REVIEW **Open access**

Interpreting Materials Data with Artificial Intelligence: From Prediction to Scientific Understanding

Fatima Zahra Amrani[1*], Youssef Benali[1], Samira El-Haddad[2]

Abstract

The integration of artificial intelligence (AI) and machine learning (ML) into materials science has fundamentally transformed how material properties are predicted, analyzed, and understood. While early data-driven approaches emphasized predictive accuracy and high-throughput screening, recent advances are increasingly focusing on interpretability and explainability, enabling AI models to contribute to mechanistic scientific insight rather than functioning as opaque black boxes. This study examines the evolution of interpretable AI in materials science and highlights the transition from property prediction to explanation-driven understanding of structure–property relationships. In this thesis, we investigate the progress in machine learning frameworks that operate with limited or implicit structural information, alongside the growing use of explainable AI (XAI) techniques to uncover physically meaningful descriptors, atomic-scale interactions, and microstructural drivers of material behavior. Methods such as graph-based learning, attention mechanisms, feature attribution, and uncertainty-aware modeling are discussed for their ability to improve model reliability, expose data bias, and guide hypothesis generation. Representative applications across alloys, perovskites, organic semiconductors, and ferroelectric materials demonstrate how interpretable models have revealed governing mechanisms spanning atomic, mesoscopic, and macroscopic length scales. Beyond individual case studies, this study examines persistent challenges in interpretable materials AI, including data quality, generalizability, explanation stability, and computational overhead. We argue that interpretability is not merely an auxiliary feature but a prerequisite for trustworthy and scientifically helpful AI in materials research. By synthesizing recent methodological and application-driven advances, this review positions interpretable AI as a critical enabler of mechanism-oriented discovery, experimental validation, and theory development, ultimately advancing AI from a predictive accelerator to an integral partner in scientific understanding.

Keywords Interpretability, Artificial intelligence, Property prediction, Machine learning, Explainable AI, Materials science

*Correspondence:
Fatima Zahra Amrani
fatima.amrani@gmail.com
[1] Department of Computational Materials Science, Faculty of Engineering, Mohammed V University, Rabat, Morocco
[2] Department of Artificial Intelligence in Engineering Systems, Faculty of Engineering, University of Fez, Fez, Morocco

Introduction

Materials science has long been driven by the objective of designing substances with tailored properties to meet the demands of applications spanning energy storage, electronics, structural engineering, and aerospace technologies. Traditionally, materials discovery and optimization relied on iterative experimentation informed by empirical heuristics and theoretical modeling. Among these approaches, density functional theory (DFT) has played a central role by providing atomistic-level insights into electronic structure and thermodynamic stability. Despite its success, DFT remains computationally intensive, limiting its applicability for exhaustive exploration of large compositional and configurational

spaces [1]. In response to these constraints, large-scale initiatives such as the Materials Genome Initiative (MGI), launched in 2011, sought to accelerate materials discovery through high-throughput computation, standardized databases, and data-sharing infrastructure [2]. While these efforts dramatically expanded the volume of available materials data—encompassing crystal structures, physical properties, and synthesis parameters—they also exposed a critical bottleneck: the exponential growth of data has outpaced the capacity of traditional analysis and modeling methodologies.

Artificial intelligence (AI) and machine learning (ML) have emerged as transformative tools for addressing this challenge by enabling data-driven inference across high-dimensional materials spaces. By the early 2020s, ML models demonstrated the ability to predict a wide range of materials properties with high accuracy, in some cases rivaling or surpassing human intuition and physics-based heuristics [3, 4]. Deep learning architectures, in particular, have been successfully applied to predict mechanical properties such as compressive strength in molecular solids and tensile performance in complex alloys, leveraging large datasets to uncover non-obvious correlations inaccessible to conventional approaches [4, 5]. These successes positioned ML as a powerful complement to first-principles methods, capable of accelerating screening and optimization processes across diverse materials domains.

However, the rapid adoption of ML has also revealed fundamental limitations. Many high-performing models function as opaque “black boxes,” producing accurate predictions without providing insight into the underlying physical or chemical mechanisms. This lack of transparency poses a significant barrier in materials science, where mechanistic understanding, reproducibility, and theoretical consistency are essential for scientific progress and experimental validation [6, 7]. Without interpretability, ML models risk being perceived as purely empirical tools, limiting their ability to inform hypothesis-driven research or guide rational materials design.

The field experienced a pivotal conceptual shift toward interpretable and explainable artificial intelligence in materials science. Interpretability refers to a model's intrinsic transparency, enabling users to trace predictions directly back to input features or learned representations. In contrast, explainability encompasses post-hoc techniques that rationalize the outputs of complex, otherwise opaque models [7, 8]. This shift reflects a growing recognition that predictive accuracy alone is insufficient; models must also provide insight into structure–property relationships to support scientific discovery. Interpretable ML frameworks have, for example, identified physically meaningful descriptors governing magnetic behavior in metallic glasses and dielectric responses in oxide materials, offering pathways to connect data-driven predictions with established materials theory [9-11]. **Figure 1** summarizes how interpretable AI closes the loop between materials data, prediction, explanation, and validation, enabling models to contribute to mechanistic understanding rather than functioning as black-box predictors.

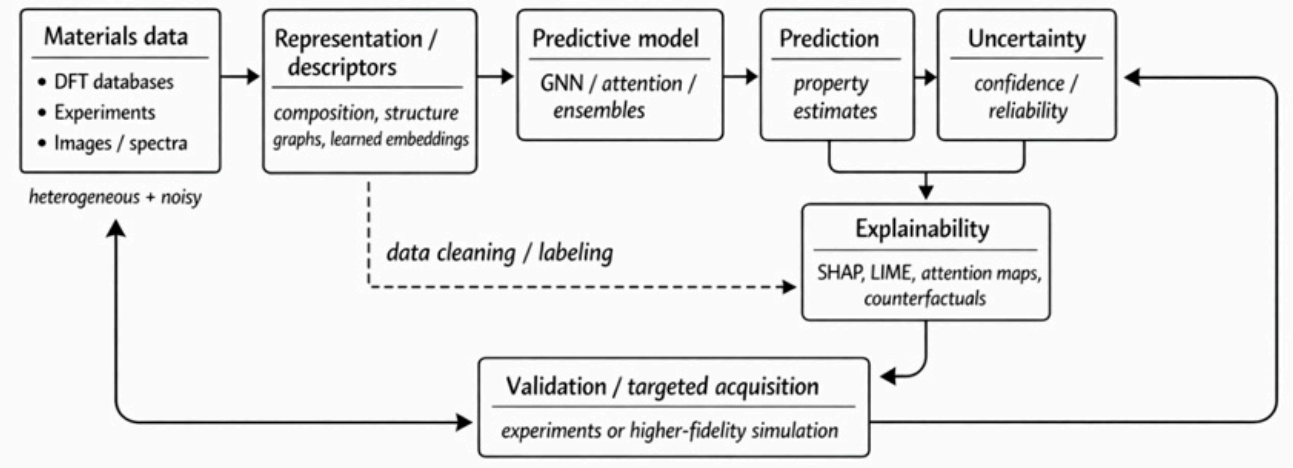

Figure 1. From prediction to scientific understanding in materials AI

A conceptual loop showing how materials data (DFT, experiments, images, spectra) are transformed into representations and predictive models, and then translated into scientific insight through explainability (feature attribution, attention, counterfactuals) and uncertainty-aware confidence assessment. Explanations and confidence guide validation and targeted data acquisition, closing the loop toward mechanism-oriented discovery rather than prediction alone.

The growing emphasis on explainability has also enabled ML to play a more active role in experimental planning and theory development. By revealing which features most strongly influence predictions, explainable models help validate hypotheses, expose biases in training data, and highlight regions of chemical space where additional data acquisition is most valuable [6, 12]. In this way, AI transitions from a passive predictive engine to an active participant in the scientific reasoning process.

The objectives of this review are threefold. First, it surveys recent advances in machine learning approaches for predicting materials properties, with an emphasis on developments reported. Second, it examines emerging techniques for enhancing model interpretability and explainability, situating them within the broader context of scientific reliability and trust. Third, it highlights representative applications across diverse material classes in which explainable AI has yielded genuine physical insights rather than mere predictive success. By focusing on this critical period, the review underscores how AI is evolving from a computational accelerator into a catalyst for deeper understanding in materials research.

Results and Discussion

Advances in machine learning for property prediction in materials

Machine learning (ML) has transformed materials property prediction by enabling models to learn complex structure–property relationships directly from data, without explicit programming of physical laws. Developments increasingly focused on addressing chemically diverse and incomplete datasets, reflecting practical limitations in experimental and computational materials databases. A notable advance was the use of deep representation learning based solely on stoichiometric information, enabling the prediction of formation energies, band gaps, and related properties without explicit crystal-structure inputs [3]. These approaches commonly employ graph neural networks (GNNs), in which compositions are represented as graphs with atoms as nodes and compositional relationships encoded as edges, allowing scalable learning across vast chemical spaces [13].

High-throughput screening has emerged as a significant application of ML-driven property prediction. In magnetic topological materials, ML-assisted electronic structure workflows enabled the screening of thousands of compounds, leading to the identification of previously unknown topological phases with high predictive accuracy [2]. Similarly, ML-based screening in alloy systems has optimized compositions for enhanced tensile strength and electrical conductivity, achieving performance gains and efficiency far beyond traditional trial-and-error or purely physics-based approaches [5]. These workflows frequently incorporate ensemble models or Bayesian optimization strategies to iteratively refine predictions, explicitly accounting for uncertainty to prioritize high-value candidates for further computation or experimental validation [14].

In organic and hybrid materials, ML models have demonstrated strong predictive performance for optoelectronic and physicochemical properties, including solubility, lipophilicity, and charge transport. Self-attention mechanisms in neural networks have been particularly effective in capturing long-range correlations between molecular substructures, improving generalization across chemical families [15]. In halide perovskites, data fusion strategies combining experimental degradation measurements with simulated descriptors have enabled more reliable predictions of environmental stability, addressing a critical bottleneck in photovoltaic applications [16, 17]. Beyond scalar property prediction, deep learning has also been applied to complex experimental observables, including inelastic neutron scattering spectra, where calibrated neural networks successfully extracted dynamic parameters from noisy, high-dimensional data [18]. To synthesize how different data sources and representation choices have shaped both predictive capability and interpretive depth in recent materials-ML studies, a consolidated overview is provided in **Table 1**.

Table 1. Data modalities, representations, and ML model families used for materials property prediction, with typical target types and interpretability levers. The table highlights how representation choices shape not only predictive performance but also the kinds of scientific insight that can be extracted

Data modality	Representation used in ML	Common model families	Typical predicted targets (examples)	What can be explained (interpretability "handle")	Common limitations
Composition-only datasets	Stoichiometric vectors; elemental embeddings	Deep nets; composition-based GNN variants	Formation energy, band gap, stability proxies	Global feature importance (element identity, fraction); compositional rules	Weak mechanistic link to bonding/structure if crystal info is absent
Crystal structures (atomic graphs)	Graphs (nodes = atoms, edges = bonds/neighbor lists)	GNNs; attention-augmented GNNs	Band gaps, formation energies, and elastic moduli	Atom/edge importance; coordination-environment contributions; attention weights	Explanations can be unstable; risk of "plausible but wrong" interpretations
Microstructure images	CNN features; latent embeddings	CNNs; vision transformers (in some works)	Strength, defects, phase fractions	Saliency/attribution maps; region-level drivers (grains/pores)	Dataset bias (imaging conditions); explanations may reflect artifacts
Spectroscopy/scattering	1D/2D signal embeddings	CNN/RNN; calibrated networks	Peaks/parameters; dynamic descriptors	Feature regions that drive predictions (peak ranges)	Noise sensitivity; preprocessing choices strongly affect explanations
Hybrid (DFT + experiments)	Data fusion features; multi-view embeddings	Multi-task nets; ensembles; Bayesian methods	Stability under stress; performance degradation trends	Which data source dominates; feature interactions, and confidence-aware screening	Domain shift between simulated and experimental distributions
Time-series functional measurements	Temporal embeddings	LSTM/temporal models	Switching dynamics; transient conductance features	Time-window importance; event-driven interpretability	Hard to map learned temporal features to physical mechanisms without careful design

Collectively, these advances demonstrate that ML has matured into a powerful predictive engine for materials discovery. However, they also highlight a growing limitation: predictive accuracy alone is insufficient for scientific progress. Without

interpretability, ML models risk functioning as opaque black boxes, limiting their ability to generate mechanistic insight or guide theory-driven materials design [12].

The importance of interpretability in materials AI

As ML models become more complex, their limited transparency poses challenges for adoption in materials science, where reproducibility, physical plausibility, and mechanistic understanding are essential [6, 7]. Interpretability addresses these challenges by enabling model decisions to be traced back to meaningful input features, thereby fostering trust and facilitating scientific insight [8].

From a scientific perspective, interpretability serves not only as a validation tool but also as a mechanism for hypothesis generation. By revealing correlations between descriptors and target properties, interpretable ML can uncover latent structure–property relationships and suggest new physical principles governing materials behavior [12]. A widely adopted taxonomy of explainable artificial intelligence (XAI) distinguishes between intrinsic interpretability—models designed to be transparent by construction—and post-hoc explainability techniques, which approximate the behavior of complex black-box models after training [7, 19].

Post-hoc methods such as Shapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) provide both global explanations of model behavior and local, instance-specific interpretations, with trade-offs between fidelity and human interpretability [19]. In materials science, interpretability has proven particularly valuable for descriptor selection and bias detection. Multi-layer feature attribution approaches have guided the selection of physically meaningful descriptors for targeted property prediction [12]. At the same time, interpretability analyses have also been used to identify dataset biases, such as those arising from uneven sampling in microstructure image datasets, improving model robustness and generalization.

By prioritizing interpretability, ML transitions from a purely predictive tool to a framework that actively contributes to scientific understanding, aligning AI-driven discovery with the fundamental goals of materials science [8].

Methods for explainable AI in materials science

Explainable AI methods in materials science can be broadly classified into model-agnostic and model-specific approaches.

Model-agnostic techniques, such as SHAP and LIME, assign feature importance scores after model training and can be applied across a wide range of ML architectures. In high-entropy and multi-principal-element alloys, SHAP analyses have identified the dominant compositional and thermodynamic factors controlling hardness and phase stability, directly informing rational alloy design strategies [20, 21]. LIME has similarly been applied to explain predictions in ferroelectric switching behavior, revealing the most influential electromechanical descriptors governing polarization dynamics [22].

Model-specific approaches embed interpretability within the learning architecture itself. Attention-based neural networks highlight the relative importance of atomic interactions or compositional features during prediction, providing intuitive insight into learned representations [23]. Graph neural networks augmented with attention mechanisms have enabled interpretable node and edge embeddings that correlate with physically meaningful quantities, such as coordination environments and angle-dependent interactions, in optical and electronic property prediction [24].

Beyond architectural transparency, informed ML approaches integrate domain knowledge directly into model construction. Physics-informed and Bayesian models incorporate physical constraints, conservation laws, or thermodynamic priors to ensure consistency with established theory [25]. For example, embedding physical knowledge into Bayesian networks has enabled layer-by-layer optimization of photovoltaic fabrication processes [26]. Unsupervised and self-supervised learning methods further enhance interpretability by extracting latent parameters from spatiotemporal datasets, such as disentangling ferroelectric domain dynamics from microscopy or spectroscopy data [27–29]. **Figure 2** summarizes how

complementary XAI families (intrinsic, post-hoc, and physics-informed) produce different forms of insight—rules, mechanisms, and confidence—supporting interpretation across diverse materials classes.

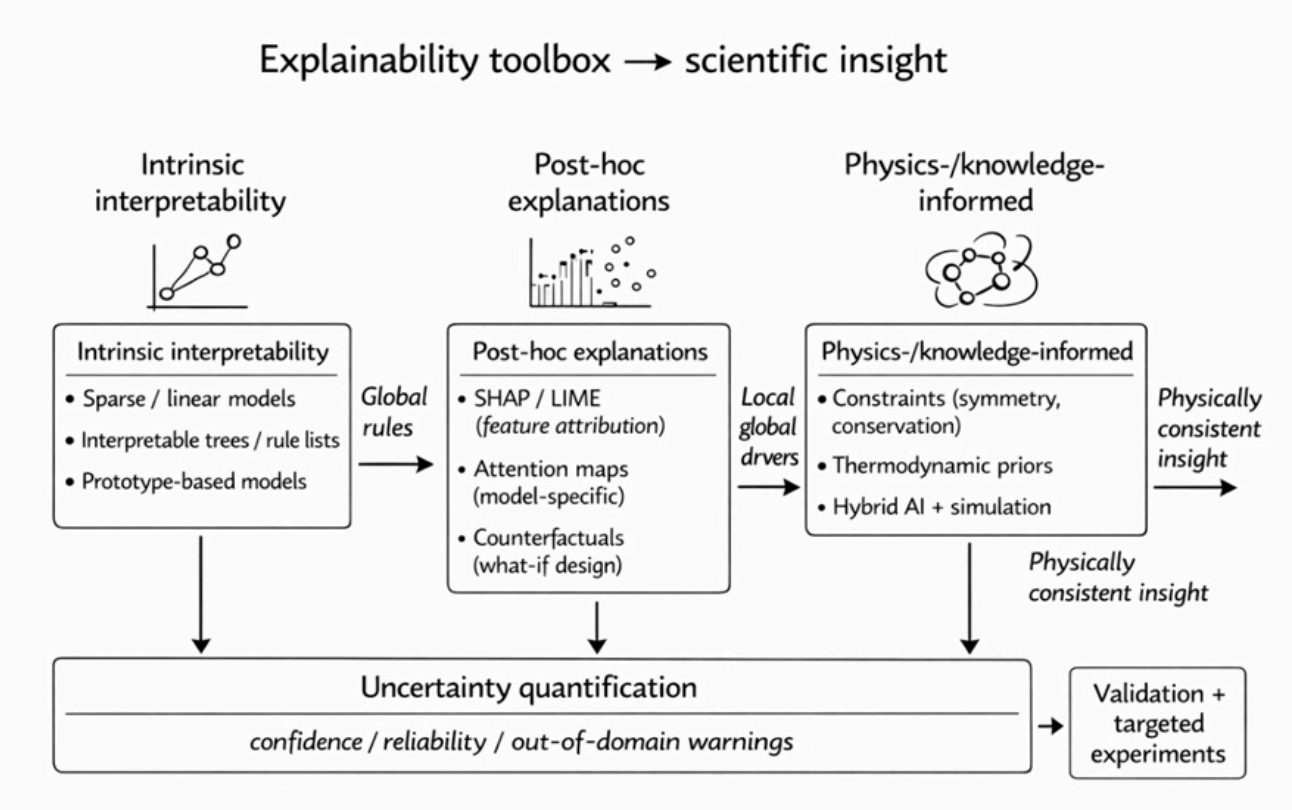

Figure 2. Explainable AI toolbox in materials science and the kinds of scientific insight each approach provides

Intrinsic interpretability (transparent-by-design models), post-hoc explanations (feature attribution and counterfactual reasoning), and physics- and knowledge-informed learning provide complementary routes to mechanistic understanding. Uncertainty quantification adds a layer of confidence that guides screening decisions and targeted validation.

Uncertainty quantification provides an additional explanatory dimension by identifying predictions with limited reliability. Techniques such as ensemble variance estimation and Bayesian neural networks flag regions of chemical space with low model confidence, guiding targeted data acquisition and reducing the risk of overconfident extrapolation [14].

Together, these explainable AI methodologies enable ML models to deliver not only accurate property predictions but also mechanistic and causal insight, strengthening the role of AI as a scientifically interpretable partner in materials discovery rather than a purely computational shortcut [29].

Applications to specific materials classes

Interpretable artificial intelligence has been successfully applied across a wide range of material classes, demonstrating its capacity not only to enhance predictive performance but also to generate scientifically meaningful insights. These applications illustrate how explainable models can bridge data-driven predictions with established materials knowledge, enabling mechanism-oriented understanding across length scales.

In metallic systems and alloys, interpretable ML approaches have expanded classical design principles. Models have predicted solid-solution formation beyond the constraints of traditional Hume–Rothery rules, revealing stabilizing factors in complex compositional spaces previously inaccessible to heuristic reasoning. SHAP-based analyses clarified the relative

importance of elemental size mismatch, electronegativity differences, and electronic parameters, providing transparent explanations for predicted phase stability [21]. In Fe-based metallic glasses, interpretable ML strategies identified key descriptors governing magnetic properties and glass-forming ability, linking compositional complexity to electronic structure and thermodynamic stability [9]. These insights have supported rational alloy design by translating high-dimensional predictions into physically interpretable rules.

Halide and oxide perovskites represent another class where explainable AI has had a substantial impact, particularly in addressing long-standing challenges related to environmental stability. Data fusion strategies combining experimental degradation data with simulated descriptors enabled ML models to predict stability trends under humidity, heat, and illumination stress. Explainability analyses revealed dominant degradation pathways, such as ion migration and surface reactions, clarifying the mechanisms underlying performance loss [16]. Furthermore, ML-guided selection of capping and passivation layers benefited from interpretability tools that explained how interfacial chemistry and barrier properties enhance environmental resilience, directly informing materials engineering decisions [17].

In organic semiconductors, deep learning models have achieved accurate predictions of optoelectronic properties, including charge mobility, band alignment, and absorption characteristics. Attribution-based explainability methods have synthesized these predictions into interpretable microstructural features, such as conjugation length, molecular packing motifs, and functional group distributions [30, 31]. By revealing how molecular architecture influences macroscopic electronic behavior, interpretable AI has supported both molecular design and the development of structure–property theory in organic electronics.

Ferroelectric materials further exemplify the explanatory power of ML when combined with time-resolved and spatially resolved data. Long short-term memory (LSTM) networks have been applied to deconvolute conductance signals associated with ferroelectric switching, enabling interpretable explanations of domain wall dynamics and transient transport behavior [32]. Complementary unsupervised learning approaches have disentangled complex domain wall geometries from microscopy data, providing insight into how local structural variations influence piezo response and electromechanical coupling [27]. Together, these studies highlight how interpretable ML can uncover mesoscale mechanisms that connect atomic-scale interactions to functional device properties.

Collectively, these examples demonstrate that AI's value in materials science extends beyond prediction toward interpretation, enabling the discovery of mechanisms spanning atomic, microstructural, and macroscopic regimes.

Challenges and limitations in interpretable AI for materials

Despite substantial progress, several challenges continue to limit the effectiveness and reliability of interpretable AI in materials science. A primary concern is data quality. Materials datasets often contain noise, inconsistencies, or sampling biases arising from experimental variability, computational approximations, or underrepresentation of certain material classes. When such issues are present, interpretable models may produce misleading explanations that reflect dataset artifacts rather than underlying physical relationships. Ensuring data curation, uncertainty awareness, and bias mitigation remains essential for trustworthy explanations.

Generalizability represents another major limitation. Many ML models are trained on narrowly defined material classes or property ranges, restricting their applicability to broader chemical spaces. When applied outside their training domain, models may generate explanations that lack physical meaning, undermining scientific reliability [1]. This challenge is particularly acute for interpretable methods, where seemingly plausible explanations may mask poor extrapolation performance.

Computational cost further constrains the scalability of explainable AI. Post-hoc explanation techniques, especially those applied to large deep learning models, often require repeated model evaluations or complex approximations, significantly increasing computational overhead [8]. This limits their routine use in high-throughput screening workflows. Moreover, the

assessment of explanations lacks standardized quantitative metrics, making it difficult to compare methods or assess explanation quality across studies [19]. To translate these conceptual limitations into actionable guidance, the principal challenges affecting interpretability in materials-AI workflows, along with corresponding mitigation practices, are organized in **Table 2**.

Table 2. Key challenges in interpretable AI for materials science and practical mitigation strategies

Challenge	Why it matters for interpretability	Typical symptom in practice	Practical mitigation (what to report/do)
Data noise + bias	Explanations may reflect artifacts rather than physics	Attributions highlight irrelevant features (e.g., measurement conditions)	Curate datasets; document preprocessing; stratify by source; bias audits; include uncertainty where possible
Limited generalizability	"Plausible" explanations can mask out-of-domain failure	Confident-looking explanations in unseen chemistries	Domain shift checks; out-of-distribution detection; external validation sets; report applicability domain
Explanation instability	Small perturbations can change feature importance	SHAP/LIME varies across runs or near decision boundaries	Stability tests (bootstrapping); aggregate explanations; compare multiple XAI methods
High computational overhead	Limits the feasibility in high-throughput workflows	Explanations are too slow for screening	Use approximations; subset analysis; prioritize explanation for top candidates and failures
Missing standards for evaluation	Hard to compare methods or claim "better explanations"	Qualitative-only interpretability claims	Report fidelity/faithfulness proxies; human-in-the-loop checks; sanity tests; transparent reporting
Over-interpretation risk	Narratives can exceed what the model supports	"Mechanism" claimed from correlation alone	Use cautious language; validate with experiments/DFT; separate correlation vs causation

The table links common failure modes (data bias, weak generalization, explanation instability) to concrete actions that improve reliability and scientific usefulness.

Hybrid approaches that integrate ML with physical models offer a promising path forward by embedding domain knowledge directly into learning architectures. However, such approaches demand interdisciplinary expertise spanning materials science, physics, and data science, which can pose practical barriers to adoption [25, 26]. Addressing these challenges will be critical for ensuring that interpretable AI fulfills its potential as a tool for scientific understanding rather than merely post-hoc rationalization.

The integration of artificial intelligence into materials science has not only accelerated property prediction but has also opened new pathways for scientific understanding through interpretable models. Nevertheless, several conceptual and practical issues warrant careful discussion, particularly the balance between predictive accuracy and interpretability, the role of domain knowledge in enhancing explainability, and the implications for experimental validation.

Balancing accuracy and interpretability

A central tension in AI-driven materials research lies in the trade-off between model complexity and transparency. Highly expressive models, such as deep neural networks, often achieve superior predictive accuracy by capturing intricate nonlinear relationships within materials data [2, 3]. For example, graph neural networks have demonstrated exceptional performance in predicting band gaps and formation energies by explicitly modeling atomic connectivity and interactions [1]. However, without interpretability tools, these predictions provide limited mechanistic insight, restricting their scientific utility [2].

Conversely, simpler models—such as linear regressions or sparsely parameterized algorithms—offer inherent interpretability but may sacrifice predictive power in complex materials systems. Recent studies have shown that with carefully engineered features, such models can achieve competitive accuracy while remaining transparent. Kostiuchenko *et al.* demonstrated that interpretable linear models could reveal physically meaningful trends, such as the influence of electron configuration and elemental identity on magnetic properties, underscoring the value of simplicity when guided by domain knowledge.

Post-hoc explanation techniques, including SHAP and LIME, have played a crucial role in bridging this gap by rendering complex models more interpretable [2, 3]. In alloy design, SHAP analyses have highlighted the dominance of compositional factors over structural descriptors in determining hardness, offering actionable insights for materials optimization [2]. However, the fidelity of these explanations remains a concern. If the explanatory approximation does not accurately reflect the underlying model behavior, it may lead to misleading interpretations and false scientific conclusions [3]. Additionally, the computational expense of post-hoc explanations in high-dimensional materials datasets poses scalability challenges, particularly for large-scale screening applications [4].

Overall, achieving a balanced integration of accuracy and interpretability remains an open challenge. Progress will likely require thoughtful combinations of model design, explainability techniques, and physical intuition to ensure that AI contributes not only predictive power but also genuine scientific understanding in materials research.

Incorporating domain knowledge for enhanced explainability

Domain-specific knowledge is crucial for making AI models more reliable and interpretable in materials science [4, 9]. The NOMAD AI Toolkit exemplifies this by integrating physical constraints into data analysis workflows, ensuring that predictions align with known laws of physics [4]. By embedding symmetry considerations or thermodynamic principles into models, researchers can reduce overfitting and improve generalizability [6]. Wang *et al.* reviewed applications in which physics-informed machine learning predicted material behavior in composites, demonstrating that prior knowledge guides feature engineering and model selection [9].

This approach also addresses data scarcity, a common challenge in materials research where experimental data is expensive to generate [1, 3]. Transfer learning and multi-task learning, informed by domain expertise, allow models trained on abundant computational data to adapt to sparse experimental datasets [2]. For example, in perovskite stability prediction, data fusion techniques combined DFT calculations with experimental observations, using explainable AI to identify key degradation factors, such as humidity sensitivity [3].

However, incorporating domain knowledge requires interdisciplinary collaboration, as materials scientists must work with AI experts to define relevant constraints [2, 4]. Misapplication of knowledge can introduce biases, such as assuming certain symmetries that do not hold for novel materials [5].

Implications for experimental validation and discovery

Interpretable AI facilitates a feedback loop between prediction and experimentation, accelerating materials discovery [1, 2, 5]. By providing mechanistic insights, these models guide targeted experiments, reducing trial-and-error [6]. For instance, explainable models have identified novel descriptors for catalytic activity in alloys, leading to the synthesis of

high-performance electrocatalysts [2]. In ferroelectric materials, attention mechanisms in neural networks revealed domain wall dynamics, informing the design of energy-efficient devices [3].

Nevertheless, challenges in validation persist. AI-derived insights must be corroborated by physical experiments or simulations to ensure reliability [4]. Discrepancies between predicted and observed behaviors often stem from incomplete data representation, such as neglecting defects or environmental effects [5]. Moreover, ethical considerations arise in AI-driven discovery, including the potential for biased datasets to perpetuate inequalities in material applications [3].

Overall, the discussion underscores that while interpretable AI holds immense promise, its success hinges on addressing these trade-offs and challenges through continued methodological advancements.

Conclusion

In conclusion, the evolution of AI in materials science from predictive tools to instruments of scientific understanding represents a transformative shift. With interpretable, explainable techniques, researchers can now extract meaningful insights from complex data, bridging the gap between computation and comprehension. Key advancements include the development of toolkits such as MAST-ML and NOMAD, which democratize access to AI methods, and the application of XAI across diverse material classes, yielding discoveries spanning mechanical strength to optoelectronic performance.

Looking forward, future directions should focus on standardizing evaluation metrics for interpretability, ensuring consistency across studies. Integrating uncertainty quantification more robustly will enhance trust in AI predictions, particularly in safety-critical applications such as battery materials. Hybrid models combining AI with quantum simulations promise even greater accuracy and insight. Additionally, expanding datasets through collaborative platforms will mitigate biases and improve model robustness.

Ultimately, interpretable AI will empower materials scientists to tackle global challenges, such as sustainable energy and advanced manufacturing, by fostering innovation grounded in fundamental understanding.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 06 Jul 2023 Revised: 13 Sep 2023 Accepted: 15 Oct 2023
Published online: 18 January 2024

References

Jacobs R, Mayeshiba T, Afflerbach B, Miles L, Williams M, Turner M, et al. The Materials Simulation Toolkit for Machine learning (MAST-ML): An automated open source toolkit to accelerate data-driven materials research. Comput Mater Sci. 2020;175:109544.

Fung V, Hu J, Ganesh P, Sumpter BG. Explainable machine learning in materials science. npj Comput Mater. 2022;8:204.

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

Ghiringhelli LM, Carbogno C, Levchenko S, Mohamed F, Scheffler M. The NOMAD Artificial-Intelligence Toolkit: turning materials science data into knowledge. npj Comput Mater. 2022;8:249.

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

Choudhary K, DeCost B. Atomistic Line Graph Neural Network for improved materials property predictions. npj Comput Mater. 2021;7:185.

Chen C, Ong SP. A universal graph deep learning interatomic potential for the periodic table. Nat Comput Sci. 2022;2:718-28.

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

Wang C, Fu H, Ma L, Yin L, Zhou M. Application of machine learning for advanced material prediction and design. eMaterials. 2022;2:e12194.

Chen L, Kim C, Batra R, Lightstone JP, Wu C, Li Z, et al. Machine-learning-based closed-loop optimization of synthesis parameters. NPJ Comput Mater. 2020;6:61.

Antunes LM, Grau-Crespo R, Butler KT. Distributed representations of atoms and materials for machine learning. NPJ Comput Mater. 2022;8:44.

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;4:1305-22.

Yuan R, Tian Y, Xue D, Zhou Y, Ding X, Sun J, et al. Machine learning in hierarchical property prediction: A case study of perovskites. Comput Mater Sci. 2020;174:109483.

Chen C, Zuo Y, Ye W, Li X, Ong SP. Learning properties of ordered and disordered materials from multi-fidelity data. Nat Comput Sci. 2021;1:46-53.

Goussard V, Duprat F, Ploix JL, Dreyfus G, Nardello-Rataj V, Aubry JM. A new machine-learning tool for fast estimation of liquid viscosity: application to cosmetic oils. J Chem Inf Model. 2020;60:2012-23.

Jablonka KM, Ong SP, Garaj S, Camp J. An ecosystem for digital reticular chemistry. ACS Cent Sci. 2021;7:1231-40.

He J, Su X, Wang C, Li J, Hou Y, Li Z, et al. Interpretable machine learning-assisted discovery of new topological materials. Acta Mater. 2022;240:118341.

Weber JK, Morrone JA, Bagchi S, Estrada JD, Pabon SK, Zhang L, et al. Simplified, interpretable graph convolutional neural networks for small molecule activity prediction. J Comput Aided Mol Des. 2022;36:173-85.

Court CJ, Yildirim B, Jain A, Cole JM. 3-D inorganic crystal structure generation and property prediction via representation learning. J Chem Inf Model. 2020;60:4518-35.

Zhong X, Gallagher B, Liu S, Kailkhura B, Hiszpanski A, Yong-Jin Han T. Explainable machine learning in materials science. npj Comput Mater. 2022;8:204.

Mishin Y. Machine-learning interatomic potentials for materials science. Acta Mater. 2021;214:116980.

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.

Pun GP, Yamakov V, Hickman J, Glaessgen E, Mishin Y. Development of a general-purpose machine-learning interatomic potential for aluminum by the physically informed neural network method. Phys Rev Mater. 2020;4:113807.

Esterhuizen JA, Goldsmith BR, Linic S. Theory-guided machine learning finds geometric structure-property relationships for chemisorption on subsurface alloys. Chem. 2020;6:3100-17.

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

Chen Y, Peng B, Kontogeorgis GM, Liang X. Machine learning for the prediction of viscosity of ionic liquid-water mixtures. J Mol Liq. 2022;351:118616.

Karamad M, Magar R, Shi Y, Wen S, Devereux M, Siahrostami S, et al. Orbital graph convolutional neural network for material property prediction. Phys Rev Mater. 2020;4:093801.

Louis SY, Zhao Y, Nasiri A, Wang X, Song Y, Liu F, et al. Graph convolutional neural networks with global attention for improved materials property prediction. Phys Rev Mater. 2020;4:053801.

Park CW, Wolverton C. Self-attention based molecule representation for predicting drug-target interaction. Phys Rev Mater. 2020;4:053603.

Lin YS, Pun GPP, Mishin Y. Development of a physically-informed neural network interatomic potential for tantalum. Comput Mater Sci. 2022;205:111180.

Kaufmann K, Maryanovsky D, Mellor WM, Zhu C, Rosengarten AS, Vecchio KS. Discovery of high-entropy ceramics via machine learning. NPJ Comput Mater. 2020;6:42.

Chan H, Cherukara MJ, Loeffler TD, Narayanan B, Sankaranarayanan SKRS. Machine learning enabled autonomous microstructural characterization in 3D samples. NPJ Comput Mater. 2020;6:173.