ORIGINAL RESEARCH **Open access**

The Curse of Optimality: When Perfect Optimization Undermines Scientific Understanding

Nour Abdelrahman[1], Karim Hassan[2], Ahmed El-Kholy[1*]

Abstract

In the rapidly expanding domain of artificial intelligence applied to materials science, the relentless pursuit of optimal predictive performance has emerged as the central organizing principle. Yet, this very imperative creates a profound paradox: models that achieve near-perfect accuracy on benchmark tasks frequently erode the scientific understanding they purport to support. When optimization dominates, systems become hyper-specialized predictors that deliver engineering-grade outputs while concealing the mechanistic pathways essential to genuine discovery. Optimization in materials AI unquestionably achieves impressive feats such as accelerated property prediction, efficient virtual screening of vast chemical spaces, and practical utility in guiding experimental synthesis; however, these gains come at the expense of interpretability, robustness, generalizability, and the capacity to generate novel hypotheses about underlying physical laws. This critical critique isolates four interlocking problems inherent to over-optimization: prediction without explanation, in which flawless forecasts provide no causal or structural insight; fragile optimality, whereby peak performance on training distributions collapses under even modest shifts in material conditions; the exploration-exploitation trap, which locks research into incremental refinement of known chemistries at the cost of venturing into truly novel territories; and optimization as epistemic closure, where the declaration of state-of-the-art accuracy prematurely terminates further inquiry. The consequences for materials science are far-reaching, manifesting as stagnant theoretical progress despite benchmark improvements, brittle knowledge bases ill-suited to real-world deployment, systematic neglect of high-potential but uncertain discoveries, and the misallocation of computational and human resources toward marginal accuracy gains rather than foundational insight. Alternative frameworks that deliberately balance predictive power with explanatory depth—ranging from explicit Pareto optimization of accuracy against interpretability to explanation-forcing model designs and satisficing strategies—are therefore not optional enhancements but necessary correctives if artificial intelligence is to fulfill its promise as a genuine partner in scientific understanding rather than a mere engineering tool. By reframing the goals of materials AI away from singular optimality. Toward epistemic multiplicity, the field can escape the curse of optimality and reclaim the generative interplay between prediction and comprehension that has historically driven materials innovation.

Keywords Epistemic costs, Optimization trade-offs, Curse of optimality, Interpretability in materials AI, Scientific understanding, Black-box optimization

*Correspondence:
Ahmed El-Kholy
ahmed.elkholy@outlook.com
[1] Department of Materials Engineering, Alexandria University, Alexandria, Egypt
[2] Department of AI Materials Systems, Ain Shams University, Cairo, Egypt

Introduction

Artificial intelligence in materials science has rapidly advanced, driven by the belief that data-driven models can accelerate materials discovery. Central to this progress is a dominant assumption: AI systems should prioritize optimal

prediction, measured by metrics such as MAE, RMSE, or accuracy. However, this emphasis creates a paradox—models that predict perfectly but lack interpretability do not advance scientific understanding. This paper critiques this "curse of optimality," arguing that highly predictive yet opaque models function as engineering tools rather than instruments of scientific knowledge.

This tension between optimization and understanding is structural. Prior work shows that machine learning excels at modeling complex structure–property relationships with minimal assumptions [1-5], enabling inverse design and automated discovery [6-8]. Even interpretability-focused studies tend to subordinate explanation to predictive performance [9, 10]. As a result, the field has converged on optimization as its dominant epistemic strategy.

Institutional incentives reinforce this trend, rewarding marginal accuracy improvements over mechanistic insight. Yet scholars have warned that interpretability is often overlooked [2], and reliance on black-box models can be problematic [11]. Alternative approaches, such as unsupervised discovery, suggest pathways that prioritize understanding [12]. The "curse of optimality" [3] highlights how perfect prediction can undermine insight.

In materials science, this issue is critical: understanding properties such as bandgaps or conductivity is essential for extrapolation and discovery. When optimization crowds out explanation, progress risks becoming narrow and fragmented. This paper, therefore, documents the optimization imperative, examines its trade-offs, and develops key critiques, including prediction without explanation and fragile optimality. The aim is not to reject optimization but to expose its epistemic costs and argue for a balance between predictive power and scientific understanding.

Figure 1 maps the manuscript's core argument by showing how the optimization imperative generates engineering gains while simultaneously producing epistemic sacrifices, four interlocking critiques, and downstream consequences for materials science.

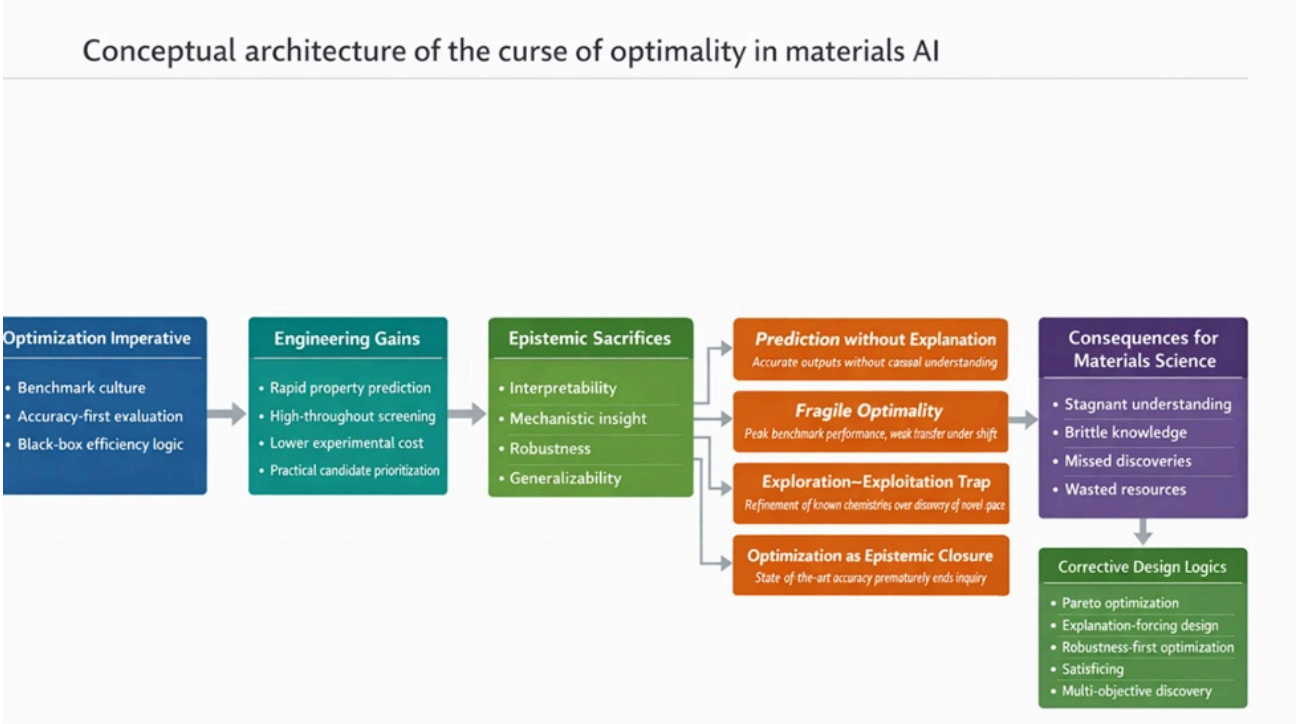

Figure 1. Conceptual architecture of the curse of optimality in materials AI

The Optimization Imperative

Optimization is the dominant logic in materials AI, with research primarily focused on maximizing predictive performance across tasks. Benchmark culture reinforces this, rewarding small accuracy gains while neglecting interpretability. Studies consistently show that supervised learning outperforms physics-based methods, driving widespread adoption [4, 5]. Inverse design and generative models further frame materials discovery as an optimization problem [6, 7].

Automated workflows extend this logic, with black-box optimization seen as the most efficient discovery strategy [8]. Even when interpretability is discussed, it is typically treated as secondary to accuracy [9, 10]. Critiques of black-box reliance [11] underscore how deeply optimization-first thinking is embedded. Alternative approaches, such as unsupervised learning, are often relegated to supporting roles [12], while data-fusion methods continue to prioritize performance gains [13].

This pattern is consistently reinforced across the literature [14-29], alongside broader concerns about model fragility [1] and interpretability myths [2]. The “curse of optimality” [3] captures this dynamic: epistemic depth is often sacrificed for marginal benchmark improvements. Consequently, research effort is heavily concentrated on tuning models rather than generating or testing scientific hypotheses.

The optimization imperative is therefore not just methodological but institutional, shaping research priorities, evaluation criteria, and definitions of progress. Establishing this dominance provides the foundation for the critiques developed in subsequent sections.

Table 1 clarifies that the curse of optimality is not merely a technical trade-off but a deeper conflict between optimization-first evaluation and science-first epistemic aims.

Table 1. Optimization-first performance criteria versus science-first epistemic criteria in materials AI

Analytical dimension	Optimization-first logic	Science-first logic	What is privileged under optimization	What is lost epistemically
Primary objective	Minimize prediction error	Build explanatory understanding	Benchmark superiority	Causal and mechanistic insight
Model evaluation	MAE, RMSE, R^2, and accuracy	Mechanistic fidelity, interpretability, and transfer validity	In-distribution performance	Theory relevance
Preferred model form	High-capacity black box	Constrained, interpretable, or hybrid architecture	Flexible function approximation	Legibility of learned structure
Training priority	Fit target labels as closely as possible	Learn relations that remain physically meaningful	Correlation capture	Causal discrimination
Error treatment	Residuals as a nuisance to minimize	Residuals as clues to unknown mechanisms	Performance polishing	Hypothesis generation
Generalization logic	Assume benchmark success implies wider usefulness	Test robustness across chemistries, structures, and process conditions	Narrow predictive confidence	Extrapolative trust

Discovery mode	Exploit dense, known regions of chemical space	Balance exploitation with uncertain, high-novelty regions	Incremental refinement	Transformative exploration
Stopping rule	Stop at state-of-the-art performance	Continue interrogation after a good prediction	Premature closure	Ongoing scientific inquiry
Resource allocation	Hyperparameter tuning, ensembling, and leaderboard competition	Mechanistic validation, symbolic extraction, physics-informed refinement	Marginal accuracy gains	Long-run knowledge accumulation

What Optimization Achieves

Optimization in materials AI delivers clear benefits. It enables high-throughput screening of millions of compounds and predicts properties such as bandgaps or formation energies with speed and accuracy often exceeding traditional methods. This translates into practical gains: accelerated discovery, reduced experimental costs, and improved targeting of synthesis efforts. Notable examples include latent-space molecular design [7], data-fusion approaches for stabilizing perovskites [13], and inverse design frameworks for multi-objective optimization [6]. Collectively, these advances have made materials AI a powerful engineering tool.

However, these gains come with significant trade-offs. Interpretability is often sacrificed, as highly optimized models become opaque [2, 11]. In practice, explainability methods are typically applied only after accuracy has been maximized [9, 10], reinforcing the secondary status of understanding. Mechanistic insight is similarly lost: models capture correlations but rarely reveal causal relationships. A network may predict thermal conductivity with high precision yet provide no insight into underlying physical mechanisms.

Generalizability and robustness also suffer. Models optimized for specific datasets often fail outside their training distributions, and highly tuned networks can be sensitive to small, non-physical perturbations [1]. These trade-offs can be conceptualized as a Pareto frontier between accuracy and interpretability: gains in one dimension often require losses in the other. Materials AI has largely operated at the extreme accuracy end of this frontier, prioritizing marginal performance improvements over explanatory value [4, 5, 8].

The long-term consequence is an erosion of scientific understanding. Black-box models limit extrapolation, hinder failure diagnosis, and constrain hypothesis generation. Although alternative approaches—such as unsupervised learning—demonstrate pathways toward insight [12], they remain marginal. Thus, optimization delivers engineering performance while systematically sacrificing interpretability, mechanistic insight, robustness, and generalizability—core attributes of scientific progress.

Critique Point 1: Prediction without Explanation

The most fundamental critique of over-optimization is that it produces prediction without explanation. Highly optimized models can function as accurate yet inscrutable oracles, offering little insight into why materials exhibit specific properties. In materials science, this absence of “why” is critical, as mechanistic understanding underpins generalization and discovery.

As Lipton [2] argued, interpretability is often superficial, while Rudin [11] emphasized that black-box predictions without explanation are scientifically problematic. In practice, materials AI provides many examples. Graph neural networks can predict bandgaps with high accuracy yet fail to reveal underlying electronic mechanisms such as orbital interactions [5]. Similarly, data-fusion models for perovskite stability achieve strong performance but obscure the physical origins of

observed trends [13]. In inverse molecular design, latent representations enable efficient optimization but lack clear mappings to chemically meaningful features [7].

Although interpretability methods exist, they are typically applied post hoc, confirming that explanation is treated as secondary [9, 10]. Black-box optimization frameworks further reinforce this pattern by focusing exclusively on performance improvement without interrogating underlying causes [8].

The result is a disconnect between prediction and understanding. Without mechanistic insight, researchers cannot distinguish causal relationships from correlations, diagnose model failures, or generate new hypotheses. Models become endpoints rather than tools for theory building. Over time, this risks producing a body of work rich in predictive capability but poor in scientific explanation. Accuracy alone, therefore, is an insufficient measure of epistemic value.

Critique Point 1: Prediction without Explanation

The first critique of over-optimization is that it produces predictions without explanation. Highly optimized models can be accurate yet inscrutable, offering no insight into why a material exhibits a given property. In materials science, this absence of mechanism limits generalization, extrapolation, and discovery. When optimization dominates, models are rewarded solely for predictive accuracy, reducing them to oracles disconnected from physical or chemical principles.

Lipton [2] describes this as the “mythos of model interpretability,” where post-hoc explanations fail to recover causal structure, while Rudin [11] argues that black-box predictions without explanation are scientifically inadequate. Materials AI provides clear examples. Graph neural networks can predict bandgaps with high accuracy yet fail to reveal underlying electronic mechanisms [5]. Data-fusion models for perovskite stability achieve strong performance but obscure the physical origins of observed trends [13]. Similarly, variational autoencoders enable efficient molecular optimization but lack interpretable mappings to chemical descriptors [7].

Although interpretability methods exist, they are typically applied after accuracy optimization, reinforcing the secondary role of explanation [9, 10]. Black-box optimization frameworks further prioritize performance over understanding [8]. The result is a growing set of high-performing models that provide little mechanistic insight.

This disconnect undermines scientific progress. Without explanation, researchers cannot distinguish causation from correlation, diagnose failures, or generate new hypotheses. Models become endpoints rather than tools for theory building. Thus, prediction alone is insufficient: any framework that decouples prediction from explanation is epistemically incomplete.

Critique Point 2: Fragile Optimality

The second critique is fragile optimality: models optimized for peak performance on specific datasets often fail under distribution shifts. This fragility arises directly from single-objective optimization, which exploits statistical patterns in training data rather than capturing underlying physical principles.

Gurney [1] showed that optimized neural networks are vulnerable to small perturbations, a problem mirrored in materials AI. Models trained on structured datasets (e.g., high-symmetry perovskites) often perform poorly on disordered or low-symmetry systems [4, 5]. Similarly, supervised models of phase transitions may achieve high accuracy within training ranges but fail under extrapolation, whereas less optimized unsupervised approaches can better capture underlying physics [12].

Fragility is also evident in automated discovery. Black-box optimization loops efficiently explore known chemical spaces but produce candidates that degrade under slight changes in experimental conditions [8]. Data-fusion models for

perovskites exhibit similar limitations when applied beyond their training environments [13]. Even inverse design approaches require additional constraints to ensure synthesizability and stability [6, 7].

This pattern reflects a deeper issue: without a mechanistic understanding, failures cannot be anticipated or corrected [2, 11]. Robust models often require sacrificing some predictive accuracy in favor of physical constraints or interpretable structures, which improve generalization. Over-optimized models, by contrast, are brittle by design.

Fragile optimality, therefore, reveals a practical dimension of the curse of optimality. Models that dominate benchmarks may be the least reliable in real-world settings, where conditions deviate from controlled datasets. Balancing accuracy with robustness and understanding is thus essential for meaningful scientific progress.

Critique Point 3: The Exploration–Exploitation Trap

The third critique identifies the exploration–exploitation trap as a structural consequence of over-optimization. Optimization inherently favors exploiting known patterns in well-sampled regions of chemical space, where uncertainty is low. However, materials discovery depends on exploring uncertain, underrepresented regions where breakthrough discoveries are most likely. When optimization dominates, exploration is systematically suppressed, narrowing the scope of discovery.

This dynamic is evident in black-box optimization frameworks. Terayama *et al.* [8] show how Bayesian optimization prioritizes candidates with great expected improvement, leading models to sample familiar regions rather than venture into high-uncertainty areas repeatedly. As a result, discovery pipelines converge on local optima instead of exploring novel chemistries.

Materials-specific examples reinforce this pattern. In halide perovskites, data-fusion models refine stability within known compositional ranges but rarely propose fundamentally new material classes [13]. Similarly, models for solid-state materials focus on well-represented oxides and chalcogenides, achieving high accuracy while failing to generalize to less-studied chemistries such as nitrides or intermetallics [5]. In molecular design, latent-space optimization favors incremental modifications of familiar scaffolds rather than generating structurally novel compounds [7].

Across these cases, optimization drives incremental refinement rather than genuine discovery. The search process becomes confined to dense regions of known data, leaving large areas of chemical space unexplored. As highlighted by the “curse of optimality” [3], the more effectively a model exploits known patterns, the less likely it is to encounter transformative unknowns.

Thus, over-optimization distorts the epistemology of materials AI by converting discovery into optimization of the familiar. Without explicit mechanisms to promote exploration, the field risks becoming an echo chamber of incremental improvements rather than a driver of breakthrough innovation.

Critique Point 4: Optimization as Epistemic Closure

The fourth critique frames optimization as epistemic closure. Once a model achieves near-optimal performance, research often halts because the dominant success criterion—benchmark accuracy—has been satisfied. This creates a premature stopping point, where inquiry into mechanisms, errors, or underlying principles is abandoned.

Scholars have noted that black-box optimization encourages this closure, as achieving state-of-the-art metrics is often treated as the endpoint rather than the beginning of deeper investigation [2, 11]. In materials AI, this pattern is widespread. For example, perovskite stability models reach near-perfect predictive accuracy, yet residual errors and their physical significance are rarely examined [13]. Similarly, studies of solid-state properties often stop at benchmark leadership without probing learned representations for mechanistic insight [4, 5].

In molecular design, once generative models achieve target performance, workflows typically end with candidate generation rather than interrogation of latent representations or underlying chemical rules [7]. Automated discovery systems reinforce this behavior by terminating optimization once uncertainty thresholds are met, leaving potential mechanistic insights unexplored [8, 14, 20]. As described by Zhou [3], this reflects the curse of optimality: inquiry ends when performance is deemed sufficient.

The consequence is a form of epistemic stagnation. Optimization provides a natural stopping rule that shifts attention away from understanding toward performance metrics. Models are treated as finished products rather than tools for continued investigation.

Thus, optimization as epistemic closure transforms AI from a driver of discovery into a constraint on inquiry. When success is defined solely by predictive accuracy, the scientific process is truncated, producing a literature rich in performance but limited in explanation.

Table 2 consolidates the manuscript's four critiques by distinguishing the mechanism that generates each problem from its practical manifestation and its cumulative scientific cost.

Table 2. The four critiques of over-optimization: generative mechanism, materials-AI manifestation, and scientific cost

Critique	Generative mechanism	Typical manifestation in materials AI	Immediate analytical effect	Long-run scientific cost
Prediction without explanation	Models are rewarded only for output fidelity, not causal transparency	Accurate bandgap, stability, or property predictions without interpretable physical relations	Correlations appear sufficient	Theory building weakens; design rules remain implicit
Fragile optimality	Single-objective training exploits statistical regularities specific to the training manifold	Strong performance on benchmark compounds, poor transfer to disordered, low-symmetry, or shifted environments	In-distribution confidence is overstated	Knowledge becomes brittle and unreliable in deployment
Exploration–exploitation trap	Acquisition functions and latent-space searches favor low-uncertainty, high-density regions	Repeated optimization within known oxides, perovskites, or familiar molecular scaffolds	Discovery narrows to local refinement	Novel chemistries and uncertain breakthroughs remain unseen
Optimization as epistemic closure	State-of-the-art performance becomes the de facto stopping rule	Research ends at benchmark leadership rather than probing residual errors or latent mechanisms	Inquiry terminates too early	Scientific investigation is truncated before a deeper explanation emerges

Consequences for Materials Science

The cumulative effect of the four critiques produces four interlocking consequences that threaten the long-term vitality of materials science. What initially appears as steady technical progress—expressed through improved benchmark performance—begins to reveal a deeper stagnation at the level of scientific understanding. Models become increasingly adept at capturing correlations, yet the underlying theoretical structure of the field remains largely unchanged. Empirical advances reported by Butler *et al.* [4] and Schmidt *et al.* [5] demonstrate clear gains in predictive capability, but they do

not translate into new physical laws or design principles. Instead, long-standing frameworks such as band theory, defect chemistry, and phonon transport continue to anchor interpretation, indicating that predictive refinement has not been accompanied by conceptual expansion.

This imbalance introduces a second, more fragile dimension of knowledge production. When optimization prioritizes in-distribution accuracy, the resulting models encode patterns that are tightly coupled to the statistical properties of their training data. Such representations often lack resilience when confronted with variation, and their apparent reliability dissolves under even modest shifts in conditions. The broader machine learning literature has long documented this phenomenon, with Gurney [1] and Rudin [11] highlighting the instability of highly optimized systems. In materials contexts, the implications are amplified: variations in synthesis conditions, temperature regimes, or impurity profiles can trigger rapid degradation in predictive performance, as shown by Sun *et al.* [13] and Gómez-Bombarelli *et al.* [7]. Under these conditions, knowledge derived from optimization proves brittle, unable to sustain its validity outside narrowly defined regimes.

A further consequence emerges through the systematic suppression of exploration. Optimization-driven workflows tend to reinforce existing knowledge structures by focusing on regions of chemical space that already exhibit favorable statistical properties. As a result, areas characterized by uncertainty—often the most promising sites for discovery—are progressively excluded from consideration. Over time, this bias narrows the horizon of inquiry, directing effort toward incremental refinement rather than genuine novelty. Terayama *et al.* [8] and Zunger [6] describe such convergence toward familiar material families, yet the opportunity cost remains largely implicit. Entire domains of potential innovation, including topological materials, high-entropy systems, and bio-inspired hybrids, risk remaining undiscovered because prevailing methodologies render them statistically unattractive.

This contraction of exploratory scope is reinforced by patterns of resource allocation that favor marginal gains in predictive accuracy over deeper forms of understanding. Significant computational and intellectual effort is invested in hyperparameter tuning, ensemble strategies, and incremental performance optimization, often yielding only negligible improvements in benchmark metrics. At the same time, approaches capable of generating mechanistic insight—such as physics-informed modeling or symbolic regression—receive comparatively less emphasis. Lipton [2] and Zhou [3] identify this asymmetry as a structural feature of optimization-centric research, where the pursuit of precision incurs a hidden epistemic cost. The consequence is a reorientation of the research agenda, in which technical refinement displaces theoretical development.

Taken together, these dynamics transform materials AI into a system that excels at prediction while remaining largely inert with respect to explanation. The resulting condition is not one of failure, but of epistemic stasis, where progress is measured without a corresponding deepening of understanding.

Alternative Approaches

A departure from this trajectory requires reconfiguring optimization within a broader epistemic framework that recognizes understanding as coequal with performance. One promising direction lies in reframing model development as a multi-objective problem, where predictive accuracy is balanced explicitly against interpretability. Rather than converging toward a single optimum, this perspective emphasizes the structure of the Pareto frontier, within which different trade-offs between performance and mechanistic transparency can be explored. The region near the inflection of this frontier—where additional gains in accuracy begin to erode interpretability—emerges as a particularly meaningful operating point. Work by Zhong *et al.* [9] and Oviedo *et al.* [10] provides a conceptual basis for such formulations, suggesting that deliberate positioning along this frontier can preserve epistemic value without abandoning predictive utility.

This shift also introduces the possibility of embedding explanatory constraints directly into model architectures, thereby requiring that predictive success be accompanied by mechanistic articulation. By incorporating physics-informed objectives or auxiliary structures that recover interpretable relationships, models can be guided toward representations

that remain legible to human reasoning. In this configuration, convergence is no longer defined solely by statistical adequacy but by the simultaneous satisfaction of accuracy and intelligibility. Wetzel [12] demonstrates the potential of unsupervised approaches to reveal latent physical structure, while Rudin [11] underscores the risks associated with unconstrained black-box optimization, reinforcing the need for such integrative designs.

A related reorientation involves redefining optimization targets to prioritize robustness rather than peak performance. By incorporating penalties for sensitivity to distributional shifts or by training across systematically perturbed representations of materials space, models can be incentivized to maintain stability under variation. This approach directly addresses the fragility observed in conventional pipelines [1], reframing predictive success as durability across conditions rather than excellence within a narrowly defined regime.

Beyond this, the assumption that optimization should proceed to its theoretical limit warrants reconsideration. In many cases, the pursuit of marginal accuracy gains yields diminishing epistemic returns while consuming disproportionate resources. Adopting a satisficing perspective—where optimization is halted once an adequate level of performance is reached—creates space for subsequent phases of mechanistic investigation. Lipton [2] and Zhou [3] argue that such an approach is not a compromise but a rational adjustment when understanding constitutes the primary objective.

Finally, restoring balance to the discovery process requires the explicit integration of exploration into optimization frameworks. By introducing incentives for novelty—whether through uncertainty-weighted acquisition functions or diversity-promoting objectives—models can be directed toward regions of chemical space that would otherwise remain neglected. Existing strategies, such as the expected improvement methods employed by Terayama *et al.* [8], provide a foundation, yet their extension toward exploration-aware formulations is necessary to counteract the entrenched bias toward exploitation.

These alternative approaches do not abandon optimization; they situate it within a richer conceptual structure in which predictive performance, interpretability, robustness, and exploratory capacity are jointly negotiated. In doing so, they offer a pathway beyond the limitations imposed by optimization-centric paradigms, enabling materials AI to function not only as a predictive tool but as a genuine engine of scientific understanding.

Table 3 translates the manuscript's normative argument into a concrete design agenda by showing how alternative optimization logics can preserve scientific understanding without abandoning predictive utility.

Table 3. Design logics for escaping the curse of optimality: from single-objective prediction to epistemically plural materials AI

Alternative approach	Revised optimization logic	What it protects or restores	Typical implementation route	Expected trade-off
Pareto optimization	Optimize predictive performance and interpretability jointly	Balance between utility and explanatory depth	Multi-objective objective functions; Pareto-front model selection	Slight reduction in peak benchmark accuracy
Explanation-forcing design	Require a mechanistic account as part of convergence	Interpretability and hypothesis generation	Physics-informed loss terms, symbolic heads, and constrained latent variables	Higher modeling complexity
Robustness-first optimization	Optimize stability under distribution shift, not just central benchmark fit	Transfer validity and deployment trustworthiness	Shift penalties, adversarial perturbation tests, cross-domain validation	Slower apparent benchmark gains
Satisficing	Stop once the prediction is good enough for	Resource reallocation toward understanding	Predefined performance threshold followed by	Forgoes final increments of

	scientific use		mechanistic interrogation	accuracy
Multi-objective discovery	Reward novelty, uncertainty, and exploration alongside performance	Search breadth and breakthrough potential	Novelty bonuses, uncertainty-weighted acquisition, and exploratory active learning	Higher short-run experimental risk

Conclusion

The curse of optimality reveals a fundamental paradox at the heart of contemporary materials AI: the more perfectly a model predicts, the less it explains, the less robust it becomes, the less it explores, and the more completely it seals off further inquiry. This critical critique has documented the optimization imperative, delineated its achievements and sacrifices, and isolated four interlocking problems—prediction without explanation, fragile optimality, the exploration-exploitation trap, and optimization as epistemic closure—each illustrated with multiple materials-specific cases drawn from perovskites, solid-state compounds, and molecular design. The consequences—stagnant understanding, brittle knowledge, missed discoveries, and wasted resources—threaten to convert AI from a scientific partner into an engineering dead-end.

Yet the diagnosis itself points toward a remedy. By embracing Pareto optimization, explanation-forcing design, robustness-first strategies, satisficing, and multi-objective discovery, the field can escape the trap and restore the generative tension between prediction and comprehension that has always driven materials innovation. Perfect prediction is not the only—or even the best—goal of materials AI. Scientific understanding must be restored to equal prominence if artificial intelligence is to fulfill its deeper promise: not merely to forecast properties faster, but to illuminate the unseen principles that govern the material world. Only then will the curse of optimality be lifted and the full potential of AI for materials science be realized.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 27 May 2022 Revised: 09 Aug 2022 Accepted: 03 Nov 2022
Published online: 18 January 2023

References

Gurney K. An introduction to neural networks. Boca Raton: CRC Press; 2018.

Lipton ZC. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue. 2018;16(3):31-57.

Zhou XY. The curse of optimality, and how to break it. In: Machine learning and data sciences for financial markets: A guide to contemporary practices. 2023. p. 354-68.

Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A. Machine learning for molecular and materials science. Nature. 2018;559(7715):547-55.

Schmidt J, Marques MR, Botti S, Marques MA. Recent advances and applications of machine learning in solid-state materials science. npj Comput Mater. 2019;5(1):83.

Zunger A. Inverse design in search of materials with target functionalities. Nat Rev Chem. 2018;2(4):0121.

Gómez-Bombarelli R, Wei JN, Duvenaud D, Hernández-Lobato JM, Sánchez-Lengeling B, Sheberla D, et al. Automatic chemical design using a data-driven continuous representation of molecules. ACS Cent Sci. 2018;4(2):268-76.

Terayama K, Sumita M, Tamura R, Tsuda K. Black-box optimization for automated discovery. Acc Chem Res. 2021;54(6):1334-46.

Zhong X, Gallagher B, Liu S, Kailkhura B, Hiszpanski A, Han TY. Explainable machine learning in materials science. npj Comput Mater. 2022;8(1):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.

Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1(5):206-15.

Wetzel SJ. Unsupervised learning of phase transitions: From principal component analysis to variational autoencoders. Phys Rev E. 2017;96(2):022140.

Sun S, Tiihonen A, Oviedo F, Liu Z, Thapa J, Zhao Y, et al. A data fusion approach to optimize compositional stability of halide perovskites. Matter. 2021;4(4):1305-22.

Tschider CA. Beyond the black box. Denv Law Rev. 2020;98:683.

Belle V, Papantonis I. Principles and practice of explainable machine learning. Front Big Data. 2021;4:688969.

Wei J, Chu X, Sun XY, Xu K, Deng HX, Chen J, et al. Machine learning in materials science. InfoMat. 2019;1(3):338-58.

Karim MR, Shajalal M, Graß A, Döhmen T, Chala SA, Boden A, et al. Interpreting black-box machine learning models for high dimensional datasets. In: 2023 IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA). New York:

IEEE; 2023. p. 1-10.

Hu W, Singh RR, Scalettar RT. Discovering phases, phase transitions, and crossovers through unsupervised machine learning: A critical examination. Phys Rev E. 2017;95(6):062122.

Farizhandi AA, Mamivand M. Processing time, temperature, and initial chemical composition prediction from materials microstructure by deep network for multiple inputs and fused data. Mater Des. 2022;219:110799.

Golovin D, Solnik B, Moitra S, Kochanski G, Karro J, Sculley D. Google Vizier: A service for black-box optimization. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2017. p. 1487-95.

Qing Y, Liu S, Song J, Zhou Y, Chen K, Wang H, Song M. A survey on explainable reinforcement learning: Concepts, algorithms, challenges. arXiv. 2022;2211.06665.

Morgan D, Jacobs R. Opportunities and challenges for machine learning in materials science. Annu Rev Mater Res. 2020;50(1):71-103.

Vasudevan R, Pilania G, Balachandran PV. Machine learning for materials design and discovery. J Appl Phys. 2021;129(7).

Li W, Qian X, Li J. Phase transitions in 2D materials. Nat Rev Mater. 2021;6(9):829-46.

Xu Q, Yang X, Song J, Ru J, Xia J, Wang S, et al. Nitrogen enrichment alters multiple dimensions of grassland functional stability via changing compositional stability. Ecol Lett. 2022;25(12):2713-25.

Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. ACM Comput Surv. 2018;51(5):1-42.

Daumé H. A course in machine learning. Alanna Maldonado; 2023. Available from: http://ciml.info

Reiser P, Neubert M, Eberhard A, Torresi L, Zhou C, Shao C, et al. Graph neural networks for materials science and chemistry. Commun Mater. 2022;3(1):93.

Sahoh B, Choksuriwong A. The role of explainable artificial intelligence in high-stakes decision-making systems: A systematic review. J Ambient Intell Humaniz Comput. 2023;14(6):7827-43.