REVIEW Open access

Interpretability in Materials AI — What "Explanation" Means and How It Should Be Evaluated Conceptually

Maria Hernandez[1*], Carlos Vega[1], Lucia Torres[2]

Abstract

The integration of artificial intelligence (AI) and machine learning (ML) into materials science has revolutionized the discovery, design, and optimization of new materials, enabling accelerated predictions of properties and behaviors previously unattainable with traditional methods. However, the "black-box" nature of many advanced AI models poses significant challenges, including a lack of transparency that hinders scientific understanding, trust, and practical adoption in materials research. This narrative review explores the concept of interpretability in materials AI, focusing on what constitutes an "explanation" and how it should be conceptually evaluated. Drawing from recent advancements in explainable AI (XAI), we delineate definitions of explanations tailored to materials informatics, emphasizing their role in bridging computational predictions with physical insights. We examine thematic aspects such as intrinsic versus post-hoc interpretability methods, the multidimensional nature of explanations (e.g., local vs. global, feature-based vs. mechanistic), and conceptual frameworks for evaluation, including criteria like fidelity, comprehensibility, robustness, and domain-specific relevance. By synthesizing the literature, we highlight how explanations can enhance materials discovery across alloy design, catalyst development, and polymer engineering, while addressing gaps in current evaluation practices. The review underscores the need for standardized conceptual metrics that go beyond quantitative benchmarks to incorporate qualitative, human-centered assessments in materials science contexts. Ultimately, this work aims to guide researchers toward developing interpretable AI systems that not only predict but also elucidate underlying material phenomena, fostering a more insightful and ethical application of AI in materials innovation.

Keywords Materials informatics, Interpretability, Explainable AI, Materials science, Explanation evaluation, Conceptual frameworks

*Correspondence:
Maria Hernandez
maria.hernandez@outlook.com
[1] Department of Materials Science and Machine Learning, Faculty of Engineering, Polytechnic University of Valencia, Valencia, Spain
[2] Department of Intelligent Materials Systems, Faculty of Engineering, University of Seville, Seville, Spain

Introduction

The field of materials science has undergone a profound transformation with the advent of artificial intelligence (AI) and machine learning (ML), collectively known as materials AI or materials informatics [1, 2]. These tools have enabled the rapid screening of vast chemical spaces, prediction of material properties, and optimization of synthesis routes, significantly shortening the traditional trial-and-error cycles that can span decades [3]. For instance, AI models have been instrumental in discovering new superconductors, catalysts for energy applications, and advanced polymers with tailored mechanical properties [4, 5]. However, as AI models grow in complexity—often employing deep neural networks (DNNs) or ensemble methods—their decision-making processes become opaque, earning them the moniker of "black boxes" [6]. This opacity raises critical concerns in materials science, where understanding the "why" behind a prediction is as

important as the prediction itself, to validate physical plausibility, guide experimental validation, and inspire new hypotheses [7].

Interpretability, broadly defined as the ability to understand and explain an AI model's inner workings in human-comprehensible terms, emerges as a pivotal requirement for trustworthy materials AI [8]. In contrast to mere accuracy, interpretability ensures that AI-driven insights align with established scientific principles, such as crystal structure-property relationships or electronic band theory [9]. The need for interpretability is amplified in materials research due to the high-stakes nature of applications, including safety-critical materials in aerospace or healthcare, and the interdisciplinary collaboration between computational scientists, experimentalists, and engineers [10]. Without clear explanations, adoption barriers persist, as stakeholders may distrust models that cannot justify predictions, potentially leading to overlooked opportunities or erroneous applications [11].

This review addresses a key gap in the literature: the conceptual underpinnings of what "explanation" means in materials AI and how it should be evaluated beyond empirical metrics. While prior reviews have generally surveyed XAI techniques [12, 13], few focus on their adaptation to material contexts, where explanations must integrate domain knowledge such as thermodynamics or quantum mechanics [14]. The objectives of this review are threefold: (1) to define "explanation" in materials AI, distinguishing it from general AI contexts; (2) to explore thematic approaches to generating explanations; and (3) to propose conceptual frameworks for their evaluation, emphasizing qualitative and interdisciplinary criteria. By limiting our scope to peer-reviewed works, we capture the rapid evolution of this nascent field and aim to provide a foundational resource for researchers seeking to integrate interpretability into materials AI workflows.

Main Text

Fundamentals of interpretability in AI and its relevance to materials science

Interpretability in AI refers to the degree to which a human can comprehend the cause of a model's decision or prediction [15]. It stands in contrast to explainability, which often implies post-hoc justifications for model behavior, though the terms are sometimes used interchangeably [16]. In general AI, interpretability is pursued to mitigate biases, ensure fairness, and comply with regulations [17]. However, in materials science, it serves a dual purpose: enhancing predictive accuracy through domain-informed constraints and facilitating scientific discovery by revealing latent patterns in material data [18].

Materials datasets are characterized by high dimensionality (e.g., atomic coordinates, electronic structures), sparsity, and noise from experimental variability [19]. Traditional ML models like random forests offer inherent interpretability via feature importance, but advanced DNNs, which excel at capturing non-linear relationships in material properties, sacrifice this for performance [20]. For example, in predicting semiconductor band gaps, a DNN might outperform simpler models but fail to explain how orbital overlaps influence the output [21]. This trade-off, often called the accuracy-interpretability dilemma, is particularly acute in materials AI, where explanations must be physically meaningful to inform synthesis or modification strategies [22].

Recent frameworks classify interpretability into intrinsic (built into the model architecture) and post-hoc (applied after training) categories [23]. Intrinsic methods, such as decision trees or linear regressions augmented with physics-based features, are favored in materials science for their alignment with interpretable descriptors such as electronegativity or lattice parameters [24]. Post-hoc methods, including SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations), approximate black-box behaviors and have been applied to interpret graph neural networks (GNNs) in molecular property prediction [25]. The relevance to materials lies in bridging the gap between data-driven predictions and theoretical models, such as density functional theory (DFT), thereby enabling hybrid approaches in which AI explanations validate or refine physical simulations [26].

Defining "Explanation" in the context of materials AI

An “explanation” in AI is a human-understandable rationale for a model’s output, but in materials AI, it must transcend generic attributions to incorporate conceptual fidelity to material phenomena [27]. We propose a multifaceted definition: an explanation is a mapping from model inputs (e.g., composition, structure) to outputs (e.g., property predictions) that elucidates causal or correlative mechanisms grounded in materials principles [28]. This differs from general AI explanations, which may prioritize user trust over scientific accuracy [29].

Types of explanations in materials AI include local (instance-specific, e.g., why a specific alloy exhibits high strength) and global (model-wide, e.g., overarching rules governing ductility) [30]. Feature-based explanations highlight input contributions, such as atomic radius in phase stability models [31], while mechanistic explanations delve into intermediate representations, akin to unveiling reaction pathways in catalysis [32]. Conceptual clarity is essential; for instance, in battery materials design, an explanation might link lithium-ion diffusion to lattice defects, providing actionable insights for doping strategies [33]. **Figure 1** provides a conceptual map of explanation types in materials AI.

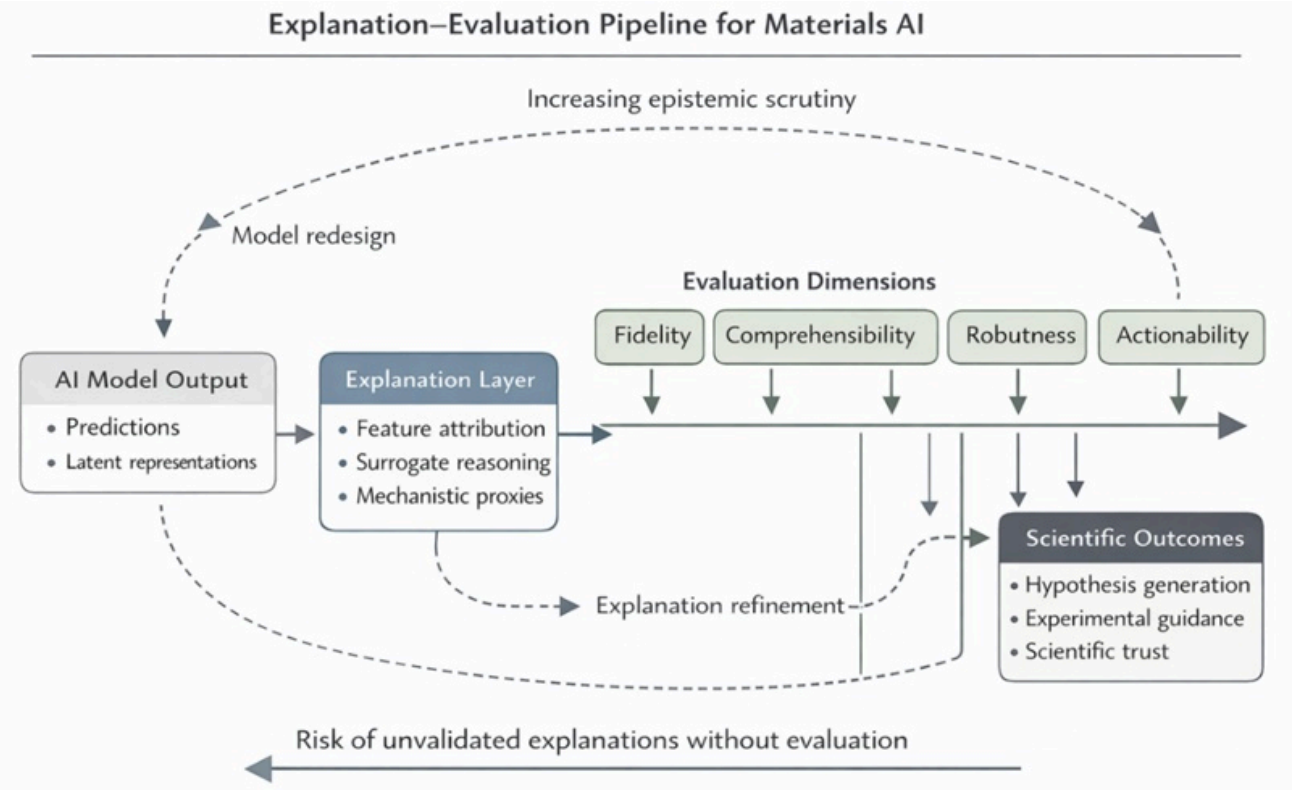

Figure 1. Conceptual landscape of explanations in materials AI

Challenges in defining explanations stem from subjectivity: what is explanatory to a computational chemist may not be to a materials engineer [12]. Recent work advocates domain-specific taxonomies, in which explanations are evaluated against benchmarks such as physical consistency or counterfactual validity (e.g., “what if” scenarios that alter material composition) [14].

Methods for generating explanations in materials AI

Generating explanations for materials-focused AI systems requires carefully adapting explainable artificial intelligence (XAI) techniques to domain-specific representations of matter. Unlike conventional tabular or image-based data, materials data are often encoded as graphs (e.g., atomic networks in crystals), strings (e.g., SMILES for molecules), or multiscale descriptors linking processing, structure, and properties [8]. As a result, explanation methods must operate not only at the level of numerical features but also at physically and chemically meaningful abstractions. **Table 1** summarizes a

taxonomy of explanation types in materials AI, organizing feature attribution, mechanistic, surrogate-based, and visualization-based approaches by scope, model dependence, typical XAI methods, representative materials tasks, and key limitations.

Table 1. Taxonomy of explanation types in materials AI

Explanation type	Scope	Model dependence	Typical XAI methods	Materials example	Key limitation
Feature attribution	Local	Model-dependent	SHAP, IG	Alloy strength	Non-causal
Mechanistic	Global	Often hybrid	Symbolic reg., PINNs	Catalysis	High complexity
Surrogate-based	Local/Global	Approximate	LIME, trees	Polymers	Fidelity loss
Visualization-based	Local	Representation-dependent	Saliency maps	Microstructures	Noise sensitivity

Explanation strategies can be broadly classified into intrinsic and post-hoc approaches. Intrinsic interpretability is achieved by designing models whose internal structure is itself human-readable. Symbolic regression exemplifies this class by directly producing explicit mathematical expressions that relate material descriptors to target properties, thereby yielding equations that can be inspected, tested, and compared with established physical laws [9]. Such approaches are particularly attractive in materials science, where explicit functional forms can be evaluated against thermodynamic or kinetic expectations. However, their applicability is often limited to relatively low-dimensional problems and may struggle with the complexity of high-throughput materials datasets.

Consequently, post-hoc explanation methods dominate applied materials AI. Attribution-based techniques—such as integrated gradients, SHAP, or attention mechanisms—assign relevance scores to input features, indicating how atomic environments, bonding motifs, or compositional features contribute to model predictions [10]. In graph neural networks (GNNs), these methods have been adapted to trace predictions back to specific atoms, bonds, or subgraphs, enabling localized interpretations in molecular and crystalline systems. While these explanations provide insight into feature importance, they remain inherently model-dependent and may not correspond to causal mechanisms.

Surrogate modeling represents another widely used explanatory strategy. In this approach, a complex black-box model (e.g., a deep neural network) is approximated locally or globally by a simpler, interpretable model such as a linear regressor or decision tree. For instance, linear surrogates have been fitted to deep learning outputs to elucidate structure–property relationships in the prediction of polymer glass transition temperatures [11]. Although surrogates can enhance transparency, they introduce approximation error and risk, obscuring non-linear interactions central to materials behavior.

Visualization-based techniques further complement explanation generation. Saliency maps and related methods highlight spatial regions or microstructural features that most strongly influence predictions, facilitating interpretability in image-based materials data such as micrographs or tomography [13]. These visual explanations are intuitively appealing to experimentalists but are sensitive to noise and resolution, raising concerns about robustness and reproducibility.

More recently, hybrid explanation approaches have emerged that explicitly integrate physical constraints. Physics-informed or constrained optimization frameworks enforce adherence to conservation laws, symmetry principles, or thermodynamic bounds, ensuring that generated explanations remain physically plausible [18]. Such hybrid methods represent a critical step toward aligning AI explanations with scientific reasoning rather than purely statistical relevance.

In practice, explanation methods have successfully supported materials discovery, for example, by identifying key descriptors governing perovskite stability or screening criteria for functional materials [20]. Nonetheless, their epistemic reliability remains contingent on data quality, representation choices, and model complexity [22]. Explanations derived from biased, sparse, or poorly characterized datasets risk reinforcing spurious correlations rather than uncovering meaningful material principles.

Conceptual frameworks for evaluating explanations

Assessing the quality of AI-generated explanations in materials science requires conceptual frameworks that extend beyond predictive accuracy. Explanations must be evaluated in terms of their scientific usefulness, epistemic validity, and practical relevance to materials research workflows [15]. To this end, a multidimensional evaluation framework is increasingly necessary.

First, fidelity concerns whether an explanation faithfully reflects the behavior of the underlying model. High-fidelity explanations accurately capture how predictions change under input perturbations, a property often assessed through sensitivity or perturbation-based tests in materials simulations [16]. Without fidelity, explanations risk becoming narrative overlays disconnected from the model's actual reasoning.

Second, comprehensibility addresses the interpretability of explanations for domain experts. An explanation that is mathematically precise but opaque to materials scientists offers limited value. Comprehensibility is commonly assessed through expert evaluation or user studies conducted in laboratory settings, examining whether explanations align with practitioners' mental models and reasoning practices [17]. **Table 2** presents conceptual evaluation criteria for explanations in materials AI, including fidelity, comprehensibility, robustness, relevance, and actionability, along with materials-specific constraints and typical failure modes.

Table 2. Conceptual evaluation criteria for explanations in materials AI

Criterion	Conceptual meaning	Materials-specific constraint	Common failure mode
Fidelity	Mirrors model behavior	Chemically valid perturbations	Narrative mismatch
Comprehensibility	Understandable to experts	Domain-aligned abstractions	Cognitive overload
Robustness	Stability under noise	Experimental uncertainty	Explanation drift
Relevance	Physical plausibility	Thermodynamics, kinetics	Spurious correlations
Actionability	Supports decisions	Synthesis feasibility	Non-actionable insights

Third, robustness evaluates the stability of explanations under noise, sampling variability, or minor perturbations in input data. This criterion is particularly critical in materials science, where experimental measurements often carry substantial uncertainty [23]. Explanations that fluctuate dramatically with small data changes undermine trust and hinder reproducibility.

Fourth, relevance measures alignment between explanations and established materials knowledge. Explanations should respect known constraints such as thermodynamic feasibility, phase stability, or mechanistic plausibility [24]. High relevance indicates that explanations are not only statistically grounded but also scientifically coherent.

Finally, actionability captures whether explanations support decision-making, such as guiding synthesis, informing experimental design, or prioritizing candidate materials. Actionability is often evaluated indirectly through downstream outcomes, including successful experimental validation or accelerated discovery cycles [28].

In practice, conceptual evaluation relies on qualitative rubrics that combine expert judgment with quantitative proxies, such as consistency of explanations across models or datasets [29]. However, the absence of standardized evaluation protocols remains a significant challenge. Emerging directions propose ontology-based and semantics-aware metrics that explicitly encode materials knowledge, offering a path toward more systematic and comparable evaluation of explanations across studies [30].

Results and Discussion

The conceptual exploration of interpretability in materials AI, as outlined in this review, underscores its pivotal role in transforming black-box models into tools for genuine scientific insight. By defining explanations as human-comprehensible mappings that integrate data-driven predictions with physical mechanisms, we have highlighted how they differ from general AI contexts, where user trust may suffice without domain-specific validity [1, 2]. This distinction is crucial in materials science, where explanations must be validated against established theories such as band theory or thermodynamics to avoid unphysical predictions [3]. However, the literature reveals persistent gaps in achieving this integration, including methodological limitations, evaluation inconsistencies, and application-specific challenges [4].

A primary challenge is the inherent trade-off between model performance and interpretability. Advanced AI models, such as deep GNNs, excel at handling high-dimensional material data, enabling breakthroughs in property prediction for complex systems, such as high-entropy alloys and 2D materials [5, 6]. Yet, their complexity often renders them opaque, limiting their utility in hypothesis generation [7]. Intrinsic interpretability methods, such as linear models with physics-informed features, offer transparency but may fail to capture non-linear interactions, as evidenced by studies on polymer chain dynamics, where simpler models underestimated glass transition temperatures [8]. Post-hoc methods, such as SHAP, have been widely adopted to dissect these models, attributing predictions to atomic or structural features in tasks such as bandgap estimation [9]. However, these approximations can be sensitive to model architecture, leading to inconsistent explanations across ensemble models [10]. In practice, this trade-off is exacerbated in low-data regimes common in materials science, where overfitting amplifies the need for robust, interpretable models [11].

The multidimensionality of explanations further complicates their application. Local explanations, focused on individual instances, are invaluable for debugging anomalous predictions, such as unexpected ductility in metals [12]. Global explanations, on the other hand, provide model-wide insights, revealing general rules like the role of electronegativity in bond strength [13]. Mechanistic explanations, which aim to uncover causal pathways, are particularly promising for materials discovery, as seen in XAI applications to reaction kinetics in catalysis [14]. Yet, achieving mechanistic depth requires fusing AI with simulation tools, such as DFT, to ensure explanations are physically plausible [15]. Challenges arise when explanations conflict with domain knowledge; for example, feature attributions in neural networks for battery electrolytes may prioritize spurious correlations over ionic mobility mechanisms [16]. This highlights the need for hybrid frameworks that constrain AI outputs to comply with physical laws, reducing the risk of erroneous insights [17].

Evaluation of explanations remains a contentious area, as conceptual frameworks must balance quantitative and qualitative criteria [18]. Fidelity, measuring how well an explanation mirrors the model's behavior, is often assessed via perturbation tests, but in materials contexts, perturbations must respect chemical validity to avoid unrealistic structures [19]. Comprehensibility, a subjective metric, demands user studies involving materials experts, as demonstrated in evaluations of visualization tools for microstructure analysis [20]. Robustness against data noise is critical, given experimental uncertainties in materials datasets, yet many XAI methods lack built-in uncertainty quantification [21]. Relevance to domain knowledge, such as thermodynamic stability, is essential to prevent explanations that are mathematically accurate but physically meaningless [22]. Actionability, the ability to guide experiments, is the ultimate test; successful examples include XAI-driven doping strategies in semiconductors that led to validated prototypes [23]. However, the lack of standardized benchmarks hinders progress, as current evaluations are often ad hoc and non-comparable across studies [24].

From an interdisciplinary perspective, interpretability facilitates collaboration between AI specialists and materials scientists, fostering trust and adoption [25]. In high-stakes applications, like nuclear materials or biomedical implants, explanations support regulatory compliance and risk assessment [26]. Ethical considerations, including bias in training data from underrepresented materials classes, can be addressed through interpretable models that expose such biases [27]. However, over-reliance on interpretability may constrain model innovation, as enforcing transparency can reduce flexibility in capturing emergent phenomena [28]. Computational efficiency is another barrier; generating explanations for large-scale screenings, such as in the Materials Genome Initiative, requires optimized algorithms to avoid bottlenecks [29].

Looking at specific subfields, interpretability has shown varying maturity. In inorganic materials, XAI has elucidated crystal stability rules, aiding inverse design [30]. For organics and polymers, feature-based explanations have highlighted molecular motifs that influence properties such as solubility [31]. In nanomaterials, visualization methods have revealed size-dependent effects, but scaling to multiscale models remains challenging [32]. Across these, a common theme is the need for ontology-based evaluations that incorporate materials semantics, ensuring explanations are not only accurate but also semantically rich [33].

In summary, while significant advances have been made, the field must address these challenges to realize the full potential of interpretable materials AI. By prioritizing domain-aligned explanations and rigorous evaluation, researchers can enhance not only prediction accuracy but also the pace of materials innovation.

Conclusion

This narrative review has synthesized the conceptual landscape of interpretability in materials AI, clarifying what constitutes an “explanation” and advocating for evaluation frameworks that encompass fidelity, comprehensibility, robustness, relevance, and actionability. Through a thematic examination of generation methods and their applications, we have demonstrated how interpretability bridges the gap between computational predictions and physical understanding, enabling advances across diverse areas, including energy materials and structural alloys. The reviewed works illustrate rapid evolution but also reveal opportunities for more integrated, user-centric approaches.

Future research should prioritize the development of standardized, materials-specific benchmarks for XAI that incorporate multiscale data and uncertainty propagation to reflect real-world scenarios. Enhancing causal inference in explanations could revolutionize inverse design, allowing AI to suggest novel compositions with justified mechanisms. Efficiency optimizations, such as model-agnostic fast attribution methods, will be key to scaling interpretability to high-throughput workflows. Interdisciplinary efforts, including collaborations with cognitive scientists to develop better comprehensibility metrics and ethical guidelines for bias-transparent AI, will ensure equitable progress. Ultimately, by embedding interpretability at the core of materials AI, the field can foster trustworthy systems that accelerate discovery and contribute to sustainable technological advancements.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 28 Mar 2024 Revised: 06 May 2024 Accepted: 11 Jun 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(1):71-103.

Batra R, Song L, Ramprasad R. Emerging materials intelligence ecosystems propelled by machine learning. Nat Rev Mater. 2020;6(8):655-78.

Zhong X, Gallagher B, Liu S, Kailkhura B, Hiszpanski A, Han TYJ. Explainable machine learning in materials science. npj Comput Mater. 2022;8(1):204.

Suh C, Fare C, Warren JA, Pyzer-Knapp EO. Evolving the materials genome: How machine learning is fueling the next generation of materials discovery. Annu Rev Mater Res. 2020;50:1-25.

Gupta V, Choudhary K, Tavazza F, Campbell C, Liao WK, Choudhary A, et al. Cross-property deep transfer learning framework for enhanced predictive analytics on small materials data. Nat Commun. 2021;12(1):6595.

Liu T, Barnard AS. The emergent role of explainable artificial intelligence in the materials sciences. Cell Rep Phys Sci. 2023;4(10):101196.

Palizhati A, Peczak N, Adcock W, Zuo Y, Deng Z, Ong SP. Interpretable and explainable machine learning for materials science and chemistry. Accounts Mater Res. 2022;3(6):597-607.

Szymanski NJ, Bartel CJ, Zeng Y, Tu Q, Gaultois MW, Johansen JM, et al. Accelerating materials discovery using artificial intelligence, high performance computing and robotics. npj Comput Mater. 2022;8(1):84.

Li S, Barnard AS. Inverse design of MXenes for high-capacity energy storage materials using multi-target machine learning. Chem Mater. 2022;34(11):4964-74.

Li S, Barnard AS. Multi-target neural network predictions of MXenes as high-capacity energy storage materials in a Rashomon set. Cell Rep Phys Sci. 2023;4(11):101633.

Gupta V, Liao W, Choudhary A, Agrawal A. Evolution of artificial intelligence for application in contemporary materials science. MRS Commun. 2023;13(5):754-63.

Bhakte A, Pakkiriswamy V, Srinivasan R. An explainable artificial intelligence based approach for interpretation of fault classification results from deep neural networks. Chem Eng Sci. 2022;250:117373.

Bhakte A, Chakane M, Srinivasan R. Alarm-based explanations of process monitoring results from deep neural networks. Comput Chem Eng. 2023;179:108442.

Naser MZ. An engineer's guide to eXplainable Artificial Intelligence and Interpretable Machine Learning: Navigating causality, forced goodness, and the false perception of inference. Autom Constr. 2021;129:103821.

Zhong Y, Tiwari A, Yamaguchi H, Lakhtakia A, Bukkapatnam STS. Identifying the influence of surface texture waveforms on colors of polished surfaces using an explainable AI approach. IISE Trans. 2023;55(7):731-45.

Karthikeyan A, Tiwari A, Zhong Y, Bukkapatnam STS. Explainable AI-infused ultrasonic inspection for internal defect detection. CIRP Ann. 2022;71(1):449-52.

Yoo J, Cho Y, Jeong B, Choi SH, Kim KK, Lim SC, et al. Explainable Artificial Intelligence Approach to Identify the Origin of Phonon-Assisted Emission in WSe2 Monolayer. Adv Intell Syst. 2023;5(7):2200463.

Mankodiya H, Jadav D, Gupta R, Tanwar S. OD-XAI: Explainable AI-based semantic object detection for autonomous vehicles. Appl Sci. 2022;12(11):5310.

Mankodiya H, Obaidat MS, Gupta R, Tanwar S. XAI-AV: Explainable artificial intelligence for trust management in autonomous vehicles. Int Conf Commun Comput Cybersecurity Inform. 2021;1-6.

Gupta V, Choudhary K, Campbell C, Liao WK, Choudhary A, Agrawal A. MPPredictor: An Artificial Intelligence-Driven Web Tool for Composition-Based Material Property Prediction. J Chem Inf Model. 2023;63(7):1865-71.

Yoo J, Cho Y, Kim DH, Kim J, Lee TG, Lee SM, et al. Unraveling the role of Raman modes in evaluating the degree of reduction in graphene oxide via explainable artificial intelligence. Nano Today. 2024;57:102366.

Naser MZ, Alavi AH. Error metrics and performance fitness indicators for artificial intelligence and machine learning in engineering and sciences. Arch Struct Constr. 2023;3(4):499-517.

Naser MZ, Tapeh ATG. Artificial intelligence, machine learning, and deep learning in structural engineering: a scientometrics review of trends and best practices. Arch Comput Methods Eng. 2023;30(1):115-59.

Oommen V, Shukla K, Goswami S, Dingreville R, Karniadakis GE. Learning two-phase microstructure evolution using neural operators and autoencoder architectures. npj Comput Mater. 2022;8(1):190.

Zhong Y, Jatav A, Afrin K, Shivaram T, Bukkapatnam STS. Enhanced SpO2 estimation using explainable machine learning and neck photoplethysmography. Artif Intell Med. 2023;145:102685.

Zhong Y, Bhattacharya A, Bukkapatnam S. EBLIME: Enhanced Bayesian Local Interpretable Model-agnostic Explanations. arXiv preprint arXiv:2305.00213.2023.

Jain S, Kirk R, Lubana ES, Dick RP, Tanaka H, Grefenstette E, et al. Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks. Int Conf Learn Represent. 2023;1-17.

Liu T, Tho ZY, Barnard AS. Understanding the importance of individual samples and their effects on materials data using explainable artificial intelligence. Digital Discovery. 2024;3(2):422-35.

Naser MZ, Alavi AH. Insights into performance fitness and error metrics for machine learning. arXiv preprint arXiv:2006.00887. 2020.

Bhakte A, Kumawat PK, Srinivasan R. Explainable AI methodology for understanding fault detection results during multi-mode operations. Chem Eng Sci. 2024;299:120493.

Wang K, Gupta V, Lee CS, Mao Y, Kilic MNT, Li Y, et al. XElemNet: towards explainable AI for deep neural networks in materials science. Sci Rep. 2024;14(1):25178.

Li Z, Li S, Birbilis N. A machine learning-driven framework for the property prediction and generative design of multiple principal element alloys. Mater Today Commun. 2024;38:107940.

Suh HC, Yoo J, Yeo K, Kim DH, Won YS, Kim T, et al. Probing nanoscale structural perturbation in a WS2 monolayer via explainable artificial intelligence. Appl Phys Rev. 2025;12(2):021304.