ORIGINAL RESEARCH **Open access**

Discovery without Understanding: A Systems Theory of Black-Box Optimization in Autonomous Materials Engineering

Daniel Brooks[1*], Amelia Carter[2], Ethan Moore[1]

Abstract

In the evolving landscape of computational and data-driven materials engineering, the integration of machine learning and high-throughput methodologies has accelerated discovery processes, yet it introduces a paradox where rapid optimization often bypasses deep scientific understanding. This manuscript presents a systems theory perspective on black-box optimization in autonomous materials engineering, emphasizing closed-loop labs where AI-driven decisions guide experimentation without explicit interpretability. Drawing from materials informatics and representation learning, we identify the discovery acceleration paradox: enhanced efficiency in inverse design and property prediction erodes traditional epistemic structures, leading to reliance on opaque models. We introduce the "Epistemic Opaque Discovery System" (EODS) framework, which conceptualizes materials discovery as a layered network of data infrastructures, model architectures, and feedback mechanisms. This framework highlights trade-offs between optimization speed and interpretability, incorporating uncertainty quantification to mitigate risks in autonomous systems. Implications extend to simulation-experiment coupling and multimodal datasets, suggesting pathways for balanced computational workflows that preserve scientific insight amid black-box dominance. By reframing discovery pipelines, EODS offers a theoretical lens for engineering resilient AI ecosystems in materials science, fostering sustainable innovation without sacrificing foundational knowledge.

Keywords Materials informatics, Representation learning, AI in materials science, Black-box optimization, Computational discovery, Data infrastructures

*Correspondence:
Daniel Brooks
daniel.brooks@gmail.com
[1] Department of Computational Materials Engineering, Faculty of Engineering, University of Manchester, Manchester, United Kingdom
[2] Department of Data-Driven Materials Science, Faculty of Engineering, University of Birmingham, Birmingham, United Kingdom

Introduction

The advent of computational and data-driven approaches has transformed materials engineering from a predominantly empirical discipline into a sophisticated ecosystem of predictive modeling and automated discovery. Over the past decade, advancements in high-throughput computation have enabled the screening of vast chemical spaces, facilitating the identification of novel materials with tailored properties for applications ranging from energy storage to advanced electronics [1]. This shift is underpinned by the convergence of materials informatics, where data-centric methodologies leverage large-scale datasets to inform design decisions [2]. Machine learning techniques, particularly deep learning architectures such as graph neural networks, have become integral, offering efficient representations of atomic structures and enabling predictions of material behaviors with unprecedented accuracy [3-5].

At the core of this transformation lies the role of AI in bridging simulation and experimentation. Autonomous discovery systems, exemplified by closed-loop labs, integrate real-time data acquisition with optimization algorithms to iteratively refine material candidates [6, 7]. These systems employ Bayesian active learning and hierarchical strategies to navigate nonequilibrium phase diagrams, accelerating synthesis processes [6, 7]. However, this efficiency comes at a cost: the increasing reliance on black-box models, where internal decision-making remains inscrutable, challenges the traditional scientific paradigm of understanding through explanation [8]. The optimization versus explanation dichotomy emerges as a critical tension, as AI-guided pipelines prioritize rapid convergence over interpretable mechanisms [9].

High-throughput infrastructures further amplify this dynamic. By coupling density functional theory with machine learning surrogates, researchers can explore expansive parameter spaces, but often at the expense of epistemic depth [1, 10]. For instance, inverse materials design utilizes neural networks to map desired properties back to structural motifs, bypassing exhaustive mechanistic analysis [11]. While this accelerates innovation, it introduces the discovery acceleration paradox: faster cycles of experimentation erode the iterative refinement of scientific models, potentially leading to overlooked uncertainties or biases in predictions [12].

Epistemic and computational constraints compound these issues. In materials AI, uncertainty quantification is essential for robust decision-making, yet black-box optimizers frequently marginalize such considerations in favor of deterministic outputs [12]. Representation learning, through architectures like crystal graph networks, enhances predictive power but obscures the linkage between data features and physical principles [13]. Moreover, multimodal datasets integrating experimental and simulated data create complex infrastructures that demand seamless coupling, yet current paradigms struggle with interpretability erosion as models scale [14].

This manuscript positions itself at the intersection of these challenges, advocating a systems theory approach to black-box optimization in autonomous materials engineering. By conceptualizing discovery as an interconnected system of data flows, model inferences, and feedback loops, we aim to illuminate the structural trade-offs inherent in modern workflows [1, 8]. Unlike empirical studies that benchmark performance, our analysis remains theoretical, focusing on interpretive insights into how opaque optimization reshapes the epistemic landscape of materials science [3, 9]. We introduce a novel framework that integrates these elements, offering a lens for steering computational pipelines toward balanced outcomes. This perspective is timely, as the field grapples with scaling AI ecosystems while preserving the foundational ethos of scientific inquiry [14].

Theoretical Background & Literature Synthesis

Materials data infrastructures

The foundation of computational materials engineering rests on robust data infrastructures that aggregate, harmonize, and operationalize information from heterogeneous scientific sources. High-throughput computation has enabled the systematic generation of extensive datasets encompassing atomic configurations, electronic structures, defect states, and thermodynamic stabilities, fundamentally altering the epistemic scale at which materials research operates [10]. These infrastructures support the transition from isolated simulation campaigns toward integrated discovery ecosystems in which density functional theory outputs, atomistic simulations, and property databases feed directly into predictive machine learning pipelines [1]. Rather than functioning as passive repositories, contemporary materials data platforms operate as dynamic substrates for algorithmic learning, continuously evolving through iterative data ingestion, model retraining, and validation cycles.

The integration of multimodal datasets further enhances the epistemic richness of these infrastructures. By combining computational outputs with experimental characterization data—such as spectroscopy, microscopy, and synthesis logs—platforms can encode complementary perspectives on material behavior [14]. This multimodal convergence improves predictive robustness while also enabling cross-domain correlation discovery that would remain inaccessible within unimodal datasets. However, this growth in volume and diversity introduces structural frictions. Data heterogeneity, inconsistent metadata standards, and fragmented storage architectures impede interoperability and reproducibility, often resulting in discovery pipelines that remain computationally powerful yet infrastructurally disjointed [2].

In autonomous discovery environments, these infrastructures must support real-time responsiveness. Closed-loop experimentation systems depend on rapid data assimilation to recalibrate optimization trajectories, requiring infrastructures capable of low-latency integration between simulation outputs, experimental readouts, and decision engines [6]. Bayesian active learning exemplifies this paradigm, where uncertainty-weighted sampling directs subsequent simulations or syntheses, effectively transforming data infrastructures into adaptive steering substrates rather than static archives [6]. Yet scalability introduces new tensions. As repositories expand, maintaining data integrity, calibration fidelity, and accessibility becomes increasingly complex. Current infrastructural paradigms often struggle to balance scale with epistemic reliability, highlighting the need for architecture-level innovation to support opaque yet high-velocity discovery systems [9]. Emerging discussions surrounding scientific foundation models further underscore this need, as pre-trained multimodal architectures require harmonized, large-scale data ecosystems to function effectively within black-box discovery regimes [14].

Representation learning architectures

Representation learning constitutes the computational backbone of data-driven materials discovery, translating raw compositional and structural inputs into machine-interpretable feature spaces. Graph neural networks have emerged as particularly powerful encoders in this domain, modeling crystal lattices and molecular assemblies as relational graphs in which atoms serve as nodes and bonds or spatial proximities function as edges [3-5]. Through message-passing operations, these architectures capture both local chemical environments and long-range structural interactions, enabling high-fidelity property prediction across diverse materials classes. Their capacity to operate directly on structural graphs reduces reliance on handcrafted descriptors, allowing representations to emerge organically from data distributions.

Empirical studies demonstrate that such graph-based systems excel in predicting material stability, electronic properties, and topological phases, frequently outperforming descriptor-driven machine learning approaches in high-dimensional search spaces [10, 13]. Extensions such as atomistic line graph neural networks further refine representational granularity by encoding edge–edge interactions, thereby incorporating angular and bond-orientation information into predictive embeddings [4]. These advances increase the resolution at which structure–property relationships can be learned, expanding the inferential reach of representation systems.

Deep learning frameworks generalize these encoding paradigms across molecular and solid-state regimes, enabling unified representation strategies that bridge chemistry and condensed matter systems [5]. Benchmarking initiatives reveal the versatility of such architectures across datasets and tasks, though they also expose persistent limitations in cross-domain generalizability and transfer learning robustness [3]. Inverse design frameworks invert the conventional predictive pipeline, deploying neural architectures to generate candidate structures conditioned on target properties, thereby collapsing design and evaluation into a single inferential step [11].

Despite their transformative predictive capacity, representation learning architectures frequently operate as epistemic black boxes. Learned embeddings encode structural information in distributed, high-dimensional manifolds that lack direct physical interpretability. As a result, scientific reasoning becomes increasingly mediated by latent proximities and attention weights rather than mechanistic descriptors, contributing to an erosion of explanatory transparency within discovery pipelines [8].

AI-Guided discovery systems

The integration of artificial intelligence into materials discovery has catalyzed a shift from sequential experimentation toward autonomous optimization ecosystems. Closed-loop laboratories exemplify this transformation, embedding machine learning models directly within experimental workflows to guide synthesis, characterization, and evaluation in real time [6, 7]. Such systems collapse traditional temporal boundaries between prediction and validation, enabling iterative refinement cycles that operate continuously rather than episodically.

Hierarchical active learning frameworks extend these capabilities into complex phase spaces, where adaptive sampling strategies enable exploration of nonequilibrium materials regimes that would be prohibitively costly under manual experimentation paradigms [7]. Bayesian optimization platforms have been benchmarked across multiple materials domains, demonstrating strong efficiency gains in navigating high-dimensional design spaces [9]. However, these gains remain contingent on initial data

distributions and embedded model assumptions, underscoring the path-dependent nature of autonomous discovery trajectories.

Ensemble learning approaches introduce additional layers of epistemic calibration by integrating uncertainty quantification directly into optimization workflows. Iteratively trained ensembles can guide automated experimentation while simultaneously estimating predictive confidence, supporting risk-aware discovery operations [12]. Complementary strategies address data scarcity, applying machine learning effectively within small-dataset regimes through transfer learning and uncertainty-aware sampling [2]. Collectively, these systems accelerate discovery throughput, yet their black-box optimization cores introduce reproducibility and epistemic validity concerns, particularly when decision pathways cannot be mechanistically interrogated [8]. Contemporary literature situates these developments within the maturation trajectory of materials informatics, framing the field as evolving toward fully integrated intelligent discovery ecosystems [1, 14].

Computational design paradigms

Inverse materials design represents a paradigmatic reorientation of computational engineering, shifting focus from forward property prediction to target-driven structure generation. Genetic algorithms, neural generative models, and hybrid optimization strategies now enable the navigation of complex structure–property manifolds with unprecedented efficiency [11, 15]. These frameworks reduce reliance on exhaustive search by steering candidate generation toward high-likelihood regions of design space.

Multifidelity learning architectures further augment design efficiency by integrating datasets of varying computational accuracy. By fusing low-cost approximations with high-precision calculations, such systems achieve improved predictive fidelity while conserving computational resources, particularly in applications involving dopant energetics and electronic bandgap estimation [16, 17]. However, reliance on probabilistic inference introduces epistemic vulnerabilities. Prediction errors, calibration biases, and uncertainty amplification can propagate through design pipelines, particularly when generative outputs are treated as deterministic recommendations [12].

Polymer informatics provides a representative case study in these dynamics. Reverse-mapping methodologies and evolutionary search strategies enable rapid traversal of compositional spaces, facilitating the discovery of functional polymer systems [15]. Yet critical assessments of regression-based design models reveal performance trade-offs, particularly in dielectric prediction contexts where data sparsity and nonlinear interactions complicate inference [18]. Broader conceptual analyses frame these developments within the emergence of machine learning–propelled intelligence ecosystems, positioning computational design as a central driver of next-generation materials innovation [14]. Still, the absence of mechanistic transparency in many generative systems reinforces the optimization–explanation tension that defines opaque discovery infrastructures [8].

Uncertainty & interpretability

Uncertainty quantification occupies a foundational role in ensuring epistemic reliability within materials AI systems. In environments characterized by noisy data, extrapolative predictions, and high-dimensional inference, quantified uncertainty functions as a critical safeguard against overconfident decision-making [12]. Ensemble learning techniques operationalize this principle by generating predictive distributions rather than point estimates, supporting automated experimentation in domains such as atom-resolved microscopy [12]. Multifidelity learning architectures similarly leverage hierarchical uncertainty structures to improve bandgap prediction fidelity, demonstrating how uncertainty modeling can enhance both accuracy and reliability [17].

Despite these advances, interpretability erosion remains a defining challenge of black-box materials informatics. As model complexity increases, the capacity to extract mechanistic rationale from predictions diminishes, raising questions regarding epistemic legitimacy and scientific accountability [4]. Efforts in explainable AI advocate for transparency across discovery workflows, from predictive modeling to autonomous design execution [8]. Complementary literature emphasizes the importance of integrating uncertainty-aware reasoning within small-data contexts and broader informatics ecosystems to strengthen epistemic robustness [1, 2]. High-throughput search initiatives targeting magnetic and topological materials further illustrate this interplay: predictive success is often contingent on quantified uncertainty, yet interpretive transparency lags behind predictive capability [10]. These literature strands map onto distinct infrastructural and epistemic functions within autonomous materials pipelines **(Table 1)**.

Table 1. Structural components underlying autonomous materials discovery and their epistemic roles, risks, and design pressures.

Dimension	What it covers in materials AI	What it enables	Sy epi
Data infrastructures	High-throughput simulation/experiment repositories; multimodal aggregation	Scale of search, reuse, cross-domain learning	Fra me sa
Closed-loop data feedback	Real-time experimental readout → model update	Rapid iteration, adaptive sampling	de “s da
Representation learning	Graph/attention encoders; latent structural manifolds	High-fidelity property prediction; non-linear structure capture	e
Optimization engines	Bayesian optimization, inverse design steering	Accelerated convergence to candidate sets	E
Uncertainty quantification	Ensemble/epistemic uncertainty signals	Risk-aware selection, calibration	Un un sil
Interpretability mechanisms	Explainable modeling; rationale recovery	Scientific credibility, mechanistic tracing	Int un
Ecosystem integration	“Intelligent” end-to-end discovery workflows	Autonomy, throughput, multi-stage coupling	ac ac

Proposed conceptual framework

The Epistemic Opaque Discovery System (EODS) framework is introduced as an original systems-theoretic construct for navigating black-box optimization in autonomous materials engineering. EODS conceptualizes discovery as a multilayered network comprising data ingestion layers, representation transformation modules, optimization engines, and feedback regulators. At the base, data infrastructures aggregate multimodal inputs from high-throughput simulations and closed-loop experiments, forming a dynamic repository that evolves through iterative updates [1, 6]. Representation learning architectures then encode these inputs into latent spaces, where graph-based transformations capture structural hierarchies without explicit physical decoding [3, 4].

Central to EODS is the black-box optimization core, which employs autonomous steering logics to guide inverse design and property exploration [7, 11]. Feedback loops integrate uncertainty signals to modulate decision paths, mitigating interpretability erosion by prioritizing epistemic risk assessment over pure efficiency [9, 12]. This structure highlights discovery workflow dynamics, where data-model-discovery pipelines interact via adaptive couplings, ensuring resilience in opaque environments [14].

A key dynamic within EODS can be conceptualized as the trade-off between optimization velocity and interpretability depth, expressed as:

$$\begin{matrix} I \\ = f(O,U) \\ = O \\ \cdot e \\ - \alpha U \end{matrix} \quad (1)$$

where I represents interpretability, O optimization output, U uncertainty measure, and α scaling factor capturing system sensitivity. This formula captures the interaction between rapid convergence and uncertainty-induced opacity, illustrating how elevated uncertainty exponentially diminishes interpretability while optimization proceeds [8, 12].

Another aspect formalizes feedback loop efficacy in steering black-box processes:

$$\begin{matrix} F \\ = \sum i \\ = 1nwi \\ \cdot (Di - Mi) \end{matrix} \quad (2)$$

with F as feedback strength, wi weights for pipeline components, Di data states, and Mi model inferences. This may be expressed as a measure of discrepancy-driven adjustment, emphasizing how deviations between

data and models propel autonomous corrections in closed-loop systems [6, 7].

Finally, representation-inference interplay is modeled as:

$$\begin{aligned} R \\ = g(S, L) \\ = S \\ \otimes L \\ + \beta\Delta \end{aligned} \quad (3)$$

where R is resultant representation, S structural input, L learning operator, ⊗ tensor convolution, β bias term, and Δ epistemic delta. This captures the convolution of raw structures with learned operators, augmented by bias to account for interpretability gaps [5, 13]. The framework is shown below in **figure 1**.

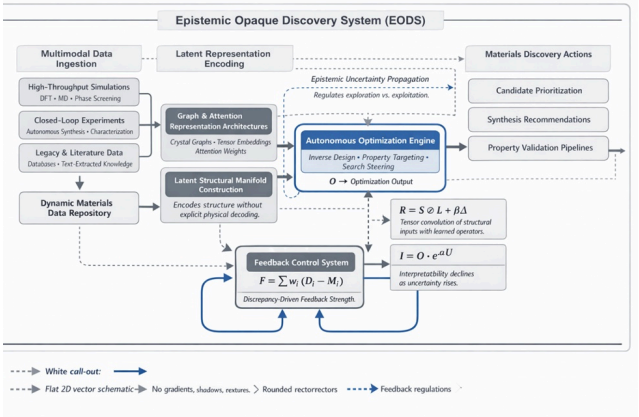

Figure 1. Epistemic Opaque Discovery System (EODS): A Systems Framework for Black-Box Optimization in Autonomous Materials Engineering

Conceptual architecture of the Epistemic Opaque Discovery System (EODS) illustrating multilayer interactions between multimodal data infrastructures, latent representation encoders, black-box optimization engines, and discrepancy-driven feedback regulators. Solid arrows denote forward discovery flows, dashed connectors indicate uncertainty propagation pathways, and curved arcs represent closed-loop corrective feedback. Embedded equations formalize interpretability–optimization trade-offs, feedback strength dynamics, and representation–inference coupling within autonomous materials discovery environments.

EODS thus provides interpretive insights into computational steering, highlighting infrastructure trade-offs in AI-driven ecosystems [2, 14].

Analytical implications

The Epistemic Opaque Discovery System (EODS) framework reveals several structural implications for how black-box optimization reshapes computational workflows in materials engineering.

First, the framework exposes a fundamental asymmetry in pipeline dynamics: optimization velocity grows superlinearly with model capacity and data volume, whereas interpretability depth tends to saturate or even degrade under the same scaling conditions [1, 8]. This asymmetry is not merely a methodological inconvenience; it is a structural feature of modern discovery pipelines in which representational power is rewarded far more directly than explanatory legibility. The implication is that incremental improvements in predictive performance—frequently achieved through deeper architectures, larger training corpora, or more aggressive hyperparameterization—do not automatically translate into commensurate gains in scientific insight. Instead, performance gains can coexist with declining intelligibility, shifting the locus of discovery control from human-guided mechanistic reasoning toward algorithmically mediated statistical inference [14]. In practice, this means that “better” models may expand the pipeline’s actionability while simultaneously reducing the field’s ability to justify why a given candidate was surfaced as promising.

Second, the feedback mechanisms formalized in EODS highlight that closed-loop systems are not merely iterative; they are self-reinforcing with respect to the representation spaces they inhabit. Once a model begins to dominate the steering logic of the pipeline, subsequent data acquisitions become increasingly conditioned on the model’s latent assumptions [6, 7, 9]. This conditioning can lead to representational lock-in, where the system preferentially explores regions of chemical space that align with existing learned patterns, potentially narrowing the effective discovery horizon even as throughput increases [2]. The feedback strength equation introduced earlier clarifies why this is structurally likely. When discrepancies between data states and model inferences are large, the loop generates strong corrective pressure and the system appears adaptive. However, when discrepancies diminish because the model has overfit, because the sampling policy has collapsed onto familiar regimes, or because data and model co-adapt around a restricted manifold, the feedback loop weakens. In that weakened state, the pipeline loses its ability to generate meaningful “surprise,” making it harder to

escape local optima in representation space even while the system remains highly productive in a narrow corridor of candidates.

Third, uncertainty propagation within EODS acts as a structural regulator rather than merely an error estimate. In black-box settings, uncertainty signals are frequently the only remaining epistemic tether to external validity [12]. When these signals are systematically underweighted in favor of exploitation-oriented acquisition functions, the pipeline becomes vulnerable to silent failure modes—cases where high-confidence but systematically biased predictions drive subsequent experiments [9]. Crucially, the interpretability–optimization trade-off equation indicates that uncertainty is not simply a measure of predictive fragility; under conditions of opacity, uncertainty becomes an amplifier of epistemic cost. Elevated uncertainty does not merely degrade interpretability in a linear way; it increases the expected scientific risk of acting on the model's outputs in a way that can compound across cycles, particularly when decision policies systematically reward confident outputs without auditing the basis of that confidence.

Fourth, the representation–inference interplay formalized in EODS suggests that modern graph-based and attention-driven architectures do not merely encode structure—they actively reshape the topology of the discovery manifold. The tensor convolution term in the representation equation indicates that learned operators can introduce non-local correlations that have no direct correspondence to classical physical descriptors [3, 13, 19]. While this capability drives predictive power, it simultaneously decouples the model's internal coordinate system from the physical coordinate system familiar to domain experts, creating a form of epistemic distance that grows with model complexity. As this distance grows, interpretive work shifts from physical reasoning to translation work, where scientists must infer meaning from latent neighborhoods, attention weights, or embedding proximities rather than from descriptors that map cleanly onto established mechanistic variables.

These implications can be organized as recurring system-level failure modes with corresponding governance levers and control variables **(Table 2)**.

Table 2. System-level failure modes in opaque autonomous discovery and the control levers implied by EODS equations and feedback structure.

Failure mode (EODS)	What it looks like in practice	Primary driver in your theory
Optimization–interpretability divergence	Pipeline gets "better" at hits but worse at explaining why	Superlinear optimization vs saturating interpretability
Representational lock-in	Exploration collapses to familiar latent neighborhoods	Closed-loop conditioning on learned priors
Silent high-confidence bias	Confident predictions steer experiments despite systematic error	Underweighted uncertainty + exploitation policies
Feedback attenuation under co-adaptation	System stops correcting because it "agrees with itself"	Data-model co-adaptation and metric complacency
Latent–physical decoupling	Model features don't map to physical descriptors	Non-local learned operators
Benchmark brittleness across datasets	Strong on one dataset, weak on another	Generalization limits in representation systems
Autonomy-induced reproducibility erosion	Hard to reproduce decisions/candidates	Opaque steering logic + evolving data

Taken together, these implications indicate that black-box optimization in autonomous materials engineering is not a neutral efficiency gain. It constitutes a qualitative reorganization of the discovery process, shifting emphasis

from explanation-centric science toward outcome-centric engineering while simultaneously altering the epistemic risk profile of the entire pipeline [8, 20]. The EODS lens therefore implies that claims of "accelerated discovery" should be evaluated not only in terms of throughput and hit rates, but also in terms of how opacity accumulates, how feedback narrows the search manifold, and how uncertainty is permitted—or not permitted—to regulate action.

Results and Discussion

The conceptual structure of EODS offers a way to reason about the long-term evolution of data-driven materials discovery ecosystems without reducing the analysis to performance metrics or empirical case studies.

One central observation is that the tension between optimization and explanation is not a temporary artifact of immature methodologies. Rather, it is an intrinsic property of scalable, representation-heavy inference systems operating under resource and time constraints [1, 21]. As autonomous pipelines mature, the fraction of discovery steps that occur inside opaque inference layers is likely to increase, not decrease. This trajectory implies that interpretability can no longer be treated as a post-hoc retrofit; it must be architected as a systemic property that competes with throughput and accuracy within the same optimization objective [8]. Under EODS, interpretability is better understood as a pipeline-level design choice—supported by interfaces, audits, and steering constraints—than as a model-level add-on.

The framework also suggests that current uncertainty quantification strategies, while valuable, remain insufficient when applied only at the prediction level. In EODS, uncertainty must propagate backward through representation layers and forward through decision layers to meaningfully modulate steering behavior [12]. This requirement points toward the need for architectures and control logics that treat uncertainty as a first-class object flowing through the pipeline, rather than an auxiliary output. Such designs would represent a departure from conventional supervised learning paradigms and align more closely with control-theoretic or active inference perspectives on autonomous systems [22]. The point is not simply to "estimate uncertainty," but to ensure that uncertainty has causal authority over what the system chooses to do next.

Another implication concerns infrastructure design. The layered structure of EODS implies that long-term progress in materials discovery will depend less on isolated model breakthroughs and more on the robustness of interfaces between layers—particularly between data infrastructures and representation modules, and between inference engines and feedback regulators [7, 18]. Weak interfaces at these junctions amplify epistemic opacity and reduce the system's ability to recover from drift or distributional shift. In that sense, epistemic reliability becomes an emergent property of the whole pipeline architecture, not something that can be guaranteed by improved predictive metrics alone.

Finally, the framework reframes the role of human oversight in autonomous discovery. Rather than serving as final validators of model outputs, humans may increasingly act as designers and curators of the systemic constraints that govern black-box behavior: defining acceptable uncertainty thresholds, shaping feedback weighting, and selecting representation biases that preserve domain-relevant structure [14, 15]. This shift does not eliminate scientific agency, but it relocates it. Scientists retain control indirectly through the governance parameters that determine how opaque inference is permitted to steer experimentation and how much epistemic risk is acceptable during accelerated exploration.

These considerations collectively suggest that the future trajectory of computational materials engineering will be determined not only by advances in algorithms and data, but by the degree to which the field can develop theoretical and infrastructural tools capable of managing opacity at scale.

Conclusion

The rise of black-box optimization in autonomous materials engineering has produced remarkable acceleration in discovery throughput, yet it simultaneously challenges the epistemic foundations that have historically underpinned materials science. By treating discovery as a systemic phenomenon rather than a sequence of isolated predictions, the Epistemic Opaque Discovery System (EODS) framework illuminates structural dynamics that are often obscured in performance-focused analyses.

The framework shows that rapid optimization is not cost-free: it systematically trades interpretability depth,

representational fidelity, and epistemic robustness for gains in cycle time and scale. Feedback loops, uncertainty flows, and representation–inference couplings emerge as critical sites where these trade-offs are negotiated, and where future resilience must be engineered.

Rather than advocating a return to fully interpretable, low-capacity models, EODS points toward a hybrid path: one in which opacity is acknowledged as an inherent feature of scalable inference, but managed through deliberate design of data–model–decision interfaces, uncertainty-aware steering logics, and meta-level governance structures. Such an approach preserves the transformative potential of autonomous systems while mitigating the risk of epistemic erosion.

Ultimately, the continued vitality of data-driven materials engineering will depend on the field's ability to integrate powerful black-box optimizers into larger epistemic architectures that remain answerable to scientific reasoning. The EODS framework provides one conceptual scaffold for reasoning about that integration.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 29 Jul 2021 Revised: 03 Sep 2021 Accepted: 03 Oct 2021
Published online: 18 March 2022

References

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(1):54.

Zhang Y, Ling C. A strategy to apply machine learning to small datasets in materials science. npj Comput Mater. 2018;4(1):25.

Fung V, Zhang J, Juarez E, Sumpter BG. Benchmarking graph neural networks for materials chemistry. npj Comput Mater. 2021;7(1):84.

Choudhary K, DeCost B. Atomistic line graph neural network for improved materials property predictions. npj Comput Mater. 2021;7(1):185.

Gong W, Yan Q. Graph-based deep learning frameworks for molecules and solid-state materials. Comput Mater Sci. 2021;195:110332.

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 Commun. 2020;11(1):5966.

Connolly AB, Sutherland DR, Amsler M, Chang M-C, Gann KR, Ament S, et al. Autonomous materials synthesis via hierarchical active learning of nonequilibrium phase diagrams. Sci Adv. 2021;7(51):abg4930.

Pilania G. Machine learning in materials science: From explainable predictions to autonomous design. Comput Mater Sci. 2021;193:110360.

Liang Q, Gongora AE, Ren Z, Tiihonen A, Liu Z, Sun S, et al. Benchmarking the performance of Bayesian optimization across multiple experimental materials science domains. npj Comput Mater. 2021;7(1):188.

Frey NC, Huddart M, Miller JM, Gaulton SM, Persson KA, Stevanović V. High-throughput search for magnetic and topological order in transition metal oxides. Sci Adv. 2020;6(50):abd1076.

Kim B, Kim J, Lee S. Inverse design of porous materials using artificial neural networks. Sci Adv. 2020;6(1):eaax9324.

Ghosh A, Sumpter BG, Dyck O, Kalinin SV, Ziatdinov M. Ensemble learning-iterative training machine learning for uncertainty quantification and automated experiment in atom-resolved microscopy. npj Comput Mater. 2021;7(1):100.

Schmidt J, Shi J, Borlido P, Chen L, Botti S, Marques MAL. Crystal graph attention networks for the prediction of stable materials. Sci Adv. 2021;7(49):abi7948.

Batra R, Song L, Ramprasad R. Emerging materials intelligence ecosystems propelled by machine learning. Nat Rev Mater. 2021;6:655-78.

Tran H, Batra R, Kim C, Huck P, Ramprasad R. Polymer design using genetic algorithm and reverse mapping. Comput Mater Sci. 2020;186:110030.

Batra R, Pilania G, Uberuaga LA, Ramprasad R. Multifidelity information fusion with machine learning: A case study of dopant formation energies in rare earth elements. Comput Mater Sci. 2019;161:414-419.

Pilania G, Gubernatis JE, Lookman T. Multi-fidelity machine learning models for accurate bandgap predictions of solids. Comput Mater Sci. 2017;129:156-163.

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(1):21. https://doi.org/10.1038/s41524-019-0153-8.

Vasudevan RK, Pilania G, Balachandran PV, Lookman T. Machine learning for materials design and discovery. J Appl Phys. 2021;129(7):070401. https://doi.org/10.1063/5.0043300.

Cai J, Chu X, Xu K, Li H, Wei J. Machine learning-driven new material discovery. Natl Sci Rev. 2020;7(6):1054–63. https://doi.org/10.1093/nsr/nwz146.

Umehara M, Stein HS, Lookman T. Analyzing machine learning models to accelerate materials discovery. npj Comput Mater. 2019;5(1):34. https://doi.org/10.1038/s41524-019-0172-5.

Kailkhura B, Gallagher B, Kim S, Hiszpanski A, Han TY-J. Reliable and explainable machine-learning methods for accelerated material discovery. npj Comput Mater. 2019;5(1):101. https://doi.org/10.1038/s41524-019-0248-2.