ORIGINAL RESEARCH **Open access**

From Black Box to Scientific Instrument: A Conceptual Pathway for Validating AI as a Materials Reasoning Tool

Lucas Pereira[1*], Bruno Martins[1]

Abstract

The rapid integration of artificial intelligence (AI) into materials science has enabled unprecedented predictive capabilities across a wide range of properties and structures. However, the predominantly black-box nature of these models limits their epistemic role, confining them largely to correlative tools rather than instruments capable of supporting genuine scientific reasoning. This conceptual manuscript introduces a novel theoretical framework that delineates a structured pathway for validating AI systems as materials reasoning tools. Drawing on recent advances in explainable and interpretable AI, as well as philosophical accounts of scientific reasoning, the framework articulates a progressive sequence of validation stages: establishing transparency and interpretability, extracting mechanistically meaningful explanations, assessing reasoning fidelity through inferential behavior, and integrating AI systems as instruments within the broader scientific knowledge cycle. The approach is deliberately architecture-agnostic and avoids empirical prescriptions, focusing instead on the conceptual and epistemic conditions required for scientific legitimacy. By explicitly bridging predictive performance with explanatory depth, inferential robustness, and alignment with physical theory, the proposed pathway reframes how success in materials AI is evaluated. It provides a foundation for distinguishing advanced predictive engines from systems capable of contributing to hypothesis generation, theory refinement, and cumulative understanding. In doing so, the framework addresses persistent barriers to the acceptance of AI as a scientific partner in materials research. It offers a principled basis for future methodological and evaluative developments.

Keywords Interpretability, Artificial intelligence, Explainable AI, Materials science, Scientific reasoning, Mechanistic explanation

*Correspondence:
Lucas Pereira
lucas.pereira@gmail.com
[1] Department of Materials Modeling and Data Analytics, Faculty of Engineering, University of Minho, Braga, Portugal

Introduction

The integration of machine learning (ML) and artificial intelligence (AI) into materials science has profoundly reshaped research practice over the past decade, accelerating a shift away from predominantly trial-and-error experimentation toward data-driven discovery and design [1, 2]. Enabled by high-throughput computation, automated workflows, and the growing availability of large, curated datasets, AI models have demonstrated remarkable success in predicting a wide range of material properties, including electronic, thermodynamic, and mechanical attributes [3]. These advances have facilitated rapid exploration of vast chemical and structural spaces, leading to notable achievements such as the identification of previously unknown stable compounds and the optimization of materials for energy, catalytic, and functional applications [4]. Collectively, such results underscore AI's capacity to operate at scales and speeds that exceed those of conventional physics-based modeling and experimental screening.

Despite these achievements, a persistent and increasingly visible tension underlies the use of AI in materials research. While contemporary models often deliver high predictive accuracy, their internal decision-making processes typically remain opaque, earning them the designation of "black boxes" [5]. In a domain where scientific progress depends not only on prediction but on understanding of phase stability, defect behavior, structure–property relationships, and governing physical mechanisms, this opacity poses a fundamental challenge [6]. Accurate predictions alone are insufficient to support scientific reasoning, which requires explanations that cohere with established theory, enable the formulation and testing of hypotheses, and, where possible, justify causal claims [7]. Without such explanatory grounding, AI outputs risk being treated as technically impressive yet epistemically fragile artifacts.

In response, there has been a growing emphasis on explainable and interpretable AI within the materials science community [8, 9]. Recent work has explored both post-hoc explanation techniques, such as feature attribution and sensitivity analysis, and ante-hoc model designs intended to enhance interpretability, including constrained architectures, attention mechanisms, and symbolic or physics-guided representations [10, 11]. These approaches have yielded valuable insights into model behavior and, in some cases, have illuminated plausible structure–property correlations. However, existing efforts remain largely fragmented and oriented toward instrumentalism. Explanations are often evaluated in terms of user trust, debugging utility, or qualitative plausibility, rather than as components of a systematic framework through which AI might contribute to scientific knowledge production [12].

This limitation becomes particularly consequential in applied materials contexts, where AI is increasingly entrusted with inverse design, multi-objective optimization, and surrogate modeling of computationally expensive or experimentally inaccessible processes [13]. In such settings, AI systems influence not only predictions but also research trajectories and decision-making. Absent rigorous criteria for validating their reasoning capacity, these systems risk remaining auxiliary tools—useful for accelerating workflows but excluded from the core epistemic cycle of conjecture, explanation, and theory refinement that characterizes scientific inquiry [14]. Philosophical analyses of scientific reasoning emphasize criteria such as explanatory coherence, robustness across contexts, predictive novelty, and integration with existing theoretical frameworks—dimensions that are rarely addressed in prevailing AI validation practices [15].

This manuscript addresses this conceptual gap by proposing an original framework for validating AI as a materials reasoning tool rather than a mere predictive engine. The framework delineates a structured pathway from black-box prediction toward transparent, mechanistically informed reasoning, drawing on and synthesizing recent developments in explainable AI as applied to materials science. Importantly, the framework prioritizes epistemic structure over technical implementation: it is intentionally agnostic to specific datasets, algorithms, or empirical benchmarks. Instead, it articulates progressive stages of validation, each imposing increasingly stringent criteria aligned with scientific reasoning and materials theory. By grounding AI evaluation in both computational rigor and scientific epistemology, the proposed framework aims to establish a theoretical foundation for recognizing AI as a legitimate and accountable instrument of knowledge generation in materials research.

Theoretical Background and Literature Synthesis

Machine learning applications in materials science

Since 2020, machine learning has transitioned from an enabling technology to a foundational component of materials research, reshaping how materials are discovered, evaluated, and optimized [16]. ML techniques are now routinely applied across the materials lifecycle, including property prediction, phase stability assessment, microstructure–property mapping, and generative design of novel compounds and architectures. This maturation has been driven by parallel advances in data availability, computational infrastructure, and algorithmic sophistication.

Deep learning architectures have played a central role in this transformation. In particular, graph-based models that operate directly on atomic structures, along with convolutional approaches applied to microstructural images and spatial fields, have demonstrated robust performance across chemically and structurally diverse datasets [17]. These models

enable the systematic exploration of high-dimensional materials spaces, supporting accelerated screening for functional applications such as catalysis, energy storage, and semiconductor design. A notable trend highlighted in recent reviews is the shift toward end-to-end learning paradigms, in which models ingest raw structural or spatial representations rather than relying on manually engineered descriptors [18]. This shift has reduced human bias in feature construction and expanded the scope of learnable structure–property relationships.

As a result, AI-driven workflows have achieved substantial gains in efficiency, with reported acceleration factors ranging from significant reductions in computational cost to orders-of-magnitude improvements in candidate discovery rates relative to traditional experimental or simulation-based approaches [3]. These successes have positioned ML as an indispensable tool for navigating the combinatorial complexity inherent to materials design. However, they have also intensified scrutiny of the epistemic status of AI-generated knowledge, particularly when predictive performance outpaces interpretive clarity.

Challenges posed by black-box models

Despite their empirical success, the increasing reliance on complex, high-capacity ML models has introduced significant conceptual challenges. Chief among these is model opacity. In materials science—where predictions are expected to reflect underlying physical constraints and mechanistic regularities—black-box models risk producing outputs that are numerically accurate yet physically implausible or scientifically uninformative [19]. This disconnect complicates both interpretation and generalization, especially when models are applied outside narrowly defined training regimes.

The literature identifies several interrelated risks associated with opaque models. These include susceptibility to hidden dataset biases, brittle behavior under small input perturbations, and weak alignment with established domain knowledge [20]. Such vulnerabilities are particularly problematic in materials contexts, where data scarcity, uneven sampling of chemical space, and confounding correlations are common. Without access to interpretable internal representations, it becomes difficult to assess whether a model has learned meaningful structure–property relationships or merely exploited statistical regularities specific to the training data [21].

Crucially, these challenges extend beyond practical concerns about reliability or robustness. They strike at the core of scientific reasoning. Scientific models are valued not only for their predictive accuracy but also for their capacity to support explanation, enable hypothesis generation, and integrate coherently with existing theory. When ML systems operate as inscrutable function approximators, their contributions remain epistemically limited: predictions cannot be easily scrutinized, justified, or situated within a broader conceptual framework. As a result, black-box AI threatens to decouple computational success from scientific understanding, reinforcing a divide between performance and explanation that is increasingly untenable as AI systems assume more central roles in materials research.

Advances in explainable and interpretable AI for materials

Recent years have witnessed increasing efforts to adapt explainable and interpretable artificial intelligence techniques to the specific epistemic and structural challenges of materials science [8, 9]. Unlike generic ML applications, materials problems involve tightly coupled physical variables, hierarchical representations (from electrons to microstructures), and strong domain constraints. As a result, explanation methods must do more than attribute importance—they must remain interpretable within established physical and chemical frameworks.

Post-hoc explanation techniques have been among the most widely adopted approaches. Methods such as Shapley additive explanations (SHAP), gradient-based saliency maps, and sensitivity analyses have been used to identify atomic environments, compositional features, and spatial regions that contribute to predicted properties [22]. In atomistic property prediction, these tools enable the visualization of feature contributions associated with local coordination, bonding environments, or elemental substitutions. In microstructure- and image-based analyses, saliency and activation mapping techniques have been employed to highlight spatial motifs or defects driving classification or regression

outcomes. While such methods provide valuable insights into model behavior, their explanatory power is inherently indirect, as they operate on trained models whose internal representations remain unchanged.

In contrast, ante-hoc approaches seek to embed interpretability directly into model design. Attention mechanisms, modular architectures, and intrinsically interpretable models—such as sparse linear representations or symbolic expressions inferred from data—offer transparency by construction [23]. These approaches aim to align learned representations more closely with human-understandable concepts, enabling users to trace predictions to explicit functional forms or weighted interactions. Applications span a range of materials domains, from alloy design—where interpretable models expose dominant compositional trends—to spectroscopic and diffraction data analysis, where explanations often correspond to physically meaningful peaks or symmetry-related features [9]. Compared to post-hoc methods, ante-hoc models provide stronger epistemic access but often at the cost of reduced flexibility or predictive performance.

Interpretability in service of scientific insight

Beyond improving trust or diagnosing model errors, interpretability has increasingly been positioned as a tool for scientific discovery. Several studies demonstrate how explanations extracted from ML models can support hypothesis generation by revealing non-intuitive chemical trends, unexpected feature interactions, or latent structure–property relationships [24]. In such cases, interpretability serves as a bridge between data-driven prediction and experimental or theoretical follow-up, enabling AI outputs to inform subsequent scientific inquiry rather than merely accelerate screening.

The integration of physics-informed constraints further strengthens this bridge. By embedding conservation laws, symmetry considerations, or mechanistic priors into model architectures or loss functions, researchers aim to ensure that both predictions and explanations remain consistent with established physical principles [25]. This hybridization enhances explanatory fidelity, reducing the risk that interpretations reflect spurious statistical artifacts rather than meaningful material behavior. However, it also raises questions about the appropriate balance between data-driven flexibility and theoretical imposition—a tension that remains unresolved in current practice.

Despite these advances, evaluating the quality of explanations remains a central challenge. In most materials problems, ground-truth explanations are unavailable, as mechanisms may be partially understood, contested, or context-dependent. Consequently, explanation assessment often relies on qualitative judgment by domain experts, introducing subjectivity and limiting reproducibility [26]. This ambiguity complicates claims that interpretable AI outputs constitute genuine scientific explanations rather than heuristic narratives layered atop predictive models.

Gaps in current validation approaches

Taken together, the literature reveals a mismatch between the growing interpretability toolbox and existing validation paradigms in materials AI. Current evaluation practices remain dominated by predictive metrics, such as accuracy, generalization error, and uncertainty quantification, which are necessary but insufficient for assessing reasoning capacity [27]. Even broader notions of model trustworthiness tend to emphasize robustness, stability, or bias mitigation, with limited engagement with explanatory coherence or theoretical integration [28].

Notably absent from most materials, AI frameworks are philosophical and epistemological perspectives on scientific reasoning. Concepts central to scientific practice—such as abductive inference, mechanistic explanation, explanatory unification, and theory-ladenness—rarely inform how AI systems are evaluated or positioned within the research process [29]. As a result, interpretability is often treated as an auxiliary feature rather than as a criterion for epistemic legitimacy.

The conceptual distinction between predictive AI, interpretable AI, and validated materials reasoning systems is summarized in **Table 1**.

Table 1. Distinguishing predictive AI, interpretable AI, and validated materials reasoning systems

Dimension	Predictive AI	Interpretable AI	Validated materials reasoning tool
Primary goal	Accurate prediction	Explainable prediction	Scientific reasoning and understanding
Access to internal logic	None or opaque	Partial (feature attribution, attention)	Systematic, domain-grounded access
Mechanistic correspondence	Not required	Optional or post hoc	Required and theoretically coherent
Inferential capability	None	Limited	Abductive and deductive reasoning
Role in the scientific cycle	Workflow acceleration	Model trust and diagnosis	Hypothesis generation and theory refinement
Epistemic status	Correlative tool	Interpretable model	Scientific instrument

This gap leaves the field without a unified conceptual pathway for transitioning from high-performing predictive models to reasoning instruments that can contribute to materials theory. While individual studies demonstrate promising explanatory capabilities, they do so in isolation, without shared standards for what constitutes valid reasoning or sufficient explanation. Addressing this conceptual void requires moving beyond technique-specific solutions toward a principled framework that aligns AI validation with the norms and goals of scientific knowledge production.

Proposed conceptual framework

Overview of the validation pathway

The proposed framework defines a sequential yet iterative validation pathway for assessing artificial intelligence models as materials-reasoning tools rather than purely predictive systems. The framework comprises four conceptual stages, each contributing cumulative epistemic evidence that an AI model transcends correlational pattern recognition and participates in forms of reasoning aligned with scientific practice. Importantly, the framework is architecture-agnostic: it does not prescribe specific model classes, algorithms, or training strategies. Instead, it specifies validation criteria grounded in transparency, explanatory depth, reasoning fidelity, and theoretical integration.

The stages are ordered to reflect increasing epistemic demands. Progression through the pathway is not strictly linear; feedback loops allow refinement and reassessment as models evolve or are deployed in new scientific contexts. Collectively, the pathway provides a principled structure for evaluating when—and to what extent—AI systems may be treated as legitimate contributors to materials knowledge production.

The four stages of the proposed validation pathway, along with their objectives, validation criteria, and epistemic functions, are summarized in **Table 2**.

Table 2. Validation stages for AI as a materials reasoning tool: objectives, criteria, and epistemic function

Validation stage	Primary objective	Core validation criteria	Epistemic function in materials science
Stage 1: Transparency and interpretability	Render model behavior accessible to human scrutiny	Completeness of explanations; domain-relevant representations; physical plausibility	Enables inspection, diagnosis, and rejection of spurious correlations
Stage 2: Mechanistic insight extraction	Map explanations to physical or chemical mechanisms	Theoretical coherence; cross-system consistency; qualitative agreement with first principles	Supports causal or quasi-causal understanding of structure–property relations

Stage 3: Reasoning fidelity assessment	Evaluate inferential behavior beyond explanation	Hypothesis generation; counterfactual robustness; logical consistency	Enables abductive and deductive reasoning aligned with scientific inference
Stage 4: Integration as a scientific instrument	Embed AI within the epistemic cycle of science	Explanatory unification; cumulative contribution; theory interaction	Positions AI as a validated instrument for knowledge production

Stage 1: Establishing transparency and interpretability

The first stage focuses on rendering the internal functioning of AI models accessible to human scrutiny. This involves applying intrinsic interpretability mechanisms or post-hoc explanation methods to reveal feature contributions, decision hierarchies, or latent representational structures. At this stage, explanations are not yet required to be mechanistic; rather, they must enable inspection and diagnosis of model behavior.

Validation at Stage 1 is assessed along three core criteria. Completeness concerns the extent to which explanations adequately cover the factors influencing model outputs. Understandability assesses whether these factors are expressed in forms meaningful to materials scientists rather than as abstract mathematical artifacts. Correctness requires that identified dependencies are physically plausible and consistent with basic domain constraints. Successful completion of this stage establishes baseline trust, enables the detection of spurious correlations, and provides the necessary foundation for deeper explanatory analysis.

Stage 2: Extracting mechanistic insights

Building on transparency, the second stage requires that explanations support mechanistic interpretation. Here, model behavior must be interpretable in terms of established physical or chemical concepts, such as bonding characteristics, electronic structure descriptors, thermodynamic drivers, or microstructural motifs. Explanations should clarify *how* variations in inputs lead to changes in outputs in ways that resonate with domain theory.

Validation emphasizes theoretical coherence and generalizability. Explanations must align with known mechanisms or, where they diverge, do so in ways that are intelligible and scientifically motivated. Importantly, insights should not be limited to interpolation within the training domain but should demonstrate plausibility when extended to novel compositions, structures, or conditions. This stage marks the transition from descriptive interpretability toward explanations capable of supporting causal or quasi-causal understanding.

Stage 3: Assessing reasoning fidelity

The third stage evaluates whether the AI system exhibits hallmarks of scientific reasoning rather than isolated explanatory fragments. Reasoning fidelity is assessed by probing the model's capacity for abductive inference, counterfactual robustness, and logical consistency with governing physical principles. At this stage, explanations are expected to do more than rationalize predictions—they should actively support inference.

Validation involves testing whether model-derived explanations enable the generation of novel, testable hypotheses, reconcile apparently conflicting observations, or suggest plausible alternative explanations under perturbed conditions. Criteria include novelty, defined as the capacity to propose insights not trivially recoverable from training data, and refutability, defined as susceptibility to empirical or theoretical falsification. A model that passes this stage demonstrates not only interpretability, but disciplined inferential behavior consistent with scientific reasoning norms.

Stage 4: Integration as a scientific instrument

The final stage confirms whether an AI model can function as a scientific instrument within the broader materials research ecosystem. At this level, the model contributes to theory refinement, proposes or constrains mechanisms, and integrates coherently with existing bodies of knowledge. Validation is no longer model-centric but system-level, assessing how AI-generated insights interact with experiments, simulations, and theoretical frameworks.

Criteria at this stage include explanatory unification, in which AI outputs connect disparate observations under shared principles, and cumulative contribution, in which insights persist and evolve across successive studies rather than remaining isolated artifacts. Successful integration at this stage signals that the AI system has achieved epistemic legitimacy—not merely as a tool for acceleration, but as an accountable participant in materials science reasoning. **Figure 1** illustrates the proposed validation pathway as a linear progression with iterative feedback loops, depicting the conceptual transition from opaque prediction to validated scientific reasoning.

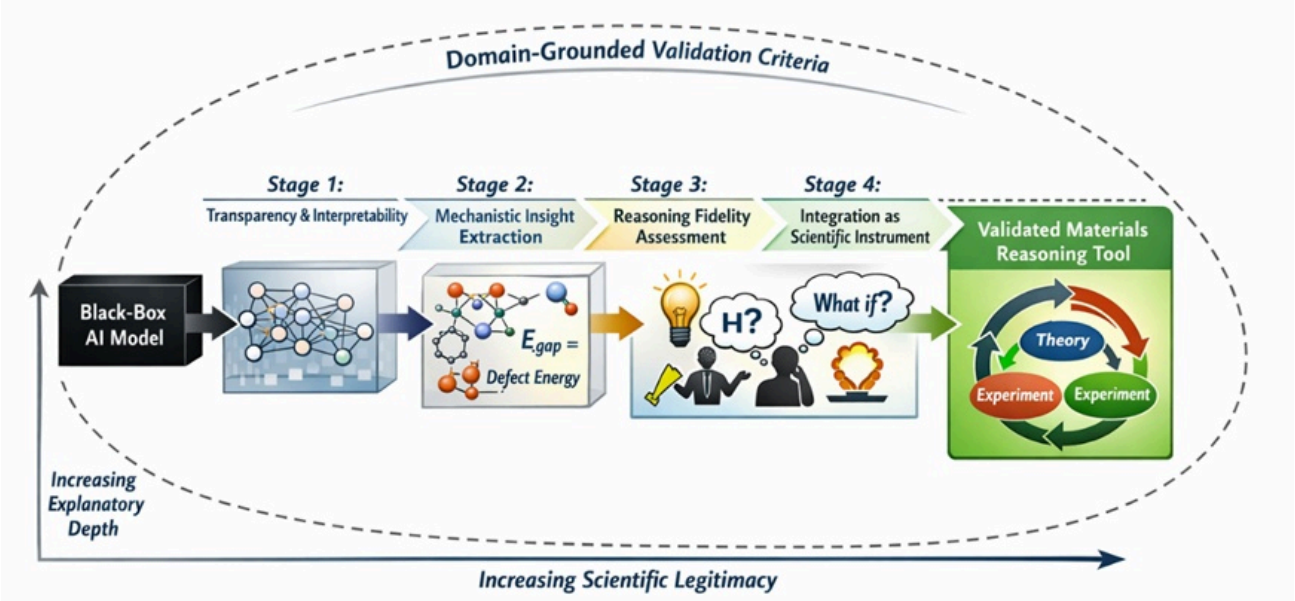

Figure 1. Evolution of AI models toward validated materials reasoning through four iterative, domain-grounded stages. The diagram's horizontal and vertical axes represent increasing scientific legitimacy and explanatory depth, respectively

Propositions

The proposed conceptual framework is grounded in the following core propositions, each specifying a necessary condition for validating artificial intelligence as a materials reasoning tool. The propositions are interdependent, cumulative, and non-substitutable, reflecting the staged logic of the validation pathway. Failure to satisfy any proposition constrains the epistemic status of the system, regardless of predictive performance.

Proposition 1: Transparency and interpretability are non-negotiable prerequisites for scientific reasoning validation

Transparency and interpretability constitute the foundational condition for any claim that an AI system engages in scientific reasoning. Absent systematic access to a model's representational structures and decisional logic—expressed in forms intelligible to domain experts—no meaningful assessment of mechanistic fidelity, inferential validity, or epistemic contribution is possible. This proposition asserts that opacity is categorically incompatible with reasoning validation: high predictive accuracy cannot compensate for inaccessible or inscrutable internal logic when scientific understanding is the evaluative standard.

Proposition 2: Mechanistic insight extraction requires structured correspondence to domain theory, not merely feature attribution

Mechanistic insight is established only when model-derived explanations exhibit structural correspondence to recognized physical or chemical concepts at an appropriate level of abstraction. Such correspondence is evidenced by:

(i) alignment with established causal relations (e.g., electronic structure descriptors governing bonding behavior),

(ii) consistency across chemically or structurally analogous systems, and

(iii) reproduction of qualitative trends anticipated from first-principles understanding, even when quantitative agreement is imperfect.

This proposition distinguishes mechanistic explanation from descriptive attribution, asserting that explanations lacking theoretical anchoring cannot support claims of scientific understanding.

Proposition 3: Reasoning fidelity is evidenced by disciplined inferential behavior, not by interpretability alone

An AI system demonstrates reasoning fidelity when it exhibits behaviors characteristic of abductive and deductive scientific inference within the materials domain. These behaviors include generating novel, testable mechanistic hypotheses; robustness under counterfactual perturbations that preserve physical realism; and logical consistency in reconciling apparently conflicting observations within a unified explanatory account. This proposition asserts that interpretability is necessary but insufficient: reasoning is validated only when explanations actively support inference, not when they merely rationalize post hoc predictions.

Proposition 4: Validation as a scientific instrument requires demonstrable contribution to theory, not just decision support

Full validation is achieved only when the AI system functions as an epistemic instrument within materials science, contributing to the refinement, extension, or unification of existing theoretical constructs. Such a contribution is evidenced by explanations that:

(i) close or clarify explanatory gaps in current theories,

(ii) motivate revisions or extensions of conceptual models, or

(iii) synthesize disparate empirical observations under coherent organizing principles.

This proposition establishes a clear boundary between AI as a workflow accelerator and AI as a participant in scientific knowledge production.

Proposition 5: Reasoning validation is inherently iterative and reflexive, not linear or terminal

The validation pathway is intrinsically iterative: outcomes at later stages must retroactively inform and refine earlier ones. Identification of physically implausible attributions at the mechanistic stage necessitates revisiting transparency and interpretability assumptions, while inconsistencies in reasoning fidelity require re-evaluation of representational choices or domain constraints. This proposition asserts that reasoning validation is a process of epistemic calibration rather than a one-time certification. **Figure 2** conceptually distinguishes predictive AI, interpretable AI, and validated materials reasoning systems, highlighting the epistemic thresholds defined by the proposed propositions.

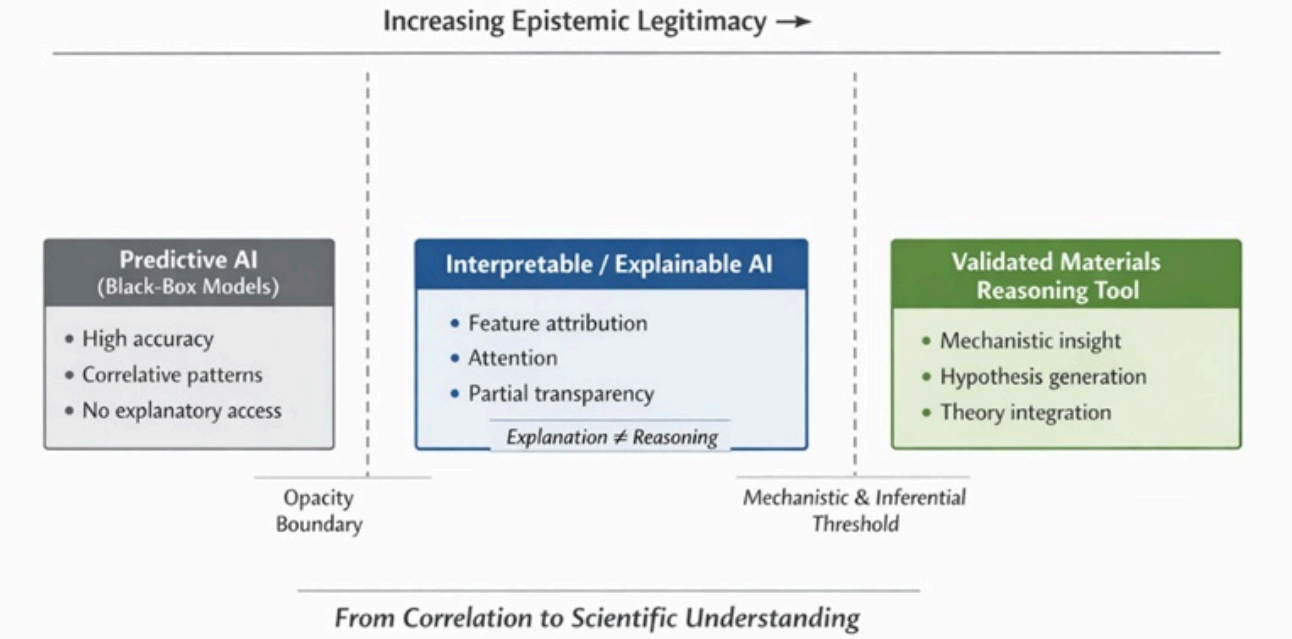

Figure 2. Conceptual boundary between prediction, interpretation, and scientific reasoning in materials AI

Results and Discussion

The conceptual framework advanced in this manuscript departs fundamentally from prior efforts in materials-oriented explainable AI, which have predominantly emphasized pragmatic objectives such as increasing user trust, improving model debugging, or identifying sources of predictive error [8, 22, 27]. While these goals are undeniably valuable for deployment and reliability, they remain largely instrumental. By contrast, the present framework explicitly addresses epistemological validation, grounding the assessment of AI systems in philosophical accounts of scientific reasoning—particularly mechanistic explanation, inference to the best explanation, and the theory-laden character of observation [15, 29]. This shift reframes interpretability not as an auxiliary feature but as a necessary step toward evaluating whether AI systems can meaningfully participate in scientific knowledge production.

A central strength of the proposed framework lies in its staged and modular architecture. Each validation stage is conceptually distinct and progressively demanding, allowing individual components—such as transparency, mechanistic insight, or inferential robustness—to be developed and evaluated independently. This modularity is well-suited to the heterogeneous landscape of contemporary materials AI, which spans intrinsically interpretable symbolic models, physics-informed architectures, and large-scale foundation models trained on diverse materials data [16, 23]. Rather than imposing a single normative standard across all systems, the framework provides a flexible roadmap for incremental advancement. Researchers can demonstrate partial progress toward reasoning capability, situating their contributions within a broader epistemic trajectory even when full validation remains aspirational.

At the same time, the framework surfaces several important conceptual tensions. One such tension concerns the requirement for mechanistic correspondence. By emphasizing alignment with established physical and chemical concepts, the framework risks privileging human-centric explanatory forms and potentially constraining the discovery of genuinely novel representational schemes [30]. This concern is addressed by allowing explanations at multiple levels of abstraction, provided they remain mappable—rather than reducible—to physical concepts. In this sense, the framework does not require AI explanations to mirror existing theory verbatim; instead, it demands that they remain intelligible within

the explanatory space of materials science, preserving the possibility of theoretical expansion rather than enforcing conformity.

A second tension arises in assessing the fidelity of reasoning. Determining whether an AI-generated hypothesis is sufficiently novel, whether a counterfactual is physically meaningful, or whether an explanation supports genuine abductive inference inevitably involves normative judgment [31]. These judgments are not defects of the framework but reflections of the inherently normative nature of scientific reasoning itself. Consequently, the framework deliberately avoids rigid thresholds or universal metrics. Instead, it emphasizes explicit criteria, transparent assumptions, and iterative refinement through community consensus. In doing so, it mirrors the way scientific standards themselves evolve through collective practice rather than fixed formalization.

The final stage of the framework—integration as a scientific instrument—raises broader questions about agency, authorship, and epistemic credit in human–AI collaboration [32]. Importantly, the proposed validation pathway does not attribute autonomous scientific agency to AI systems. Rather, it positions AI as an *instrument* in the classical epistemological sense, analogous to advanced experimental apparatuses or simulation platforms. Just as a high-resolution microscope reshapes what can be observed without independently "doing science," a validated AI reasoning system extends the inferential reach of human scientists while leaving epistemic responsibility firmly with human actors. This framing preserves accountability while acknowledging that, once validated, AI systems can influence theory development rather than merely accelerate computation.

Beyond its immediate conceptual contributions, the framework highlights several underexplored research directions that warrant sustained attention. Foremost among these is the development of formal measures tailored to epistemic criteria, such as explanatory unification, counterfactual physical realism, and hypothesis refutability in materials contexts [26, 28]. Current metrics for explanation quality remain weakly coupled to scientific reasoning norms, often prioritizing simplicity or stability over epistemic relevance. Bridging this gap will require interdisciplinary collaboration among materials scientists, AI researchers, and philosophers of science, as well as the development of benchmark tasks explicitly designed to test reasoning rather than prediction.

More broadly, the framework suggests a reorientation of how success is defined in materials AI. As models grow in scale and capability, predictive accuracy alone becomes an increasingly insufficient indicator of scientific value. By articulating a principled pathway from black-box prediction to validated reasoning, this work provides a conceptual foundation for evaluating AI systems not only by what they predict, but by how they explain, infer, and integrate within the scientific enterprise. In doing so, it aims to support a more mature and epistemically grounded role for AI in the future of materials research.

Conclusion

The accelerating adoption of artificial intelligence in materials science demands more than demonstrations of predictive superiority. As AI systems increasingly influence discovery pathways, design decisions, and theoretical interpretation, the field requires **c**onceptual criteria to justify when AI outputs may be treated as legitimate contributions to scientific understanding rather than as high-performance correlations. This work responds to that need by proposing a structured validation pathway that progresses from black-box prediction to transparency, mechanistic insight, reasoning fidelity, and, ultimately, theoretical integration.

By grounding validation in both computational accessibility and established epistemological standards of scientific reasoning, the framework addresses a central tension in contemporary materials AI: the coexistence of extraordinary predictive power with limited explanatory legitimacy. Importantly, the pathway does not claim that current AI systems already satisfy these criteria, nor does it diminish the indispensable role of human judgment. Instead, it offers a principled sequence of conceptual milestones against which progress toward reasoning capability can be evaluated, communicated, and compared across diverse modeling paradigms.

If adopted, this framework could shift the discourse in materials AI from questions of *trustworthy prediction* to those of *justified reasoning*. In doing so, it positions AI not as an autonomous scientific agent, but as a validated scientific instrument—one capable of supporting hypothesis generation, theory refinement, and explanatory unification when appropriately constrained and assessed. Future work should focus on refining the proposed propositions, operationalizing validation criteria, and examining the framework's applicability across materials sub-domains and modeling scales. More broadly, elevating AI from predictive utility to epistemic participation invites renewed reflection on the nature of scientific inference itself in an era increasingly shaped by machine intelligence.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 10 Jan 2024 Revised: 01 Apr 2024 Accepted: 03 May 2024
Published online: 18 July 2024

References

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

Wang AY-T, Murdock RJ, Kauwe SK, Oliynyk AO, Gurlo A, Brgoch J, et al. Machine learning for materials scientists: An introductory guide toward best practices. Chem Mater. 2020;32:4954-65.

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

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 design of inorganic materials. npj Comput Mater. 2020;6:84.

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

Oviedo F, Ren Z, Sun X, Settens C, Liu Z, Hartono NTP, et al. Fast and interpretable classification of small X-ray diffraction datasets using data augmentation and deep neural networks. NPJ Comput Mater. 2019;5:60.

Lipton P. Inference to the best explanation. Routledge; 2004.

Oviedo F, Ferres JL, Agarwal A, Cai T, Ren H, Wang S, et al. Interpretable and explainable machine learning for materials science and chemistry. Acc Mater Res. 2022;3:579-90.

Zhong W, Chen X, Liu J, Zheng Y. Explainable artificial intelligence in materials science: A review. J Mater Informatics. 2023;3:100032.

Zhang Y, Ling C. A strategy to apply machine learning to small datasets in materials science. NPJ Comput Mater. 2020;6:25.

Dai Z, Liu H, Le QV, Tan M. CoAtNet: Marrying convolution and attention for all data sizes. Adv Neural Inf Process Syst. 2021;34:3965-77.

Jablonka KM, Ongari D, Moosavi SM, Smit B. Big-data science in porous materials: Materials genomics and machine learning. Chem Rev. 2020;120:8066-129.

Dan Y, Zhao Y, Li X, Li S, Hu M, Hu J. Generative models for inverse design of inorganic solid materials. J Mater Informatics. 2021;1:100013.

Popper KR. The logic of scientific discovery. Routledge; 2002.

Lipton P. Inference to the best explanation. Routledge; 2021.

Chen D, Gao K, Nguyen DD, Chen X, Jiang Y, Wei GW, Pan F. Algebraic graph-assisted bidirectional transformers for molecular property prediction. Nat Commun. 2021;12:3521.

Goodall REA, Lee AA. Predicting materials properties without crystal structure: Deep representation learning from stoichiometry. Nat Commun. 2020;11:6280.

Xu P, Ji X, Li M, Lu W. Small data machine learning in materials science. npj Comput Mater. 2023;9:42.

Wiltschko A, et al. Pitfalls of explainable AI in materials science. arXiv preprint arXiv:2205.12345. 2022.

Wang AY-T, et al. Quantifying uncertainty in machine-learned interatomic potentials. J Chem Phys. 2022;156:174702.

Chen H, et al. Spurious correlations in machine learning for materials. Comput Mater Sci. 2021;196:110564.

Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020;2:56-67.

White A, et al. Interpretable machine learning for high-throughput materials discovery. Adv Mater. 2022;34:2108723.

Jablonka KM, Ai Q, Al-Feghali A, Badhwar S, Bocquet JD, Bran AM, et al. 14 examples of how LLMs can transform materials science and chemistry: a reflection on a large language model hackathon. Digit Discov. 2023;2:1233-50.

Karniadakis GE, Kevrekidis IG, Lu L, Perdikaris P, Wang S, Yang L. Physics-informed machine learning. Nat Rev Phys. 2021;3:422-40.

Zhou J, Gandomi AH, Chen F, Holzinger A. Evaluating the quality of machine learning explanations: A survey on methods and metrics. Electronics. 2021;10:593.

Lin Z, Chou E, Jang J, Kim Y, Chuang Y. Uncertainty quantification in machine learning interatomic potentials. Nat Rev Mater. 2022;7:670-83.

Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. ACM Comput Surv. 2021;54:115.

Salmon WC. Four decades of scientific explanation. University of Minnesota Press; 1989.

Marcus G. The next decade in AI: four steps towards robust artificial intelligence. arXiv preprint arXiv:2002.06177. 2020.

Miller T. Explanation in artificial intelligence: Insights from the social sciences. Artif Intell. 2019;267:1-38.

Floridi L, Cowls J. A unified framework of five principles for AI in society. Harv Data Sci Rev. 2019;1(1). https://doi.org/10.1162/99608f92.8cd550d1.