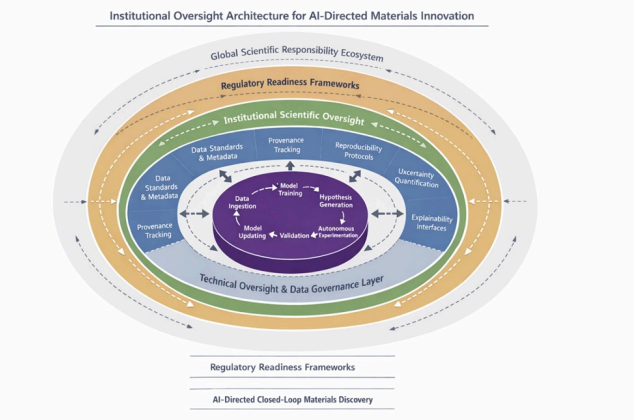

Figure 1. Institutional oversight architecture for AI-directed closed-loop materials discovery.

Figure 1 illustrates a multi-layer institutional oversight architecture embedded within AI-directed materials innovation systems. At the core lies a closed-loop discovery engine encompassing data ingestion, model training, hypothesis generation, autonomous experimentation, validation, and iterative model updating. Surrounding this operational nucleus are four concentric governance strata. The technical infrastructure layer governs data standards, provenance, reproducibility, and uncertainty documentation. Institutional scientific oversight encompasses laboratory safety boards and review committees responsible for experimental authorization and risk evaluation. Regulatory readiness interfaces address export control, intellectual property, and environmental compliance considerations. The outermost societal governance layer includes standards organizations, funding agencies, and multi-stakeholder ethics councils. Radial audit pathways and escalation channels ensure continuous monitoring, accountability, and compliance across all discovery phases, positioning oversight as an infrastructure-native component of autonomous materials innovation.

Taken together, these technological and institutional developments signal that AI-driven materials discovery has entered a phase of infrastructural maturity. Closed-loop optimization, robotic autonomy, foundation-model cognition, and governance architectures are no longer separable trajectories but interdependent system components. The acceleration of discovery capability therefore necessitates a parallel acceleration in oversight design, ensuring that innovation velocity remains aligned with safety, accountability, and societal responsibility.

Results and Discussion

The remarkable technical achievements in computational and data-driven materials engineering have simultaneously exposed a cluster of systemic vulnerabilities that cannot be resolved through further algorithmic refinement alone. These challenges—spanning data governance, model interpretability, safety assurance, regulatory preparedness, and equitable access—collectively demonstrate why institutional oversight models must be treated as core infrastructure rather than peripheral add-ons. The literature consistently shows that the same features enabling rapid discovery (autonomy, scale, and minimal human intervention) are precisely those that amplify risk when left ungoverned.

Data governance and reproducibility remain foundational weaknesses. Community surveys of autonomous experimentation platforms repeatedly identify the lack of unified metadata schemas, provenance tracking, and standardized reporting formats as the primary barrier to cross-institutional validation and regulatory acceptance [10, 15, 20]. In closed-loop systems, each iteration of active learning or Bayesian optimization depends on the integrity of prior data; undocumented uncertainty or bias introduced at any step propagates through the entire pipeline, rendering downstream results non-reproducible [7, 16]. Multimodal integration of simulation and experiment data exacerbates the issue, as disparate data types are fused without common ontologies or audit trails [19]. The consequence is not merely scientific inefficiency but a fundamental erosion of trust: when an autonomous laboratory proposes a new catalyst or alloy, regulators, funders, and downstream users have no reliable mechanism to verify the chain of evidence.

Explainability deficits compound these problems and directly undermine scientific responsibility. Graph neural networks, representation learning models, and emerging foundation models routinely achieve high predictive accuracy, yet their internal decision pathways remain opaque [5, 6, 13, 14]. Reviews of explainable machine learning in materials science emphasize that without interpretable outputs, it is impossible to assign accountability for erroneous predictions or to perform the mechanistic validation required for safety-critical applications [13, 14]. In autonomous laboratories, this opacity is particularly acute: a self-driving system may select an experiment, execute it, and update its surrogate model without any human-readable rationale [8-10]. The literature therefore positions explainability not as an academic luxury but as a prerequisite for regulatory readiness and ethical deployment.

Safety and unintended consequences introduce another layer of urgency. On-the-fly closed-loop platforms have demonstrated the capacity to discover novel compositions orders of magnitude faster than traditional methods [7, 8]. However, the same speed that enables breakthrough also compresses the window for hazard assessment. Materials proposed by generative models or active-learning loops may exhibit previously unknown toxicity, instability, or environmental persistence that standard computational screening fails to capture [6]. Low-cost “frugal twin” systems further democratize access but simultaneously distribute the responsibility for safety validation across a wider, less coordinated set of operators [20]. The absence of standardized ethical review processes or mandatory safety metadata in autonomous workflows leaves a regulatory vacuum that no single institution can fill.

Equity and infrastructure concentration complete the challenge landscape. Large-scale materials informatics platforms, high-performance computing resources, and fully autonomous laboratories remain concentrated in a handful of well-resourced institutions and consortia [1, 17]. This geographic and economic asymmetry risks creating a two-tier global materials innovation ecosystem in which only privileged actors can participate in AI-directed discovery. Community perspectives on autonomous systems explicitly warn that without deliberate mechanisms for data and model sharing, the technology will widen rather than narrow capability gaps [10, 21].

These challenges are not independent; they form a tightly coupled system. Poor data standards undermine explainability; opaque models hinder safety assessment; concentrated infrastructure limits the diversity of oversight perspectives. The literature therefore converges on a single conclusion: oversight cannot be retrofitted after deployment. It must be engineered into the discovery infrastructure itself—through standardized metadata layers, embedded explainability requirements, automated auditing loops, and multi-stakeholder governance protocols that operate at the same cadence as the autonomous cycles they oversee.

Future research directions

Future work must shift from documenting challenges to co-designing institutional oversight models that are native to the computational and autonomous ecosystems described in the reviewed literature. Four priority directions emerge from cross-study synthesis.

First, research should focus on the development and validation of machine-readable governance ontologies that can be embedded directly into active-learning and closed-loop platforms. Extending existing standards initiatives, these ontologies would automatically capture provenance, uncertainty estimates, and ethical metadata at every iteration, making results auditable by design rather than by retrospective effort [15].

Second, interdisciplinary studies are needed to create hybrid human–AI oversight frameworks tailored to autonomous laboratories. These frameworks would define clear escalation thresholds—based on model uncertainty, predicted material novelty, or potential hazard scores—where human review boards or regulatory interfaces are automatically triggered [10, 20]. Pilot implementations in both high-end and frugal-twin settings would provide empirical evidence for scalable governance.

Third, the community should establish and test regulatory sandboxes specifically for AI-directed materials discovery. Such controlled environments would allow autonomous systems to propose and synthesize candidates under simulated regulatory scrutiny, generating the data required to update safety assessment protocols and intellectual-property guidelines for AI-generated inventions.

Fourth, research into explainable foundation models and representation-learning architectures must prioritize regulatory-grade interpretability. Rather than treating explainability as a post-hoc analysis, future models should incorporate built-in mechanisms for mechanistic traceability and uncertainty decomposition that align with the needs of standards bodies and oversight committees [6, 13, 14].

Collectively, these directions move the field from technology-centric to governance-native development, ensuring that the infrastructure-level advances already achieved in materials informatics and autonomous experimentation are matched by equally sophisticated institutional oversight structures.

Conclusion

This narrative review has synthesized the rapid evolution of computational and data-driven materials engineering, with particular emphasis on materials informatics, representation learning, active learning, multimodal integration, and the emergence of autonomous closed-loop discovery systems. The literature demonstrates that these capabilities have transformed materials innovation from a primarily human-driven to an AI-directed process. At the same time, the same body of work reveals consistent, cross-cutting calls for standards, explainability, reproducibility, and scientific responsibility.

Institutional oversight models—spanning governance regimes, regulatory readiness, and standards bodies—are not external constraints on innovation but essential enablers that must be embedded at the infrastructure level. By integrating accountability, auditability, and ethical review directly into discovery pipelines, the field can accelerate safe, equitable, and societally beneficial materials solutions while mitigating the systemic risks that autonomy and scale inevitably introduce.

The reviewed studies provide both the technical foundation and the explicit community mandate for this next phase. The challenge now is to translate these insights into operational oversight frameworks that evolve in lockstep with the technology they govern. Only through such deliberate, infrastructure-native integration can AI-directed materials innovation fulfil its transformative promise responsibly.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 14 Nov 2024 Revised: 04 Apr 2025 Accepted: 15 Jul 2025

Published online: 18 September 2025

References

Schmidt J, Marques MRG, Botti S, Marques MAL. Recent advances and applications of machine learning in solid-state materials science. npj Comput Mater. 2019;5:83. https://doi.org/10.1038/s41524-019-0221-0.

Ramprasad R, Batra R, Pilania G, Mannodi-Kanakkithodi A, Kim C. Machine learning in materials informatics: recent applications and prospects. npj Comput Mater. 2017;3:54. https://doi.org/10.1038/s41524-017-0056-5.

Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A. Machine learning for molecular and materials science. Nature. 2018;559(7715):547-55. https://doi.org/10.1038/s41586-018-0337-2.

Lookman T, Balachandran PV, Xue D, Yuan R. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. npj Comput Mater. 2019;5:21. https://doi.org/10.1038/s41524-019-0153-8.

Merchant A, Batzner S, Schoenholz SS, Aykol M, Cheon G, Cubuk ED. Scaling deep learning for materials discovery. Nature. 2023;624:80-5. https://doi.org/10.1038/s41586-023-06735-9.

Sun T, Wang Z, Feng G. Identifying MOFs for electrochemical energy storage via density functional theory and machine learning. npj Comput Mater. 2025;11(90). https://doi.org/10.1038/s41524-025-01590-w.

Kusne AG, Yu H, Wu C, Zhang H, Hattrick-Simpers J, DeCost B, et al. On-the-fly closed-loop materials discovery via Bayesian active learning. Nat Communs. 2020;11:5966. https://doi.org/10.1038/s41467-020-19597-w.

Szymanski NJ, Rendy B, Fei Y, Kumar RE, He T, Milsted D, et al. An autonomous laboratory for the accelerated synthesis of novel materials. Nature. 2023;624:86-91. https://doi.org/10.1038/s41586-023-06734-w.

MacLeod BP, Parlane FGL, Morrissey TD, Häse F, Roch LM, Dettelbach KE, et al. Self-driving laboratory for accelerated discovery of thin-film materials. Sci Adv. 2020;6:eaaz8867. https://doi.org/10.1126/sciadv.aaz8867.

Foster M, Sumpter BG, Hattrick-Simpers J, Kusne AG, Kalinin SV. Autonomous experimentation systems for materials development: A community perspective. Matter. 2021;4(9):2702-26. https://doi.org/10.1016/j.matt.2021.06.036.

Pyzer-Knapp EO, Manica M, Curioni A. Foundation models for materials discovery – current state and future directions. npj Comput Mater. 2025;11(61). https://doi.org/10.1038/s41524-025-01538-0.

Lei G, Docherty R, Cooper SJ. Materials science in the era of large language models: a perspective. Digit Discov. 2024;3:1257-72. https://doi.org/10.1039/D4DD00074A.

Jablonka KM, Ongari D, Moosavi SM, Smit B. Explainable machine learning in materials science. npj Comput Mater. 2022;8:204. https://doi.org/10.1038/s41524-022-00884-7.

Maier G, Hamaekers J, Martilotti DS, Ziebarth B. Predicting properties of oxide glasses using informed neural networks. In Informed Machine Learning . Cham: Springer Nature Switzerland; 2025. p. 161-85.

Himanen L, Geurts A, Foster AS, Rinke P. Data-driven materials science: status, challenges, and perspectives. Adv Sci. 2019;6(21):1900808.

Shi J, Albreiki F, Colon YJ, Srivastava S, Whitmer JK. Transfer learning facilitates the prediction of polymer–surface adhesion

strength. J Chem Theory Comput. 2023;19(14):4631-40.

Shen L, Zhou J, Yang T, Yang M, Feng YP. High-throughput computational discovery and intelligent design of two-dimensional functional materials for various applications. Acc Mater Res. 2022;3(6):572-83.

De Breuck PP, Wang HC, Rignanese GM, Botti S, Marques MAL. Generative AI for crystal structures: a review. npj Comput Mater. 2025;11(370). https://doi.org/10.1038/s41524-025-01881-2.

Kalinin SV, Mukherjee D, Roccapriore K, Blaiszik BJ, Ghosh A, Ziatdinov MA, et al. Machine learning for automated experimentation in scanning transmission electron microscopy. npj Comput Mater. 2023;9(227). https://doi.org/10.1038/s41524-023-01142-0.

Volk AA, Abolhasani M. Performance metrics to unleash the power of self-driving labs in chemistry and materials science. Nat communs. 2024;15(1):1378.

Minocha N, Pandey P. self-driving labs for materials sciences. Artificial intelligence in polymer science and nanotechnology: Autonomous labs, polymer design, materials simulation; 2025. p. 69.