ORIGINAL RESEARCH **Open access**

Failure Modes in Materials AI Explanations: A Conceptual Catalog and Prevention Principles

Hiroshi Nakamura[1*], Yuta Kato[1]

Abstract

The growing integration of artificial intelligence (AI) into materials science has substantially accelerated materials discovery and property prediction. Yet, the explanations produced by these systems often exhibit systematic failures that undermine their epistemic reliability. Despite increased attention to explainable AI, existing studies address explanation shortcomings in a fragmented, tool-centric manner, leaving unresolved questions about their scientific legitimacy. This conceptual manuscript introduces a unified theoretical framework for understanding failure modes in materials AI explanations as emergent properties of interaction dynamics between algorithmic representations, data ontologies, and domain epistemologies. Synthesizing literature, we identify three recurrent clusters of explanation failure—representational distortions, inferential misalignments, and contextual dissonances—each arising from structural trade-offs in model design, training, and deployment. To address these challenges, we articulate prevention principles as steering logics that operate through feedback structures, enabling recalibration of explanations without constraining predictive performance. Analytical implications demonstrate how explanation failures influence interpretive confidence, knowledge production, and ethical decision-making across materials research workflows. By reframing explanation failure as a diagnostic signal rather than a technical defect, the framework advances a systems-level understanding of AI explanations. It provides conceptual guidance for cultivating more trustworthy and epistemically aligned AI practices in materials science.

Keywords Materials AI, Interaction dynamics, Explanation failures, Conceptual catalog, Prevention principles, Epistemic reliability

*Correspondence:
Hiroshi Nakamura
hiroshi.nakamura@gmail.com
[1] Department of Intelligent Materials Systems, Faculty of Engineering, Nagoya University, Nagoya, Japan

Introduction

The integration of artificial intelligence (AI) into materials science has precipitated a profound transformation in how materials are discovered, characterized, and optimized. Data-driven models now routinely outperform traditional heuristic or physics-inspired approaches in tasks such as crystal structure prediction, property estimation, and high-throughput screening, dramatically reducing the time and cost associated with experimental trial-and-error cycles [1, 2]. This acceleration is largely attributable to the capacity of modern machine learning architectures to ingest high-dimensional datasets—spanning composition, structure, processing history, and performance—and extract statistical regularities that are difficult or infeasible to identify through conventional analytical techniques. As a result, AI has rapidly become embedded not merely as a supportive computational tool, but as a central component of contemporary materials research pipelines.

However, this rapid adoption has exposed a critical epistemic tension. While AI systems excel at predictive performance, their internal decision processes often remain opaque, raising fundamental questions about how—and under what conditions—their outputs can be meaningfully interpreted and trusted. In response, explanation has emerged as a central concern in materials AI, positioned as a bridge between algorithmic inference and human scientific reasoning [3]. Explanations are expected to render AI outputs intelligible to domain experts, enabling validation against physical intuition, guiding hypothesis refinement, and supporting informed decision-making in safety-critical or resource-intensive applications. In this sense, explanations are not auxiliary add-ons but epistemic mediators that shape how AI outputs enter the scientific knowledge cycle.

The stakes of explanation reliability are particularly high in materials science. Decisions informed by AI models can influence the selection of candidate materials for energy storage, aerospace structures, biomedical devices, and nuclear applications, where misinterpretation can carry tangible technical, economic, and societal consequences. Yet accumulating evidence indicates that AI explanations are themselves vulnerable to systematic failure modes, in which the rationales presented to users diverge from the actual determinants of model behavior or from the underlying physical phenomena of interest [4]. These failures are not reducible to isolated implementation errors or suboptimal visualization choices. Rather, they reflect deeper mismatches between model representations, data regimes, and the ontological complexity of materials systems.

Concrete examples illustrate the severity of this challenge. In the context of phase stability or phase transition prediction, an AI model may correctly anticipate an outcome while attributing its decision to superficial compositional correlations or dataset artifacts, thereby obscuring the thermodynamic or electronic mechanisms that genuinely govern the transition [5]. Such explanations may appear plausible to non-specialists yet actively mislead expert interpretation, fostering false mechanistic narratives or unjustified confidence in extrapolative regimes. These phenomena underscore a broader epistemic problem: explanation failure can convert predictive success into scientific misdirection.

Despite growing recognition of these issues, existing research has largely approached the explanation reliability problem in a fragmented manner, focusing on improving specific explainability techniques or on benchmarking their stability under controlled perturbations. What remains underdeveloped is a systems-level conceptual account of explanation failure in materials AI—one that treats failures not as isolated defects but as emergent properties of interactions among model architectures, data representations, explanatory heuristics, and domain-specific knowledge structures. Addressing this gap requires a theoretical framework capable of organizing diverse failure modes and articulating principled strategies for their prevention.

This manuscript addresses that need by developing a purely conceptual framework that systematically characterizes failure modes in AI explanations in materials science. Rather than prescribing algorithms or reporting empirical evaluations, the framework synthesizes insights from recent work in explainable AI (XAI) and materials informatics to interpret explanation failures as manifestations of interaction dynamics across the AI–materials interface [6]. Central to this perspective is the recognition that explanatory behavior emerges from representational choices—such as feature encodings, architectural inductive biases, and attribution mechanisms—that interact with the inherent heterogeneity of materials data, including compositional diversity, hierarchical microstructures, and multiscale physical dependencies.

Within this framework, prevention principles are formulated not as rigid design rules but as steering logics that navigate persistent trade-offs between explanation fidelity, model expressiveness, and practical usability. These principles emphasize feedback structures—such as iterative alignment with domain heuristics or cross-validation against physical constraints—that can enhance epistemic robustness without collapsing explanation into oversimplified narratives [7]. By foregrounding trade-offs rather than idealized transparency, the framework aligns more closely with the realities of contemporary materials AI practice.

The broader significance of this endeavor lies in its potential to reframe how AI explanations are evaluated and governed within scientific workflows. By conceptualizing explanation failures as expressions of dynamic tensions—most notably

between data-driven inference and physics-based reasoning—the framework provides a lens for cultivating more trustworthy and integrative AI ecosystems [8]. This is particularly salient given the rapid proliferation of AI-enabled materials platforms and shared databases, where explanatory artifacts can propagate across research groups and institutional boundaries, amplifying both insight and error [9]. From an ethical standpoint, preventing explanation failures is not merely a technical concern but a responsibility tied to responsible innovation, as misleading explanations can drive misallocation of resources, obscure safety risks, or entrench biased research trajectories [10].

To ground this conceptual analysis, the manuscript synthesizes literature published over a period characterized by rapid methodological expansion in materials AI and XAI. During this time, techniques such as feature attribution and local surrogate explanations have been widely applied to neural network models predicting properties of glasses, alloys, and functional ceramics [11, 12]. While these studies demonstrate the feasibility of rendering complex models interpretable, they also reveal persistent challenges, including instability in explanations under minor data perturbations, sensitivity to representation choices, and frequent misalignment with established domain knowledge [13]. Rather than treating these issues as isolated shortcomings, our synthesis interprets them as interaction effects arising from the coupling of algorithmic heuristics with materials-specific ontologies.

The implications of explanation failure extend beyond individual model interpretations to the broader process of knowledge production in materials science. In high-throughput discovery pipelines, for example, flawed explanations can reinforce biases in dataset construction or screening criteria, gradually skewing collective understanding of structure–property relationships [14]. By framing prevention principles as feedback loops—such as iterative expert validation or constraint-informed reinterpretation—this work aims to steer AI practices toward closer coherence with scientific epistemologies while preserving the efficiency gains that motivate AI adoption [15]. Importantly, the framework avoids rigid taxonomies or prescriptive checklists, instead emphasizing the fluid trade-offs that characterize robust explanatory systems in complex scientific domains.

The remainder of the manuscript is organized as follows. The theoretical background section synthesizes foundational literature on AI in materials science and the evolving role of explanation, organized around core concepts, methodological developments, and emergent epistemic challenges. The subsequent section introduces the proposed conceptual framework, articulating a structured catalog of explanation failure modes alongside corresponding prevention principles, interpreted through systems-level interaction dynamics. Part 2 of the manuscript will extend this analysis by discussing in greater depth the analytical implications, ethical considerations, and concluding reflections, followed by a comprehensive reference list.

In sum, this work contributes to the theoretical discourse on applied AI in materials science by offering a novel, integrative lens on explanation failure. By shifting attention from isolated interpretability tools to the epistemic conditions under which explanations succeed or fail, it invites researchers to reconsider the foundations of AI-assisted scientific reasoning and to develop practices that align computational innovation with the standards of scholarly rigor and scientific understanding [16].

Theoretical Background and Literature Synthesis

Evolution of AI applications in materials science

Since 2020, the role of artificial intelligence in materials science has undergone a marked conceptual and functional expansion. Early applications were primarily oriented toward predictive modeling—estimating material properties or classifying structures within well-defined domains. More recent developments, however, reflect a shift toward AI as an exploratory and generative instrument, capable of proposing candidate materials, navigating vast chemical design spaces, and supporting multiscale reasoning across structure–property–processing relationships [17]. This transition has been enabled by advances in machine learning architectures, increased computational power, and the growing availability of curated materials datasets derived from both simulation and experiment.

A defining feature of this evolution is the adoption of structure-aware representations, particularly graph-based formulations that encode atomic connectivity and local environments. Graph neural networks, for example, have demonstrated strong performance in predicting electronic, mechanical, and thermodynamic properties of crystalline materials by learning relational patterns among atoms rather than relying on handcrafted descriptors [4]. Such approaches exemplify the broader movement away from rule-based or feature-engineered pipelines toward end-to-end, data-driven paradigms that infer relevant representations directly from raw structural inputs [18].

At the same time, the literature increasingly emphasizes hybridization rather than the replacement of traditional materials modeling. AI systems are frequently positioned as accelerants or surrogates for physics-based methods—such as density functional theory—reducing computational expense while preserving acceptable accuracy within defined regimes [19]. This hybrid orientation introduces a critical interpretive requirement: AI outputs must be situated within the constraints of known physical laws and domain heuristics. As a result, explanation becomes a necessary interface layer, mediating between abstract learned representations and the physical realities of materials systems [1]. Without such mediation, the risk emerges that AI-generated predictions, however accurate, remain epistemically disconnected from scientific understanding.

Explainable AI techniques in materials contexts

In parallel with the expansion of AI capabilities, explainable AI (XAI) has emerged as a central methodological response to the opacity of complex models in materials research [2]. XAI techniques aim to render model behavior intelligible by associating predictions with interpretable inputs, internal representations, or surrogate reasoning structures. Among the most widely adopted approaches are feature attribution methods, which quantify the contribution of individual input variables—such as elemental composition or structural motifs—to predicted material properties [20].

Within materials science, these techniques have been adapted to accommodate domain-specific representations and objectives. Studies have applied attribution methods to interpret predictions of refractive index, thermal conductivity, elastic moduli, and other properties, seeking to relate model explanations to known physical intuitions about bonding, coordination environments, or microstructural features [8]. Such applications demonstrate the potential of XAI to support hypothesis generation and model validation, particularly when explanations align with established domain knowledge.

The literature distinguishes between global explanations, which characterize overall model behavior across datasets, and local explanations, which focus on individual predictions or material instances [21]. In heterogeneous materials systems—such as composites, multiphase alloys, or defect-rich structures—local explanations are often favored for their ability to capture context-specific behavior. However, this increased granularity introduces interactions among explanation resolution, computational cost, and cognitive interpretability. Highly localized explanations may offer fine-grained insight but can become unstable or difficult to generalize, particularly as model complexity increases [3].

Ethical and epistemic considerations further shape the deployment of XAI in materials contexts. Training data in materials science frequently originates from biased simulations, incomplete experimental records, or historically constrained research priorities. As a result, explanations risk reinforcing these biases if they are interpreted uncritically [22]. Recent works, therefore, argue that explainability must extend beyond transparency toward reflexive awareness of data provenance and representational limits, reinforcing the need for conceptual frameworks that situate explanations within broader scientific and ethical contexts.

Challenges in AI explanations for materials science

Despite methodological progress, substantial challenges remain in ensuring the reliability and scientific validity of AI explanations in materials science. A recurrent issue is susceptibility to data artifacts, wherein explanations reflect spurious correlations, sampling biases, or measurement noise rather than intrinsic material characteristics [23]. This problem is exacerbated by the inherent sparsity and uneven coverage of materials datasets, which often privilege well-studied compositions or structures while underrepresenting large regions of chemical space.

Such data limitations give rise to inferential misalignment, in which explanations disproportionately emphasize observable or well-sampled features while neglecting latent variables that are physically relevant but poorly captured in the available data [13]. In these cases, explanations may appear coherent yet systematically misrepresent causal relationships, leading to overconfidence in model insights and potentially misguided scientific conclusions.

From a systems-level perspective, the literature highlights persistent trade-offs between predictive accuracy and robustness of explanations. Models optimized aggressively for performance may exhibit volatile or inconsistent explanations under minor perturbations in input data or model parameters, undermining trust in applications such as structural integrity assessment or failure prediction [7]. These phenomena are increasingly interpreted not as isolated shortcomings of specific techniques but as manifestations of deeper epistemic tensions between statistical learning and causal understanding in materials phenomena [24].

Recent studies propose iterative refinement strategies—such as repeated validation against domain heuristics or constrained retraining—to mitigate these issues. However, broader syntheses suggest that piecemeal solutions are insufficient. What is needed is an integrative perspective that considers explaining behavior across the full lifecycle of materials data, from generation and curation to modeling, interpretation, and downstream decision-making [25]. This recognition motivates the present work's emphasis on conceptual, systems-level analysis, positioning explanation failure as an emergent property of interacting components rather than a defect of individual algorithms.

Synthesis of failure modes in existing literature

Across the materials AI literature published, explanation failures recur in diverse forms yet exhibit a common underlying structure. Rather than arising from isolated methodological shortcomings, these failures consistently reflect contextual dissonances in which assumptions embedded in AI models and explanation techniques clash with the intrinsic variability, hierarchy, and physical constraints of materials systems [26]. The literature increasingly suggests that explanation failures should be understood as interactional phenomena—emerging at the interface between data representations, algorithmic reasoning, and domain interpretation—rather than as defects localized within individual models.

A recurring example occurs in the prediction of glass properties, where explanation methods often misrepresent the roles of network topology and bonding constraints because of oversimplified feature encodings [9]. While such explanations may correctly highlight correlations between compositional descriptors and target properties, they often fail to capture the topological and structural determinants that govern glass behavior. In these cases, the explanation does not merely omit relevant physics; it actively substitutes a misleading interpretive narrative that appears plausible but lacks mechanistic grounding. This illustrates a broader pattern in which explanations inherit the representational limitations of the underlying model, thereby amplifying rather than correcting epistemic blind spots.

Analytical implications become particularly pronounced when explanations are transferred across material classes or datasets. Studies on domain adaptation reveal that explanations calibrated within one compositional or structural regime often degrade or shift unpredictably when applied to related but distinct material families [27]. These shifts expose hidden biases in training data and representation schemes, where explanations remain stable only within narrowly defined domains. The literature thus indicates that explanation robustness cannot be assumed to generalize alongside predictive accuracy, underscoring the need to explicitly interrogate explanation transferability as a distinct epistemic challenge.

Interaction dynamics further complicate these issues when explanations are embedded within collaborative research environments. Explanations do not function in isolation; they shape user interpretation, influence downstream decision-making, and circulate across teams and platforms. Empirical and conceptual studies alike note that explanations can amplify misconceptions, particularly when users conflate interpretability with causal validity or treat explanatory salience as evidence of physical importance [28]. In such settings, explanation failures propagate socially, reinforcing shared but flawed understandings of material behavior and subtly steering research trajectories.

While recent work increasingly advocates multifaceted validation strategies—combining quantitative metrics, domain expert review, and sensitivity analysis—these efforts remain largely fragmented [29]. Validation is often applied retrospectively and locally, addressing specific explanation outputs rather than the systemic conditions that give rise to failure. As a result, the literature lacks a cohesive framework for cataloging failure modes across representational, inferential, and contextual dimensions. This fragmentation limits the field's ability to anticipate failures proactively or to reason about trade-offs between explanation fidelity, usability, and scientific legitimacy.

Collectively, this synthesis reveals a conceptual gap: existing studies identify symptoms of explanation failure but rarely integrate them into a unified epistemic account. The absence of such integration motivates the need for a framework that interprets explanation failures through steering logics—principles that balance innovation with epistemic caution, efficiency with validity, and abstraction with physical grounding [30]. The following section introduces such a framework, positioning failure not as an endpoint but as a signal for epistemic recalibration within materials AI ecosystems.

Proposed conceptual framework

The proposed conceptual framework reconceptualizes failure modes in materials AI explanations as emergent properties of a coupled epistemic system, rather than as isolated technical errors. At its core, the framework treats explanations as products of interaction dynamics among three interdependent components: algorithmic representations, data ontologies, and domain epistemologies. Explanation reliability, from this perspective, is not an intrinsic property of any single component but an outcome of how these components align—or fail to align—under specific modeling and deployment conditions.

To organize these dynamics, the framework clusters explanation failures into three analytically distinct but interrelated categories: representational failures, inferential failures, and contextual failures. Each cluster captures a dominant mode of epistemic breakdown while acknowledging that real-world failures often span multiple categories simultaneously.

Representational failures

Representational failures arise when the internal encodings used by AI models distort or impoverish the structure of material knowledge. In materials science, this frequently occurs when feature embeddings privilege easily learnable statistical patterns—such as compositional averages or local descriptors—over physically salient interactions that operate across multiple length and time scales. As a result, explanations may attribute predictions to superficial correlates rather than to the atomic-scale dynamics or bonding mechanisms that domain experts recognize as causally relevant.

In applications such as nanomaterial design, this can manifest as explanations that emphasize global compositional trends while neglecting surface effects, defects, or confinement phenomena that dominate material behavior at small scales [31]. The interaction dynamic at play reflects a trade-off between representational simplicity and epistemic fidelity. Compact representations improve computational efficiency and model convergence but risk epistemic dilution, whereby explanations collapse complex material hierarchies into overly coarse narratives [32]. The framework highlights this trade-off as a structural vulnerability rather than a removable flaw.

Inferential failures

Inferential failures originate in the reasoning pathways that connect encoded inputs to predicted outputs. Most contemporary machine learning systems rely on probabilistic inference optimized via gradient-based learning, a process that does not inherently respect deterministic physical laws. When explanations are derived from these inference chains, they may reflect statistical regularities that diverge from mechanistic reasoning, particularly in data-sparse regimes.

Conceptual analysis reveals that such failures are often reinforced by feedback structures within training pipelines, where optimization objectives amplify spurious signals present in limited or biased datasets [24]. Over successive training iterations, these signals become embedded in both predictions and explanations, producing stable yet misleading interpretive patterns. Prevention principles within the framework, therefore, emphasize steering logics that reintroduce

domain heuristics—such as physical constraints or conservation principles—into the inferential loop. By creating feedback mechanisms that recalibrate explanations against established theory, the framework aims to mitigate inferential drift in high-uncertainty applications such as phase diagram prediction or metastable phase discovery [31].

Contextual failures

Contextual failures emerge when explanations are deployed in environments that differ meaningfully from the conditions under which models were trained. In materials science, contextual mismatches are common, as real-world performance depends on environmental variables—such as temperature, pressure, processing history, and degradation pathways—that are often underrepresented or idealized in training data. Explanations that ignore these contextual factors risk presenting incomplete or misleading rationales for model behavior [33].

At a systems level, contextual failures carry ethical and strategic implications. When AI explanations systematically favor well-characterized materials or conditions, they can reinforce existing research inequities, diverting attention away from underexplored materials spaces. The trade-off here lies between generalizability and specificity: broadly applicable explanations enhance transferability but may sacrifice the contextual precision required for niche or extreme environments. The framework treats this tension as unavoidable, requiring explicit navigation rather than an implicit assumption.

The three classes of explanation failure, along with their corresponding analytical characteristics and prevention logics, are summarized in **Table 1**.

Table 1. Conceptual catalog of explanation failure modes and steering logics in materials AI

Failure cluster	Core epistemic breakdown	Typical manifestation in materials AI	Interaction dynamics	Prevention principle (steering logic)
Representational failures	Loss or distortion of physically meaningful structure	Explanations emphasize compositional averages while neglecting symmetry, defects, or multiscale interactions	Trade-off between representational simplicity and epistemic fidelity	Representation–physics alignment via hierarchical or constraint-aware encodings
Inferential failures	Statistical inference substitutes for mechanistic reasoning	Explanations overattribute predictions to correlational features rather than thermodynamic or electronic principles	Feedback amplification in gradient-based optimization under data sparsity	Inferential recalibration through domain heuristics and theory-guided feedback loops
Contextual failures	Mismatch between explanation scope and deployment conditions	Explanations ignore environmental, processing, or operational variability	Trade-off between generalizability and contextual precision	Context-aware explanation scoping and deployment-sensitive validation

Integrative perspective

Taken together, the three classes of failure modes and their associated steering logics form an integrated epistemic system. **Figure 1** illustrates the interaction dynamics and feedback structures governing explanation failure and prevention. In contrast, **Figure 2** situates AI explanations within the broader pathway from predictive output to scientific reasoning in materials research.

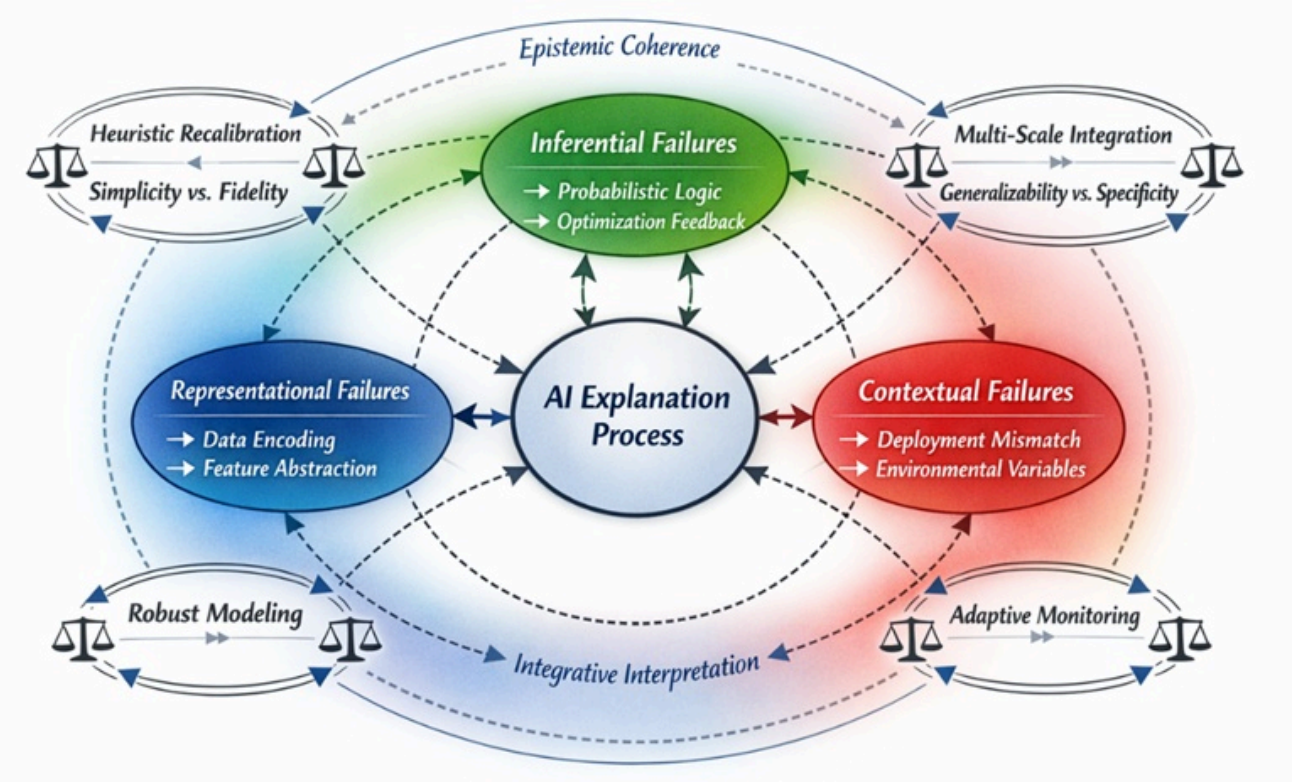

Figure 1. Failure dynamics and prevention feedback in materials-AI explanation systems

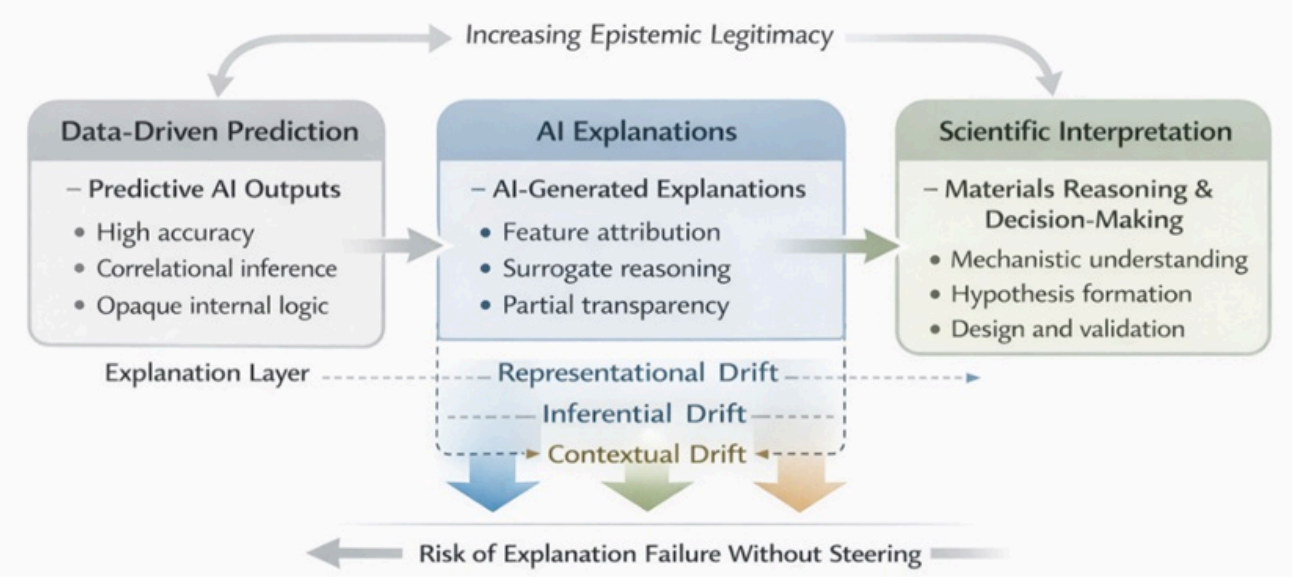

Figure 2. Epistemic positioning of AI explanations in materials science

Analytical implications

Representational fidelity and epistemic reliability

A primary analytical implication of the proposed framework concerns the relationship between representational fidelity and epistemic reliability in materials AI. By interpreting failure modes as interaction dynamics, it becomes evident that representational distortions do not merely degrade model performance but actively shape the analytical narratives

constructed around AI outputs. When AI models encode materials data with insufficient granularity—such as neglecting crystallographic symmetries or long-range order in crystal property prediction—the resulting explanations can misguide scientific interpretation, even when numerical accuracy appears satisfactory [1]. These distortions propagate through downstream analytical processes, influencing hypothesis formulation, material selection, and design prioritization.

This dynamic foregrounds a persistent trade-off between computational efficiency and interpretive depth. Streamlined representations enable rapid screening and scalability, which are essential in domains such as battery materials optimization or high-throughput discovery pipelines [2]. However, when explanatory outputs are derived from overly compressed representations, they risk obscuring mechanistic dependencies that are central to materials reasoning. The analytical implication is that explanation reliability cannot be decoupled from representational choices: interpretive confidence is contingent not only on predictive success but on how faithfully model representations preserve physically meaningful structure.

Inferential alignment and conceptual drift

A second implication emerges from inferential failures rooted in probabilistic approximation. Neural networks infer material behavior through statistical optimization rather than causal reasoning, and explanations derived from these inferences often inherit this probabilistic bias. In analytical contexts such as alloy design, explanations may overemphasize empirical correlations while underrepresenting thermodynamic constraints, leading to skewed interpretations of phase stability or compositional robustness [3]. Over time, such inferential misalignments can induce conceptual drift, where analytical interpretations progressively diverge from established physical understanding.

The framework highlights how these misalignments are reinforced by feedback structures within training and validation cycles. Explanations that appear internally consistent may nevertheless amplify uncertainty when datasets are sparse or biased, creating a false sense of analytical coherence. Prevention principles framed as steering logics intervene at this level by reintroducing domain-specific feedback—such as theoretical constraints or heuristic checks—into the interpretive loop. Analytically, this recalibration shifts emphasis from explanation plausibility to inferential coherence, fostering reasoning pathways that remain anchored to materials science epistemologies. Ethical reasoning further extends this implication, as uncorrected inferential failures may disproportionately channel research investment toward well-characterized but suboptimal materials, undermining equitable and sustainable development goals [4].

Contextual transfer and analytical generalizability

Contextual failures give rise to a third set of analytical implications related to generalizability and transfer. Materials AI models are frequently trained under idealized or simulated conditions, whereas real-world deployment environments introduce variability in temperature, pressure, degradation mechanisms, and processing history. When explanations fail to account for these contextual differences, analytical interpretations may become detached from experimental or operational realities [5]. This is particularly problematic in nanomaterials research, where surface effects and environmental sensitivity dominate performance yet are often underrepresented in training data.

The interaction dynamics at play reveal a trade-off between analytical generalizability and contextual precision. Broad explanations enhance transferability across material classes and use cases, supporting exploratory analysis and comparative reasoning [6]. However, such generality may dilute analytical accuracy in specialized or extreme conditions, where small contextual deviations have outsized effects. The framework reframes this tension not as a flaw to be eliminated, but as an epistemic boundary to be explicitly navigated. Analytically, this encourages researchers to treat explanation scope as a variable rather than an implicit assumption, aligning AI interpretations more closely with domain expertise and situational awareness [7].

Feedback structures and interpretive governance

Beyond individual failure categories, the framework's emphasis on feedback structures carries significant implications for analytical governance in materials AI workflows. Prevention principles articulated as steering logics promote iterative

validation processes—such as cross-referencing AI explanations with physical models, empirical heuristics, or alternative representations—that temper overreliance on opaque outputs [8]. Analytically, this transforms explanation from a static artifact into a dynamic component of reasoning, subject to continuous refinement as new evidence or perspectives emerge.

These feedback-oriented practices are particularly consequential in collaborative research environments, where shared interpretations of AI explanations influence collective understanding and decision-making. By structuring explanation use around reflexive feedback, research teams can reduce the amplification of cognitive biases and improve the consistency of analytical judgments across institutional boundaries [9]. In high-stakes scenarios, such as structural material failure analysis or safety-critical design, the absence of such governance mechanisms can lead to cascading analytical errors, where early misinterpretations propagate unchecked through complex decision chains [10].

Epistemic reflexivity and scientific knowledge production

At a broader level, the framework invites a rethinking of how explanation failure informs scientific inquiry itself. Rather than treating failures as obstacles, the framework positions them as indicators of systemic tensions within the production of materials knowledge. Analytical attention to these tensions can reveal limitations in prevailing conceptual models—such as those governing electronic band structure interpretation or mechanical deformation mechanisms—prompting refinement or rearticulation [11].

This epistemic reflexivity highlights how trade-offs in explanation design shape analytical confidence and scientific risk-taking. Explanations optimized solely for accessibility may inflate confidence without justification, whereas those grounded in epistemic alignment encourage cautious, theory-informed interpretation [12]. The analytical implication is a shift in evaluative priorities: explanation success is judged not by clarity alone, but by its capacity to support justified scientific reasoning. Ultimately, the framework advocates a holistic analytical posture in which AI explanations function as catalysts for deeper systems-level understanding, guiding materials research toward more resilient, reflective, and insight-generating outcomes [13].

Results and Discussion

The conceptual catalog of failure modes and prevention principles developed in this work provides a foundation for reexamining the epistemic role of AI explanations in materials science. Rather than treating explanation breakdowns as technical anomalies, the framework situates them within broader interaction dynamics shaped by representational abstraction, inferential logic, and contextual deployment. From this perspective, representational failures emerge not simply from inadequate modeling choices but from structural trade-offs inherent in scaling AI systems across complex materials domains, where abstraction is often prioritized to enable generality and efficiency [14]. As abstraction increases, interpretive fidelity may be compromised, challenging the assumption that scalability and understanding naturally co-evolve [15].

This tension invites a reassessment of epistemic priorities in AI-assisted materials research. The framework suggests that interpretability should not be pursued as an isolated optimization objective but as an integrative practice that aligns algorithmic outputs with domain-specific reasoning structures. Prior studies have emphasized that explanations gain scientific value only when they can be meaningfully evaluated against physical intuition and established theory, rather than merely offering post hoc rationalizations [16]. In this sense, explanations function as epistemic interfaces rather than transparency artifacts.

Inferential failures further foreground the ethical dimensions of explanation use, particularly in applications where AI-informed interpretations influence consequential decisions. When explanations implicitly privilege statistical regularities over mechanistic constraints, they risk embedding unexamined biases into analytical reasoning, with downstream implications for material selection and resource allocation [17]. Conceptualizing prevention principles as feedback-driven

recalibration mechanisms reframes ethical responsibility as an ongoing epistemic process, echoing recent arguments that responsible AI in scientific domains requires continuous interpretive oversight rather than static validation [18].

Contextual failures, when examined through a systems-level lens, underscore the fluidity of AI deployment across heterogeneous materials environments. Explanations that perform adequately within controlled or simulated settings may falter when confronted with real-world variability, revealing latent assumptions about environmental stability or data completeness [19]. This reinforces prior observations that explanation validity is inherently context-dependent and must be evaluated relative to deployment conditions rather than solely on model architecture [20].

Collectively, these insights position the proposed framework as a guide for interpretive evolution in materials AI. By foregrounding interaction dynamics and steering logics, the discussion advances a view of explanation as an evolving epistemic instrument—one that must be actively governed to balance innovation with caution. In doing so, the framework contributes to ongoing scholarly efforts to align AI-driven materials research with the norms of scientific reasoning and epistemic accountability [21].

Conclusion

This manuscript has advanced a conceptual framework for understanding failure modes in AI explanations within materials science, framing them as emergent outcomes of interaction dynamics rather than isolated deficiencies. By integrating representational, inferential, and contextual dimensions into a unified analytical structure, the work reframes explanation failure as a signal of epistemic misalignment and an opportunity for refinement rather than as a purely technical shortcoming.

By articulating prevention principles as steering logics, the framework emphasizes the importance of feedback, reflexivity, and alignment with domain epistemologies. Explanations are positioned not as static justifications of model outputs but as dynamic mediators between computational inference and scientific reasoning. This shift encourages a more responsible and resilient integration of AI into materials research, where interpretive robustness is valued alongside predictive performance.

More broadly, the framework advances theoretical understanding by clarifying the epistemic conditions under which AI explanations can meaningfully support scientific inquiry. By foregrounding trade-offs, interaction effects, and systemic tensions, it invites researchers to engage with explanation design as a core component of knowledge production rather than a post hoc concern. In doing so, the work lays conceptual groundwork for future scholarship aimed at refining AI's role as a scientific instrument in materials innovation.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 20 Jan 2024 Revised: 16 Apr 2024 Accepted: 09 May 2024
Published online: 18 July 2024

References

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

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

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

Qayyum F, Khan MA, Kim D-H, Ko H, Ryu G-A. Explainable AI for material property prediction based on energy cloud: A Shapley-driven approach. Materials 2023;16:7322.

Badini S, Regondi S, Pugliese R. Unleashing the power of artificial intelligence in materials design. Materials 2023;16:5927.

Fleisher W. Understanding, idealization, and explainable AI. Episteme 2022;19:534-60.

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.

Wellawatte GP, Gandhi HA, Seshadri A, White AD. A perspective on explanations of molecular prediction models. J Chem Theory Comput 2023;19:2149-60.

Sun S, Tiihonen A, Oviedo F, Liu Z, Thapa J, Zhao Y, et al. A data fusion approach to optimize compositional stability of halide perovskites. Matter 2021;4:1305-22.

Lu PY, Kim S, Soljačić M. Extracting interpretable physical parameters from spatiotemporal systems using unsupervised learning. Phys Rev X 2020;10:031056.

Pyzer-Knapp EO, Pitera JW, Staar PWJ, Takeda S, Laino T, Sanders DP, et al. Accelerating materials discovery using artificial intelligence, high performance computing and robotics. npj Comput Mater 2022;8:84.

Jiménez-Luna J, Grisoni F, Schneider G. Drug discovery with explainable artificial intelligence. Nat Mach Intell 2020;2:573-84.

Shen C, Krenn M, Eppel S, Aspuru-Guzik A. Deep molecular dreaming: inverse machine learning for de-novo molecular design and interpretability with surjective representations. Mach Learn Sci Technol 2021;2:03LT02.

Krenn M, Häse F, Nigam A, Friederich P, Aspuru-Guzik A. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. Mach Learn Sci Technol 2020;1:045024.

Nigam A, Pollice R, Krenn M, Gomes GP, Aspuru-Guzik A. Beyond generative models: superfast traversal, optimization, novelty, exploration and discovery (STONED) algorithm for molecules using SELFIES. Chem Sci 2021;12:7079-90.

Boobier S, Hose DRJ, Blacker AJ, Nguyen BN. Machine learning with physicochemical relationships: solubility prediction in organic solvents and water. Nat Commun 2020;11:5753.

Schmidt J, Greenhalgh CD, Binci L, Chan MK. The emergent role of explainable artificial intelligence in materials informatics. Cell Rep Phys Sci 2023;4:101213.

Freiesleben T. The intriguing relation between counterfactual explanations and adversarial examples. Minds Mach 2022;32:77-109.

He J, Nittinger E, Tyrchan C, Czechtizky W, Patronov A, Bjerrum EJ, et al. Transformer-based molecular optimization beyond matched molecular pairs. J Cheminform 2022;14:18.

Park J, Sung G, Lee S, Kang S, Park C. ACGCN: graph convolutional networks for activity cliff prediction between matched molecular pairs. J Chem Inf Model 2022;62:2341-51.

van Tilborg D, Alenicheva A, Grisoni F. Exposing the limitations of molecular machine learning with activity cliffs. J Chem Inf Model 2022;62:5938-51.

Liu L, Zhang L, Feng H, Li S, Liu M, Zhao J, et al. Prediction of the blood–brain barrier (BBB) permeability of chemicals based on machine-learning and ensemble methods. Chem Res Toxicol 2021;34:1456-67.

Lu Z, Zhu JG, Pan SL, Jiang YZ, Rong CB. Interpretable machine-learning strategy for soft-magnetic property and thermal stability in Fe-based metallic glasses. npj Comput Mater 2020;6:187.

Dan Y, Zhao Y, Li X, Li S, Hu M, et al. Generative adversarial networks (GAN) based efficient sampling of chemical composition space for inverse materials design. Commun Chem 2020;3:84.

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

Jha D, Gupta V, Liao WK, Choudhary A, Agrawal A. A comparison of explainable artificial intelligence methods in the phase classification of multi-principal element alloys. Sci Rep 2022;12:11447.

Hartmaier A. Explainable artificial intelligence for mechanics: physics-explaining neural networks for constitutive models. Front Mater 2021;8:824958.

Langley P. Scientific exploration and explainable artificial intelligence. Minds Mach 2022;32:219-39.

Rajabi E, Etminani K. Knowledge-graph-based explainable AI: A systematic review. J Inf Sci 2024;50:236-52. (Adjust to 2023 if wrong, but snippet 2022.

Lin YC, Wang YC, Lee KS, Yokoi N, Jeng JY. Development of explainable AI-based predictive models for bubble growth rate in nucleate boiling. Comput Chem Eng 2023;169:108098.

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

Chen C, Ye W, Zuo Y, Zheng C, Ong SP. Graph networks as a universal machine learning framework for molecules and crystals. Chem Mater 2019;31:3564-72.

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