ORIGINAL RESEARCH **Open access**

The Microstructure–Property "Explanation Gap": A Conceptual Anatomy of Why AI Explanations Often Fail

Ahmed Mansour[1*], Omar Saeed[1], Lina Hassan[2]

Abstract

Artificial intelligence (AI) has become increasingly effective at predicting material properties from microstructure-informed representations, enabling rapid screening and accelerated decision-making. Yet, the "explanations" attached to these predictive systems frequently fail to support the kind of understanding required in microstructure–property science—namely, transferable mechanisms, intervention-relevant guidance, and defensible generalization under realistic shifts in processing, measurement, and operating regimes. This conceptual paper argues that explanation failure in materials AI is often structural rather than incidental: many popular explanation toolkits are optimized for interpreting model behavior rather than for producing scientifically legitimate accounts of why a microstructure yields a property outcome. We define the microstructure–property explanation gap as the persistent mismatch between what explainability tools can formally justify and what materials reasoning demands for action. To anatomize this gap, we identify four recurring causes: representational non-identifiability, confounding by processing history, multi-scale emergence, and instability under distribution shift. Building on this anatomy, we propose a novel theoretical framework—the Explanation Integrity Triad (EIT)—which evaluates any AI explanation along three axes: Representational Integrity, Causal Integrity, and Operational Integrity. The EIT provides a domain-specific vocabulary to prevent mechanistic overclaims and align explanation practices with scientific accountability in applied materials informatics.

Keywords Materials informatics, Interpretability, Explainable AI, Microstructure–property relations, Counterfactual explanations, Graph learning

*Correspondence:
Ahmed Mansour
ahmed.mansour@gmail.com
[1] Department of Materials Engineering and AI Applications, Faculty of Engineering, Cairo University, Cairo, Egypt
[2] Department of Intelligent Materials Systems, Faculty of Engineering, Alexandria University, Alexandria, Egypt

Introduction

Microstructure–property relations form the core explanatory substrate of materials science: they connect structural organization across length scales to measurable macroscopic behavior, linking processing histories to performance-critical outcomes. In applied materials informatics, this explanatory program increasingly intersects with machine learning systems that predict properties from microstructure-derived representations, enabling accelerated screening, surrogate modeling, and decision support in high-dimensional design spaces [1–4]. This predictive success has helped establish AI as a practical instrument within materials research pipelines [3–5].

However, the increasing use of AI predictions in design and engineering has amplified an unresolved epistemic tension: predictions do not automatically yield explanations. A model may predict accurately while offering explanations that are scientifically shallow, unstable, or misleading when read as mechanisms. This tension is particularly pronounced in microstructure–property learning, where the domain's causal structure is mediated by thermodynamic and kinetic

constraints, spatial topology, and multi-scale emergence—properties that are difficult to compress into the independent-variable narratives implicitly encouraged by many interpretability tools [1, 3, 6].

In contemporary practice, explainability in applied machine learning is commonly operationalized through post-hoc methods such as feature attributions, saliency maps, and local surrogate approximations. These methods aim to summarize how a trained predictor responds to inputs, and they have become widely adopted because they are computationally accessible and produce intuitive outputs (importance rankings, heatmaps, local linear models) [7–9]. Yet, interpretability research has repeatedly cautioned that such explanations can be fragile, manipulable, or overly sensitive to modeling degrees of freedom—especially when users treat associational explanations as causal accounts [10–12]. For microstructure–property science, this danger is not marginal: the scientific language of "drivers," "mechanisms," and "control variables" invites interpretive upgrading of explanation outputs into intervention claims, even when the explanation method does not warrant such upgrading.

This paper introduces and develops a conceptual diagnosis of this problem: the microstructure–property explanation gap. We define the explanation gap as the mismatch between (a) what explainability tools typically provide—model-behavior decomposition or local sensitivity narratives—and (b) what materials science requires for legitimate action: explanations that are transferable, intervention-relevant, and stable under realistic shifts in microstructure acquisition and processing regimes [7, 13]. The gap persists because microstructure–property systems violate multiple implicit assumptions embedded in mainstream explainability, including feature independence, local linearity, and the stability of the semantics of the representation space.

The explanation gap has practical consequences. In low-stakes screening, a weak explanation may still provide helpful intuition for prioritization, because the action is reversible and the consequence of error is modest. But in higher-stakes contexts—such as processing optimization, failure-sensitive certification, deployment-critical components, or safety-limited performance—explanatory legitimacy becomes part of decision defensibility. In those settings, the field cannot equate a visually plausible importance map with a mechanism claim. Explanation must be treated as a scientific object requiring integrity checks rather than as a rhetorical add-on to prediction.

We argue that explanation failure in materials AI is often structural rather than incidental, and can be traced to four predictable sources. The first is representational non-identifiability, in which microstructural information is encoded in descriptors or embeddings that collapse distinct morphologies into similar feature values, rendering explanations non-unique and semantically unstable [1, 2]. The second is confounded by processing history: microstructure variables frequently co-vary with fabrication parameters and latent constraints, which undermines the causal interpretation of feature importance [13–15]. The third is multi-scale emergence, since material properties arise from coupled spatial structures across scales that rarely conform to additive "feature contribution" logic [3, 6]. The fourth is instability under distribution shifts, where microstructure measurements and derived descriptors depend on acquisition pipelines, segmentation procedures, and regime changes, leading to explanations that drift even when predictive accuracy appears intact [16]. **Table 1** summarizes the structural anatomy of the microstructure–property explanation gap and shows how each failure mode maps to a specific integrity deficiency in the proposed Explanation Integrity Triad (EIT).

Table 1. The "microstructure–property explanation gap" — structural failure modes and how the EIT diagnoses them

Explanation-gap cause (structural)	What it looks like in microstructure → property AI	Why common XAI fails (core mismatch)	Typical false upgrade (what authors wrongly claim)	The EIT integrity axis that fails first	What "responsible explanation" must include (minimum conditions)
1) Representational	Multiple distinct microstructures map to similar	Post-hoc attributions explain	"Feature X is the driver mechanism"/"this	R-Integrity (Representational Integrity)	Explicit semantic mapping from explanation units →

non-identifiability	descriptor values/embeddings (different morphologies become indistinguishable in feature space).	coordinates of an encoding, not a uniquely defined microstructural entity. Feature importance is faithful to the model, but not uniquely tied to microstructure meaning.	region of latent space corresponds to dislocation pinning / grain boundary strengthening.”	collapses first because the explanation object has unstable or non-unique semantics.	microstructural concepts; identifiability statement (“what is lost in compression”); avoid mechanism language unless the representation supports stable, interpretable microstructural meaning.
2) Confounding by processing history	Microstructural variables co-vary with upstream fabrication/thermal history (e.g., grain size changes with both heat treatment and composition).	XAI treats input variables as separable “knobs,” but microstructure features are often downstream proxies of unobserved process constraints—attribution ≠ intervention.	“Changing porosity will improve strength” (when porosity is only a proxy for processing path).	C-Integrity (Causal Integrity) fails because the explanation lacks intervention semantics and causal identifiability.	Explicit statement of what intervention is meaningful (do-operator logic); separate proxy predictors from intervention levers; causal assumptions disclosed; do not read importance as causality.
3) Multi-scale emergence	Properties arise from coupled structures across length scales (topology, connectivity, interactions), not from additive independent features.	Additive explanations (e.g., SHAP-style) force decompositions that ignore interaction-driven emergence. Explanations become locally plausible but scientifically thin.	“Feature contributions sum to the mechanism” / “the model discovered the physical law.”	Usually, both R-integrity and C-integrity weaken because the explanatory unit is too local/simple to represent multi-scale coupling and cannot support causal narratives.	Interaction-aware explanatory framing (non-additive, relational statements); acknowledge emergence; require mechanism claims to include scale-bridging logic and coupling structure.
4) Instability under distribution shift	Explanations drift with new alloys, imaging settings, segmentation pipelines, or unseen processing regimes—even when accuracy	XAI often assumes stable input semantics, but microstructure “inputs” are partly constructed by	“This explanation generalizes” / “important features remain the same across domains.”	O-Integrity (Operational Integrity) fails because the explanation is not robust under realistic	Stress explanation stability under plausible shifts (new imaging/segmentation regimes); report explanation sensitivity; define validity boundaries

	appears acceptable.	the pipeline. Explanation stability is not guaranteed under a shift.		deployment shifts.	(scope + failure modes).
Net result: the Explanation Gap (definition)	Explanations are treated as mechanisms, but they remain model summaries tied to the dataset regime and encoding choices.	The system explains "what the predictor used," not "why the material behaves that way."	"Mechanism confirmed by saliency map."	EIT mismatch across axes	Explanations must be integrity-graded: R-only (debugging/intuition), R+C (conditional intervention guidance), R+C+O (high-stakes defensible explanation).

Because these sources are intrinsic to the structure–property reasoning process, simply attaching explainability modules to trained models cannot close the interpretive gap. What is needed instead is a domain-specific framework that classifies what an explanation is allowed to claim and identifies the conditions under which it can be legitimately upgraded from a mere model summary to a mechanism-relevant warrant.

To meet this need, we propose the Explanation Integrity Triad (EIT)—a theory-and framework that distinguishes three forms of integrity within an explanation. Representational Integrity (R-Integrity) concerns whether the explanatory object corresponds to stable, interpretable microstructural semantics rather than opaque or collapsing encodings [1, 2]. Causal integrity (C-integrity) asks whether the explanation supports claims that are meaningful under interventions rather than mere associative decompositions, which depend on whether the explanatory variables have coherent causal semantics [13–15]. Operational integrity (O-integrity) evaluates whether the explanation remains stable and functionally relevant under plausible distribution or measurement shifts, such as dataset drift or pipeline variation [16].

By decomposing validity along these three dimensions, the EIT prevents one of the most common scientific missteps in materials AI: the unwarranted elevation of mathematically faithful but semantically unstable explanations into mechanistic claims. Rather than asking whether a model is "explainable," the EIT reframes the question to: Which integrity conditions does this explanation genuinely satisfy, and what level of scientific or causal language is therefore justified?

The remainder of Part 1 develops the theoretical background underlying this diagnosis and introduces the EIT as a practical conceptual tool for explanation auditing in microstructure–property AI.

Theoretical Background & Literature Synthesis

Microstructure–property science is an explanatory discipline, not merely a predictive mapping

Materials science traditionally treats microstructure–property relations as explanatory linkages: microstructures are not passive correlates but structured mediators produced by processing and constrained by thermodynamic–kinetic pathways. When AI enters this domain, the explanation request is therefore not cosmetic; it concerns whether the model can support reasoning about why a property outcome occurs and how to manipulate the system to achieve desired performance [3–5]. Machine learning in materials has expanded rapidly, but the interpretability of microstructure-derived representations remains an open conceptual challenge [3, 4].

Descriptor learning and semantic fragility: Why “important features” may not be meaningful mechanisms

A dominant approach in materials AI is to represent structure using descriptor families that map complex atomic or microstructural configurations into vectors amenable to learning. DScribe formalizes this representational layer through descriptors such as SOAP-like and many-body encodings, enabling scalable materials prediction but also introducing a compression step that shapes what models can learn and what explanations can reference [1]. Continued updates to descriptor libraries further emphasize how representational design choices (including derivatives and refined feature constructions) influence model behavior and sensitivity [2].

From an explanation standpoint, descriptorization creates a risk: explanations may refer to engineered statistics rather than stable microstructural entities. If multiple distinct microstructures share similar descriptor values, then feature-attribution explanations become non-identifiable: they cannot uniquely pick out a mechanism because the representation erases degrees of freedom that may be causally relevant. In microstructure–property science, this produces semantic fragility—the explanation is faithful to the descriptor space but not necessarily faithful to the material phenomenon.

Post-hoc explanation methods: What they guarantee (and what they do not)

Many popular explanation methods are designed to summarize model behavior rather than scientific causality. SHAP-style methods are an influential family because they provide an additive attribution structure with a principled foundation and produce interpretable “feature contributions” for individual predictions or global summaries [9]. However, even within the SHAP ecosystem, the validity of attributions depends on assumptions about feature dependence, coalition structure, and baseline behavior—conditions that are rarely aligned with correlated, constrained microstructure variables. Moreover, post-hoc explanations are vulnerable to manipulation and can appear plausible even when they are not epistemically grounded [12].

A broader interpretability literature emphasizes that explanation outputs can be unstable, fragile, or misleading if users infer more than the method warrants [7, 8, 10]. In microstructure–property learning, such misuse is particularly likely because materials scientists naturally interpret “importance” as “driver,” and “driver” as “cause.” Yet these are not equivalent categories.

Fragility, adversarial explanation failure, and the illusion of understanding

Multiple lines of research demonstrate that explanation methods can be fragile: small changes can shift the salience of explanations, and post-hoc explanation systems can be attacked or engineered to provide comforting narratives without altering the underlying predictive logic [10–12]. This matters for materials AI because explanatory artifacts often serve as a bridge between prediction and scientific interpretation. If that bridge is unstable, it cannot support strong claims such as mechanism inference or process-optimization guidance.

Thus, the explanation gap is partly a reliability problem: explanations are treated as scientific anchors even when their stability has not been verified.

Counterfactual explanations: Contrastive appeal, feasibility risk

Counterfactual explanations attempt to provide “what would need to change for the prediction to change,” and are often framed as actionable because they suggest interventions [14, 15]. Yet the literature on algorithmic recourse and counterfactual explanations also emphasizes feasibility constraints: counterfactuals are meaningful only if the proposed changes correspond to plausible, realizable actions [14, 15].

In microstructure–property settings, this feasibility requirement is stringent. Microstructure is not a freely editable input; it is the endpoint of a processing path. Therefore, counterfactual explanations that propose isolated microstructural edits

("reduce porosity but keep everything else fixed") can violate physical realizability. Such counterfactuals might be valid in representation space but invalid in materials space. This directly undermines C-Integrity and O-Integrity, even if the counterfactual is formally correct as a minimal-change solution.

Causal representation learning: Toward intervention-relevant explanations

Recent scholarship argues that scientific explanation requires causal structure, and that representation learning should aim to capture causal factors rather than purely predictive signals [13]. This direction is especially relevant to microstructure–property learning because causal variables may not align with the most predictive correlates in observational datasets. When processing pathways shape microstructure descriptors, causal structure clarifies which variables are true intervention levers and which are downstream proxies.

Therefore, without causal assumptions (explicit or implicit), explanations remain associational: they describe what the model used, not what would remain stable under intervention.

Distribution shift and explanation stability as a first-class integrity requirement

Operational deployments of materials AI occur under shifts: new alloy families, new imaging pipelines, altered segmentation protocols, or changed processing windows. Research on predictive uncertainty under dataset shift highlights that model reliability can degrade under shift and that uncertainty evaluation must be part of trustworthy deployment [16]. Explanations are subject to similar instabilities: even if accuracy appears acceptable, the explanatory mapping can drift, making explanations unreliable guides for action. This motivates O-Integrity as a primary explanation criterion rather than a secondary "nice-to-have."

Proposed conceptual framework

The Explanation Integrity Triad (EIT): A three-axis integrity model for explanation legitimacy

We propose the Explanation Integrity Triad (EIT) as a domain-specific framework to classify when and why AI explanations fail in microstructure–property learning. The EIT treats an "AI explanation" as a scientific object whose legitimacy depends on three separable integrity properties.

Representational Integrity (R-Integrity)

R-Integrity asks whether the explanation's units correspond to stable, interpretable microstructural semantics rather than opaque encodings. If explanation refers to descriptor coordinates, latent dimensions, or compressed features that lack identifiable microstructural meaning, then R-Integrity is weak—even if the explanation is faithful to the model [1, 2]. This integrity dimension, therefore, distinguishes explainability that remains internal to the model from that which can be translated into scientific microstructural language.

Causal Integrity (C-Integrity)

C-Integrity asks whether the explanation supports intervention-relevant claims. Feature attributions may identify predictive correlates without licensing causal statements, especially when inputs are correlated due to processing constraints [9, 12]. Counterfactual explanations may appear action-like but fail C-Integrity if the suggested edit violates feasibility or ignores upstream processing levers [14, 15]. C-Integrity, therefore, treats explanations as action-claims that require causal semantics and intervention coherence [13].

Operational Integrity (O-Integrity)

O-Integrity asks whether the explanation remains stable and decision-relevant under realistic shifts in data acquisition, segmentation, processing windows, and deployment regimes. Explanations that are persuasive under one dataset regime

may collapse under a shift, just as predictive reliability can degrade under a shift in datasets [16]. O-Integrity prevents explanations from being treated as robust scientific narratives when they are only locally valid artifacts.

Why EIT closes the explanation gap without requiring "perfect interpretability"

The EIT does not claim that all materials AI explanations must become fully causal. Instead, it allows explanation claims to be graded by integrity type. Explanations that satisfy only R-level integrity are acceptable for building human intuition, model inspection, and debugging, but they should not be promoted to mechanistic accounts. Explanations that satisfy both R and C integrity can support tentative intervention guidance, provided the underlying assumptions and limitations are explicitly stated. Explanations that satisfy the integrity of R, C, and O together are candidates for high-stakes scientific reasoning, because they remain stable under distributional shifts and coherent under intervention logic. In this way, the EIT reduces the explanation gap by preventing epistemic overreach: it forces authors and practitioners to specify which integrity conditions are satisfied before deploying mechanistic language or causal claims **(Figure 1)**.

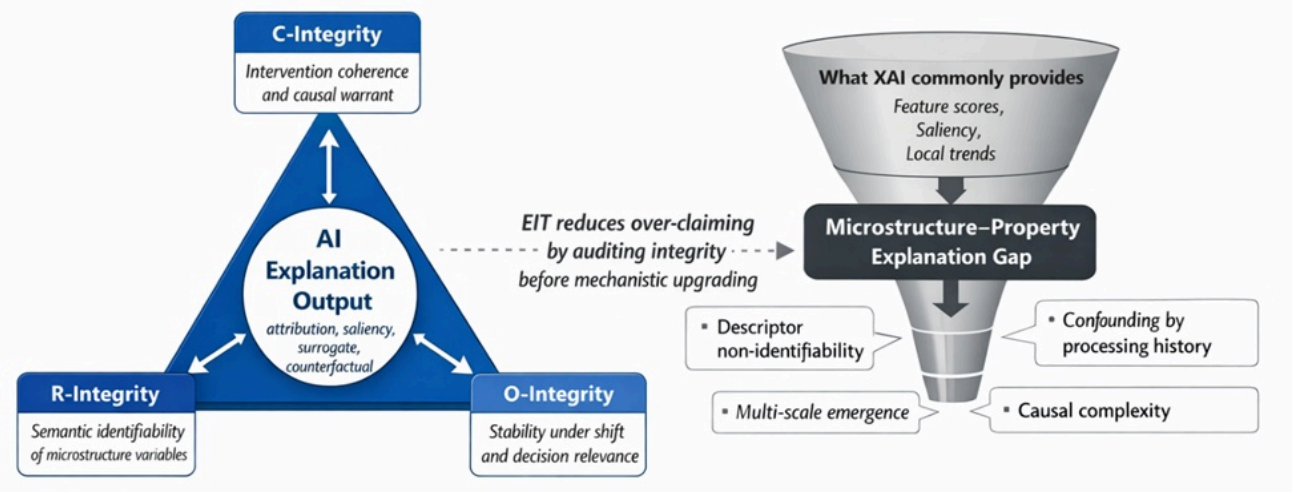

Figure 1. The Explanation Integrity Triad (EIT) and the microstructure–property explanation gap.

Propositions

This section states explicit theoretical propositions that follow from the microstructure–property explanation gap and from the proposed Explanation Integrity Triad (EIT). These are not empirical hypotheses; they are epistemic–normative claims about what kinds of scientific statements AI explanations can responsibly support in microstructure–property learning.

Proposition 1 — Explanations are validity claims, not visual artifacts

An explanation is not the presence of a saliency map, feature ranking, or counterfactual suggestion. It is a validity claim: a statement about what the system is justified in asserting about a microstructure–property relation. Therefore, explainability in materials AI must be evaluated as a claim about warrant, not as a plot or post-hoc add-on [7, 8].

Proposition 2 — The explanation gap is primarily caused by misaligned semantics, not missing algorithms

The dominant cause of explanation failure in microstructure–property AI is not that explanation algorithms are underdeveloped, but that the semantics of microstructure representations do not align with scientific explanation requirements. When microstructure is compressed into descriptor space, explanations frequently attach to features that lack unique microstructural meaning, producing non-identifiable explanation objects even when predictions remain accurate [1, 2].

Proposition 3 — Representational Integrity is a necessary precondition for scientific interpretation

If an explanation refers to variables without stable microstructural semantics, the explanation cannot be interpreted as a microstructure-based claim. Thus, R-integrity is necessary for any explanation that aspires to scientific meaning. However, R-Integrity alone does not justify causal language, because meaningful descriptors can remain confounded by processing history and data selection effects [1, 13].

Proposition 4 — Causal Integrity is equivalent to intervention coherence

In microstructure–property reasoning, causal integrity is not achieved by "importance" scoring; it is achieved when the explanation can support coherent intervention semantics (i.e., "what would happen if we did X?") under explicit causal assumptions [13]. Explanations that do not specify the intervention's meaning remain associational, regardless of their intuitive plausibility [9, 12].

Proposition 5 — Counterfactual explanations are only action-guiding when feasibility is explicitly constrained

Counterfactual explanations appear to provide "what-if" guidance, but in microstructure–property settings, they can generate impossible microstructures or physically unrealizable edits if feasibility constraints are omitted. Therefore, counterfactual explanations should be treated as scientifically valid only if the counterfactual space is restricted to plausible and actionable changes—otherwise they fail both C-Integrity and O-Integrity [14, 15].

Proposition 6 — Operational Integrity is the decisive requirement for high-stakes usage

Explanations can be persuasive yet brittle. Because microstructure is measured through pipelines that shift (e.g., imaging parameters, segmentation, descriptor construction), the stability of explanations under these shifts must be a primary criterion for scientific trust [16]. Hence, O-Integrity differentiates explanations suitable for exploratory insight from explanations suitable for deployment decisions.

Proposition 7 — Explanation legitimacy should be graded, not binary

The EIT implies that explanation legitimacy is not "explainable vs non-explainable." Instead, explanation claims should be graded by their integrity profile (R only; R+C; R+C+O). This grading prevents mechanistic upgrading of only valid explanations as model behavior summaries [7, 8].

Proposition 8 — Mechanistic language is licensed only when all three integrities are satisfied

Mechanistic explanation in materials science implies causal structure, manipulability, and robustness across realistic conditions. Therefore, the mechanistic interpretation of AI explanations is justified only when R-integrity, C-integrity, and O-integrity are all satisfied to an explicit threshold. Without this triadic satisfaction, mechanistic language becomes epistemic overreach [13, 16-19].

Results and Discussion

Why microstructure–property explanation failure is predictable

The microstructure–property explanation gap can be treated as a predictable outcome of four structural tensions:

Microstructure is relational, but explanations are often feature-additive.

Many explanation frameworks encourage additive decomposition ("feature contributions"), even though microstructural effects often arise from topology, connectivity, and interaction terms rather than independent factors [3, 6]. This yields explanations that are locally descriptive but scientifically thin.

The microstructure is constrained by processing history

Microstructure variables are not independent knobs; they are outcomes of process pathways. This undermines causal inference from observational importance scoring: a microstructure feature can appear "important" because it proxies for processing and selection constraints, not because it is itself a causal lever [13].

Measurement pipelines partly constitute the microstructure seen by AI

The observed microstructure is shaped by imaging conditions, segmentation, and feature extraction. Explanations can therefore become explanations of pipeline artifacts rather than explanations of material mechanisms, degrading operational reliability under realistic shift.

Shifts are inevitable in applied materials AI

New alloy families, new processing windows, and new microstructure-imaging regimes shift the distribution. Research on uncertainty under dataset shift demonstrates that trust cannot be assumed stable when the regime changes; explanation stability must be treated analogously [16-22].

These conditions show why explanation failure is not a matter of "using the wrong explanation package" but of interpreting explanation outputs beyond their epistemic capacity.

Reframing SHAP-style explanations: valuable but integrity-limited

SHAP-style approaches are widely adopted because they provide systematic feature attribution and can support global interpretability summaries [9]. Yet attribution is an allocation of predictive responsibility, not a causal claim. Moreover, explanation outputs can be engineered or manipulated, emphasizing that post-hoc explanations are not automatically faithful scientific evidence [12, 23-26].

Under the EIT lens, SHAP often provides *some* level of interpretability but is usually limited to partial integrity: R-Integrity depends on whether the input features correspond to stable microstructural semantics, while C-Integrity is typically absent unless causal assumptions are explicitly modeled. Therefore, SHAP explanations should be treated as "what the model used" rather than "what the material mechanism is."

Saliency and sensitivity: the risk of explanation brittleness

Saliency maps and gradient-based explanations are attractive because they appear fine-grained and locally precise. However, interpretability research has shown that saliency methods can fail "sanity checks," meaning their outputs may not reliably reflect the model's learned reasoning [10]. This supports the argument that explanation outputs can be persuasive without being epistemically stable.

In microstructure contexts, where spatial saliency maps may resemble microstructural "hotspots," the risk is amplified: users can easily read saliency as defect-mechanism localization, even when the explanation method is not warranted to support such claims.

Counterfactuals and recourse: design guidance versus feasibility illusion

Counterfactual explanations are conceptually aligned with materials design because they present contrastive reasoning: "what change yields improvement?" Yet surveys of algorithmic recourse emphasize feasibility, actionability, and consequence-aware reasoning as essential constraints for counterfactual validity [14]. Diverse counterfactual generation further highlights that counterfactual outputs must be interpreted as a set of plausible alternatives, not a single "true minimal cause" [15, 27-30].

In microstructure–property AI, feasibility becomes scientifically binding: microstructure edits must correspond to realizable processing interventions. Without such constraints, counterfactual explanations create a feasibility illusion: they propose a microstructure that exists only in representation space. Under EIT, such explanations fail C-Integrity and O-Integrity, even if they satisfy a formal counterfactual criterion.

Causal representation learning as a pathway toward explanation legitimacy

A central implication of the EIT is that scientific explanation requires aligning representations with causal structure. Causal representation learning argues that representations should reflect causal factors that remain stable under interventions and shifts [13, 31-33]. In microstructure–property settings, this suggests that explanation should not merely interpret a trained predictor, but should be co-designed with representational commitments that support intervention semantics.

This does not mean every material model must become a full causal graph. It means that explanation claims must disclose their causal status: when causal structure is absent, explanations must remain explicitly associational and must not be mechanistically upgraded.

Operational Integrity: Explanation as decision infrastructure

Operational integrity extends explainability beyond interpretation toward decision reliability. Work on uncertainty under dataset shift shows that even well-calibrated systems can degrade under shift; thus, reliability must be evaluated under regime change [2, 6, 16]. Analogously, explanations must be treated as part of the decision infrastructure: if they shift under plausible changes in measurement or data regime, they cannot support high-stakes action.

The EIT, therefore, pushes the field toward a maturity criterion: explanations should not only be available but also disclose stability conditions and failure boundaries.

Implications for scientific writing, peer review, and “mechanistic claim hygiene”

A recurring failure pattern in materials AI publications is a narrative escalation from prediction to attribution plot to mechanistic language. Models generate accurate predictions, which are followed by saliency or attribution visualizations, and the discussion then slides into causal or physical interpretation that the method itself does not warrant. The Explanation Integrity Test (EIT) offers a conceptual corrective by requiring authors to commit to the status of their explanations explicitly. Specifically, authors must state which integrity profile the explanation satisfies—whether it is merely robust, conditionally informative, or operationally grounded (O). They must also clarify what kind of claim is being made, distinguishing a model summary or interpretive aid from a genuine mechanistic warrant. Finally, they must specify the conditions under which the explanation remains valid, including model scope, data regime, and assumptions. Together, these requirements establish a basis for mechanistic claim hygiene, in which the use of scientific and causal language is constrained by the integrity level of the explanation rather than by rhetorical momentum.

Conclusion

This conceptual manuscript defined the microstructure–property explanation gap: the persistent mismatch between what explainability tools typically provide and what microstructure–property science demands for legitimate understanding and action. We argued that explanation failure in materials AI is frequently structural rather than accidental, arising from representation compression, non-identifiability of microstructural encodings, confounding by processing histories, multi-scale emergence, and instability under inevitable distribution shifts.

To address this problem, we introduced the Explanation Integrity Triad (EIT), a theory-first framework that evaluates explanation legitimacy along three independent axes: Representational Integrity, Causal Integrity, and Operational Integrity. The EIT reframes explainability as an integrity-graded scientific claim rather than a post-hoc visualization. Representational Integrity ensures that explanation objects correspond to stable microstructural semantics; Causal

Integrity ensures that explanations support coherent intervention reasoning rather than associational decompositions; and Operational Integrity ensures that explanations remain stable and decision-relevant under plausible measurement and regime shifts.

The key contribution of the EIT is not to demand “perfect interpretability,” but to prevent epistemic overreach—especially the common escalation from feature importance to mechanistic claims. By distinguishing explanations that merely summarize model behavior from explanations that can support intervention-relevant and shift-stable reasoning, the EIT provides a materials-specific vocabulary for responsible scientific interpretation. This allows explainability methods to be used productively without being assigned unwarranted explanatory authority.

Ultimately, the microstructure–property explanation gap should be viewed not as a temporary inconvenience but as a fundamental scientific boundary condition: explanation claims must be constrained by representation, causal structure, and operational stability. The EIT offers a conceptual roadmap for closing the gap—not by decorating prediction with explanation artifacts, but by auditing and aligning explanation practices with the epistemic standards of microstructure–property science.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 25 Aug 2023 Revised: 18 Nov 2023 Accepted: 28 Dec 2023
Published online: 18 January 2024

References

Himanen L, Jäger MOJ, Morooka EV, Federici Canova F, Ranawat YS, Gao DZ, et al. DScribe: library of descriptors for machine learning in materials science. Comput Phys Commun. 2020;247:106949.

Laakso J, Himanen L, Pouillon Y, Jäger MOJ, Foster AS. Updates to the DScribe library: new descriptors and derivatives. J Chem Phys. 2023;158(23):234802.

Schmidt J, Marques MRG, Botti S, Marques MAL. Recent advances and applications of machine learning in solid-state materials science. npj Comput Mater. 2020;6:83.

Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A. Machine learning for molecular and materials science. Nature. 2020;559:547–55.

Wei J, Chu X, Sun X, Xu K, Deng H, Chen J, et al. Machine learning in materials science. InfoMat. 2020;2(3):338–58.

Chen C, Zuo Y, Ye W, Li X, Ong SP. Graph networks as a universal machine learning framework for molecules and crystals. Chem Mater. 2020;32(2):309–18.

Roscher R, Bohn B, Duarte MF, Garcke J. Explainable machine learning for scientific insights and discoveries. IEEE Access. 2020;8:42200–16.

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

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

Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity checks for saliency maps. Adv Neural Inf Process Syst. 2020;33:9505–15.

Ghorbani A, Abid A, Zou J. Interpretation of neural networks is fragile. AAAI Conf Artif Intell. 2021;35(7):5841–8.

Slack D, Hilgard S, Jia E, Singh S, Lakkaraju H. Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods. AAAI Conf Artif Intell. 2021;35(14):11856–63.

Schölkopf B, Locatello F, Bauer S, Ke N, Kalchbrenner N, Goyal A, et al. Toward causal representation learning. Proc IEEE. 2021;109(5):612–34.

Karimi AH, Barthe G, Schölkopf B, Valera I. A survey of algorithmic recourse: Contrastive explanations and consequential recommendations. ACM Comput Surv. 2021;54(8):1–36.

Mothilal RK, Sharma A, Tan C. Explaining machine learning classifiers through diverse counterfactual explanations. Proc ACM FAT. 2020;607–17.

Ovadia Y, Fertig E, Ren J, Nado Z, Sculley D, Nowozin S, et al. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. Adv Neural Inf Process Syst. 2021;34:13991–4005.

Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L. Explaining explanations: an overview of interpretability of machine learning. Proc IEEE. 2020;108(3):369–92.

Lipton ZC. The mythos of model interpretability. Commun ACM. 2022;65(10):36–43.

Samek W, Wiegand T, Müller KR. Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models. ITU J ICT Discov. 2021;1(1):39–48.

Covert I, Lundberg S, Lee SI. Explaining by removing: a unified framework for model explanation. J Mach Learn Res. 2021;22(209):1–90.

Frye C, de Mijolla D, Cowton L, Stanley M, Feige I. Shapley explainability on the data manifold. Adv Neural Inf Process Syst. 2021;34:17410–23.

Guidotti R. Counterfactual explanations and how to find them: literature review and benchmarking. Inf Syst. 2022;101:101–11.

Wachter S, Mittelstadt B, Russell C. Counterfactual explanations without opening the black box: automated decisions and the GDPR. Harv J Law Technol. 2022;31(2):841–87.

Glymour C, Zhang K, Spirtes P. Review of causal discovery methods based on graphical models. Front Genet. 2020;11:81.

Mooij J, Janzing D, Schölkopf B. From ordinary differential equations to causal discovery. Nat Commun. 2020;11:5104.

Runge J, Nowack P, Kretschmer M, Flaxman S, Sejdinovic D. Detecting and quantifying causal associations in large nonlinear time series datasets. Sci Adv. 2020;6(10):eaau4996.

Imbens GW. Potential outcome and directed acyclic graph approaches to causality: relevance for empirical practice. J Econ Lit. 2020;58(4):1129–79.

Yuan H, Yu H, Wang J, Li K, Ji S. Explainability in graph neural networks: a taxonomy and survey. ACM Trans Knowl Discov Data. 2022;16(5):1–35.

Pope PE, Kolouri S, Rostami M, Martin CE, Hoffmann H. Explainability methods for graph convolutional neural networks. Proc CVPR. 2020;10772–81.

Zhang Y, Tiňo P, Leonardis A, Tang K. A survey on neural network interpretability. IEEE Trans Emerg Top Comput Intell. 2021;5(4):726–42.

Koh PW, Liang P. Understanding black-box predictions via influence functions. Proc ICML. 2020;188–97.

Jha D, Ward L, Paul A, Liao W, Choudhary A, Wolverton C, et al. ElemNet: deep learning the chemistry of materials from only elemental composition. Sci Rep. 2020;10:1477.

Xie T, Grossman JC. Crystal graph convolutional neural networks for an accurate prediction of material properties. Phys Rev Lett. 2020;120:145301.