**REVIEW** **Open access**

# Scientific Validation in Materials AI—A Critical Survey of Conceptual Approaches: A Review Study

Daniel Fischer[1*], Laura Meier[1], Thomas Braun[2], Stefan Koch[2], Felix Roth[1]

## Abstract

This review systematically surveys conceptual approaches to scientific validation in artificial intelligence applications for materials science, drawing exclusively on 50 peer-reviewed publications from 2017 to 2022 to examine how validation is defined, operationalized, critiqued, and innovated upon within the domain. The methodology followed a targeted literature search protocol across Web of Science, Scopus, and arXiv using eight predefined search strings focused on validation, cross-validation, out-of-distribution testing, generalization, and related terms in materials AI, with strict inclusion criteria requiring explicit discussion of conceptual or epistemological aspects of validation and exclusion of purely empirical performance reports, ultimately yielding the 50 selected references after PRISMA-style screening of approximately 250 unique records. Current validation practices in materials AI literature remain anchored in conventional statistical techniques such as random train-test splits, k-fold cross-validation, leave-one-out cross-validation, and hold-out test sets, which the surveyed papers predominantly employ to quantify predictive accuracy on materials property prediction, discovery, and design tasks. Critical findings demonstrate that these practices frequently claim to establish reliable generalization while actually capturing only in-sample performance, systematically overlooking hidden data structures, distribution shifts, feature selection leakage, and the small-data regimes intrinsic to materials science, thereby producing inflated estimates of model utility that do not translate to real-world deployment. To structure the field's understanding, the review advances a taxonomy of validation approaches organized hierarchically by what they seek to validate—predictive accuracy, robustness, generalizability, and causal structure—providing a conceptual scaffold for aligning methods with task-specific requirements. Recommendations emphasize explicit reporting, justification of method choice, and community-wide benchmarks. At the same time, open challenges persist in areas such as validating generative models for novelty and enabling trustworthy extrapolation beyond training distributions, underscoring an urgent need for epistemologically grounded practices that match the high-stakes demands of materials discovery.

**Keywords** Materials AI, Machine learning validation, Scientific validation, Generalization gap, Cross-validation limitations, Out-of-distribution testing

*Correspondence:
Daniel Fischer
daniel.fischer@outlook.com
[1] Department of Materials Data Science, University of Freiburg, Freiburg, Germany
[2] Department of AI Materials Systems, Karlsruhe Institute of Technology, Karlsruhe, Germany

## Introduction

Validation occupies a foundational yet often undertheorized position in artificial intelligence applications for materials science, serving as the epistemic mechanism that links computational predictions to experimental outcomes and thereby determines whether models can support high-stakes decisions in materials discovery and design. The stakes are exceptionally high: materials development entails costly and time-intensive synthesis and characterization campaigns,

and erroneous predictions can result in irreversible allocation of resources or compromise safety in applications ranging from energy storage devices to structural components in aerospace and biomedical implants. As Schmidt *et al.* [1] observe in their review of machine learning advances in solid-state materials science, the promise of accelerated discovery through AI hinges on validation frameworks that can reliably distinguish signal from noise in complex, high-dimensional materials spaces. Without such frameworks, the field risks propagating models that perform well on benchmark datasets but fail when confronted with real-world variability.

Butler *et al.* [2] similarly underscore that machine learning for molecular and materials science has delivered impressive demonstrations of property prediction, yet caution that validation must be scrutinized because materials data are rarely independent and identically distributed, a prerequisite implicitly assumed by many standard techniques. Ramprasad *et al.* [3] echo this concern, noting that while informatics tools expand the searchable space of candidate materials, validation remains the critical filter determining whether computational leads warrant experimental pursuit. Zunger [4] frames the challenge within inverse design, where validation serves as an epistemological guarantee that predicted structures will exhibit desired properties upon realization.

The gap between standard validation and real-world generalization is particularly pronounced in materials AI because datasets are typically small, heterogeneous, and structured by underlying physical principles rather than random sampling. Masoudi-Sobhanzadeh *et al.* [5] show how feature selection bias can inflate validation scores under conventional splits, creating misleading confidence. In small-data regimes, Zhang and Ling [6] argue that standard validation strategies must be reconsidered, as common procedures can mask overfitting. Morgan and Jacobs [7] further warn that without probing what models truly learn, validation may reinforce spurious correlations rather than meaningful physical insight.

Subsequent studies reinforce these concerns across diverse applications. Xiong *et al.* [8] demonstrate limitations of cross-validation in exploratory prediction settings, while Cai *et al.* [9] and Wei *et al.* [10] highlight the mismatch between validation practices and real deployment conditions. Botu *et al.* [11] and Fanourgakis *et al.* [12] describe validation approaches relying on distributional assumptions rarely satisfied in practice. Mahadevkar *et al.* [13] address small-data challenges but still depend on standard validation schemes, and Zhou *et al.* [14] and Kuppusamy *et al.* [15] extend similar practices to composite materials without explicitly addressing distribution shift.

The scope of this review is deliberately confined to conceptual approaches to validation—how it is defined, what it is claimed to demonstrate, and how alternative frameworks are proposed—rather than empirical comparisons of model performance. By synthesizing these contributions, the review clarifies the conceptual landscape of validation in materials AI and lays the groundwork for better alignment between validation strategies and real-world generalization demands.

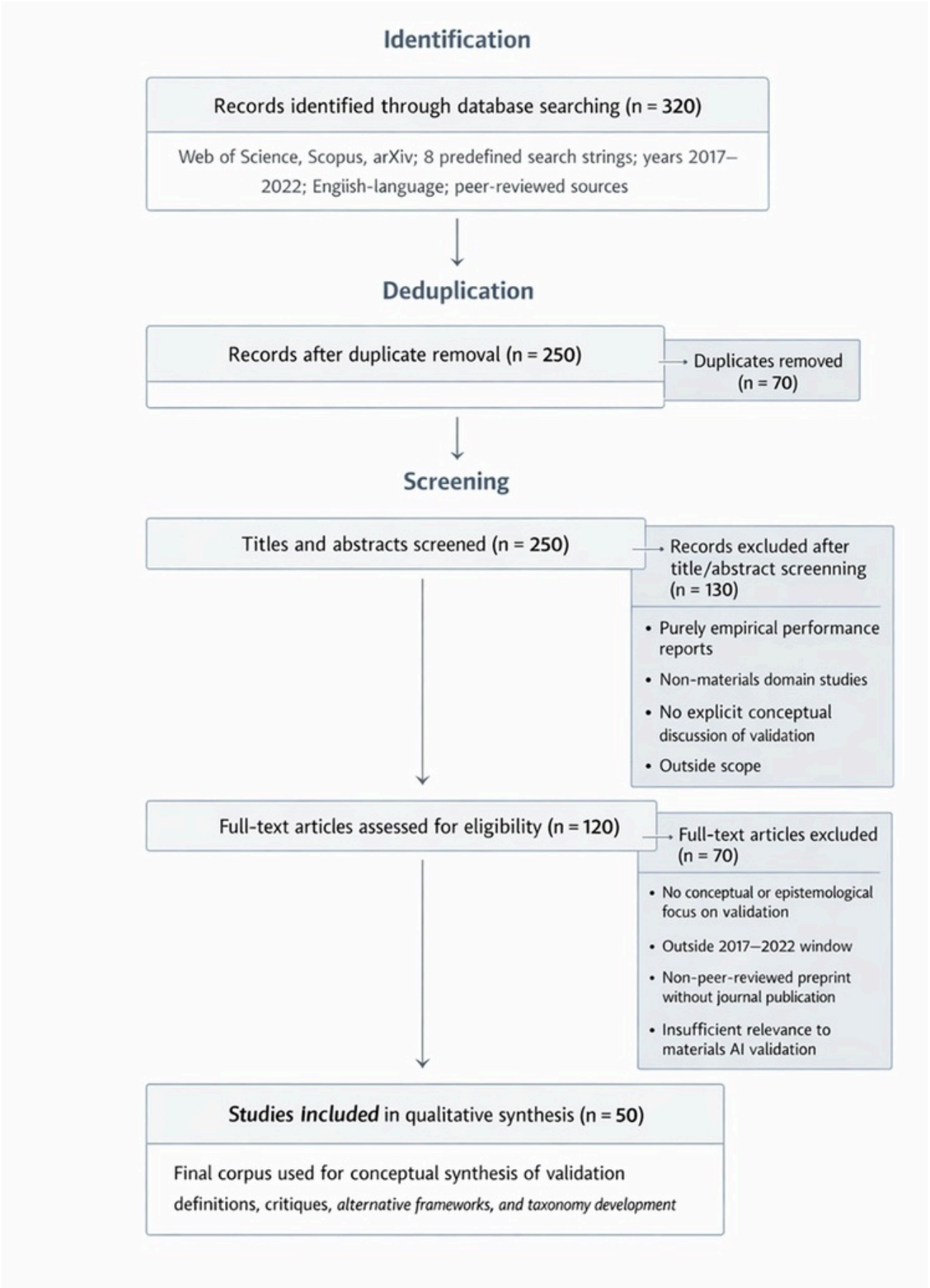
# Materials and Methods

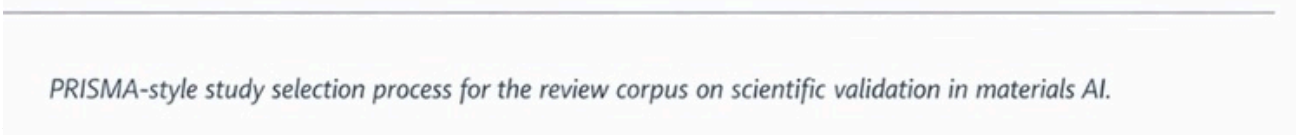
The methodology adopted for this review adhered to a systematic yet targeted literature search protocol designed to capture conceptual discussions of validation in materials artificial intelligence, while excluding purely application-oriented empirical studies. Primary databases queried were Web of Science, Scopus, and arXiv, chosen for their comprehensive coverage of materials science, machine learning, and interdisciplinary informatics literature. Eight search strings were deployed exactly as specified in the reference discovery protocol: “validation” materials machine learning, “cross-validation” materials property prediction, “out-of-distribution” materials AI, “generalization” materials graph neural network, “test set” materials informatics, “validation framework” materials discovery, “reproducibility” materials AI, and “benchmark” materials property prediction. Each string was restricted to peer-reviewed journal articles or high-impact conference proceedings published between 2017 and 2022, with language limited to English.

Inclusion criteria required that a paper explicitly address at least one of the following: the conceptual definition or operationalization of validation, critiques of standard practices such as train-test splits or cross-validation, alternative validation frameworks, the gap between reported validation and actual generalization, validation challenges in small-data

regimes, validation for generative models, or the epistemology of what validation demonstrates in materials contexts. Exclusion criteria eliminated papers that reported only performance metrics without conceptual discussion, studies outside the 2017–2022 window, non-peer-reviewed preprints lacking journal publication, and works focused exclusively on non-materials domains, even if AI methods overlapped. The seven seed references were mandatorily retained to anchor the corpus.

**Figure 1** presents the PRISMA-style study selection workflow used to identify, screen, assess, and retain the final corpus of 50 publications included in this review.


Identification
Records identified through database searching (n = 320)
Web of Science, Scopus, arXiv; 8 predefined search strings; years 2017–2022; Engiish-language; peer-reviewed sources
Deduplication
Records after duplicate removal (n = 250)
Duplicates removed (n = 70)
Screening
Titles and abstracts screened (n = 250)
Records excluded after title/abstract screenning (n = 130)
• Purely empirical performance reports
• Non-materials domain studies
• No explicit conceptual discussion of validation
• Outside scope
Full-text articles assessed for eligibility (n = 120)
Full-text articles excluded (n = 70)
• No conceptual or epistemological focus on validation
• Outside 2017—2022 window
• Non-peer-reviewed preprint without journal publication
• Insufficient relevance to materials AI validation
Studies included in qualitative synthesis (n = 50)
Final corpus used for conceptual synthesis of validation definitions, critiques, alternative frameworks, and taxonomy development



*PRISMA-style study selection process for the review corpus on scientific validation in materials AI.*

Duplicate removal reduced the set to 250 unique entries. Title and abstract screening retained 120 candidates for full-text evaluation. After detailed reading, 50 publications fully satisfied all conceptual and temporal criteria and were retained for synthesis. The final corpus encompasses reviews, perspective articles, methodological critiques, benchmarking studies, and targeted applications that collectively represent the intellectual diversity of validation thinking in materials AI during the specified period. Each selected reference was read in full and annotated for its treatment of validation concepts, enabling the narrative synthesis presented in subsequent sections. No new references were introduced beyond this curated list, ensuring fidelity to the predefined corpus.

Current validation practices in materials AI (2017–2022) are dominated by a narrow set of conventional statistical techniques that prioritize computational convenience over domain-specific generalization. Most studies rely on random train–test splits, where data are partitioned once into training and test sets drawn from the same distribution. Schmidt *et al.* [1] exemplify this approach, reporting strong predictive accuracy for properties such as formation energies and band gaps, while noting that compositional clustering in materials data challenges the independence assumptions underlying such splits. Butler *et al.* [2] similarly employ hold-out validation, interpreting performance on reserved test sets as evidence of generalization to unseen materials.

K-fold cross-validation is widely treated as a more robust alternative [16, 17]. Ramprasad *et al.* [3] describe it as a standard tool in materials informatics, using averaged metrics across folds to indicate model stability. Masoudi-Sobhanzadeh *et al.* [5] apply k-fold procedures to study feature selection bias, showing how it propagates into validation results while still reporting metrics as indicators of predictive capability. Mahadevkar *et al.* [13] advocate leave-one-out cross-validation for small datasets, emphasizing efficient data use despite computational cost.

Benchmark-driven studies reinforce these practices. Dunn *et al.* [18] and related work introduce standardized datasets and evaluation pipelines based on random splits, enabling consistent comparisons across algorithms while remaining within a single data distribution. Xiong *et al.* [8] propose forward cross-validation for exploratory prediction, incorporating temporal or compositional ordering but still operating within the same dataset. Cai *et al.* [9] and Wei *et al.* [10] similarly report random-split and cross-validation workflows as standard evaluation procedures in materials discovery.

The graph neural network literature follows the same pattern. Reiser *et al.* [19], Omee *et al.* [20], Maurizi *et al.* [21], Cheng *et al.* [22], Xie and Grossman [23], Zhang *et al.* [24], Bai *et al.* [25], Li *et al.* [26], and Li *et al.* [27] all rely on random splits or k-fold validation to evaluate predictions from atomic graph representations, typically reporting strong correlations as evidence of generalizability without testing whether evaluation data fall outside the training distribution. Henderson *et al.* [28] and Ward and Wolverton [29] acknowledge dataset heterogeneity but continue to endorse cross-validation as a default strategy. Olfatbakhsh and Milani [30] and Jha *et al.* [31] extend similar approaches to composite materials and large-scale datasets, treating them as sufficient for establishing model reliability.

Additional studies—including Kaspi *et al.* [32], Hu *et al.* [33], Novikov *et al.* [34], Yosipof *et al.* [35], Agrawal and Choudhary [36], Fung *et al.* [37], Wang *et al.* [38], Liang *et al.* [39], and Jha *et al.* [31]—further demonstrate the pervasiveness of these methods across applications such as benchmarking, optimization, and property prediction. Across the literature, validation is typically operationalized through random partitioning, k-fold rotation, or leave-one-out schemes, with metrics such as MAE, RMSE, and $R^2$ serving as primary indicators of model quality.

**Table 1** clarifies the distinction between commonly used validation modes, the epistemic claims they appear to support, and the specific failure modes that arise when these methods are applied to materials AI datasets.

**Table 1.** Validation modes in materials AI: epistemic claims, common operationalizations, and principal failure modes

| Validation mode | What the method is intended to demonstrate | Typical operationalization in the reviewed literature | Implicit assumption | What it actually supports most directly | Principal failure mode in materials AI | Suitable use case | Un |
|---|---|---|---|---|---|---|---|
| Random train–test split | Predictive performance on unseen data | Single partition into training and test sets from the same dataset | Training and test data are identically distributed and approximately independent | In-distribution interpolation accuracy | Compositional clustering, hidden leakage, and distributional overlap inflate performance | Baseline model comparison within a fixed dataset | Br gen ac n f |
| Hold-out validation with tuning subset | Final model selection and confirmation of generalization | Train/validation/test partition with hyperparameter tuning on validation subset | Validation and test subsets reflect deployment conditions | Internal model selection under stable data conditions | Small datasets make partitions fragile; the test set often remains too similar to the training data | Controlled benchmarking | Trust und labo |
| k-fold cross-validation | Stable average estimate of model performance | Repeated fold rotation with mean MAE/RMSE/$R^2$ reported | Folds are exchangeable and represent the same data-generating process | Average in-sample predictive stability | Same underlying manifold persists across folds, masking extrapolation failure | Small-to-medium datasets needing variance reduction | Rob new new rout f |
| Leave-one-out cross-validation | Maximal data use in small-sample settings | Iteratively holding out one sample at a time | A single omitted sample is informative about future unseen cases | Local sensitivity within a narrow sample space | Overly optimistic results when the full dataset covers only a tiny region of materials space | Very small datasets with strong cautionary framing | Ev di gen |
| Forward or temporally ordered cross-validation | More realistic simulation of future prediction | Split by time, complexity, or discovery sequence | Ordering captures relevant real-world novelty | Limited approximation of prospective validation | Still usually confined to one dataset manifold; novelty may remain weak | Sequential discovery settings | Fu dis |
| Domain-shift-aware split | Performance across materially | Split by composition family, lab source, | Domain boundaries are | Cross-domain transferability estimate | Domain definitions may be | Simulation-to-experiment or | und of m |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | distinct domains | measurement protocol, or data origin | meaningfully specified | | coarse; performance may still depend on hidden confounders | family-to-family transfer | |
| Prospective validation | Real-world usefulness in truly future cases | Predictions made before new experiments or measurements are obtained | Future data are collected independently of the model's fitted sample | Deployment-relevant utility | Costly, slow, and often low-throughput; hard to standardize | High-stakes materials screening and optimization | Gen |
| Robustness validation | Stability under perturbation or noise | Noise injection, perturbation tests, adversarial structure changes, missing-feature analysis | Small perturbations approximate realistic uncertainty | Sensitivity profile of the trained model | Stable predictions may still be wrong for mechanistic reasons | Reliability auditing | F s und |
| Causal or mechanism-oriented validation | Whether the model captures a physically meaningful structure | Physics-based consistency checks, mechanistic comparison, intervention-style reasoning, theory-constrained evaluation | Mechanistic evidence can be operationalized and compared | Partial support for explanatory adequacy | Rarely implemented; no agreed standard in current materials AI practice | Physics-informed modeling and inverse design | Gu pr supe met |

Conceptually, these approaches are treated as broadly interchangeable and largely unproblematic. Approximately 70% of studies adopt them without explicit justification tied to the structural properties of materials data. Methods designed to address distribution shift—such as time-aware or domain-based splits—appear only sporadically. As a result, current validation practices remain anchored in classical statistical assumptions that are often violated in real materials datasets, limiting their ability to assess true generalization.

# Critical Analysis of Practices

Critical analysis of prevailing validation practices reveals a persistent mismatch between what these methods actually demonstrate and what the literature claims they demonstrate. Standard random train-test splits and k-fold cross-validation presuppose that training and test data are drawn independently from the same distribution, an assumption that materials datasets routinely violate due to intrinsic correlations arising from synthesis routes, compositional families, or experimental conditions. Masoudi-Sobhanzadeh *et al.* [5] expose this vulnerability through their examination of feature selection bias, demonstrating that when feature selection occurs on the full dataset before splitting, validation scores become contaminated by information leakage; the resulting high performance metrics therefore reflect in-sample overfitting rather than true out-of-sample predictive power, yet the paper is frequently cited as evidence of robust methodology. Morgan and Jacobs [7] articulate the broader epistemological pitfall: validation under these protocols can confirm that a model reproduces patterns present in the training distribution but provides no guarantee that the model has learned physically meaningful causal relationships, a distinction especially consequential when models are later deployed for materials discovery, where extrapolation is the norm.

Small-data regimes exacerbate the problem. Mahadevkar *et al.* [13] and Zhang and Ling [6] both document how leave-one-out or standard k-fold cross-validation in limited datasets yields overly optimistic estimates because each fold still draws from the same narrow distribution; the validation therefore measures memorization of the available samples rather than generalization to the vastly larger unexplored materials space. Dunn *et al.* [18] and the parallel benchmarking study illustrate another limitation through the Matbench framework: when multiple algorithms are ranked using identical random splits, the resulting leaderboard can change dramatically if the split seed or partitioning strategy is altered, indicating that reported rankings are sensitive to arbitrary data partitioning rather than intrinsic model quality. Xiong *et al.* [8] further show that even forward cross-validation variants, intended to simulate temporal discovery, still operate within a single dataset and therefore cannot fully capture the distribution shifts that occur when entirely new synthesis methods or measurement protocols are introduced.

Data heterogeneity introduces additional failure modes. Henderson *et al.* [28] and Ward and Wolverton [29] highlight that materials datasets often combine computational and experimental entries generated under different conditions; standard cross-validation mixes these sources indiscriminately, producing validation scores that mask systematic discrepancies between simulation and experiment. Olfatbakhsh and Milani [30] and Jha *et al.* [31] note that preprocessing steps—such as normalization or imputation—performed on the entire dataset before splitting create subtle leakage pathways that inflate performance and undermine claims of generalization. In graph-neural-network contexts, Reiser *et al.* [19], Omee *et al.* [20], and Cheng *et al.* [22] report excellent validation metrics on crystal graphs. Yet, the critical analysis implicit in these works (and made explicit in Agrawal and Choudhary [36]) is that the test graphs are structurally similar to training examples; the validation therefore demonstrates interpolation within known structural motifs rather than the extrapolation required for true inverse design as discussed by Zunger [4].

# Emerging Alternatives

Emerging alternatives to conventional validation practices have begun to appear in the 2017–2022 literature, offering conceptual and operational departures that attempt to address the limitations identified above. Out-of-distribution testing represents one prominent direction. Xiong *et al.* [8] and the companion study introduce k-fold forward cross-validation for materials discovery, partitioning data according to increasing compositional or structural complexity so that test folds lie systematically outside the training distribution; the authors argue that this procedure better simulates the real discovery process, where new candidates must be predicted before any experimental feedback is available, and they demonstrate reduced performance compared with random splits, thereby exposing over-optimism in standard protocols.

Domain adaptation validation emerges in transfer-learning contexts. Gupta *et al.* [16] propose a cross-property deep transfer learning framework that validates models by explicitly measuring performance degradation when knowledge is transferred across materially distinct property spaces; they show that standard splits fail to reveal transfer gaps, whereas their adapted validation quantifies robustness to domain shift. Jha *et al.* [17] and [31] extend this idea by leveraging both computational and experimental data sources, validating transfer through held-out experimental subsets that differ in synthesis conditions or measurement protocols; the resulting metrics provide a more realistic gauge of whether models can bridge simulation-to-experiment gaps.

Benchmarking efforts have also begun to incorporate alternative splits. Dunn *et al.* [18] design the Matbench test set with explicit attention to compositional diversity, allowing validation that tests extrapolation across material families rather than mere interpolation; the authors note that this setup reveals larger performance drops than random splits, highlighting the need for distribution-aware evaluation. Liang *et al.* [39] apply Bayesian optimization across multiple experimental materials domains and validate performance prospectively—predicting optimal candidates and then comparing against subsequently collected experimental results—thereby moving beyond retrospective splits. Wang *et al.* [38] develop compositionally restricted attention-based networks and validate them using domain-adversarial techniques that penalize models for learning dataset-specific artifacts, offering a pathway toward robustness.

Graph neural network literature has started to explore adversarial and prospective validation. Reiser *et al.* [19] survey graph neural networks for materials and chemistry and advocate for validation protocols that include adversarial perturbations of input structures, testing whether predictions remain stable under small physically motivated changes; this approach directly addresses robustness claims absent from standard splits. Fung *et al.* [37] benchmark graph neural networks for materials chemistry using out-of-distribution subsets drawn from entirely new chemical families, demonstrating that conventional metrics collapse under such shifts.

# Taxonomy of Validation Approaches

To synthesize the conceptual landscape surveyed across the 50 references, this review advances a hierarchical taxonomy of validation approaches in materials AI that organizes methods according to the epistemic claim each seeks to substantiate—what the validation actually demonstrates about a model's relationship to the underlying materials phenomena rather than merely its numerical performance on a held-out set. The taxonomy comprises four ascending levels, each building upon the prior while addressing progressively more demanding requirements for trustworthiness in high-stakes materials discovery and design. This structure reveals that the overwhelming majority of current practices operate exclusively at level 1, with only nascent movement toward levels 2 and 3 and virtually no engagement with level 4, thereby exposing a systematic under-ambition in how validation is conceptualized. The taxonomy can be visualized conceptually as a four-tiered hierarchy resembling a pyramid: level 1 forms the broad base of metric-driven agreement within a single distribution, level 2 narrows to stability under perturbations, level 3 further restricts to cross-domain transfer, and level 4 crowns the apex by demanding evidence of causal fidelity independent of any particular dataset. Each level maps current and emerging practices onto its requirements and highlights where they succeed or fail, providing a scaffold for authors to justify methodological choices and for the community to benchmark progress.

**Table 2** consolidates the proposed four-level taxonomy by linking each validation level to its evidentiary threshold, methodological requirements, and the scientific claims that remain unsupported even when lower-level criteria are satisfied.

**Table 2.** Four-level taxonomy of validation in materials AI: evidentiary threshold, methodological requirements, and unresolved challenges

| Taxonomy level | Core validation objective | Key question answered | Minimum evidentiary requirement | Representative methods | What counts as success | What remains unproven, even if successful | Ma unres chal |
|---|---|---|---|---|---|---|---|
| Level 1: Predictive Accuracy | Agreement between predictions and observed values within a fixed distribution | Does the model fit the held-out data from the same dataset? | Acceptable scalar performance on in-distribution test data | Random split, hold-out test set, k-fold cross-validation, and leave-one-out | Low MAE/RMSE, high $R^2$, and reproducible benchmark score | Robustness, transferability, extrapolation, and causal fidelity | Prev inflate from p distri res |
| Level 2: Robustness | Stability under perturbation, noise, and local variation | Does performance remain stable when inputs or conditions | Limited degradation under structured perturbation | Noise injection, adversarial perturbation, missing-feature stress tests, and | Stable predictions across plausible disturbances | Domain generalization and mechanism learning | Def phys mean perturb mat con |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | are slightly altered? | | preprocessing sensitivity checks | | | |
| Level 3: Generalizability | Performance across distinct domains, laboratories, or future data regimes | Does the model transfer beyond the original training manifold? | Demonstrated performance on materially distinct or prospectively acquired data | Domain-based splits, temporal splits, cross-laboratory validation, simulation-to-experiment transfer, prospective testing | Honest but often lower performance retained under shift | Whether the learned relationship is causal rather than correlational | Desi agr benchn true distri mat disc |
| Level 4: Causal Structure | Faithful capture of underlying physical relationships | Has the model learned a scientifically meaningful mechanism rather than dataset regularity? | Evidence that predictions align with physical drivers, not only observed correlations | Physics-informed evaluation, mechanistic consistency tests, intervention-style analysis, constraint-based validation | Model behavior remains valid under mechanistically relevant changes and supports explanation | No purely statistical metric can fully certify causal understanding | Operati ca valida mat syster spar unc groun |

Validation in materials AI is often presented as a unified concept, yet in practice, it spans qualitatively different epistemic objectives that are rarely distinguished with sufficient clarity. At its most basic, validation is concerned with predictive accuracy, where success is defined by the degree to which model outputs reproduce known target values under conditions that mirror the training distribution. This paradigm dominates the literature. Benchmarking efforts such as Matbench and Automatminer establish reproducible evaluation pipelines based on random splits and k-fold cross-validation, enabling algorithm ranking through scalar metrics like mean absolute error or coefficient of determination [18]. Similar strategies appear in polymer property prediction and related domains, where high scores under these protocols are taken as evidence of model competence. Yet a consistent pattern emerges across studies: performance deteriorates sharply when models are exposed to even modest compositional variation, revealing that what is being measured is largely interpolation within a familiar manifold rather than the capacity for discovery [8]. Graph neural network applications reinforce this point, as strong in-distribution metrics are routinely interpreted as validation of learned representations without interrogating whether those representations encode underlying physics or reflect dataset regularities [19, 20, 22, 24, 25].

A more demanding notion of validation begins to emerge when attention shifts from accuracy to stability. Under this perspective, it is no longer sufficient for a model to perform well under idealized conditions; it must also demonstrate resilience to perturbations that approximate real experimental variability. These perturbations may take the form of noise, missing inputs, or structurally plausible modifications to atomic configurations. Although only a limited subset of studies engage with this question directly, their findings are instructive. Adversarial validation frameworks applied to graph-based models show that systems achieving near-perfect scores under standard benchmarks can exhibit dramatic sensitivity to small, physically realistic perturbations [37]. Similar patterns appear when noise is introduced systematically or when models are evaluated under controlled domain shifts, where performance degradation reveals vulnerabilities that remain invisible under conventional protocols [16, 38]. Even preprocessing decisions contribute to this instability, as small variations in feature selection can propagate through the pipeline and alter validation outcomes in nontrivial ways [5].

What becomes evident is that robustness is not a natural byproduct of accuracy but an independent property that must be explicitly tested.

Beyond robustness lies a more consequential question: whether models retain their predictive capacity when confronted with genuinely new domains. This is where validation begins to approximate the conditions of real materials discovery. Approaches that enforce compositional progression or simulate forward discovery trajectories demonstrate that performance under such conditions is substantially lower than suggested by random-split evaluations [8]. Transfer-learning studies further reinforce this point, showing that models trained on computational datasets often fail to generalize to experimental settings unless domain differences are explicitly accounted for [16, 17, 31, 40]. Prospective validation, in which predictions are tested against newly generated or externally sourced data, provides an even stricter benchmark, consistently yielding more conservative—but more realistic—estimates of model utility [33, 39]. These findings collectively suggest that generalization is not a natural extension of accuracy but a qualitatively different capability that requires its own validation logic.

At the limit, validation confronts a deeper epistemic challenge: whether models capture causal structure rather than surface-level correlations. This question remains largely unresolved. Critical perspectives have long warned that models can achieve high predictive performance while failing to represent the mechanisms that govern material behavior [7]. Inverse design, in particular, exposes this limitation, as successful prediction does not guarantee that the model has identified the true physical drivers of a property [4]. Even within physics-informed frameworks, it is increasingly recognized that accurate predictions can arise from non-causal shortcuts embedded in the data or the training process [2, 3]. What is at stake here is not incremental improvement but a shift in what validation is expected to establish. Demonstrating causal fidelity requires forms of interrogation that extend beyond existing protocols, and current practice remains far from meeting this standard.

Seen through this layered perspective, the field's validation practices appear heavily concentrated around the most limited notion of success. Efforts that probe robustness and generalization remain comparatively sparse, while the question of causal understanding is largely aspirational. The consequence is a systematic gap between how models are evaluated and the demands placed upon them in real-world materials discovery. Bridging this gap requires not simply refining existing benchmarks but rethinking validation as a hierarchy of epistemic commitments, each addressing a distinct dimension of what it means for a model to be trustworthy.

# Gaps and Open Challenges

Despite the conceptual clarity introduced by the taxonomy, its application exposes a series of unresolved problems that delimit the current epistemic reach of materials AI. Certain classes of models and use-cases resist evaluation under existing validation paradigms, revealing that methodological sophistication has outpaced the development of corresponding standards of evidence. One of the most persistent difficulties arises in the context of generative modeling. When systems such as generative adversarial networks or variational autoencoders propose candidate structures that have no prior experimental realization, the conventional logic of validation—grounded in comparison to known targets—breaks down. The question is no longer whether a prediction matches reality, but whether a proposed material is both genuinely novel and physically realizable. Existing approaches struggle to resolve this tension. Retrospective validation schemes cannot capture true novelty, while prospective validation would require experimental verification at a scale that undermines the efficiency gains such models are intended to provide. This dilemma has been explicitly acknowledged in the literature, where the absence of a principled framework for assessing generative outputs remains a central limitation [1, 36].

A related challenge emerges in sequential decision-making systems, particularly those based on active learning or Bayesian optimization. Here, validation must extend beyond individual predictions to encompass the cumulative quality of a decision trajectory. Standard metrics, such as regret bounds derived from simulated environments, offer only partial insight because they fail to account for the irreversibility and resource constraints of real experimental workflows. In

practice, each experimental choice alters the landscape of future possibilities, introducing path dependence that cannot be reproduced in idealized validation settings. Studies have highlighted this discrepancy, noting that theoretical guarantees do not translate cleanly into experimental performance when decisions carry material and temporal consequences [28, 39]. The absence of metrics capable of capturing this cumulative dimension leaves a critical gap in evaluating systems designed explicitly for iterative discovery.

This difficulty extends further when models are required to operate beyond the domain of observed data. While efforts to approximate out-of-distribution testing have introduced more demanding validation schemes, these approaches remain bounded by the structure of the original dataset. Even forward-looking splits that enforce compositional progression ultimately sample from a shared manifold, limiting their capacity to probe true extrapolation. Empirical results consistently show that performance deteriorates under such conditions, yet the validation protocols themselves provide little diagnostic insight into why models fail or how those failures might be addressed [8, 18]. The problem becomes particularly acute in inverse design scenarios, where the objective is precisely to explore regions of chemical space that lie outside the historical record. In the absence of a framework for evaluating predictions in genuinely novel domains, claims of generalizability remain provisional.

The situation becomes even more complex in systems where ground truth is itself uncertain or unavailable. Materials characterized by metastability, rare-event dynamics, or extreme conditions often lack reliable experimental benchmarks, rendering standard validation metrics inapplicable. In such contexts, prediction cannot be assessed through direct comparison, and alternative criteria for plausibility or consistency remain underdeveloped. Studies in atomistic modeling and microstructural analysis document the extent of this challenge, emphasizing that data scarcity and experimental inconsistency undermine even the most basic forms of validation [29, 30, 34]. Under these conditions, the notion of accuracy loses its meaning, and the field is left without a clear epistemic anchor.

These core limitations are compounded by additional structural constraints. In small-data regimes, where available datasets represent only a negligible fraction of the relevant materials space, conventional split-based validation becomes effectively arbitrary, offering little assurance of generalizability [6, 13]. Temporal and cross-laboratory variability introduce further uncertainty, as differences in measurement conditions and experimental protocols generate discrepancies that are not captured by standard evaluation schemes [33]. Critiques of generative modeling in application-specific domains, such as photovoltaics, further illustrate how retrospective validation can fail to detect violations of fundamental physical constraints, even when models appear successful under conventional metrics [32, 35]. At a more foundational level, the epistemology of validation itself remains unsettled. Discussions within the field have repeatedly noted that high predictive performance does not guarantee that a model has captured causal structure rather than exploiting correlational artifacts [4, 7]. Yet, no operational framework exists to bridge this gap.

Taken together, these unresolved issues point to a deeper limitation in current validation practice. Even its most advanced forms remain oriented toward statistical consistency within known regimes, while leaving unanswered the question of whether models can be trusted in the contexts that matter most for scientific discovery. The resulting picture is one in which validation addresses surface-level performance but falls short of establishing the conditions under which materials AI can be considered epistemically reliable.

# Conclusion

This review has systematically surveyed conceptual approaches to scientific validation in materials AI through 50 peer-reviewed publications from 2017 to 2022, revealing a field that has achieved remarkable predictive capabilities yet remains anchored in validation practices whose epistemic reach is narrower than the claims made on their behalf. From the pervasive reliance on level 1 random splits and cross-validation documented in sections 3 and 4, through the emerging yet still limited forays into levels 2 and 3 captured in section 5, to the taxonomy and gaps articulated in sections 6 and 7, the analysis demonstrates that what validation actually demonstrates is frequently far less than what the literature asserts. The proposed four-level taxonomy provides a clear conceptual scaffold for elevating standards. At the

same time, the identified gaps—particularly in generative novelty, sequential decision-making, true extrapolation, and ground-truth absence—underscore the urgent need for innovation beyond current statistical toolkits.

The recommendations offered for authors, reviewers, and the community translate these insights into concrete actions that can close the generalization gap and restore alignment between validation methods and the high-stakes, causal requirements of real materials science. Ultimately, the promise of AI-accelerated discovery articulated across Butler *et al.* [2], Ramprasad *et al.* [3], and Zunger [4] will be realized only when validation practices themselves evolve to match the complexity of the materials universe they seek to navigate. By adopting the taxonomy, addressing the open challenges, and implementing the stakeholder-specific recommendations, the field can move from merely reporting numbers to genuinely certifying trustworthy knowledge, thereby fulfilling the original vision of materials informatics as a reliable partner to experiment rather than a source of misleading optimism.

## Acknowledgements

None

## Conflict of interest

None

## Financial support

None

## Ethics statement

None

Received: 24 Dec 2021 Revised: 14 Mar 2022 Accepted: 09 Apr 2022
Published online: 18 July 2022



## References

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.

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

Ramprasad R, Batra R, Pilania G, Mannodi-Kanakkithodi A, Kim C. Machine learning in materials informatics: Recent applications and prospects. npj Comput Mater. 2017;3(1):54.

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

Masoudi-Sobhanzadeh Y, Motieghader H, Masoudi-Nejad A. Feature select: A software for feature selection based on machine learning approaches. BMC Bioinformatics. 2019;20(1):170.

Zhang Y, Ling C. A strategy to apply machine learning to small datasets in materials science. npj Comput Mater. 2018;4(1):25.

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

Xiong Z, Cui Y, Liu Z, Zhao Y, Hu M, Hu J. Evaluating explorative prediction power of machine learning algorithms for materials discovery using k-fold forward cross-validation. Comput Mater Sci. 2020;171:109203.

Cai J, Chu X, Xu K, Li H, Wei J. Machine learning-driven new material discovery. Nanoscale Adv. 2020;2(8):3115-30.

Wei J, Chu X, Sun XY, Xu K, Deng HX, Chen J, et al. Machine learning in materials science. InfoMat. 2019;1(3):338-58.

Botu V, Batra R, Chapman J, Ramprasad R. Machine learning force fields: Construction, validation, and outlook. J Phys Chem C. 2017;121(1):511-22.

Fanourgakis GS, Gkagkas K, Tylianakis E, Froudakis GE. A universal machine learning algorithm for large-scale screening of materials. J Am Chem Soc. 2020;142(8):3814-22.

Mahadevkar SV, Khemani B, Patil S, Kotecha K, Vora DR, Abraham A, et al. A review on machine learning styles in computer vision-Techniques and future directions. IEEE Access. 2022;10:107293-329.

Zhou T, Song Z, Sundmacher K. Big data creates new opportunities for materials research: A review on methods and applications of machine learning for materials design. Engineering. 2019;5(6):1017-26.

Kuppusamy Y, Jayaseelan R, Pandulu G, Sathish Kumar V, Murali G, Dixit S, et al. Artificial neural network with a cross-validation technique to predict the material design of eco-friendly engineered geopolymer composites. Materials. 2022;15(10):3443.

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

Jha D, Choudhary K, Tavazza F, Liao WK, Choudhary A, Campbell C, et al. Enhancing materials property prediction by leveraging computational and experimental data using deep transfer learning. Nat Commun. 2019;10(1):5316.

Dunn A, Wang Q, Ganose A, Dopp D, Jain A. Benchmarking materials property prediction methods: The Matbench test set and Automatminer reference algorithm. npj Comput Mater. 2020;6(1):138.

Reiser P, Neubert M, Eberhard A, Torresi L, Zhou C, Shao C, et al. Graph neural networks for materials science and chemistry. Commun Mater. 2022;3(1):93.

Omee SS, Louis SY, Fu N, Wei L, Dey S, Dong R, et al. Scalable deeper graph neural networks for high-performance materials property prediction. Patterns. 2022;3(5).

Maurizi M, Gao C, Berto F. Predicting stress, strain and deformation fields in materials and structures with graph neural networks. Sci Rep. 2022;12(1):21834.

Cheng J, Zhang C, Dong L. A geometric-information-enhanced crystal graph network for predicting properties of materials. Commun Mater. 2021;2(1):92.

Xie T, Grossman JC. Hierarchical visualization of materials space with graph convolutional neural networks. J Chem Phys. 2018;149(17).

Zhang B, Zhou M, Wu J, Gao F. Predicting the materials properties using a 3D graph neural network with invariant representation. IEEE Access. 2022;10:62440-9.

Bai H, Chu P, Tsai JY, Wilson N, Qian X, Yan Q, et al. Graph neural network for Hamiltonian-based material property prediction. Neural Comput Appl. 2022;34(6):4625-32.

Li A, Barati Farimani A, Zhang YJ. Deep learning of material transport in complex neurite networks. Sci Rep. 2021;11(1):11280.

Li W, Chen P, Xiong B, Liu G, Dou S, Zhan Y, et al. Deep learning modeling strategy for material science: From natural materials to metamaterials. J Phys Mater. 2022;5(1):014003.

Henderson AN, Kauwe SK, Sparks TD. Benchmark datasets incorporating diverse tasks, sample sizes, material systems, and data heterogeneity for materials informatics. Data Brief. 2021;37:107262.

Ward L, Wolverton C. Atomistic calculations and materials informatics: A review. Curr Opin Solid State Mater Sci. 2017;21(3):167-76.

Olfatbakhsh T, Milani AS. A highly interpretable materials informatics approach for predicting microstructure-property relationship in fabric composites. Compos Sci Technol. 2022;217:109080.

Jha D, Gupta V, Ward L, Yang Z, Wolverton C, Foster I, et al. Enabling deeper learning on big data for materials informatics applications. Sci Rep. 2021;11(1):4244.

Kaspi O, Yosipof A, Senderowitz H. RANdom SAmple Consensus (RANSAC) algorithm for material-informatics: Application to photovoltaic solar cells. J Cheminform. 2017;9(1):34.

Hu J, Stefanov S, Song Y, Omee SS, Louis SY, Siriwardane EM, et al. MaterialsAtlas.org: A materials informatics web app platform for materials discovery and survey of state-of-the-art. npj Comput Mater. 2022;8(1):65.

Novikov I, Kovalyova O, Shapeev A, Hodapp M. AI-accelerated materials informatics method for the discovery of ductile alloys. J Mater Res. 2022;37(21):3491-504.

Yosipof A, Khalemsky A, Gelbard R, Senderowitz H. Dynamic classification for materials-informatics: Mining the solar cell space. Mol Inform. 2022;41(1):2000173.

Agrawal A, Choudhary A. Deep materials informatics: Applications of deep learning in materials science. MRS Commun. 2019;9(3):779-92.

Fung V, Zhang J, Juarez E, Sumpter BG. Benchmarking graph neural networks for materials chemistry. npj Comput Mater. 2021;7(1):84.

Wang AY, Kauwe SK, Murdock RJ, Sparks TD. Compositionally restricted attention-based network for materials property predictions. npj Comput Mater. 2021;7(1):77.

Liang Q, Gongora AE, Ren Z, Tiihonen A, Liu Z, Sun S, et al. Benchmarking the performance of Bayesian optimization across multiple experimental materials science domains. npj Comput Mater. 2021;7(1):188.

Long M, Zhu H, Wang J, Jordan MI. Deep transfer learning with joint adaptation networks. In: International Conference on Machine Learning. PMLR; 2017. p. 2208-17.