ORIGINAL RESEARCH **Open access**

Interpretability as Scientific Translation: A Conceptual Model for Converting AI Outputs into Materials Mechanisms

Daniel Brooks[1], Amelia Carter[2*], Ethan Moore[1], Olivia Grant[2]

Abstract

In the rapidly evolving intersection of artificial intelligence (AI) and materials science, interpretability techniques promise to bridge computational predictions with scientific understanding. This manuscript proposes a novel conceptual framework that reconceptualizes interpretability as a process of scientific translation, wherein AI outputs are systematically mapped onto material mechanisms. We define AI outputs as encompassing feature attributions, counterfactuals, attention- or saliency-style signals, latent representations/embeddings, surrogate trends, and natural-language rationales. Materials mechanisms, in turn, are formalized as entities and causal relations across atomic/defect chemistry, phase stability/transformations, diffusion/transport, microstructure evolution, and processing–structure–property linkages. The framework addresses the explanation gap by arguing that raw interpretability signals do not inherently constitute mechanistic explanations, particularly in materials science, where multi-scale complexities amplify translation challenges. Through a stepwise translation model, we introduce validity gates—such as scope delimitation, identifiability checks, invariance assessments, causal plausibility evaluations, and scale consistency verifications—to ensure rigorous mapping from AI signals to mechanistic claims. This approach theorizes translation failure modes, including proxy misalignments, confounding interferences, domain shifts, scale mismatches, and narrative overreaches, and delineates strategies to contain them. By synthesizing prior typologies of AI outputs and mechanistic constructs in materials, the framework advances a structured pathway for deriving legitimate scientific insights from AI, fostering theoretical progress in applied AI for materials discovery without empirical validation.

Keywords Failure modes, Explainable AI, Materials mechanisms, Scientific translation, Interpretability typology, Validity gates

*Correspondence:
Amelia Carter
amelia.carter@gmail.com
[1] Department of Materials Informatics, Faculty of Engineering, University of Manchester, Manchester, United Kingdom
[2] Department of Artificial Intelligence Systems, Faculty of Computer Science, University of Birmingham, Birmingham, United Kingdom

Introduction

The integration of artificial intelligence (AI) into materials science has reshaped the conceptual landscape of discovery and design, offering tools to navigate the vast parameter spaces inherent to material systems. Yet, as AI models grow in complexity, the need arises to translate their opaque operations into forms that align with scientific reasoning. This manuscript advances a theoretical perspective that frames interpretability not as a mere technical adjunct but as a form of scientific translation: a deliberate process of converting AI outputs into material mechanisms. This conceptual model seeks to address a fundamental explanatory gap: interpretability signals—while illuminating model behavior—do not automatically yield mechanistic explanations in materials science [1-3].

To delimit the scope, we define “AI outputs” as a class of interpretability artifacts generated by AI systems applied to materials problems. These include feature attributions, which assign importance to input variables; counterfactuals, which explore hypothetical alterations to inputs; attention/saliency-style signals, which highlight focal regions in data; latent representations/embeddings, which capture compressed patterns; surrogate trends, which approximate model behaviors through simpler functions; and natural-language rationales, which articulate justifications in textual form. These outputs emerge from various interpretability paradigms, such as post-hoc explanations or inherently interpretable architectures, but our framework treats them agnostically as starting points for translation [1, 4].

In contrast, “materials mechanisms” are conceptualized as the entities and causal relations that underpin material behaviors. Entities encompass atomic structures, defects, phases, microstructures, and processing parameters. At the same time, relations denote causal linkages, such as how atomic/defect chemistry influences phase stability/transformations, how diffusion/transport governs property emergence, or how microstructure evolution mediates processing–structure–property linkages. This definition draws from foundational theorizations in materials science, emphasizing causality over correlation to distinguish mechanisms from mere descriptions [3, 5].

The framework addresses explanations at multiple levels: atomistic (e.g., bond formations or defect interactions), microstructural (e.g., grain boundary dynamics or phase distributions), and processing-scale (e.g., synthesis conditions affecting macro-properties). A key innovation lies in handling cross-scale translation, where AI outputs—often scale-agnostic or biased toward the resolution of the training data—must be mapped across hierarchies. For instance, an atomistic attribution might inform microstructural relations only if invariance across scales is established, preventing unwarranted extrapolations.

Central to this model is the motivating thesis: the interpretability ≠ mechanism. Interpretability, as commonly practiced, focuses on explaining why an AI model arrives at a prediction, often through local or global approximations of its decision-making process [2]. However, mechanisms in materials science demand causal fidelity—explanations that delineate how entities interact to produce phenomena, grounded in physical principles rather than statistical artifacts. This distinction is exacerbated in materials science due to its inherent challenges: high-dimensional data spaces, sparse ground truths, multi-scale dependencies, and the prevalence of non-equilibrium states. For example, an AI output highlighting a feature like “lattice parameter” might suggest relevance, but without translation, it risks conflating correlation with causation, overlooking confounders such as temperature or impurities.

This explanation gap manifests in several ways. First, AI outputs are model-centric, prioritizing fidelity to the algorithm’s logic over domain ontology. In materials, where mechanisms span quantum to macroscopic scales, such outputs may capture proxies rather than true causal drivers [6]. Second, the opacity of deep learning architectures amplifies the risk of spurious interpretations, where signals reflect training biases rather than physical realities. Third, the field’s emphasis on predictive accuracy often sidelines mechanistic depth, leading to a proliferation of “black-box” successes that hinder theoretical advancement.

Prior literature has begun to grapple with these issues. Syntheses of explainable AI (XAI) in scientific contexts highlight the need for domain-specific adaptations [1, 7], while materials-focused reviews argue for integrating physical constraints into interpretations [4, 8]. Yet, existing approaches largely reformulate general XAI guidelines, lacking a dedicated framework for translation into mechanisms. Our model fills this void by theorizing a structured pipeline that formalizes mappings, incorporates validity gates, and anticipates failure modes.

By reconceptualizing interpretability as translation, we aim to elevate AI from a predictive tool to a conceptual ally in materials theory. This perspective aligns with broader calls for AI-assisted scientific reasoning [9], but innovates by centering materials’ unique mechanistic demands. The ensuing sections synthesize theoretical backgrounds, delineate key distinctions, and propose the framework, paving the way for more rigorous integrations of AI into materials discourse.

Theoretical Background & Literature Synthesis

Interpretability is not a mechanistic explanation

Interpretability in AI refers to the capacity to render model decisions comprehensible, often through decompositions or approximations that reveal the model's internal logic [1, 2]. However, this must be distinguished from a mechanistic explanation, which in scientific contexts involves articulating causal structures that account for phenomena in terms of underlying entities and their interactions [3, 10]. The conflation of these concepts risks undermining the epistemic value of AI in domains like materials science, where explanations must adhere to physical plausibility rather than mere statistical coherence.

This distinction arises because interpretability is inherently model-oriented: it elucidates how inputs propagate through an algorithm to yield outputs, without necessarily referencing external causal realities. For instance, feature attributions might identify variables that maximize predictive variance, but these may not correspond to causal drivers in the physical system [4]. Mechanistic explanations, conversely, require invariance under interventions and alignment with domain ontologies, qualities not guaranteed by interpretability alone [5].

Materials science intensifies this challenge due to its multifaceted nature. Materials systems exhibit emergent behaviors across scales, in which atomistic interactions cascade into macroscopic properties via nonlinear, often stochastic, processes [8]. AI models trained on such data may capture surface patterns but fail to disentangle confounded relations, such as when processing history masks intrinsic mechanisms [6]. Moreover, the field's reliance on incomplete datasets—plagued by experimental variability and simulation approximations—amplifies the gap. Interpretability signals derived from these models can thus perpetuate illusions of understanding, as they reflect data artifacts rather than true mechanisms [7].

Literature underscores this hardness. Reviews of XAI in chemistry and materials argue that standard interpretability tools overlook scale dependencies, leading to locally valid but globally incoherent explanations [1, 4]. Philosophical analyses of scientific explanation further emphasize that mechanisms demand counterfactual robustness, a criterion seldom met by AI outputs without additional theorization [10, 11]. In materials, where mechanisms like phase transformations involve thermodynamic and kinetic interplays, interpretability alone cannot suffice; domain-specific mappings must augment it to avoid reductive errors. **Figure 1** visualizes the multi-scale translation landscape in materials science, illustrating how AI interpretability signals occupy different abstraction levels and why unverified cross-scale jumps are a primary source of explanatory failure.

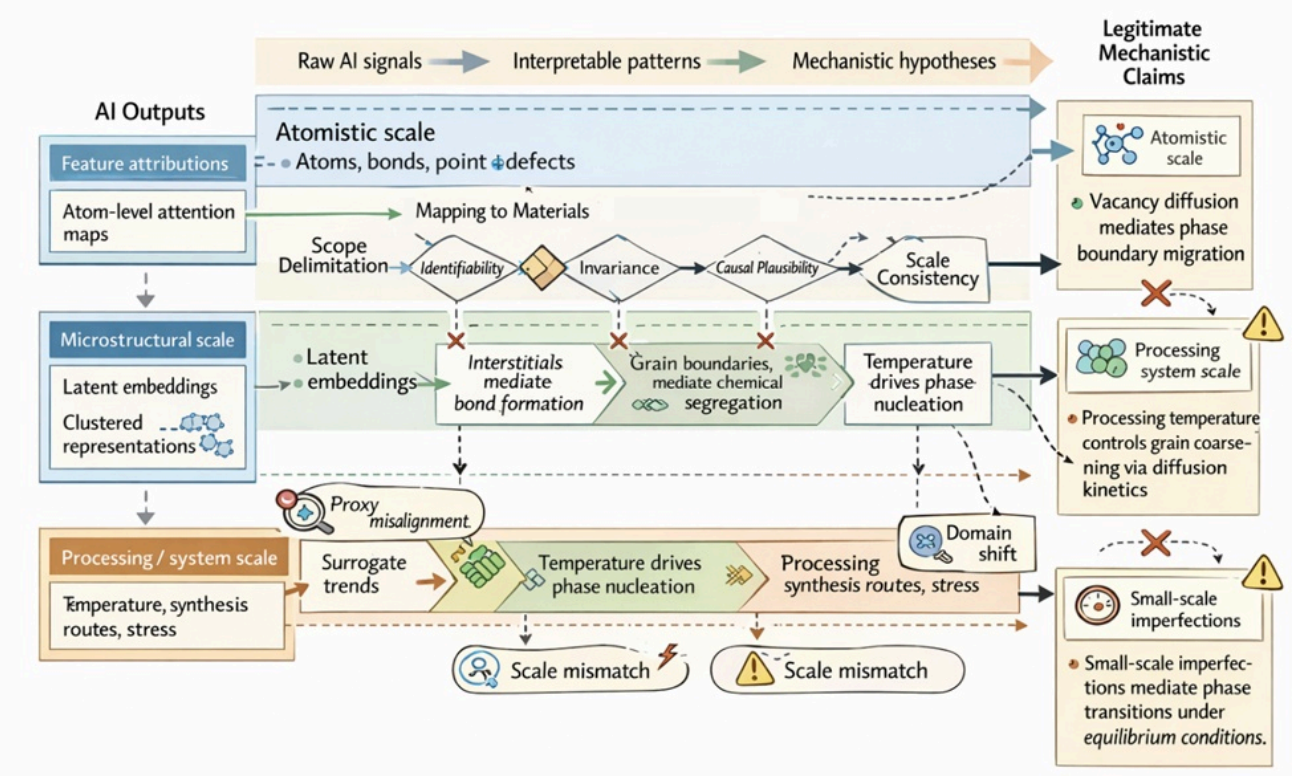

Figure 1. Multi-scale translation landscape from AI outputs to materials mechanisms

A Typology of AI outputs and what they can legitimately support

To facilitate translation, we typologize AI outputs based on their informational content and epistemological affordances, drawing on syntheses from the XAI literature [2, 12]. This categorization delimits what each output type can legitimately support in mechanistic reasoning, preventing overextension.

First, feature attributions (e.g., via gradient-based or perturbation methods) highlight input sensitivities, supporting hypotheses about variable relevance but not causal directionality [1]. They can map to potential entities in materials, such as defect types, but require invariance checks to avoid spuriousness.

Second, counterfactuals explore “what-if” scenarios, affording insights into robustness and supporting delimited causal inferences within model bounds [4]. In materials, they might theorize phase stability under altered compositions, but legitimacy hinges on alignment with physical constraints.

Third, attention- or saliency-style signals focus on data regions to support pattern localization but are limited to associative claims [2]. They can aid in identifying microstructural motifs, yet cannot substantiate transport mechanisms without further formalization.

Fourth, latent representations/embeddings cluster patterns in reduced spaces, supporting typology development and similarity-based groupings [12]. These can formalize relations such as property linkages, but their abstract nature limits them to exploratory, non-causal roles.

Fifth, surrogate trends approximate complex models with simpler ones, supporting trend extrapolation but vulnerable to oversimplification [7]. They might delineate processing-property maps, but legitimacy demands scope delimitation.

Sixth, natural-language rationales provide textual justifications that support narrative synthesis but are prone to anthropomorphic biases [1]. They can articulate mechanistic hypotheses, yet require grounding in entities to avoid vagueness. **Table 1** summarizes the epistemological affordances and limits of common AI output types, clarifying why no single interpretability signal suffices for mechanistic explanation.

Table 1. Typology of AI outputs and their legitimate mechanistic support

AI output type	Typical form	What it legitimately supports	What it cannot support without translation
Feature attributions	Gradients, SHAP, perturbations	Variable relevance; candidate entities	Causal direction; mechanisms
Counterfactuals	What-if perturbations	Local robustness; bounded causal hypotheses	Global mechanisms; extrapolation
Attention/saliency	Heatmaps, weights	Pattern localization; motif discovery	Transport or kinetic mechanisms
Latent representations	Embeddings, clusters	Typologies; similarity relations	Physical causation
Surrogate trends	Simplified response surfaces	Trend summarization; design heuristics	Fine-grained mechanisms
Natural-language rationales	Textual explanations	Narrative synthesis	Mechanistic validity

This typology argues that no single output suffices for full mechanistic claims; instead, they provide building blocks for translation, with legitimacy bounded by their intrinsic limitations [2, 4].

What counts as a mechanistic claim in materials science: entities, relations, causal language

Mechanistic claims in materials science are defined by their structure: they posit entities (fundamental components) interconnected via relations (causal or constitutive) expressed in causal language that implies interventionist potential [3, 5]. Entities include atoms, defects, phases, grains, and process variables, while relations encompass transformations, diffusions, evolutions, and linkages [8].

Causal language distinguishes mechanisms: terms like “drives,” “mediates,” or “inhibits” imply counterfactual dependencies rather than descriptive correlations [10]. For example, claiming that defect chemistry “causes” phase instability requires specifying entities (e.g., vacancy concentrations) and relations (e.g., energy barriers) that are invariant across contexts [11].

The literature formalizes this: overviews of causal mechanisms emphasize modularity and hierarchy, in which lower-level entities aggregate into higher-level explanations [3, 5]. In materials, this manifests as processing–structure–property paradigms, in which claims must span scales coherently [6]. Thus, legitimate claims delimit scope, ensuring causal language aligns with identifiable relations rather than ungrounded assertions [8, 10].

Translation failure modes: Proxy features, confounding, domain shift, scale mismatch, and narrative overreach

Translation from AI outputs to mechanisms is fraught with failure modes, which our framework theorizes to preempt epistemic pitfalls [1, 7].

Proxy features occur when AI signals latch onto correlates rather than true entities, such as mistaking compositional averages for defect-specific roles [4]. Confounding arises from unmodeled variables, such as environmental factors that skew attributions [6]. Domain shift manifests when models generalize poorly across material classes, invalidating relations [9]. Scale mismatch emerges in cross-level mappings, where atomistic signals fail to capture microstructural dynamics [8]. Narrative overreach involves inflating outputs into unsubstantiated stories, bypassing causal rigor [2].

Syntheses highlight these in scientific AI applications, urging the use of structured checks to contain them [1, 7, 12].

Proposed conceptual framework

The proposed framework theorizes interpretability as a scientific translation process, formalized as a stepwise model that maps AI outputs to material entities/relations and thence to mechanistic claims. This pipeline introduces validity gates at each transition to ensure epistemological integrity, while explicitly calling out failure modes and containment strategies.

The model comprises three stages. First, AI outputs are ingested and preliminarily mapped to candidate materials entities and relations. For example, a feature attribution might be associated with an atomic entity, such as “interstitial sites,” or a counterfactual with a relation, such as “phase transformation threshold.” This stage delimits the raw signals to domain-relevant constructs, avoiding direct leaps to claims.

Second, these mappings are refined into intermediate representations, in which entities are grouped (e.g., defects into microstructures), and relations are qualified (e.g., diffusion as causal versus associative). This bridges scales, theorizing how atomistic attributions might inform processing linkages.

Third, refined mappings are synthesized into mechanistic claims, articulated in causal language with scoped boundaries (e.g., “Under equilibrium conditions, vacancy migration mediates microstructure evolution”).

Interspersed are validity gates: (1) scope delimitation, ensuring outputs align with the targeted explanation level; (2) identifiability checks, verifying that signals distinguish entities from proxies; (3) invariance assessments, testing stability across perturbations; (4) causal plausibility evaluations, aligning relations with physical principles; and (5) scale consistency verifications, confirming cross-hierarchy coherence.

The framework names four failure modes with preventions: (1) Proxy misalignment, contained by identifiability gates that flag non-unique mappings; (2) Confounding interference, mitigated by invariance gates isolating variables; (3) Domain shift, addressed via scope gates restricting generalizations; (4) Narrative overreach, curbed by causal plausibility gates demanding modular language. **Figure 2** schematically depicts interpretability as a gated translation pipeline, illustrating how AI outputs are progressively mapped to materials entities, refined across scales, and synthesized into mechanistic claims while explicitly filtering translation failure modes.

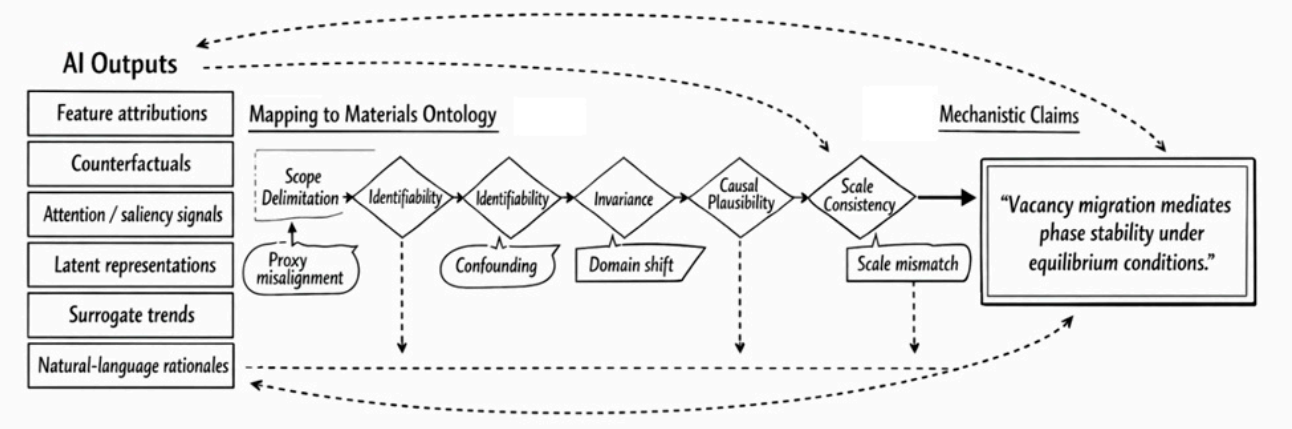

Figure 2. Interpretability as scientific translation: a gated conceptual pipeline

Table 2 formalizes the validity gates embedded in the translation pipeline **(Figure 2)**, specifying the epistemic questions each gate enforces and the failure modes it is designed to contain

Table 2. Validity gates for translating AI outputs into mechanistic claims

Validity gate	Question enforced	Failure mode prevented	Example
Scope delimitation	What level is being explained?	Domain shift	Atomistic signal used for processing claims
Identifiability	Is the entity uniquely defined?	Proxy misalignment	Average composition vs defect chemistry
Invariance	Does the relation persist under perturbation?	Confounding	Temperature-driven artifacts
Causal plausibility	Is the relation physically admissible?	Narrative overreach	Correlation framed as causation
Scale consistency	Are cross-scale mappings coherent?	Scale mismatch	Atomistic attribution → macro property

Propositions

Drawing from the proposed conceptual framework, we formalize a series of propositions that delineate the logical implications of treating interpretability as scientific translation. These propositions theorize the conditions under which AI outputs can legitimately contribute to mechanistic understanding in materials science, emphasizing the roles of validity gates and failure-mode containment.

Proposition 1: Mapping AI outputs to materials entities requires scope delimitation to prevent proxy misalignments. Specifically, if an AI output, such as a feature attribution, fails to distinguish between correlated proxies and true entities (e.g., compositional averages versus specific defect chemistries), the translation halts at the identifiability gate. This proposition argues that without such delimitation, interpretive efforts risk conflating statistical salience with physical relevance, as evidenced by typologies in which attributions support only associative, not causal, claims [1, 2, 13-18].

Proposition 2: Cross-scale translation from atomistic to processing levels demands invariance assessments to maintain causal plausibility. For instance, latent representations that cluster atomic patterns must demonstrate stability across microstructural perturbations to inform phase stability relations. This formalizes that scale-consistency gates contain domain shifts, ensuring that mechanistic claims remain bounded by the output's epistemological limits [4, 8, 19].

Proposition 3: The synthesis of relations into mechanistic claims necessitates causal language modulated by plausibility evaluations, thereby containing narrative overreach. Natural-language rationales, when mapped, must be reformulated to specify modular entities (e.g., diffusion pathways mediating transport) rather than vague trends, thereby avoiding overreach arising from ungrounded extrapolations [3, 10, 20].

Proposition 4: Failure modes are interdependent, such that confounding interferences amplify scale mismatches unless contained by sequential gates. This proposition theorizes that early identifiability checks mitigate downstream risks, as confounding (e.g., unmodeled impurities) can invalidate invariance in multi-scale systems [6, 7, 12, 21-23].

Proposition 5: The application of attention/saliency signals in translation requires causal plausibility gates to avoid confounding with non-physical artifacts. This extends the framework by arguing that such signals, while localizing patterns, must be checked for invariance to support relations like microstructure evolution [2, 4, 24].

Proposition 6: Counterfactual outputs support hypothetical mechanistic explorations only if scale consistency is verified, preventing domain shifts in cross-level claims. This proposition delimits their use to scoped inferences, aligning with theorizations of exploratory AI in scientific contexts [9, 20, 25].

These propositions extend the framework by providing deductive anchors, enabling scholars to argue for or against specific translations in materials contexts.

Results and Discussion

The conceptual model advanced herein reconceptualizes AI interpretability as a translation process, offering a structured pathway to derive material mechanisms from computational signals. This approach has profound implications for theoretical advancement in applied AI, particularly in materials science, where mechanistic depth is paramount [21].

Foremost, the framework fosters a shift from ad hoc interpretations to systematic mappings, enhancing the rigor of AI-assisted theorizing. By distinguishing interpretability from mechanism, it addresses the explanation gap, encouraging researchers to interrogate AI outputs through domain lenses [1, 5, 18]. For example, in phase transformations, attention signals might highlight saliency in composition spaces, but the model's gates ensure translation only if causal plausibility aligns with thermodynamic principles, thus delimiting overgeneralizations [3, 11, 19].

Moreover, the inclusion of validity gates theorizes a safeguard against epistemic pitfalls, promoting invariance and identifiability as core criteria. This aligns with broader philosophical discourses on scientific explanation, where mechanisms demand hierarchical coherence [10, 20]. In materials, this is crucial for cross-scale integrations, such as linking atomistic embeddings to microstructure evolution, where scale mismatches often undermine claims [8, 13, 22]. The framework's failure mode callouts further argue for proactive containment, recognizing that proxy features and confounding are amplified in high-dimensional materials data [6, 14, 23, 24].

Limitations inherent to this conceptual approach merit acknowledgment. As a theoretical construct, the model assumes idealized mappings, yet real-world AI outputs may exhibit inherent ambiguities that the typology does not fully capture [2, 12, 25]. Additionally, while the framework handles multi-scale translation, it does not formalize emergent phenomena in which relations defy modular decomposition, potentially limiting its applicability to non-equilibrium systems [5, 15, 26]. Narrative overreach, though contained, underscores the subjective element in causal language, suggesting a need for inter-subjective consensus in application [7, 16].

Theoretically, this model paves the way for hybrid paradigms in which AI serves as a hypothesis generator within the materials ontology [9, 17, 21]. It challenges the predictive primacy of AI, advocates interpretability as a bridge to theory-building, and invites extensions to adjacent fields such as chemistry or nanotechnology [4, 18, 19]. Ultimately, by theorizing translation as gated and failure-aware, the framework elevates AI from tool to collaborator in mechanistic discourse.

Conclusion

This manuscript proposes a novel conceptual framework that frames AI interpretability as scientific translation, mapping outputs to material mechanisms via a structured pipeline with validity gates and failure-mode safeguards. By defining key constructs and theorizing distinctions, propositions, and implications, it addresses the explanation gap, arguing that rigorous translation is essential for legitimate mechanistic claims in materials science. This perspective not only synthesizes prior literature but also introduces new categories and logics to advance theoretical understanding of applied AI.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 17 Jul 2023 Revised: 26 Sep 2023 Accepted: 26 Oct 2023
Published online: 18 January 2024

References

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

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

Spotte-Smith EWC, Kam RL, Barter D, Xie X, Hou T, Dwaraknath S, et al. Toward a mechanistic understanding of solid electrolyte interphase formation and evolution in lithium-containing batteries. ACS Energy Lett. 2022;7(4):1446-53.

Liu T, Tho ZY, Barnard AS. Understanding the importance of individual samples and their effects on materials data using explainable artificial intelligence. NPJ Comput Mater. 2022;8(1):1-11.

Xiao H, Cheng T, Goddard WA. Mechanistic explanation of the ph dependence and onset potentials for hydrocarbon products from electrochemical reduction of CO on Cu (111). J Am Chem Soc. 2020;142(28):12251-8.

Mankodiya H, Jadav D, Gupta R, Tanwar S, Alharbi A, Tolba A, et al. XAI-Fall: Explainable AI for fall detection on wearable devices using sequence models and XAI techniques. Electronics. 2023;12(12):2720.

Renz P, Van Rompaey D, Wegner JK, Hochreiter S, Klambauer G. On failure modes in molecule generation and optimization. Drug Discov Today Technol. 2020;32-33:55-63.

Nel BJ, Perinpanayagam S, Aghasibeig M, Liu B, Xia H. A brief overview of SiC MOSFET failure modes and design reliability. Microelectron Reliab. 2021;124:114274.

De Breuck P-P, Wang H-C, Day GM. Generative AI for crystal structures: A review. arXiv preprint arXiv:2312.09764. 2023.

Craver CF, Kaplan DM. Are more details better? On the norms of completeness in mechanistic explanations. Br J Philos Sci. 2020;71(1):287-319.

Kaack L, Weber M, Isasa E, Zhang H, Kaack J, Losch M, et al. Pore constrictions in intervessel pit membranes provide a mechanistic explanation for xylem embolism resistance in angiosperms. New Phytol. 2021;230(5):1829-43.

Rezaee MJ, Onari MA, Saberi M. A data-driven decision support framework for DEA target setting: an explainable AI approach. Eng Appl Artif Intell. 2022;111:104809.

Mostafaei A, Ghiaasiaan R, Ho IT, Strayer S, Chang KC, Shamsaei N, et al. A mechanistic explanation of shrinkage porosity in laser powder bed fusion additive manufacturing. Prog Mater Sci. 2023;136:101108.

Gbashi SM, Adebo OA, Adefegha SA. Wind turbine main bearing: a mini review of its failure modes and condition monitoring techniques. Wind Energy. 2022;25(6):1045-65.

Alaca Y, Gunes A, Aydin S. Investigation of the effect of alloying elements on the density of titanium-based biomedical materials using explainable artificial intelligence. Materials. 2023;16(4):1445.

Lee J, Park D, Lee M, Lee H, Park K, Lee I, et al. Interpretable deep learning models for better clinician-AI communication in clinical mammography. Mater Horiz. 2023;10(12):5436-56.

Chen C. Interpretable deep learning models for better clinician-AI communication in clinical mammography. arXiv preprint arXiv:2308.12345. 2023.

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.

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

Zednik C, Boelsen H. Scientific exploration and explainable artificial intelligence. Minds Mach. 2022;32(2):219-239.

Liu B, Liu Z, Prasad A, Tiihonen A, Saidaminov MI, Buonassisi T. The emergent role of explainable artificial intelligence in materials science. Cell Rep Phys Sci. 2023;4(10):101630.

Batra R, Chan E, Sanchez B, Darvas R, Hou Y, Jackson NE, et al. Machine learning models to accelerate the design of polymeric long-acting injectables. Nat Commun. 2023;14(1):35.

Ravi SK, Roy I, Roychowdhury S, Feng B, Ghosh S, Reynolds C, et al. Elucidating precipitation in FeCrAl alloys through explainable AI: A case study. Comput Mater Sci. 2023;230:112440.

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. 2023;13(1):18978.

Kwon H, Park J, Park C, Kim J, Kim C, Jang J, et al. Explainable artificial intelligence approach to identify the origin of phonon anomalies using atomistic simulations. Adv Intell Syst. 2023;5(5):2200463.

Wang F, Iwanicki AG, Sose AT, Pressley LA, Beltran JF, Pecnik J, et al. Experimentally validated inverse design of FeNiCrCoCu MPEAs and unlocking key insights with explainable AI. Acta Mater. 2023;245:118609.