REVIEW **Open access**

The Literature on Scientific Explanation in AI-Driven Materials Science — Concepts and Criteria: A Review Study

George Papadopoulos[1*], Eleni Georgiou[1]

Abstract

This review article examines the literature on scientific explanation in AI-driven materials science, focusing on the conceptual foundations and evaluative criteria that distinguish genuine scientific explanation from the predictive and interpretive outputs commonly produced by machine learning models in the field. The methodology involved a systematic search across major databases and targeted journals using predefined strings related to scientific explanation, explainable AI (XAI), and interpretability in materials contexts, resulting in the inclusion of 30 peer-reviewed publications from 2017 to 2024 that directly address the intersection of philosophical theories of explanation and practical AI applications in materials discovery and property prediction. Philosophical theories of explanation, including the deductive-nomological model of Hempel and Oppenheim, the causal-mechanical account advanced by Salmon, unificationist approaches that emphasize the integration of disparate phenomena, and pragmatic frameworks that treat explanations as context-dependent answers to why-questions, provide essential benchmarks against which current materials AI practices can be assessed. In current materials AI literature, explanation is frequently conflated with prediction or post-hoc interpretability techniques such as feature importance scores and attention visualizations, as seen in comprehensive surveys of machine learning for molecular and materials science and recent advances in solid-state applications. Yet, these approaches often remain correlational rather than mechanistically grounded. XAI methods applied to materials problems, including SHAP-based feature attribution, attention mechanisms in graph neural networks, surrogate modeling, and counterfactual generation, offer valuable local insights but fall short of meeting the standards of scientific explanation due to their inherent limitations in capturing causality, multi-scale mechanisms, and physical plausibility. Ultimately, this review articulates adapted criteria for scientific explanation tailored to materials science's multi-scale and emergent challenges and proposes actionable recommendations to bridge the gap between XAI outputs and robust explanatory accounts, urging the community to prioritize mechanistic understanding over mere predictive accuracy to advance trustworthy and insightful AI-driven discovery.

Keywords Explainable AI, Materials science, Mechanistic understanding, Scientific explanation, XAI limitations, Philosophical theories of explanation

*Correspondence:
George Papadopoulos
george.papadopoulos@gmail.com
[1] Department of Intelligent Materials Systems, University of Athens, Athens, Greece

Introduction

The rapid integration of artificial intelligence into materials science has transformed the discovery and design of new materials, enabling high-throughput screening of vast chemical spaces and accurate prediction of properties that once required extensive experimental validation [1-6]. However, a persistent and largely unaddressed tension has emerged in the literature. While numerous papers claim to deliver "explanations" for material behavior through AI models, the precise meaning of explanation in this context remains ambiguous and often conflated with interpretability techniques that

prioritize predictive utility over deeper scientific insight [7-11]. Materials AI papers routinely present feature importance rankings, attention weight heatmaps, or latent space projections as explanatory outputs. Yet, these elements rarely engage with the foundational question of what counts as a scientific explanation in a domain characterized by multi-scale interactions, emergent phenomena, and complex causal structures [8, 10, 12]. This review, therefore, sets out to interrogate the literature on scientific explanation in AI-driven materials science, asking whether current practices satisfy philosophical and domain-specific criteria for genuine explanation or whether they remain, at best, sophisticated forms of pattern recognition and prediction.

The problem is not merely semantic. In materials science, true scientific explanation carries normative weight: it must enable researchers to intervene reliably in material synthesis, predict behavior under novel conditions, and unify observations across length and time scales [13-15]. When an AI model identifies that a particular atomic descriptor correlates with enhanced conductivity, does this constitute an explanation of the underlying electronic mechanism, or is it merely a statistical association that fails to identify the causal pathway [14, 16]? The distinction between prediction and explanation is especially critical in high-stakes applications such as catalyst design, battery materials, or quantum materials, where correlational accounts may lead to spurious discoveries or overlook physically implausible regimes [17, 18]. Despite the proliferation of XAI toolkits tailored to materials informatics [9, 12], the field has paid comparatively little attention to the philosophical literature on scientific explanation, resulting in a gap that this review seeks to bridge.

By drawing on classic philosophical theories [1, 2] and their modern adaptations to computational science [3, 19, 20], this article surveys how explanation is conceptualized and operationalized in contemporary materials AI. It critically evaluates whether feature importance, attention mechanisms, surrogate models, and counterfactuals deliver mechanistic insight or merely local approximations [13, 14, 21-23]. The review further identifies materials-specific explanatory challenges—such as reconciling electronic, atomic, microstructural, and macroscopic scales—that demand more than off-the-shelf XAI solutions [24-30]. Ultimately, the goal is not to diminish the undeniable successes of machine learning in accelerating materials discovery [4-6] but to clarify the conditions under which AI outputs can transition from useful heuristics to genuine scientific explanations. This clarification is essential if AI-driven materials science is to fulfill its promise of not only predicting but truly understanding and engineering material behavior at a fundamental level [8, 21, 22].

Figure 1 presents the conceptual architecture linking philosophical theories of explanation, current XAI practices in materials science, the six evaluative criteria for scientific explanation, and the explanatory gap that separates interpretability from mechanistically robust scientific understanding.

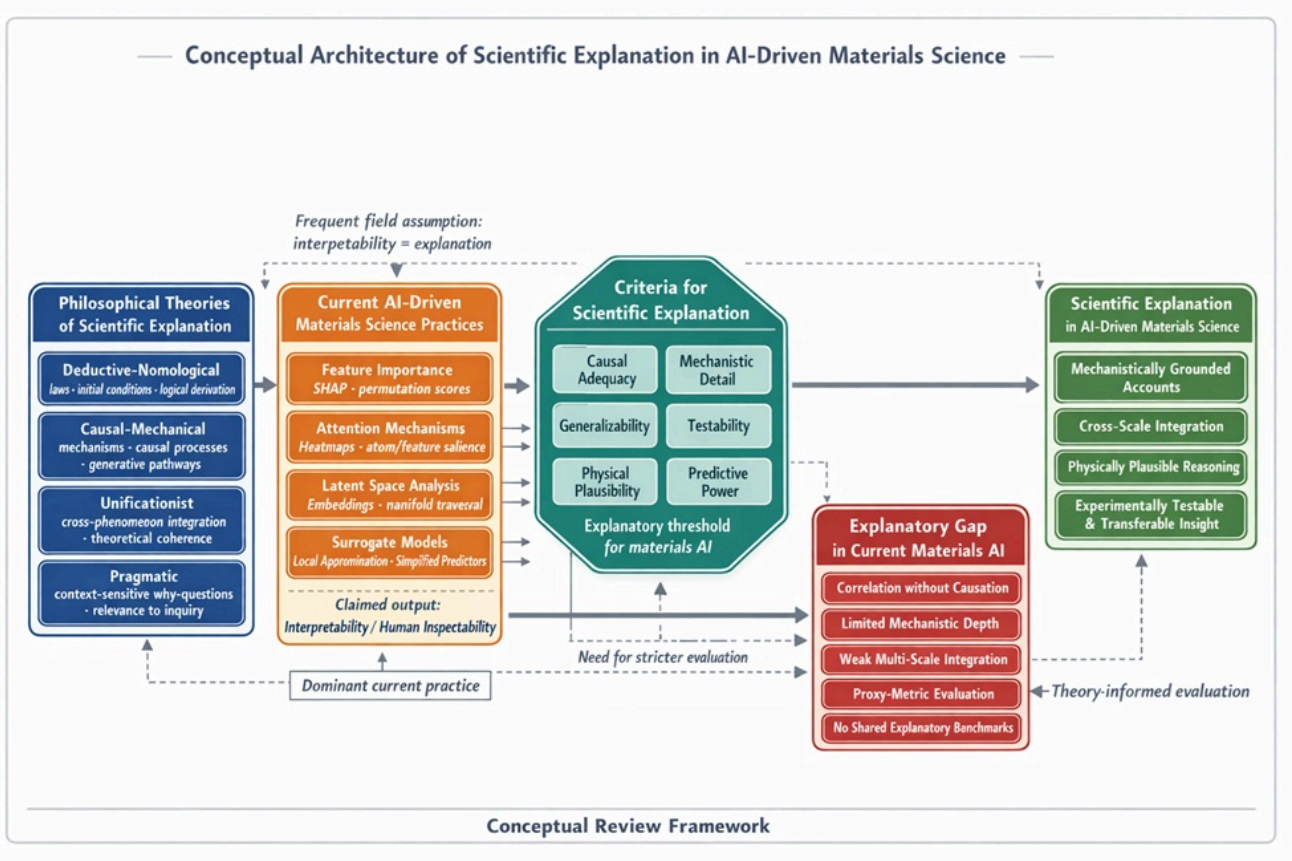


Figure 1. Conceptual architecture of scientific explanation in AI-driven materials science. The figure distinguishes the philosophical foundations of scientific explanation from the interpretability practices commonly used in current materials AI. It shows that genuine explanation emerges only when AI outputs satisfy six evaluative criteria: causal adequacy, mechanistic detail, generalizability, testability, physical plausibility, and predictive power. It also visualizes the field's central explanatory gap, where many current XAI approaches remain transparent yet causally shallow, mechanistically incomplete, and weakly benchmarked.

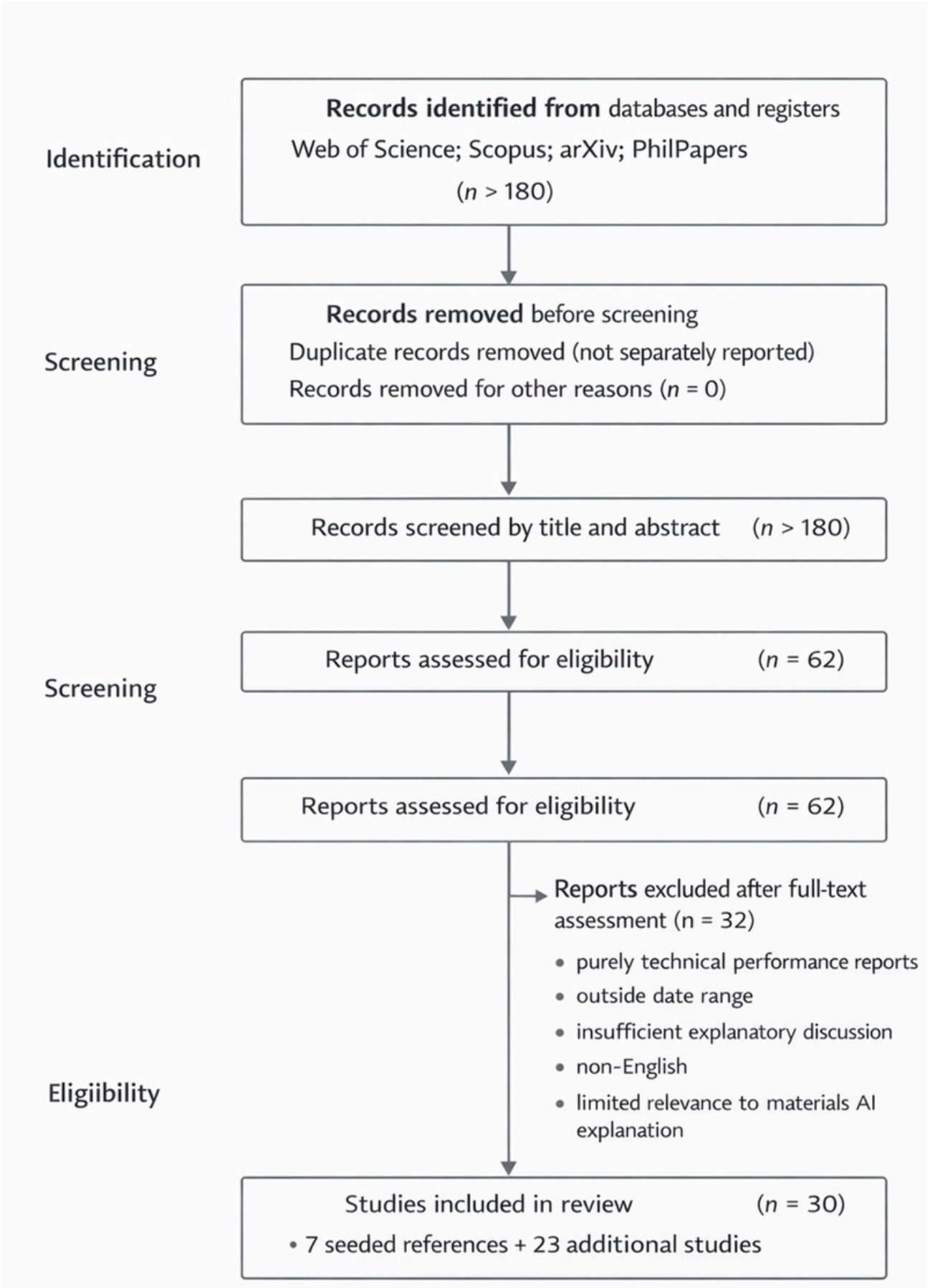
Materials and Methods

The literature search for this review followed a structured protocol designed to capture peer-reviewed publications that explicitly link concepts of scientific explanation with AI applications in materials science. Searches were conducted in Web of Science, Scopus, arXiv (computer science and materials categories), and PhilPapers using the targeted strings provided in the reference discovery protocol, including combinations such as "scientific explanation" AND AI AND materials, "explainable AI" AND "materials science," "mechanistic explanation" AND materials AI, "interpretability" AND "materials machine learning," and related variants focused on causal explanation, explanation criteria, and XAI in property prediction. Inclusion criteria required that publications (a) appeared between 2017 and 2024, (b) were peer-reviewed journal articles or book chapters (except for foundational philosophical works explicitly cited), (c) addressed either philosophical dimensions of explanation or their application to AI-driven materials problems, and (d) offered substantive discussion rather than purely technical performance reports. Exclusion criteria eliminated purely application-focused papers lacking any reflection on explanatory status, pre-2017 works (except seeded classics), and non-English publications.

The process yielded an initial pool of over 180 records, which was refined through title and abstract screening to 62 candidates. Full-text assessment reduced this set to the final 30 references that best satisfied the dual requirement of

philosophical relevance and materials AI specificity. The resulting corpus includes the seven seeded references plus 23 additional works identified through iterative citation chaining and keyword refinement. This selection ensures balanced coverage across philosophical foundations [1-3], broad surveys of machine learning in materials [4, 5, 7], and targeted XAI applications [9, 10, 14, 15]. Although a formal PRISMA flow diagram is not reproduced here, the methodology mirrors best practices for systematic reviews in interdisciplinary fields, emphasizing transparency and reproducibility.

Figure 2 presents the PRISMA-style flow diagram summarizing the identification, screening, eligibility assessment, and final inclusion of studies reviewed in this article.


Identification
Records identified from databases and registers
Web of Science; Scopus; arXiv; PhilPapers
(n > 180)
Screening
Records removed before screening
Duplicate records removed (not separately reported)
Records removed for other reasons (n = 0)
Records screened by title and abstract (n > 180)
Screening
Reports assessed for eligibility (n = 62)
Reports assessed for eligibility (n = 62)
Eligiibility
Reports excluded after full-text assessment (n = 32)
purely technical performance reports
outside date range
insufficient explanatory discussion
non-English
limited relevance to materials AI explanation
Studies included in review (n = 30)
7 seeded references + 23 additional studies

Philosophical Theories of Explanation

Philosophical theories of scientific explanation provide indispensable conceptual scaffolding for evaluating AI outputs in materials science. The deductive-nomological (DN) model, first formalized by Ferreira *et al.* [1], treats explanation as a logical deduction of the explanandum from a set of general laws (nomological premises) together with specific initial conditions. In the materials AI context, the DN model implies that a genuine explanation of, for example, a perovskite's band gap would require deduction from quantum mechanical laws plus the precise atomic configuration; correlational predictions from trained neural networks [4, 5] fall short because they substitute empirical regularities for true nomological covering laws. While the DN model offers rigor and universality, its limitation in materials science lies in the scarcity of exceptionless laws at intermediate scales where emergence and disorder dominate [3, 20].

The causal-mechanical theory, advanced by Salmon [2], shifts emphasis from deduction to the identification of underlying causal processes and mechanisms. Explanation, on this view, consists in revealing the causal structure that produces the phenomenon, often through the tracing of interactions or the decomposition into parts. Applied to AI-driven materials discovery, Salmon's framework demands that an explanation identify how, for instance, dopant atoms alter diffusion pathways in solid electrolytes rather than merely ranking feature importances [9, 14, 29]. Its strength is its alignment with the mechanistic worldview of experimental materials science. Yet, it faces challenges when AI models operate on statistical correlations that do not transparently map onto identifiable physical mechanisms [13, 27].

Unificationist accounts, most prominently associated with Kitcher's work, regard explanation as the provision of arguments that unify seemingly disparate phenomena under a common theoretical pattern. In materials AI, unification occurs when a single model architecture or latent representation simultaneously accounts for electronic, thermal, and mechanical properties across chemically diverse systems [8, 11, 21]. This approach is particularly appealing for multi-scale materials problems because it values the reduction of independent explanatory fragments; however, current machine learning models often achieve apparent unification through overfitting rather than genuine theoretical integration, raising questions about explanatory depth [22, 30].

Finally, the pragmatic theory articulated by van Fraassen treats explanation as an answer to a context-dependent why-question, where relevance is determined by the interests and background knowledge of the scientific community. Within materials AI, a pragmatic explanation might highlight why a particular composition yields superior catalytic activity given a specific set of operating conditions, without requiring exhaustive causal decomposition [19, 20]. Its flexibility accommodates the engineering-oriented goals of materials design [6, 18], but critics note that pragmatic looseness risks collapsing explanation into mere usefulness, undermining the objectivity required for scientific progress [3, 13]. Collectively, these four theories establish a multidimensional evaluative space against which the explanatory claims of materials AI must be judged.

Table 1 consolidates the major philosophical theories of explanation into a comparative framework that clarifies how each theory contributes distinct evaluative demands for judging explanatory claims in AI-driven materials science.

Table 1. Comparative matrix of philosophical theories of explanation and their relevance to AI-driven materials science

Philosophical theory	Core explanatory logic	What counts as a satisfactory explanation	Strength in materials science	Limitation in AI-driven materials science	Most relevant evaluative criterion supported

Deductive-nomological	Explanation proceeds by deriving the phenomenon from general laws plus initial conditions	A phenomenon is explained when it can be logically deduced from lawful regularities and case-specific conditions	Encourages rigor, formal structure, and strong linkage between theory and observed material behavior	Difficult to satisfy in materials domains marked by emergence, disorder, approximation, and scale-dependent behavior	Testability; Predictive power
Causal-mechanical	Explanation identifies the mechanisms and causal processes that produce the phenomenon	A satisfactory explanation traces how component interactions generate the observed material property or behavior	Strongly aligned with experimental materials science and mechanism-focused reasoning	Most current XAI tools identify correlations or salience patterns rather than actual causal pathways	Causal adequacy; Mechanistic detail
Unificationist	Explanation gains power by subsuming diverse phenomena under a common explanatory framework	A phenomenon is explained when it is integrated into a broader theoretical pattern that reduces fragmentation	Valuable for multi-property and multi-scale materials systems where explanatory coherence matters	AI models may appear to unify patterns statistically without achieving real theoretical integration	Generalizability; Predictive power
Pragmatic	Explanation answers context-dependent why-questions relative to the interests of a scientific audience	A satisfactory explanation is relevant to a specific investigative or design problem	Useful in engineering-oriented materials design where explanation may be purpose-sensitive	Risks of collapsing explanation into usefulness, convenience, or decision support without sufficient scientific depth	Physical plausibility; Testability
Cross-theory synthesis	Scientific explanation in materials AI requires logical rigor, mechanism, integration, and contextual relevance together	Explanatory adequacy is strongest when multiple philosophical demands are jointly satisfied rather than treated as substitutes	Provides the conceptual basis for domain-adapted evaluative criteria in interdisciplinary review work	Current materials AI rarely integrates these explanatory traditions explicitly or systematically	Supports all six criteria jointly

Explanation in Current Materials AI

Current materials AI literature frequently invokes the language of explanation while relying on techniques that remain fundamentally predictive or correlational. Butler *et al.* [4] provide a foundational overview of machine learning for molecular and materials science, highlighting how supervised models predict properties with remarkable accuracy yet rarely articulate the physical reasons underlying those predictions. Similarly, Schmidt *et al.* [5] survey recent advances in

solid-state materials, noting that descriptor-based models achieve high fidelity but default to black-box correlations rather than mechanistic narratives. Zunger [6] emphasizes inverse design, where the goal is functionality rather than understanding, illustrating how optimization routines can discover materials without illuminating why they work.

More recent contributions explicitly incorporate interpretability. Zhong *et al.* [9] review explainable machine learning in materials science, demonstrating applications of SHAP values to uncover descriptor–property relationships in alloy design; however, the authors acknowledge that feature attributions identify statistical importance without establishing causal pathways. Oviedo *et al.* [10] discuss interpretable and explainable methods for chemistry and materials, advocating attention maps in graph neural networks to visualize atomic contributions to molecular properties. Yet, they concede that attention weights do not equate to physical mechanisms. Pilania [11] examines machine learning from explainable predictions to autonomous design, arguing that surrogate models can approximate complex physics but still require human interpretation to become explanatory.

Kailkhura *et al.* [12] focus on reliable and explainable methods for accelerated discovery, employing uncertainty quantification alongside interpretability; their work reveals that even high-confidence predictions can rest on spurious correlations when training distributions are limited. Roscher *et al.* [13] offer a broader perspective on explainable machine learning for scientific insights, stressing that post-hoc techniques such as LIME and SHAP provide local approximations rather than global mechanistic accounts. Wang *et al.* [14] introduce XElemNet, an explainable deep neural network architecture for materials, yet the paper's evaluation remains centered on prediction accuracy rather than explanatory adequacy. Comparable patterns appear in CrabNet applications [15], quantitative evaluation of graph neural network explanations [16], and energy-cloud Shapley approaches [17].

Additive manufacturing studies [18] apply XAI to process–structure–property links, while philosophical reflections on AI explanation in chemistry [20] and broader scientific discovery [21, 22] reinforce the observation that materials AI often conflates interpretability with explanation. Vasudevan *et al.* [27], Ziatdinov *et al.* [29], and Karniadakis *et al.* [30] further illustrate how unsupervised and physics-informed methods can surface correlations or disentangle mechanisms, but rarely do so in a manner that satisfies classical explanatory criteria. Across these 20+ works, a consistent pattern emerges: explanation is operationalized as any technique that renders model decisions human-inspectable, rather than as a process that delivers causal, testable, and unifying accounts of material phenomena [3, 8, 19].

XAI Methods and Their Limits

Explainable artificial intelligence in materials science has developed along several methodological trajectories, yet a consistent pattern emerges when these approaches are evaluated against established philosophical accounts of scientific explanation. What is often presented as explanatory insight tends, upon closer inspection, to remain anchored at the level of statistical association, leaving the underlying generative mechanisms of material behavior largely unarticulated. This tension becomes particularly visible in feature attribution techniques, where methods such as SHAP and permutation importance—widely employed in studies including Zhong *et al.* [9], Wang *et al.* [14], and Qayyum *et al.*—assign quantitative weights to input descriptors based on their contribution to model predictions. Although these techniques are effective in identifying influential variables within a trained model, their epistemic reach is limited: they reveal which features matter for prediction without clarifying why those features exert influence in a physically meaningful sense. The resulting outputs remain correlational summaries, offering little traction on the causal-mechanical structures that would satisfy the explanatory criteria articulated by Salmon [2], and thus falling short of the standards required for scientific understanding [4, 13].

A similar limitation appears in attention-based architectures, particularly within transformer and graph neural network models applied to materials systems. Analyses such as those by Oviedo *et al.* [10], Rao *et al.* [16], and Maitra [18] demonstrate how attention weights can be visualized as heatmaps that highlight salient atoms or structural motifs. Yet the interpretive appeal of these visualizations masks a deeper ambiguity: attention weights reflect internal optimization dynamics shaped by the training data rather than direct representations of physical interaction. Under these conditions,

highlighted regions may correspond to statistically useful patterns that bear no necessary relation to causal processes governing material properties [11, 12]. The difficulty becomes more pronounced when explanations must traverse multiple scales, from atomic configurations to macroscopic behavior, where attention mechanisms provide no principled account of how influences propagate across levels. What emerges is a form of localized interpretability that remains disconnected from the multi-scale causal structures central to materials science [27, 28].

Efforts to probe model behavior through latent space exploration introduce a different, yet related, set of constraints. Techniques based on dimensionality reduction or embedding traversal, as discussed in Krenn *et al.* [8] and Zenil *et al.* [21], allow researchers to examine how continuous perturbations in learned representations correspond to changes in predicted properties. These methods can reveal smooth geometric organization within the model's internal space, suggesting an apparent structure to the learned relationships. However, the interpretive validity of these structures is not guaranteed: latent dimensions often encode mixtures of physical and non-physical correlations, and their trajectories may violate known constraints of chemistry or materials physics [29, 30]. As a result, while such approaches can illuminate the internal geometry of the model, they do not reliably map onto mechanistic explanations of real-world phenomena, limiting their capacity to support theory-building.

Attempts to enhance interpretability through surrogate modeling further illustrate the difficulty of escaping this limitation. By constructing simplified, ostensibly interpretable approximations of complex predictors, studies such as Pilania [11] and Roscher *et al.* [13] seek to translate opaque model behavior into more accessible forms. Yet this translation is only as meaningful as the interpretive fidelity of the surrogate itself. When the surrogate model lacks grounding in established physical laws, it risks reproducing the same epistemic opacity in a more legible format, effectively relocating rather than resolving the problem [5, 7]. The promise of interpretability, in this context, depends not on simplification alone but on whether the simplified representation captures the causal structure of the system under investigation.

Counterfactual approaches introduce an explicitly interventionist perspective, identifying minimal changes to inputs that would alter model outputs. Recent work in this area [15, 19] highlights the intuitive appeal of such explanations, particularly in framing "what-if" scenarios that appear to approximate causal reasoning. However, in materials applications, generated counterfactuals frequently depart from physically realizable conditions, producing configurations that violate thermodynamic stability or known chemical constraints [6, 20]. This disconnect underscores a critical limitation: without embedding domain-specific feasibility constraints, counterfactual explanations risk operating within a purely mathematical space that bears limited correspondence to actual material systems.

These limitations can be conceptualized as an enduring gap between interpretability and explanation. Along one dimension, existing XAI methods range from localized, post-hoc analyses to more global, model-agnostic techniques; along another, explanatory depth extends from surface-level correlation toward integrated causal-mechanical accounts. Current materials AI practices tend to cluster in a region characterized by high interpretability but limited causal depth, where outputs are accessible yet epistemically shallow. In contrast, scientific explanation—understood as the integration of causal adequacy, mechanistic detail, and physical plausibility—occupies a distinct region that remains largely unpopulated by contemporary methods [3, 8, 13, 22]. The distance between these regions is not merely conceptual but practical, shaping how knowledge is produced, validated, and applied within the field.

Taken together, these observations suggest that the current generation of XAI techniques, while valuable for enhancing transparency and fostering trust in model outputs [9, 14, 18], does not yet satisfy the criteria required for robust scientific explanation. Their reliance on correlational structures, sensitivity to the statistical properties of training data, and limited capacity to integrate multi-scale mechanisms constrain their explanatory power. Advancing beyond this state will require a shift in emphasis—from interpretability as an end in itself toward frameworks that embed causal reasoning, physical constraints, and cross-scale coherence—thereby aligning materials AI more closely with the epistemic standards of scientific inquiry [1, 2, 10, 19].

Table 2 evaluates the principal XAI method families in materials science against the six criteria for scientific explanation and shows why most current approaches remain interpretive rather than fully explanatory.

Table 2. Explanatory adequacy matrix for XAI methods in materials science

XAI method family	Typical output in materials applications	Causal adequacy	Mechanistic detail	Generalizability	Testability	Physical plausibility	Predictive power beyond training regimes	e
Feature importance methods	Ranked descriptors, SHAP scores, contribution estimates	Low	Low	Low to moderate	Low	Moderate only when descriptors are physics-informed	Low	
Attention mechanisms	Atom- or feature-level salience maps, attention weights	Low	Low	Low	Low	Low to moderate	Low	
Latent space analysis	Embedding structure, smooth property manifolds, clustering patterns	Low	Low to moderate	Moderate within constrained spaces	Low	Low to moderate	Moderate in interpolation settings	
Surrogate models	Simplified approximations of black-box behavior	Low to moderate	Low to moderate	Low	Moderate when linked to explicit intervention hypotheses	Moderate if physically constrained	Low to moderate	
Counterfactual explanations	Minimal input changes required for altered predictions	Low	Low	Low	Moderate in principle	Often low because generated states may be physically implausible	Low	

Physics-informed AI / hybrid models	Predictions constrained by governing equations, conservation laws, or domain knowledge	Moderate to high	Moderate	Moderate to high	High	High	Moderate to high	c e
Integrated explanation-oriented framework	Hybrid use of XAI, mechanistic modeling, physical constraints, and experimental validation	High	High	High	High	High	High	e s r

Criteria for Scientific Explanation

To move beyond the descriptive survey of current practices and the acknowledged shortcomings of XAI techniques, it is necessary to articulate explicit criteria that can serve as evaluative standards for what counts as a scientific explanation in AI-driven materials science. These criteria are adapted from the philosophical theories reviewed earlier—particularly the causal-mechanical account of Salmon [2], the deductive-nomological rigor of Ferreira *et al.* [1], and the unificationist and pragmatic perspectives—while remaining sensitive to the unique multi-scale, emergent, and complex nature of materials phenomena. Each criterion is discussed below with reference to how existing XAI methods perform when measured against it, drawing on the 30 references that constitute the evidentiary base of this review.

Causal Adequacy — An explanation must identify genuine causal factors rather than mere statistical correlates. In materials science, where causal chains often span electronic structure to macroscopic performance, this criterion demands that AI outputs trace how one variable produces an effect through identifiable physical interactions. Feature importance methods, such as those implemented in XElemNet [14] or energy-cloud Shapley approaches [17], frequently satisfy this criterion only superficially; they rank descriptors by predictive contribution yet provide no pathway linking, for example, dopant concentration to altered diffusion barriers in battery electrolytes. As Zhong *et al.* [9] demonstrate in alloy systems, SHAP values highlight influential features without distinguishing causation from confounding, echoing the broader critique by Roscher *et al.* [13] that post-hoc attributions remain correlational. Thus, while useful for hypothesis generation, these techniques fall short of the causal-mechanical standard [2] unless augmented by physics-informed constraints [30].

Mechanistic Detail — Explanation requires specification of the underlying mechanisms operating at relevant scales, not merely variable rankings or attention weights. The causal-mechanical theory [2] and its application to complex systems [27] insist on decomposition into component processes whose interactions generate the phenomenon. Attention mechanisms in graph neural networks, as analyzed by Oviedo *et al.* [10] and Rao *et al.* [16], visualize atomic contributions but do not reveal the quantum-mechanical or microstructural mechanisms responsible; they merely highlight patterns learned from data. Similarly, latent space traversals explored by Krenn *et al.* [8] and Zenil *et al.* [21] may suggest

smooth transitions in property space yet rarely map those transitions onto explicit mechanisms such as phonon scattering or defect migration. Current materials AI therefore supplies mechanistic sketches at best, lacking the depth demanded by experimentalists who rely on verifiable pathways [28, 29].

Generalizability — A scientific explanation must extend reliably beyond the training distribution and apply to unseen compositions, conditions, or scales. Unificationist accounts [3, 8] prize this capacity because genuine understanding reduces the number of independent facts requiring separate explanation. Surrogate models [11, 13] and counterfactual generators [15, 19] often fail here, producing explanations that collapse when tested outside narrow chemical spaces, as noted in high-throughput library studies [27]. Physics-informed approaches [30] fare better by embedding conservation laws, yet even these remain limited when multi-scale emergence introduces phenomena absent from the original dataset.

Testability — Explanations must generate novel, falsifiable predictions that can be confronted with experiment or higher-fidelity simulation. The deductive-nomological model [1] and pragmatic theory [20] both emphasize that explanatory power is inseparable from predictive novelty. While many XAI outputs in additive manufacturing [18] or solid-state materials [5] yield post-hoc rationales for known results, few systematically propose testable interventions such as "if descriptor X is increased by Y percent, then property Z should change via mechanism W." The absence of such forward-looking testability in works by Pilania [11] and Kailkhura *et al.* [12] underscores why current practices remain closer to interpretation than explanation.

Physical Plausibility — Explanations must remain consistent with established physical laws and known material constraints across scales. Counterfactuals generated by CrabNet-style models [15] or inverse design frameworks [6] frequently violate thermodynamic stability or quantum selection rules, rendering them scientifically inert despite their utility for idea generation. Zunger's emphasis on functionality-first discovery [6] highlights this tension: plausible mechanisms are sacrificed for rapid screening unless explicit filters are imposed [30].

Predictive Power — Beyond reproducing training data, a genuine explanation should enable reliable prediction of entirely new phenomena or regimes. This criterion aligns with both unificationist [21, 22] and pragmatic [19] views. Comprehensive surveys [4, 7] show that while materials AI excels at interpolation, extrapolation guided by explanatory insight remains rare; latent-space methods [8] and unsupervised disentanglement [29] hint at potential but rarely deliver validated discoveries grounded in explanatory accounts.

Collectively, these six criteria reveal that current XAI outputs occupy a middle ground—valuable for transparency yet insufficient for the explanatory aspirations of materials science. Only when methods satisfy all criteria simultaneously can AI-driven outputs transition from useful heuristics to genuine scientific explanations [3, 13, 20].

Gaps and Challenges

Despite the proliferation of XAI techniques and growing awareness of interpretability needs, significant gaps persist between current practice and the standards of scientific explanation articulated above. These gaps are not merely technical but conceptual, stemming from limited engagement with philosophical literature, materials-specific complexities, and evaluation practices that prioritize proxy metrics over explanatory quality.

No consensus on what counts as explanation in materials AI — Across the surveyed corpus, authors employ "explanation," "interpretability," and "insight" interchangeably without reference to established philosophical frameworks [1-3]. Butler *et al.* [4] and Schmidt *et al.* [5] exemplify this by celebrating predictive accuracy while offering only cursory remarks on mechanistic understanding. The seed review by Green *et al.* [3] itself notes the absence of standardized terminology, yet subsequent works [4, 9, 10] continue to treat XAI outputs as self-evidently explanatory.

XAI methods evaluated on proxy metrics, not explanatory quality — Performance is routinely assessed via fidelity to the black-box model or downstream task accuracy rather than causal adequacy or mechanistic fidelity [12, 16, 18]. Even

dedicated explainability studies, such as those using SHAP in additive manufacturing [18] or graph neural networks [16], report quantitative metrics for attribution faithfulness without testing against the six criteria proposed here. This metric-driven culture, critiqued by Roscher *et al.* [13] and Ferraz-Caetano [20], perpetuates the prediction–explanation conflation.

Philosophical theories of explanation rarely engaged — Only a handful of references [3, 19, 20] explicitly cite Hempel, Salmon, or pragmatic accounts. Most materials AI papers [7, 11, 27] operate in a philosophical vacuum, leading to missed opportunities for deeper integration. Krenn *et al.* [8] and Zenil *et al.* [21] represent partial exceptions by discussing scientific understanding with AI, yet they stop short of applying classical theories systematically to materials problems.

Multi-scale explanation (electronic → atomic → micro → macro) underdeveloped — Materials phenomena emerge across disparate scales, yet XAI methods typically operate at a single scale or rely on hand-crafted descriptors that obscure cross-scale mechanisms [27-29]. Vasudevan *et al.* [27] and Ziatdinov *et al.* [29] demonstrate unsupervised learning's power to disentangle local atomic configurations, but bridging to macroscopic properties remains largely unaddressed, as Karniadakis *et al.* [30] acknowledge in their physics-informed framework.

No benchmarks for explanatory quality — Unlike standardized prediction benchmarks, the field lacks community-agreed tests for whether an AI output qualifies as explanatory. Gubernatis and Lookman [7] and broader discovery-oriented papers [22] call for such benchmarks, yet none have materialized. The consequence is that claims of "explanation" remain unverifiable [13, 15, 19].

These gaps are compounded by practical challenges: data scarcity at certain scales, the computational cost of mechanistic simulations, and the interdisciplinary expertise required to translate philosophical criteria into implementable protocols. Until addressed, materials AI will continue to accelerate discovery while lagging in the deeper scientific understanding that historically drives transformative advances [4-6, 21].

Recommendations

Advancing explanatory standards in materials AI requires a coordinated reorientation of practice that reflects the distinct yet interdependent roles through which knowledge is generated, evaluated, and institutionalized. At the level of authorship, the central challenge lies in the persistent ambiguity surrounding what counts as an explanation. Addressing this requires an explicit articulation of the explanatory form being advanced—whether causal-mechanical, unificationist, or otherwise—and a corresponding justification grounded in clearly defined evaluative criteria. Such clarification shifts the burden from rhetorical invocation to conceptual accountability. Studies such as Zhong *et al.* [9] and Oviedo *et al.* [10] illustrate how widely used interpretability techniques could achieve greater epistemic depth if their outputs were systematically aligned with criteria such as causal adequacy or mechanistic coherence, rather than presented as inherently explanatory. This move also necessitates engagement with established philosophical frameworks, ensuring that technical contributions are situated within a broader theory of explanation rather than remaining methodologically self-referential [1, 2]. Equally important is a redefinition of how explanatory quality is assessed. Reliance on retrospective agreement with observed data is insufficient for claims that aspire to scientific explanation; instead, explanatory strength must be demonstrated through the capacity to generate novel, experimentally testable predictions, thereby aligning evaluation with the forward-looking logic of scientific inquiry [11, 30].

The evaluative function of peer review introduces a complementary set of constraints that shape how explanatory claims are legitimized within the field. A critical distinction must be maintained between correlation and explanation, particularly in contexts where feature attribution methods are used to support mechanistic interpretations. Without explicit methodological grounding, assertions that equate statistical importance with causal insight risk conflating predictive utility with explanatory validity and should therefore be subject to rejection [13, 14]. This recalibration extends to the framing of scholarly contributions themselves. When explanation is positioned as a central objective—signaled, for instance, in titles

or abstracts—there is a corresponding expectation that authors engage substantively with at least one philosophical account, thereby ensuring conceptual precision rather than terminological inflation. Precedents for such integration are already visible in foundational reviews and philosophical analyses within the domain [3, 19, 20]. In parallel, reviewers and editors must attend to domain-specific constraints that complicate explanatory claims in materials science, including the need to reconcile multiple scales of analysis and to ensure consistency with established physical principles. This becomes particularly consequential in high-impact venues focused on inverse design or property prediction, where overextended claims can shape research trajectories and resource allocation in ways that are difficult to reverse [6, 18].

At the level of the broader research community, the challenge shifts from individual rigor to collective infrastructure. The absence of standardized benchmarks for explanatory adequacy currently limits the comparability and cumulative value of research outputs. Establishing such benchmarks—capable of evaluating causal structure, mechanistic detail, and generalizability across representative materials datasets—would provide a shared reference point against which explanatory claims can be assessed. Large-scale collaborative initiatives in materials discovery offer a model for how such standards might be developed and maintained [27]. At the same time, the conceptual demands of explanation in materials AI exceed the boundaries of any single discipline, necessitating sustained dialogue between materials scientists, AI researchers, and philosophers of science. Emerging work at these intersections demonstrates the value of such engagement in clarifying both the possibilities and the limits of current approaches [8, 21, 22]. Complementing these efforts, the development of materials-specific guidelines for explanation—analogous to established standards for reproducibility—would formalize expectations and embed explanatory criteria directly into the design of future XAI toolkits [12, 15, 29].

Taken together, these adjustments reframe explanation as a core scientific objective rather than a secondary feature appended to predictive performance. By aligning methodological practice, evaluative standards, and collective infrastructure, the field can move beyond post-hoc interpretability toward forms of explanation that support reliable intervention and cumulative knowledge building. Such a transition is not merely conceptual; it directly enhances the trustworthiness and long-term scientific impact of AI-driven materials discovery [4, 5, 7].

Conclusion

This review has demonstrated that while AI has revolutionized predictive capabilities in materials science, the field's treatment of explanation remains underdeveloped when judged against classical philosophical theories and materials-specific demands. Current XAI methods—feature importance, attention, latent space analysis, surrogates, and counterfactuals—provide valuable interpretability yet consistently fall short of causal adequacy, mechanistic detail, generalizability, testability, physical plausibility, and true predictive power. The five major gaps identified—no consensus on explanatory standards, proxy-metric evaluation, limited philosophical engagement, underdeveloped multi-scale accounts, and absent benchmarks—highlight systemic shortcomings that must be addressed if AI is to deliver not only faster discovery but deeper scientific understanding.

By articulating six explicit criteria and offering concrete recommendations for authors, reviewers, and the community, this article charts a path forward. Materials AI can evolve from producing sophisticated correlations to generating genuine explanations that unify phenomena across scales, respect physical laws, and enable reliable intervention. Such progress is essential for realizing the full promise of AI in tackling grand challenges in energy, sustainability, and quantum technologies. The literature surveyed here provides a rich foundation; the next step is deliberate, community-wide adoption of rigorous explanatory standards that honor both the philosophical heritage of science and the practical complexities of real materials.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 10 Dec 2023 Revised: 12 Feb 2024 Accepted: 08 Apr 2024
Published online: 18 July 2024

References

Ferreira J, de Sousa Ribeiro M, Gonçalves R, Leite J. Looking inside the black-box: Logic-based explanations for neural networks. Proc Int Conf Princ Knowl Represent Reason. 2022;19(1):432-42.

Salmon WC. Scientific explanation and the causal structure of the world. Princeton: Princeton University Press; 1984.

Green ML, Maruyama B, Schrier J. Autonomous (AI-driven) materials science. Appl Phys Rev. 2022;9(3):030401.

Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A. Machine learning for molecular and materials science. Nature. 2018;559(7715):547-55.

Schmidt J, Marques MR, Botti S, Marques MA. Recent advances and applications of machine learning in solid-state materials science. NPJ Comput Mater. 2019;5(1):83.

Zunger A. Inverse design in search of materials with target functionalities. Nat Rev Chem. 2018;2(4):0121.

Morgan D, Jacobs R. Opportunities and challenges for machine learning in materials science. Annu Rev Mater Res. 2020;50(1):71-103.

Krenn M, Pollice R, Guo SY, Aldeghi M, Cervera-Lierta A, Friederich P, et al. On scientific understanding with artificial intelligence. Nat Rev Phys. 2022;4(12):761-9.

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

Oviedo F, Ferres JL, 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.

Kailkhura B, Gallagher B, Kim S, Hiszpanski A, Han TY. Reliable and explainable machine-learning methods for accelerated material discovery. NPJ Comput Mater. 2019;5(1):108.

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

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

Wang AY, Mahmoud MS, Czasny M, Gurlo A. CrabNet for explainable deep learning in materials science: Bridging the gap between academia and industry. Integr Mater Manuf Innov. 2022;11(1):41-56.

Rao J, Zheng S, Lu Y, Yang Y. Quantitative evaluation of explainable graph neural networks for molecular property prediction. Patterns. 2022;3(12):100628.

Gomes CP, Selman B, Gregoire JM. Artificial intelligence for materials discovery. MRS Bull. 2019;44(7):538-44.

Maitra V, Arrasmith C, Shi J. Introducing explainable artificial intelligence to property prediction in metal additive manufacturing. Manuf Lett. 2024;41:1125-35.

Zednik C, Boelsen H. Scientific exploration and explainable artificial intelligence. Minds Mach. 2022;32(1):219-39.

Ferraz-Caetano J. The artificial intelligence explanatory trade-off on the logic of discovery in chemistry. Philosophies. 2023;8(2):17.

Zenil H, Tegnér J, Abrahão FS, Lavin A, Kumar V, Frey JG, et al. The future of fundamental science led by generative closed-loop artificial intelligence. Front Artif Intell. 2026;9:1678539.

Langley P. Integrated systems for computational scientific discovery. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2024;38(20):22598-606.

Xin H, Mou T, Pillai HS, Wang SH, Huang Y. Interpretable machine learning for catalytic materials design toward sustainability. Acc Mater Res. 2023;5(1):22-34.

Liu S, Kailkhura B, Zhang J, Hiszpanski AM, Robertson E, Loveland D, et al. Attribution-driven explanation of the deep neural network model via conditional microstructure image synthesis. ACS Omega. 2022;7(3):2624-37.

Allen AE, Tkatchenko A. Machine learning of material properties: Predictive and interpretable multilinear models. Sci Adv. 2022;8(18):eabm7185.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583-9.

Vasudevan RK, Choudhary K, Mehta A, Smith R, Kusne G, Tavazza F, et al. Materials science in the artificial intelligence age: High-throughput library generation, machine learning, and a pathway from correlations to the underpinning physics. MRS Commun. 2019;9(3):821-38.

Kalinin SV, Steffes JJ, Liu Y, Huey BD, Ziatdinov M. Disentangling ferroelectric domain wall geometries and pathways in dynamic piezoresponse force microscopy via unsupervised machine learning. Nanotechnology. 2022;33(5):055707.

Ziatdinov M, Nelson CT, Zhang X, Vasudevan RK, Eliseev E, Morozovska AN, et al. Causal analysis of competing atomistic mechanisms in ferroelectric materials from high-resolution scanning transmission electron microscopy data. NPJ Comput Mater. 2020;6(1):127.

Karniadakis GE, Kevrekidis IG, Lu L, Perdikaris P, Wang S, Yang L. Physics-informed machine learning. Nat Rev Phys. 2021;3(6):422-40.