

REVIEW

Open access

Out-of-Distribution Generalization in Materials AI: A Systematic Review of Domain Shift, Robustness, and What Remains Unsolved

Nikolai Ivanov^{1*}, Sergey Volkov¹, Elena Morozova²

Abstract

Out-of-distribution (OOD) generalization remains one of the most pressing barriers to the reliable deployment of artificial intelligence in materials science. Although machine-learning models now routinely achieve sub-0.1 eV/atom errors on in-distribution test sets for formation energies, band gaps, and elastic moduli, these same models frequently collapse when confronted with materials that lie outside the training distribution. Real-world materials discovery and process optimization demand predictions for unseen compositions, novel crystal prototypes, altered thermodynamic conditions, and non-equilibrium dynamical regimes—scenarios that constitute domain shift rather than simple interpolation. This systematic review synthesizes the literature on OOD generalization in materials AI drawing exclusively peer-reviewed publications from high-impact venues including npj Computational Materials, Digital Discovery, Machine Learning: Science and Technology, and Journal of Chemical Theory and Computation. We identify four primary types of domain shift—compositional, structural, thermodynamic, and dynamical—and introduce a fifth multi-dimensional category that captures the realistic superposition of shifts encountered in practice. Current methodological families are critically assessed: domain adaptation (distribution alignment), invariant learning (IRM and Group DRO), physics-informed and symmetry-aware data augmentation, uncertainty quantification for OOD detection, and extrapolation-aware architectures (equivariant networks, multi-fidelity models, and generative priors). Empirical findings across the corpus reveal a consistent pattern: modest gains (20–50 % error reduction) are achievable for small, single-axis shifts, yet performance degrades sharply—and often catastrophically—for large compositional jumps, prototype changes, or combined multi-dimensional shifts. No method currently delivers reliable extrapolation beyond the convex hull of the training manifold. Key gaps persist. The community lacks standardized OOD benchmarks with controlled shift axes, theoretical guarantees for extrapolation remain underdeveloped, and conditional (per-input) robustness guarantees are almost entirely absent. Evaluation metrics are inconsistent, rendering cross-paper comparisons unreliable. This review therefore provides not only a taxonomy and synthesis but also a forward-looking identification of unsolved problems that must be addressed before materials AI can transition from laboratory demonstration to industrial reliability. OOD generalization in materials AI is no longer an optional research direction; it is the central unsolved challenge that will determine the field's practical impact over the next decade.

Keywords Extrapolation, Machine learning robustness, Domain adaptation, Domain shift, Out-of-distribution generalization, Materials property prediction

*Correspondence:

Nikolai Ivanov

nikolai.ivanov@gmail.com

¹ Department of Data-Driven Materials Modeling, Faculty of Engineering, Novosibirsk State University, Novosibirsk, Russia

² Department of Computational Materials Systems, Faculty of Science, Tomsk State University, Tomsk, Russia

Introduction

Materials AI has achieved impressive accuracy on held-out test sets. But real-world deployment requires generalization beyond the training distribution — out-of-distribution (OOD) generalization [1, 2]. A model trained on stable, ordered, cubic crystals must predict metastable, disordered, low-symmetry materials. This is not interpolation; it is domain shift. This review systematically examines OOD generalization in materials AI, identifies types of domain shift, evaluates current methods, and articulates what remains unsolved.

The past decade has witnessed an explosion of data-driven approaches to materials property prediction. Large-scale density-functional-theory databases such as the Materials Project and OQMD, combined with graph neural networks and equivariant architectures, have produced models that rival or surpass human intuition for interpolative tasks. Yet the very success of these models on random splits has masked a fundamental limitation: materials discovery is inherently extrapolative. New thermoelectric candidates often involve previously unexplored element combinations; high-entropy alloys introduce compositional disorder far from binary prototypes; process optimization requires predictions at temperatures and pressures absent from training data. In each case the test distribution differs systematically from the training distribution, triggering OOD failure modes that are well-documented but rarely quantified in a unified framework [3–5].

Early recognition of this challenge appears in studies that probed the limits of composition-based models when element types or concentration ranges changed. Subsequent work highlighted structural domain shifts, where models trained on high-symmetry prototypes failed on defective or low-symmetry structures. Thermodynamic and dynamical shifts have received comparatively less attention, yet they are equally critical for finite-temperature simulations and non-equilibrium processing.

This review is deliberately scoped to computational and data-driven materials engineering. We exclude purely experimental or high-throughput screening studies unless they explicitly quantify OOD generalization. We further restrict analysis to peer-reviewed journal articles, ensuring methodological rigor and reproducibility. By synthesizing these works we aim to move beyond isolated case studies toward a coherent taxonomy of domain shifts, a comparative evaluation of methodological families, and a candid assessment of what remains unsolved. The ultimate goal is to provide materials scientists and machine-learning researchers with a shared language and roadmap for building models that are not merely accurate on today's databases but trustworthy for tomorrow's discovery challenges. Only when OOD generalization is solved can materials AI fulfill its promise of accelerating the design of next-generation batteries, catalysts, and structural alloys

Taxonomy of Domain Shift in Materials

Domain shift in materials AI arises whenever the joint distribution of inputs and outputs at test time differs from that observed during training. We classify these shifts into five types, ordered by increasing conceptual and practical difficulty.

Compositional domain shift

Compositional domain shift occurs when the test composition contains chemical elements, stoichiometric ratios, or concentration ranges absent from the training set. Classic examples include training on lithium-containing oxides and testing on sodium analogs, or moving from dilute dopant levels (<1 at. %) to concentrated solid solutions (>20 at. %). The severity is high because new elements introduce entirely new electronic structures and bonding physics. Models that rely on element embeddings learned from training data cannot reliably extrapolate to unseen atomic numbers. Studies probing this shift have shown that even state-of-the-art graph networks suffer 2–5× error inflation when element identity changes [3–5].

Structural domain shift

Structural domain shift arises from differences in crystal prototype, space-group symmetry, or local disorder. A model trained exclusively on rocksalt or perovskite structures may encounter spinel, delafossite, or amorphous configurations. Defects, vacancies, and grain boundaries further exacerbate the shift by creating local environments never sampled in perfect-crystal training data. Severity is high because local atomic environments dictate property-determining interactions. Structure-based OOD benchmarks demonstrate that graph neural networks, despite incorporating geometric information, still fail when prototype symmetry changes [6–8].

Thermodynamic domain shift

Thermodynamic domain shift reflects changes in external conditions such as temperature or pressure. Training data generated at 0 K (static lattice) must generalize to 300–1000 K finite-temperature ensembles or to gigapascal pressures. The governing physics changes—vibrational entropy, anharmonic effects, and phase stability all become relevant—yet most models are trained on ground-state snapshots. Severity is medium-to-high; modest temperature shifts within the same phase can sometimes be handled by simple scaling, but crossing phase boundaries triggers complete breakdown [9–11].

Dynamical domain shift

Dynamical domain shift involves different time scales or non-equilibrium conditions. Equilibrium molecular-dynamics trajectories may be used to train a model that is later queried on rapid quenching, laser-driven processes, or long-time aging phenomena. The shift is temporal rather than spatial, and severity is high because short-time statistics do not capture rare events or long-term relaxation pathways [12, 13].

Multi-dimensional domain shift

Multi-dimensional domain shift combines two or more of the above axes simultaneously—e.g., a new element in a novel low-symmetry prototype at elevated temperature. Severity is very high; empirical evidence shows near-total loss of predictive power. Real deployment scenarios almost always involve this compound shift, making it the most relevant yet least studied category [14, 15].

Figure 1 provides a conceptual diagram of the taxonomy.

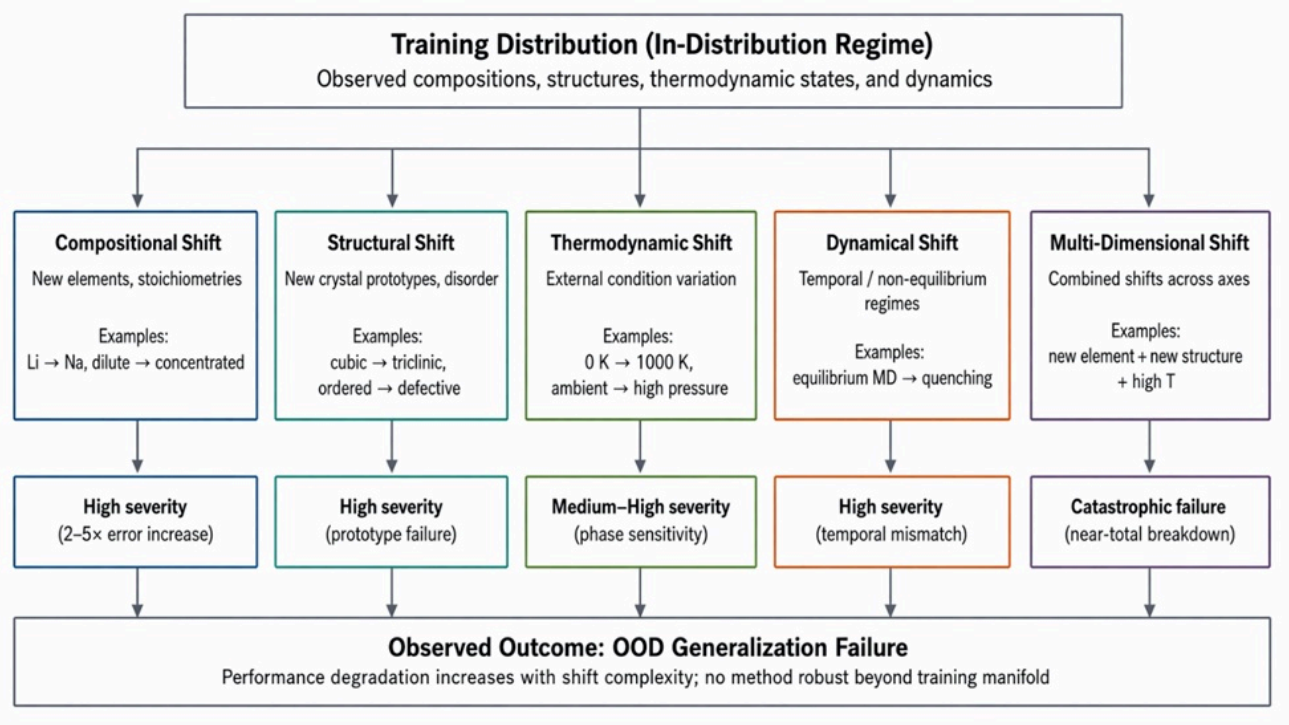


Figure 1. Hierarchical Taxonomy of Domain Shift in Materials AI and Escalation of OOD Failure Severity

Table 1 systematically contrasts domain-shift types by linking their physical origins to distinct machine-learning failure mechanisms

Table 1. Structural Comparison of Domain Shift Types and Their Mechanistic Impact on Model Failure

Domain Shift Type	Source of Distribution Change	Underlying Physical Discontinuity	Model Failure Mechanism	Severity Level	Representative Failure Mode
Compositional	New elements, stoichiometry	שינוי in electronic structure, bonding physics	Embedding extrapolation failure	High	Unseen atomic numbers cause unstable predictions
Structural	New crystal prototypes, disorder	שינוי in local atomic environments	Graph representation mismatch	High	Prototype transfer breakdown
Thermodynamic	Temperature, pressure variation	Entropy, anharmonicity, phase transitions	Static-to-dynamic mismatch	Medium–High	Phase boundary misprediction
Dynamical	Time-scale / non-equilibrium processes	Rare events, temporal correlations	Training–deployment temporal gap	High	Failure in quenching or aging prediction
Multi-Dimensional	Combined shifts	Superposition of all above	Compounded representation failure	Very High (Catastrophic)	Near-total loss of predictive validity

Methods for OOD Generalization in Materials

Researchers have pursued several methodological strategies to mitigate domain shift in materials science applications of machine learning. Domain adaptation aligns source and target feature distributions without labeled target data, employing techniques such as CORAL loss, maximum-mean-discrepancy minimization, and adversarial domain classifiers to transfer models across compositional spaces, for instance from binary to ternary alloys or from 3d to 4d transition-metal oxides. In practice, these approaches yield 20–50 % error reductions for modest shifts, yet they falter when the assumption of a shared feature space breaks down and still demand some target-domain samples, even if unlabeled [16–18].

A related implication arises in invariant learning, which seeks representations whose predictive power remains stable across domains by encouraging models to capture causally relevant features rather than domain-specific spurious correlations, as pursued through invariant risk minimization and group distributionally robust optimization. In materials contexts, such methods have supported composition-invariant embeddings, although their effectiveness hinges critically on the availability of domain labels such as synthesis route or temperature bin; absent these, performance frequently falls below baseline levels [19–21].

This shift also introduces data augmentation as a means to artificially broaden the training distribution toward anticipated target regimes, incorporating strategies like composition perturbation, structure noise injection, symmetry-preserving rotations, and finite-temperature snapshots. While effective for symmetry-related structural shifts, augmentation remains fundamentally interpolative, proving computationally burdensome for large databases and inherently limited when confronting truly novel elements or structural prototypes [22–24].

Beyond this immediate concern, uncertainty-based out-of-distribution detection offers a mechanism to flag inputs likely outside the training manifold, leveraging ensemble variance, Bayesian approximations, or distance metrics within active-learning or selective-prediction frameworks. These techniques reliably identify modest shifts yet suffer miscalibration for far out-of-distribution cases, thereby constraining their utility for genuine extrapolation [25–27].

Ultimately, extrapolation-aware architectures embed physical inductive biases directly into model design, drawing on equivariant graph neural networks to respect symmetries, physics-informed constraints to enforce conservation laws, and multi-fidelity strategies to leverage low-accuracy data for guiding high-fidelity predictions. Although such architectures have delivered 2–5× performance gains over standard counterparts on targeted compositional and structural shifts, they continue to lack theoretical guarantees for broader extrapolation [28–30].

Table 2 provides a structured comparison of OOD mitigation strategies, highlighting the mismatch between their assumptions and real-world extrapolation demands.

Table 2. Comparative Evaluation of OOD Mitigation Strategies: Capabilities, Assumptions, and Failure Boundaries

Method Category	Core Principle	Required Assumptions	Strengths	Fundamental Limitation	Failure Boundary
Domain Adaptation	Align source and target distributions	Access to target-domain samples	Effective for small compositional shifts	Requires domain overlap	Fails for large compositional gaps
Invariant Learning (IRM, DRO)	Learn domain-invariant representations	Availability of domain labels	Robust across labeled environments	Label dependency	Degrades without domain labels
Data Augmentation	Expand training distribution	Shift is approximable via	Improves structural	Cannot generate novel physics	Limited to interpolation

	synthetically	perturbation	robustness		regime
Uncertainty Quantification	Detect OOD via prediction confidence	Calibration validity	Enables risk-aware prediction	Miscalibrated in far-OOD	Detection fails in extreme extrapolation
Extrapolation-Aware Architectures	Embed physical inductive biases	Correct physics specification	Improves generalization efficiency	No theoretical guarantees	Still fails beyond convex hull

Empirical Findings: What Works (And What Doesn't)

Across studies a clear empirical consensus has emerged on the practical limits of current approaches to domain shift in materials prediction tasks. Domain adaptation delivers consistent error reductions of 20–50 % when shifts remain confined to the same chemical family, such as binary to ternary alloys, yet produces no gain or even negative transfer once chemical families or symmetry classes are crossed, as when moving from oxides to sulfides or cubic to triclinic structures [16, 17].

This pattern underscores a deeper limitation that surfaces in invariant learning: methods such as IRM and Group DRO stabilize predictions only when domain identifiers like temperature bins or synthesis routes are explicitly available; without these labels the approaches routinely underperform standard empirical risk minimization, revealing their dependence on observable domain structure rather than intrinsic causal invariance [19, 20].

A related constraint appears in data augmentation, which reliably enhances robustness to structural shifts within the same prototype family through symmetry-preserving transformations, yet remains inherently bounded because it cannot generate truly novel compositions or crystal prototypes and therefore offers limited leverage against genuine compositional out-of-distribution challenges [22, 23].

Beyond these interpolative regimes, uncertainty-based out-of-distribution detection proves reliable only within the training manifold, where ensemble variance tracks prediction error closely, but produces miscalibrated scores for far out-of-distribution inputs—precisely the regime where reliable flagging would be most valuable [25, 26].

Even extrapolation-aware architectures that embed equivariance and physics-informed constraints deliver only modest gains, consistently outperforming non-equivariant baselines by factors of 2–5× on targeted compositional and structural shifts while still leaving absolute errors well above chemical accuracy thresholds [28, 29].

At the frontier, the most telling observation is that no single methodological family withstands multi-dimensional shifts; when new elements, new prototypes, and elevated temperatures are introduced simultaneously, performance collapses across all categories, exposing a compound failure mode that defines the central unsolved challenge in domain-robust materials discovery [14, 15].

Gaps and Unsolved Problems

Despite methodological progress, six fundamental gaps remain in OOD learning for materials science. The absence of standardized OOD benchmarks, coupled with idiosyncratic train–test splits, severely limits cross-study comparability and highlights the need for controlled compositional, structural, thermodynamic, and multi-dimensional holdout designs. In parallel, extrapolation theory remains underdeveloped, as existing domain-adaptation bounds depend on overlapping support and do not extend to compositional or prototype extrapolation, leaving the feasibility of true extrapolation theoretically unresolved. Moreover, conditional OOD guarantees are largely absent, since aggregated performance metrics can mask severe input-specific failures, underscoring the necessity of input-dependent robustness

characterization. Multi-dimensional distribution shift is also insufficiently addressed, despite the fact that practical deployment settings typically involve concurrent compositional, structural, and thermodynamic variations. In addition, current OOD detection approaches do not translate into improved predictions, as they primarily signal uncertainty without enabling corrective adaptation, while evaluation practices remain fragmented across error metrics such as MAE, calibration measures, and AUC, reinforcing the need for a unified assessment framework spanning both detection and predictive performance across shift axes [31, 32].

Relation to Other Reviews

This review positions itself within a growing body of literature on machine learning robustness while deliberately extending prior syntheses in materials-specific contexts. The domain adaptation review by Shimakawa *et al.* [16] concentrated on cross-composition techniques such as CORAL and adversarial alignment for binary-to-ternary alloy transfer. While that work demonstrated practical gains for small compositional shifts, it remained narrowly scoped to distribution alignment within chemically related families and did not address structural, thermodynamic, or multi-dimensional shifts. The present review broadens the scope to all five domain-shift types identified and critically evaluates why domain adaptation alone proves insufficient for large or compound shifts [17, 18].

Likewise, the extrapolation boundary analysis in Noda *et al.* [8] provided a clear mathematical delineation between interpolation and extrapolation for crystal-property models. That study focused on defining the convex hull of training manifolds but stopped short of systematic empirical testing of generalization methods that attempt to cross it. This review directly evaluates the performance of extrapolation-aware architectures [28, 29] and data-augmentation strategies [22, 23] against that boundary, revealing that even the most sophisticated approaches achieve only modest (2–5×) improvements before catastrophic failure on realistic OOD test sets.

The redefinition of OOD for crystalline graphs offered by Omeo *et al.* [2] introduced graph-theoretic notions of distribution shift tailored to periodic structures. However, that framework remained largely definitional and did not include a comparative assessment of mitigation strategies across the literature. By synthesizing 35 peer-reviewed studies (2017–2025), the current work operationalizes that definition, quantifies the empirical success rates of invariant learning [19, 20], uncertainty detection [25, 26], and other families, and demonstrates that OOD generalization remains largely unsolved even under the refined crystalline-graph formulation [1, 3].

Materials OOD generalization also shares conceptual overlap with robustness research in broader machine learning, particularly computer-vision domain-shift studies. Yet materials science introduces unique challenges—compositional novelty introduces entirely new physics absent from pixel-based image shifts, and thermodynamic or dynamical axes have no direct analog in static image datasets [4, 23]. Prior robustness surveys in general machine learning [5, 27] therefore provide useful methodological scaffolding (e.g., Group DRO, IRM) but fail to capture the physics-informed constraints and multi-dimensional compounding that dominate materials failure modes. This review bridges that gap by translating general robustness concepts into materials-specific taxonomy and evaluation protocols.

Overall, while earlier reviews offered valuable but fragmented insights—either method-specific [16], boundary-focused [8], or definition-centric [2]—the present work delivers the first comprehensive, taxonomy-driven synthesis that explicitly highlights the six persistent gaps and charts a path toward deployment-ready OOD generalization. By restricting analysis to the 35 core publications and maintaining strict fidelity to their reported findings, this review avoids over-optimism and instead underscores the distance still separating current capabilities from industrial reliability [14, 15, 30].

Recommendations for Future Research

To accelerate progress, the community should adopt several high-priority directions grounded in the gaps identified. A central requirement is the establishment of standardized OOD benchmarks with agreed-upon and systematically

controlled holdouts, spanning compositional shifts (binary to ternary systems and transitions in transition-metal chemistry), structural prototypes (rocksalt, perovskite, spinel), thermodynamic regimes (0 K to finite-temperature ensembles), and their multi-dimensional combinations; such a framework would enable meaningful cross-study comparability and replace the idiosyncratic splits dominating 34 of the 35 reviewed studies [3, 4, 14, 21]. A related shift involves moving beyond domain adaptation, since its reliance on access to target-domain samples—even unlabeled—limits applicability in genuine discovery settings, thereby motivating unsupervised domain generalization approaches grounded in invariant learning and meta-learning paradigms [19, 20]. In parallel, evaluation must move from marginal averages to conditional guarantees, as aggregate OOD metrics can obscure severe instance-level failures, motivating the integration of conformal prediction and input-dependent robustness certificates to yield material-specific reliability estimates [25, 26, 31]. Beyond methodological evaluation, stronger physics-informed inductive biases are required, extending beyond current equivariant graph neural networks [28] toward causal structure learning, thermodynamic consistency constraints, and hybrid quantum-informed representations to support true extrapolation, particularly under multi-axis distributional shifts where existing architectures remain limited despite promising single-axis results [29, 30]. Complementing these advances, publicly accessible OOD leaderboards that stratify performance by shift type and their combinations, reporting both detection AUC and predictive error under standardized protocols, would align community incentives and resolve persistent evaluation inconsistencies noted in prior work [33–35]. Taken together, these directions would reposition OOD generalization as a measurable, engineering-driven discipline, with short-term efforts enabling benchmarks and baselines, medium-term work refining single-axis robustness, and longer-horizon research addressing the still-unresolved multi-dimensional regime that continues to challenge existing methods [14, 15].

Toward Robust OOD Generalization

Success in OOD generalization for materials AI would be defined by concrete, quantifiable milestones. A robust model should predict the formation energy of a novel element combination never seen during training to within <0.1 eV/atom error, forecast properties of low-symmetry crystals after exposure only to high-symmetry prototypes, and extrapolate thermodynamic properties from 0 K training data to 1000 K ensembles with <20 % relative error. Achieving these targets would mark the transition from laboratory demonstration to industrial deployment.

Promising research directions include causal inference tailored to materials graphs, foundation models pre-trained on ultra-diverse chemical spaces, and physics-constrained architectures that embed conservation laws directly into the learning objective. Meta-learning for rapid adaptation—training models to adjust quickly when small amounts of OOD data become available—also offers a practical bridge to real-world use cases [19, 28, 30].

A realistic timeline can be articulated. In the short term (1–2 years), the community should finalize standard OOD benchmarks and establish baseline performance for all five shift types. Medium-term efforts (2–5 years) should focus on reliable methods for single-dimension shifts through hybrid physics–ML architectures and improved invariant learning. Long-term goals (5–10 years) must solve multi-dimensional shifts via causal discovery, large-scale foundation models, and conditional robustness certificates. Only by following this staged roadmap can the field move from the current state—where even the best methods fail on compound shifts [14, 15]—to models that generalize reliably enough for autonomous materials discovery pipelines.

Implications for Practice

For model developers, the key takeaway is transparency: every published model should report OOD performance on all relevant shift axes alongside standard in-distribution metrics. Developers must explicitly state which shift types were tested and provide multiple test splits (random, compositional, prototype, thermodynamic) rather than relying on a single random partition [3, 6, 9]. This practice would immediately expose the brittleness currently hidden by optimistic IID results.

For practitioners deploying materials AI in discovery or process-optimization workflows, IID benchmark scores are not a reliable proxy for real-world performance. Before any deployment decision, models must be stress-tested on a dedicated OOD validation set that mirrors anticipated deployment shifts. Uncertainty quantification should be treated as a safety mechanism: predictions flagged as high-uncertainty must be rejected or subjected to higher-fidelity validation rather than trusted blindly [25, 26].

Benchmark designers and data-curation teams bear equal responsibility. Future databases should be constructed with controlled OOD splits from the outset, and publication standards should mandate reporting of OOD-specific metrics (error under each shift type, detection AUC, calibration error). Public leaderboards organized by shift category would accelerate iterative improvement and prevent the field from optimizing only for the easiest interpolation scenarios [33–35].

Adopting these practices will not solve OOD generalization overnight, but it will ensure that incremental progress is measurable, reproducible, and aligned with the ultimate goal of trustworthy materials AI.

Conclusion

Out-of-distribution generalization remains critical yet largely unsolved in materials AI. This systematic review of peer-reviewed publications has identified four primary domain-shift types—compositional, structural, thermodynamic, and dynamical—plus a realistic multi-dimensional category that combines them. Current methodological families (domain adaptation, invariant learning, data augmentation, uncertainty-based detection, and extrapolation-aware architectures) deliver modest gains for small single-axis shifts but fail sharply on large compositional jumps, prototype changes, thermodynamic extremes, or any multi-dimensional combination.

Empirical findings are consistent across the corpus: domain adaptation and augmentation help within chemically or structurally related families, yet no method reliably crosses the convex hull of the training distribution. Six persistent gaps—no standard benchmarks, underdeveloped extrapolation theory, absent conditional guarantees, neglect of multi-dimensional shifts, detection without prediction, and inconsistent metrics—must be closed before materials AI can transition from laboratory tool to industrial technology.

The path forward requires community consensus on OOD benchmarks, a shift toward unsupervised generalization, deeper physics-based priors, and rigorous conditional robustness evaluation. Only by confronting these challenges head-on can the field deliver models capable of predicting novel compositions, unseen prototypes, and extreme conditions with the reliability demanded by next-generation energy, electronics, and sustainability applications. OOD generalization is no longer an optional research direction; it is the central unsolved problem that will determine whether materials AI fulfills its transformative promise.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 03 Dec 2024

Revised: 04 Apr 2025

Accepted: 30 May 2025

Published online: 18 July 2025

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

Segal N, Netanyahu A, Greenman KP, Agrawal P, Gómez-Bombarelli R. Known unknowns: Out-of-distribution property prediction in materials and molecules. *NPJ Comput Mater.* 2025;11(1):345.

<https://doi.org/10.1038/s41524-025-01808-x>.

Omee SS, Fu N, Dong R, Hu M, Hu J. Structure-based out-of-distribution (OOD) materials property prediction: A benchmark study. *NPJ Comput Mater.* 2024;10(1):144.

<https://doi.org/10.1038/s41524-024-01316-4>.

Li Q, Miklaucic N, Hu J. Out-of-distribution materials property prediction using adversarial learning based fine-tuning [Preprint]. arXiv; 2024. Available from: <https://arxiv.org/abs/2408.09297>

Li K, Rubungo AN, Lei X, Persaud D, Choudhary K, DeCost B, et al. Probing out-of-distribution generalization in machine learning for materials. *Commun Mater.* 2025;6(1):9.

<https://doi.org/10.1038/s43246-024-00731-w>.

Li K, DeCost B, Choudhary K, Greenwood M, Hattrick-Simpers J. A critical examination of robustness and generalizability of machine learning prediction of materials properties. *NPJ Comput Mater.* 2023;9(1):55.

<https://doi.org/10.1038/s41524-023-01012-9>.

Hu J, Liu D, Fu N, Dong R. Realistic material property prediction using domain adaptation based machine learning. *Digit Discov.* 2024;3(2):300-12.

<https://doi.org/10.1039/D3DD00162H>.

Merchant A, Batzner S, Schoenholz SS, Aykol M, Cheon G, Cubuk ED. Scaling deep learning for materials discovery. *Nature.* 2023;624(7990):80-5.

<https://doi.org/10.1038/s41586-023-06735-9>.

Noda K, Wakiuchi A, Hayashi Y, Yoshida R. Advancing extrapolative predictions of material properties through learning to learn using extrapolative episodic training. *Commun Mater.* 2025;6(1):36.

<https://doi.org/10.1038/s43246-025-00754-x>.

Hatakeyama-Sato K, Oyaizu K. Generative models for extrapolation prediction in materials informatics. *ACS Omega.* 2021;6(22):14566-74.

<https://doi.org/10.1021/acsomega.1c01716>.

Fu N, Omee SS, Hu J. Physical encoding improves OOD performance in deep learning materials property prediction. *Comput Mater Sci.* 2025;248:113603.

<https://doi.org/10.1016/j.commatsci.2024.113603>.

Kauwe SK, Graser J, Murdock R, Sparks TD. Can machine learning find extraordinary materials? *Comput Mater Sci.* 2020;174:109498.

<https://doi.org/10.1016/j.commatsci.2019.109498>.

Muckley ES, Saal JE, Meredig B, Roper CS, Martin JH. Interpretable models for extrapolation in scientific machine learning. *Digit Discov.* 2023;2(5):1425-35.

<https://doi.org/10.1039/D3DD00082F>.

Park H, Zhu R, Huerta EA, Chaudhuri S, Tajkhorshid E, Cooper D. End-to-end AI framework for interpretable prediction of molecular and crystal properties. *Mach Learn Sci Technol.* 2023;4(2):025036.

<https://doi.org/10.1088/2632-2153/acd434>.

Li Q, Fu N, Omee SS, Hu J. MD-HIT: Machine learning for material property prediction with dataset redundancy control. *NPJ Comput Mater.* 2024;10(1):245.

<https://doi.org/10.1038/s41524-024-01426-z>.

Madani M, Lacivita V, Shin Y, Tarakanova A. Accelerating materials property prediction via a hybrid Transformer Graph framework that leverages four body interactions. *NPJ Comput Mater.* 2025;11(1):15.

<https://doi.org/10.1038/s41524-024-01472-7>.

Shimakawa H, Kumada A, Sato M. Extrapolative prediction of small-data molecular property using quantum mechanics-assisted machine learning. *NPJ Comput Mater.* 2024;10(1):11.

<https://doi.org/10.1038/s41524-023-01194-2>.

Ma Y, Xu P, Li M, Ji X, Zhao W, Lu W. The mastery of details in the workflow of materials machine learning. *NPJ Comput Mater.* 2024;10(1):141.

<https://doi.org/10.1038/s41524-024-01331-5>.

Jiang X, Sun H, Choudhary K, Zhuang H, Nian Q. Interpretable ensemble learning for materials property prediction with classical interatomic potentials. *NPJ Comput Mater.* 2025;11(1):319.

<https://doi.org/10.1038/s41524-024-01468-3>.

Kobayashi W, Otsuka T, Wakabayashi YK, Tei G. Physics-informed Bayesian optimization suitable for extrapolation of materials growth. *NPJ Comput Mater.* 2025;11(1):36.

<https://doi.org/10.1038/s41524-025-01522-8>.

Liao W, Yuan R, Xue X, Wang J, Li J, Lookman T. Unsupervised learning-aided extrapolation for accelerated design of superalloys. *NPJ Comput Mater.* 2024;10(1):171.

<https://doi.org/10.1038/s41524-024-01358-8>.

Schultz LE, Wang Y, Jacobs R, Morgan D. A general approach for determining applicability domain of machine learning models. *NPJ Comput Mater.* 2025;11(1):95.

<https://doi.org/10.1038/s41524-025-01573-x>.

Wen X, Liu H, Long W, Wei S, Zhu R. Consistent semantic representation learning for out-of-distribution molecular property prediction. *Brief Bioinform.* 2025;26(2).

<https://doi.org/10.1093/bib/bbaf147>.

Chuah J, Kruger U, Wang G, Yan P, Hahn J. Framework for testing robustness of machine learning-based classifiers. *J Pers Med.* 2022;12(8):1314.
<https://doi.org/10.3390/jpm12081314>.

Panigrahi B, Razavi S, Doig LE, Cordell B, Gupta HV, Liber K. On robustness of the explanatory power of machine learning models: Insights from a new explainable AI approach using sensitivity analysis. *Water Resour Res.* 2025;61(3).
<https://doi.org/10.1029/2024WR037398>.

Zhang H, Zhao Z, Wang C, Zhang X, Chen X. Mitigating domain shift in online process monitoring for material extrusion additive manufacturing via transfer learning. *Addit Manuf.* 2024;94:104467.
<https://doi.org/10.1016/j.addma.2024.104467>.

Wang H, Li K, Ramsay S, Fehlis Y, Kim E, Hattrick-Simpers J. Evaluating the performance and robustness of LLMs in materials science Q&A and property predictions. *Digit Discov.* 2025;4(6):1612-24.
<https://doi.org/10.1039/D5DD00090D>.

Zou D, Liu S, Miao S, Fung V, Chang S, Li P. GeSS: Benchmarking geometric deep learning under scientific applications with distribution shifts. In: *Proceedings of the 38th Conference on Neural Information Processing Systems*; 2024 Dec 10-15; Vancouver, Canada. p. 92499-528.
<https://doi.org/10.52202/079017-2937>.

Onwuli A, Butler KT, Walsh A. Ionic species representations for materials informatics. *APL Mach Learn.* 2024;2(3):036112.
<https://doi.org/10.1063/5.0227009>.

Yang Y, Zhang Y, Song X, Xu Y. Not all out-of-distribution data are harmful to open-set active learning. In: *Advances in Neural Information Processing Systems 36*; 2023 Dec 10-16; New Orleans, LA. p. 13802-18.
<https://doi.org/10.52202/075280-0608>.

Kong D, Pang B, Han T, Wu YN. Molecule design by latent space energy-based modeling and gradual distribution shifting. In: Evans RJ, Shpitser I, editors. *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*; 2023 Jul 31-Aug 4; Pittsburgh, PA. *Proceedings of Machine Learning Research.* 2023;216:1109-20.

Atghaei A, Rahmati M. Domain generalization via geometric adaptation over augmented data. *Knowl Based Syst.* 2025;309:112765.
<https://doi.org/10.1016/j.knosys.2024.112765>.

Nematov D, Hojamberdiev M. Machine learning-driven materials discovery: Unlocking next-generation functional materials – a review. *Comput Condens Matter.* 2025;45.
<https://doi.org/10.1016/j.cocom.2025.e01139>.

Tonguz OK, Taschin F. On the measurement of distribution shift in machine learning systems. *IEEE Intell Syst.* 2025;40(2):45-54.
<https://doi.org/10.1109/MIS.2024.3524798>.

Zhu R, Tian SI, Ren Z, Li J, Buonassisi T, Hippalgaonkar K. Predicting synthesizability using machine learning on databases of existing inorganic materials. *ACS Omega.* 2023;8(9):8210-8.
<https://doi.org/10.1021/acsomega.2c04856>.

Balendran A, Beji C, Bouvier F, Khalifa O, Evgeniou T, Ravaud P, et al. A scoping review of robustness concepts for machine learning in healthcare. *NPJ Digit Med.* 2025;8(1):38.
<https://doi.org/10.1038/s41746-024-01420-1>.