ORIGINAL RESEARCH Open access

Exploration without Explanation: A Conceptual Critique of Black-Box Materials Discovery

Hiroshi Tanaka[1], Yuki Sato[1*], Kenji Mori[2], Rina Okabe[1]

Abstract

The integration of machine learning into materials discovery has accelerated exploratory processes, yet it often privileges predictive accuracy over interpretive clarity. This manuscript examines the conceptual tensions arising from black-box approaches in materials science, where algorithmic opacity obscures the underlying logics of material behaviors and interactions. By synthesizing recent literature on explainable artificial intelligence within computational materials contexts, the analysis highlights epistemic trade-offs between rapid exploration and the need for explanatory depth. Conceptual interpretations reveal how opaque models may reinforce feedback loops of uncertainty, limiting the integrative understanding of material systems. The framework interprets these dynamics through steering logics that balance algorithmic efficiency with interpretive accessibility, emphasizing ethical considerations in knowledge production. Systems-level insights underscore the interplay between data-driven discovery and human-centric reasoning, suggesting that unexamined opacity could constrain the broader interpretive landscape of materials innovation. This critique advocates for a reflective integration of explainability, not as a corrective add-on, but as an intrinsic dimension of exploratory practices. Ultimately, the discussion fosters a nuanced appreciation of how explanation shapes the conceptual boundaries of discovery, urging a reevaluation of priorities in computational materials paradigms.

Keywords Conceptual critique, Epistemic trade-offs, Materials discovery, Black-box models, Explainable AI, Interpretive integration

*Correspondence:
Yuki Sato
yuki.sato@gmail.com
[1] Department of Intelligent Materials Engineering, Faculty of Engineering, University of Tokyo, Tokyo, Japan
[2] Department of AI-Driven Materials Discovery, Faculty of Information Science, Kyoto University, Kyoto, Japan

Introduction

In the evolving landscape of materials science, the advent of advanced computational techniques has reshaped the paradigms of discovery. Traditional approaches, rooted in empirical experimentation and theoretical modeling, have increasingly intersected with data-driven methodologies, particularly those leveraging machine learning algorithms. These tools facilitate vast explorations of material spaces, enabling the identification of novel compounds and properties through pattern recognition in large datasets. However, this shift introduces conceptual complexities, especially when models operate as black boxes—systems where internal decision-making processes remain inaccessible to human interpreters [1, 2]. The critique presented here focuses on the implications of such opacity, interpreting how it influences the conceptual foundations of knowledge generation in materials discovery.

The allure of black-box models lies in their capacity to navigate complex, high-dimensional spaces efficiently, often outperforming interpretable alternatives in predictive tasks [3, 4]. Yet, this efficiency comes at the cost of explanatory insight, raising questions about the nature of understanding in scientific inquiry. In materials contexts, where properties emerge from intricate atomic and molecular interactions, the absence of clear rationales for predictions can hinder the integration of new findings into broader theoretical frameworks [5, 6]. This interpretive gap not only affects individual researchers but also permeates collaborative ecosystems, where shared comprehension is essential for advancing collective knowledge.

Epistemic considerations further complicate this picture. Black-box approaches may inadvertently prioritize correlative patterns over causal relationships, fostering a form of knowledge that is performative rather than elucidative [7, 8]. In materials discovery, this manifests as accelerated identification of candidates without concomitant articulation of why certain structures yield specific functionalities. Such dynamics invite reflection on the steering logics that guide research priorities: does the pursuit of speed and scale overshadow the value of conceptual depth? Ethical dimensions emerge here, as opaque processes could perpetuate biases embedded in training data, subtly directing discovery pathways in unintended ways [9, 10].

Moreover, the systems-level interactions between human experts and algorithmic tools warrant scrutiny. Materials science inherently involves multidisciplinary integration, spanning chemistry, physics, and engineering. When black-box models dominate, the feedback structures between computational outputs and human interpretation may become asymmetrical, with algorithms dictating exploratory directions without reciprocal accountability [11, 12]. This imbalance prompts an examination of how interpretive practices can be reinstated, not as barriers to progress, but as enhancers of robust discovery.

Historical precedents in scientific methodology underscore these concerns. Earlier shifts toward computational simulation in materials research emphasized transparency to validate outcomes [13, 14]. In contrast, contemporary black-box methods challenge this norm, potentially fragmenting the cohesive narrative of material behaviors. Conceptual interpretations of this evolution reveal a tension between exploration as an end in itself and exploration as a means to explanatory synthesis. The former risks reducing discovery to a probabilistic exercise, while the latter integrates findings into enduring conceptual schemas.

Trade-offs in this domain extend to practical realms, such as the reproducibility of discoveries. Without explanatory mechanisms, replicating or extending black-box-derived insights becomes challenging, as the underlying logics remain veiled [15, 16]. This opacity can create epistemic silos, in which advancements in one subfield fail to inform others because of inaccessible rationales. In response, recent scholarly discourse has begun to advocate for hybrid approaches that blend predictive power with interpretive tools, though these remain conceptually underexplored [17, 18].

The manuscript's analytical lens interprets these issues through a framework that elucidates the dynamics of opacity in materials discovery. By focusing on interaction patterns and feedback loops, it highlights how unexamined reliance on black boxes may constrain the field's interpretive horizon. This perspective encourages a reevaluation of discovery as an integrative process, where explanation serves as a conceptual bridge between data and theory [19, 20].

Ultimately, this critique contributes to ongoing dialogues by interpreting the conceptual costs of opacity. It posits that sustainable progress in materials science demands a balanced orchestration of exploratory zeal and explanatory rigor, fostering systems where knowledge is not merely accumulated but meaningfully interconnected [1, 21].

Theoretical Background and Literature Synthesis

Emergence of machine learning in materials discovery

The incorporation of machine learning into materials science has transformed exploratory paradigms, enabling the navigation of vast compositional and structural landscapes. This evolution interprets data as a primary driver of insight, where algorithms discern patterns from extensive repositories of material properties [1, 3]. Conceptual analyses reveal how these methods facilitate a shift from hypothesis-driven to data-centric inquiry, altering the interaction dynamics between researchers and material systems [5, 7]. In this context, black-box models emerge as efficient explorers, yet their opacity introduces interpretive challenges, as the logics governing predictions often elude direct scrutiny [9, 11].

Recent literature interprets this integration as a double-edged sword: while accelerating discovery, it complicates the synthesis of findings into coherent theoretical narratives [13, 15]. Systems-level insights suggest that without transparent linkages, discoveries risk isolation from broader material principles, affecting the feedback structures that sustain scientific advancement [17, 19].

Conceptual dimensions of black-box opacity

Black-box opacity in machine learning refers to the inaccessibility of internal decision-making pathways, a feature prevalent in deep neural networks used for materials prediction [2, 4]. Interpretive examinations highlight how this characteristic prioritizes outcome over process, potentially undermining epistemic reliability in discovery contexts [6, 8]. Ethical reasoning underscores the implications for knowledge equity, as opaque models may amplify data biases, steering exploratory trajectories in subtle, unaccountable ways [10, 12].

Synthesis of contemporary works interprets opacity as a barrier to integrative understanding, where material behaviors are reduced to probabilistic outputs rather than explained through interaction dynamics [14, 16]. This perspective reveals trade-offs between computational scalability and conceptual accessibility, prompting reflections on how opacity shapes the interpretive landscape of materials science [18, 20].

Explainability paradigms in computational materials

Efforts to address opacity have led to the development of explainable artificial intelligence (XAI) techniques tailored to materials domains [3, 21]. Conceptual interpretations of these paradigms emphasize post-hoc methods that approximate black-box decisions, offering partial insights into feature importance or decision boundaries [2, 4, 22, 23]. However, literature syntheses caution that such approximations may introduce their own interpretive ambiguities, complicating the integration of explanations into discovery workflows [1, 2].

Systems-level analyses interpret XAI as a mediator in human-algorithm interactions, facilitating feedback loops that enhance conceptual coherence [9, 22]. Yet, the challenge lies in balancing explanatory depth with exploratory breadth, as overly intricate interpretations could constrain the efficiency of black-box approaches [8, 10].

Epistemic and ethical trade-offs

The epistemic trade-offs of black-box discovery manifest in the tension between rapid iteration and sustained understanding [5, 24]. Analytical implications suggest that without explanatory anchors, material knowledge may fragment, hindering cross-disciplinary integrations essential for innovation [1, 2]. Ethical considerations interpret this as a question of stewardship: opaque processes risk entrenching power imbalances in knowledge production, where algorithmic authority overshadows human interpretive agency [3, 4].

Literature integrations reveal how these trade-offs influence steering logics in research, potentially prioritizing short-term gains over long-term conceptual robustness [5, 6]. This synthesis underscores the need for reflective practices that interpret opacity not as an inevitable byproduct, but as a conceptual pivot for reimagining discovery ethics [7, 8].

Integration challenges in materials systems

Materials discovery involves complex systems where properties arise from multifaceted interactions [9, 10]. Conceptual critiques interpret black-box models as disruptors of integrative synthesis, as their outputs often lack the connective tissue to link predictions with underlying dynamics [11, 12]. Feedback structures in literature highlight how this disconnection can perpetuate cycles of uncertainty, limiting the field's interpretive evolution [13, 14].

Synthesis of recent perspectives emphasizes the role of hybrid frameworks in mitigating these challenges, interpreting them as bridges between opacity and clarity [15, 16]. Such integrations foster a more holistic conceptual environment in which exploration and explanation coexist in dynamic equilibrium [17, 18].

Proposed conceptual framework

Epistemic steering logics in black-box materials discovery

1. Conceptual orientation: Discovery as an epistemically steered system

The proposed framework interprets black-box materials discovery through the lens of epistemic steering logics, positioning discovery not as an autonomous algorithmic act but as a dynamically guided knowledge process. Within this perspective, artificial intelligence systems function as navigational engines operating across vast material possibility spaces, yet their trajectories are shaped by the interpretive infrastructures that surround them. Opacity is therefore conceptualized not merely as a technical limitation but as an epistemic modulator that conditions how knowledge is generated, transmitted, and integrated.

Discovery emerges as a networked system of interactions in which algorithmic exploration, human interpretation, and conceptual validation operate as interdependent strata. The framework foregrounds the tension between exploratory efficiency—often maximized through black-box optimization—and interpretive integration, which sustains scientific coherence. In doing so, it reframes reliance on black boxes as a steering condition that reshapes the topology of discovery pathways rather than simply accelerating them.

2. Opacity as a modulator of knowledge flows

At the structural core of the framework lies the interpretation of opacity as a regulatory filter governing epistemic circulation. Black-box architectures, while computationally powerful, introduce gradients of visibility that shape which forms of knowledge are amplified and which remain latent.

Analytical implications reveal that unexamined reliance on opaque predictions can generate asymmetric feedback loops. Predictive outputs become recursively privileged, steering subsequent exploration toward performance optimization while marginalizing conceptual interrogation. Over time, such asymmetry risks narrowing the interpretive bandwidth of discovery ecosystems.

Opacity thus functions as a directional force—not halting discovery, but bending its epistemic trajectory toward outcome-dominant reasoning rather than mechanism-based understanding.

3. Systems-level mapping of conceptual tensions

From a systems perspective, the framework maps materials discovery as a landscape of intersecting tensions. Algorithmic pathways traverse high-dimensional material spaces, yet these routes intersect with human interpretive domains where meaning, plausibility, and scientific value are negotiated.

Steering logics operate within this intersection. When opacity intensifies, integrative processes fragment, producing localized predictive successes without systemic coherence. Conversely, interpretive enrichment fosters cross-domain synthesis, linking predictions to theory, mechanisms, and design principles.

This systems-level interpretation positions opacity as a potential disruptor of epistemic continuity, capable of segmenting discovery into computational silos unless counterbalanced by integrative infrastructures.

4. Recursive interaction structures

The framework conceptualizes interaction patterns among data, models, and interpreters as recursive rather than sequential. Predictive outputs inform subsequent exploratory cycles, shaping dataset prioritization, parameter navigation, and experimental validation strategies.

Within opaque systems, these recursive loops become epistemically vulnerable. Biases embedded within training corpora or optimization logics may propagate undetected, reinforcing skewed discovery trajectories. The absence of interpretive checkpoints allows such biases to sediment into the exploratory architecture itself.

Trade-offs, therefore, emerge in the allocation of conceptual resources:

- Prioritizing black-box speed enhances exploratory throughput but compresses interpretive depth.
- Emphasizing explainability enriches systems understanding but introduces cognitive and computational overhead.

5. Ethical reasoning and accountability trade-offs

Ethical reasoning is embedded within the framework as a structural dimension rather than an external evaluation layer. Opacity reshapes accountability architectures by obscuring causal traceability between model inference and scientific conclusion.

Rapid exploration may yield accelerated innovation, yet it risks attenuating long-term epistemic cohesion. Without interpretive transparency, responsibility for error propagation, design misdirection, or material risk becomes diffused across human–machine assemblages.

Thus, the framework views black-box discovery as an ethical negotiation space, in which efficiency gains must be weighed against epistemic responsibility and scientific accountability.

6. Explanation as a connective epistemic medium

Rejecting linear discovery metaphors, the framework advances the notion of a conceptual ecology. Within this ecology, explanation operates as the connective substrate linking predictive signals to scientific meaning.

Explanatory interfaces—particularly those enabled through Explainable AI (XAI)—are interpreted as integrative bridges. They recalibrate discovery dynamics by reinserting mechanistic reasoning, causal mapping, and uncertainty contextualization into otherwise opaque pipelines.

Analytical implications suggest that deliberate integration of explanatory infrastructures can:

- Reinforce conceptual robustness
- Enhance cross-domain interpretability
- Prevent exploratory isolation
- Sustain theoretical coherence

7. Integrative knowledge ecosystems

The framework ultimately envisions materials discovery as an adaptive epistemic ecosystem capable of accommodating diverse material contexts and modeling paradigms. Opacity influences the ecosystem's structural plasticity—either constraining adaptive learning or enabling reflective recalibration when mediated through interpretive systems.

Rather than framing black-box AI as inherently flawed, the framework positions it as an opportunity structure—a catalyst for developing more resilient integrative architectures that balance predictive power with epistemic depth. **Figure 1** illustrates the conceptual framework of epistemic steering, depicting black-box discovery as a layered vertical ecology in which an algorithmic, exploratory base interacts recursively through an opacity-modulating filter with an upper stratum of human interpretive integration.

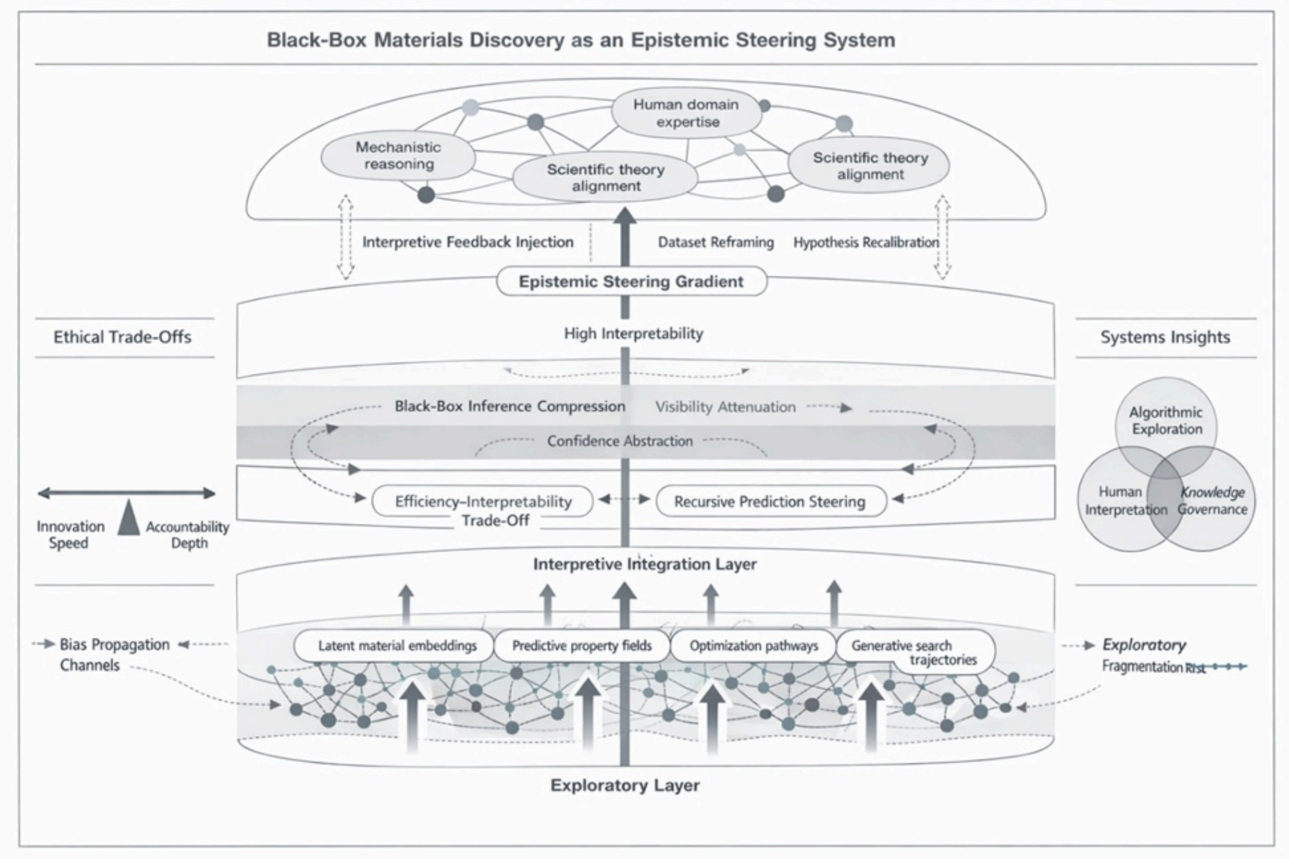

Figure 1. Layered conceptual ecology of epistemic steering in black-box materials discovery, depicting the recursive integration of algorithmic exploration, opacity modulation, and human interpretive synthesis along a gradient of interpretability.

Analytical implications

The interpretive lens applied to black-box materials discovery reveals several layered implications for the field's epistemic architecture. One central dynamic concerns the reconfiguration of knowledge authority within exploratory processes. When algorithmic opacity dominates, interpretive agency shifts toward the model's latent patterns, subtly repositioning human researchers as interpreters of outputs rather than co-authors of conceptual frameworks [1, 3]. This shift involves a form of epistemic delegation, in which the steering logic of discovery increasingly relies on unexamined computational heuristics, potentially eroding the integrative capacity that has historically characterized materials science [5, 7].

Systems-level analysis further interprets how opacity propagates through feedback structures. Black-box predictions, by virtue of their lack of transparent rationales, generate recursive cycles in which subsequent explorations are conditioned by prior opaque decisions [9, 11]. Such cycles can amplify certain material subspaces while marginalizing others, creating path-dependent trajectories that are difficult to interrogate conceptually [13, 15]. The implication here is not merely inefficiency but a structural constraint on the conceptual plasticity of the discovery process—where the field's ability to adaptively reinterpret material behaviors becomes tethered to algorithmic momentum rather than reflective synthesis [17, 19].

Ethical reasoning interprets these dynamics as introducing asymmetries in accountability. In opaque systems, responsibility for erroneous or biased discoveries disperses across data provenance, model architecture, and interpretive practices, complicating attribution and remediation [3, 21]. This dispersion interprets a form of distributed opacity that challenges traditional norms of scientific stewardship, particularly in domains where material innovations carry societal consequences [4, 23]. The analytical implication is that unaddressed reliance on black boxes may inadvertently institutionalize forms of knowledge production that prioritize scale over traceability [5, 10].

Another layer of implication emerges in the interaction between disciplinary boundaries. Materials discovery inherently spans chemistry, physics, and engineering; black-box approaches, by compressing complex interactions into abstract representations, can obscure the cross-domain conceptual bridges essential for holistic understanding [2, 4]. Interpretive consequences include the potential fragmentation of knowledge ecosystems, where subfields advance in parallel without sufficient conceptual interoperability [6, 8]. This fragmentation interprets a risk to the field's long-term coherence, as explanatory voids hinder the emergence of unifying theoretical narratives [10, 12].

Trade-offs in resource allocation also warrant analytical attention. The pursuit of black-box efficiency often demands substantial computational infrastructure and data curation, diverting conceptual effort from interpretive refinement toward optimization of predictive performance [14, 16]. This redirection interprets a subtle redefinition of discovery priorities, where conceptual depth is subordinated to exploratory throughput [18, 20]. The implication is that sustained reliance on opacity may reshape the intellectual culture of materials science, privileging performative metrics over integrative insight [1, 2]. The layered epistemic trade-offs and steering dynamics emerging from black-box materials discovery are synthesized in **Table 1**.

Table 1. Epistemic trade-offs and steering dynamics in black-box materials discovery

Framework dimension	Conceptual role in discovery	Epistemic risks introduced by opacity	Systems-level consequences	Ethical/interpretive implications	Integrative mitigation pathways
Algorithmic exploration	Enables high-throughput navigation of vast compositional and structural material spaces through predictive modeling	Overreliance on correlative pattern recognition; limited mechanistic transparency	Accelerated candidate identification without parallel theory formation	Shifts discovery authority toward model inference rather than scientific reasoning	Hybrid modeling integrating physics-based constraints and interpretable layers
Predictive efficiency	Optimizes property prediction accuracy and screening speed	Privileges performance metrics over explanatory depth	Reinforces outcome-dominant discovery cultures	Normalizes performative knowledge production	Multi-objective optimization incorporating interpretability metrics
Opacity modulation	Filters knowledge flows between model outputs and human interpretation	Attenuates causal traceability; obscures inference pathways	Produces fragmented epistemic circulation	Diffuses accountability across human–machine systems	Layered explainability interfaces and transparency audits

Feedback loop recursion	Guides iterative discovery cycles through model-informed exploration	Bias propagation across training and inference cycles	Path-dependent exploration trajectories	Reinforces embedded dataset inequities	Bias diagnostics and recursive interpretive checkpoints
Interpretive integration	Assimilates predictions into scientific theory and design logic	Reduced integration capacity when explanations are absent	Weak cross-domain conceptual synthesis	Marginalizes disciplinary dialogue	Human-AI co-interpretation frameworks
Knowledge authority redistribution	Repositions epistemic agency within discovery ecosystems	Delegation of reasoning to opaque architectures	Erosion of the researcher’s interpretive primacy	Raises questions of epistemic justice and governance	Participatory interpretive validation structures
Reproducibility infrastructure	Supports validation and extension of discoveries	Limited reproducibility due to inaccessible reasoning pathways	Epistemic silos across subfields	Constrains cumulative knowledge building	Transparent model documentation and interpretive reporting standards
Cross-disciplinary connectivity	Links chemistry, physics, and engineering insights	Compression of complex interactions into abstract latent spaces	Fragmentation of interdisciplinary synthesis	Weakens shared scientific vocabularies	Mechanistic mapping and causal representation overlays
Resource allocation logics	Directs computational and intellectual investment	Prioritizes infrastructure over conceptual reasoning	Cultural shift toward scale-centric discovery	Devaluation of interpretive scholarship	Balanced funding of interpretive and exploratory infrastructures
Ethical accountability structures	Governs responsibility in discovery outcomes	Distributed opacity obscures error attribution	Institutionalization of untraceable decision chains	Risks of societal and industrial misapplication	Ethical XAI frameworks and traceability protocols
Explanatory interface development	Provides interpretive bridges between prediction and theory	Post-hoc explanations may introduce interpretive distortions	Partial restoration of epistemic visibility	Creates new interpretive dependencies	Intrinsic explainability embedded in model design
Epistemic steering logics	Directs discovery trajectories through exploration–explanation balance	Skew toward speed and scale over conceptual depth	Systemic narrowing of interpretive horizons	Reconfigures norms of scientific progress	Reflexive steering audits and meta-interpretive governance

Conceptual ecology formation	Frames discovery as an interconnected knowledge ecosystem	Disruption of ecological coherence under opacity	Isolated predictive islands	Weak theoretical consolidation	Integrative discovery platforms linking AI, theory, and experiment

Finally, the framework interprets black-box discovery as a site of conceptual contingency. Opacity does not inherently invalidate findings, but it renders their epistemic status provisional—dependent on post-hoc validation or serendipitous alignment with existing theory [9, 22]. This contingency interprets an opportunity for reflexive practice. By acknowledging opacity as a structural feature rather than a flaw, the field can cultivate interpretive strategies that treat black-box outputs as prompts for conceptual reconstruction rather than final truths [8, 24].

Results and Discussion

The conceptual critique articulated through the proposed framework invites reflection on the broader implications of opacity in contemporary materials discovery. At its heart lies the tension between the accelerative promise of black-box methods and the integrative demands of scientific understanding. While these approaches have undeniably expanded the accessible frontiers of material space, their interpretive costs manifest as fragmented knowledge flows and attenuated explanatory depth [1, 2, 5, 6]. This tension does not call for abandoning black-box techniques but rather for their situated reconfiguration within a wider epistemic ecology.

Interpretive integration emerges as a pivotal response. Rather than treating explainability as an afterthought, the framework suggests embedding reflective practices that treat opacity as a diagnostic signal. Such practices would interpret black-box outputs not as endpoints but as inflection points—invitations to reexamine material interactions through complementary lenses of theory, simulation, and experiment [3, 7, 9, 11]. This approach interprets discovery as a dialogic process in which algorithmic exploration and human conceptualization mutually enrich one another.

Ethical considerations further enrich this dialogue. The distributed nature of opacity in black-box systems raises questions of epistemic justice: whose conceptual priorities are amplified or silenced by algorithmic steering? [3, 13, 15, 21] Addressing this requires deliberate mechanisms that foreground transparency in data curation, model selection, and output interpretation, ensuring that exploratory trajectories remain accountable to diverse scientific communities [4, 5, 10, 23].

The systems-level perspective also highlights the recursive character of discovery. Feedback loops between opaque predictions and subsequent explorations can entrench certain conceptual commitments while foreclosing others [4, 8, 10, 12]. Recognizing this recursivity interprets the need for meta-level steering—practices that periodically interrogate the conceptual assumptions embedded in black-box pipelines [14, 16, 18, 20]. Such steering would not seek to eliminate opacity but to modulate its influence, preserving exploratory agility while safeguarding interpretive resilience.

Cross-disciplinary implications warrant attention as well. Materials science's reliance on black-box methods intersects with parallel developments in chemistry, physics, and data science, where similar tensions between prediction and explanation prevail [1, 2, 17, 19]. Conceptual dialogues across these domains could foster shared interpretive vocabularies, mitigating the risk of isolated advancement [8, 9, 22, 24]. This convergence interprets an opportunity to reimagine discovery as a transdisciplinary endeavor, where opacity serves as a shared problem space rather than a disciplinary barrier.

Ultimately, the critique reframes black-box materials discovery not as a technological inevitability but as a conceptual choice—one that invites ongoing reflection on the values that guide scientific progress. By interpreting opacity as a site of productive tension, the field can cultivate practices that harmonize the scale of exploration with the depth of understanding, ensuring that materials innovation remains both expansive and epistemically grounded.

Conclusion

This manuscript has interpreted the conceptual landscape of black-box materials discovery through a critical and integrative lens. By synthesizing recent scholarship on machine learning applications in materials science and situating it within broader epistemic and ethical considerations, the analysis has illuminated the trade-offs inherent in privileging predictive opacity over explanatory clarity. The proposed framework further interprets these trade-offs as dynamic steering logics, where interaction patterns, feedback structures, and accountability asymmetries shape the contours of knowledge production.

The central interpretive insight is that opacity, while enabling rapid exploration, simultaneously constrains the integrative potential of discovery. Rather than viewing this constraint as an insurmountable barrier, the critique reframes it as an invitation to rebalance exploratory and explanatory priorities. Conceptual recalibration—through deliberate incorporation of interpretive interfaces, reflexive meta-practices, and cross-disciplinary dialogue—offers pathways toward more resilient and equitable knowledge ecosystems.

As materials science continues to navigate the convergence of computation and experimentation, sustained attention to the conceptual costs of opacity will remain essential. The future vitality of the field depends not only on the scale of discoveries but on their capacity to enrich shared interpretive frameworks. By treating explanation as an intrinsic dimension of exploration, rather than an optional enhancement, materials discovery can aspire to a more cohesive and reflexive form of scientific advancement.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 11 Aug 2022 Revised: 24 Sep 2022 Accepted: 01 Dec 2022
Published online: 18 January 2023

References

Chen X, Yu D, Zhao L, Liu F. Explainable machine learning in materials science. npj Comput Mater. 2022;8:204.

Oviedo F, Ferres JL, Buonassisi T, Butler KT. Interpretable and explainable machine learning for materials science and chemistry. Acc Mater Res. 2022;3(6):597-607.

Allen AEA, Tkatchenko A. Machine learning of material properties: Predictive and interpretable multilinear models. Sci Adv. 2022;8(18):eabm7185.

Lu Z, Pu J, Wang F, He Z, Hu C, Wang L. Interpretable machine-learning strategy for soft-magnetic property and thermal stability in Fe-based metallic glasses. npj Comput Mater. 2020;6:187.

Braham EJ, Ruddock JM, Hardin JO. Generating and leveraging explanations of AI/ML models in materials and manufacturing research. Patterns (N Y). 2025;6(9):101340.

Nematov D, Hojamberdiev M. Machine learning-driven materials discovery: unlocking next-generation functional materials – a review. J Mater Inform. 2024;4:14.

Li Z, Yang Y, Nematov D, Hojamberdiev M. Advancing materials discovery through artificial intelligence. Appl Mater Today. 2025;41:101659.

Semnani P, Vormittag J, Haghighatlari M, Hüser F, Müller KR. A machine learning and explainable AI framework tailored for unbalanced experimental catalyst discovery. J Phys Chem C. 2024;128(49):35634-46.

Merchant A, Batzner S, Schoenholz SS, Aykol M, Cheon G, Cubuk ED. Scaling deep learning for materials discovery. Nature. 2023;624(7990):80-5.

Chen C, Zuo Y, Ye W, Li XG, Deng Z, Ong SP. A critical review of machine learning of energy materials. Adv Energy Mater. 2020;10(8):1903242.

Longo L, Brcic M, Cabitza F, Choi J, Confalonieri R, Del Ser J, et al. Explainable artificial intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. Inf Fusion. 2024;106:102301.

Rudin C, Chen C, Chen Z, Huang H, Semenova L, Zhong C. Interpretable machine learning: fundamental principles and 10 grand challenges. Stat Surv. 2022;16:1-85.

Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: a review of machine learning interpretability methods. Entropy. 2021;23(1):18.

Dan Y, Zhao Y, Li X, Li S, Hu M, Hu J. Generative adversarial networks (GAN) based efficient sampling of chemical composition space for inverse materials design. npj Comput Mater. 2020;6:84.

Lu S, Zhou Q, Ouyang Y, Guo Y, Li Q, Wang J. Accelerated discovery of metamaterial design rules via neural architecture search. Nat Commun. 2021;12:3884.

Esterhuizen J, Goldsmith BR, Linic S. Interpretable machine learning for catalytic materials design. Chem. 2020;6(12):3100-17.

Schilter O, Vaucher A, Schwaller P, Laino T. Interpretable machine learning for chemistry and materials science. Digit Discov. 2023;2(3):728-35.

Zhang Z, Chen X. Neural scaling laws from large-N field theory: solvable model beyond the ridgeless limit. Mach Learn Sci Technol. 2024;5(1):015012.

Deshmukh SS, Ghosh S, Das D, Mallipeddi R, Das AK. Researchers develop new metallic materials using data-driven frameworks and explainable AI. Adv Mater. 2025;37(20):2400123.

Yuan J, Li Z, Yang Y, Nematov D, Hojamberdiev M. Applications of machine learning method in high-performance materials design: a review. J Mater Inform. 2024;4:14.

Adel T, Ghahramani Z, Weller A. Explainable artificial intelligence for spectroscopy data: a quantitative computational framework. Chemom Intell Lab Syst. 2024;234:104809.

Zhang H, Bao H, Li Y, Chen L, Ou J. Dramatically enhanced combination of ultimate tensile strength and electric conductivity of alloys via machine learning screening. Acta Mater. 2020;200:803-10.

Longo L, Goebel R, Lecue F, Kieseberg P, Holzinger A. Explainable artificial intelligence: concepts, applications, challenges and visions. Mach Learn Knowl Extr. 2020;4(3):956-83.

Nematov D, Hojamberdiev M. Machine learning-enabled optoelectronic material discovery: a comprehensive review. Opto-Electron Adv. 2025;8:13.