

ORIGINAL RESEARCH

Open access

Redefining “Out-of-Distribution” for Crystalline Materials Graphs: A Boundary for Domain Shift Detection

Bruno Martins^{1*}, Lucas Pereira¹, Renata Azevedo², Pedro Costa¹

Abstract

Out-of-distribution (OOD) generalization has become a central claim in machine learning for materials discovery, yet its meaning remains unstable when crystalline materials are represented as graphs. In current practice, the term is applied to qualitatively different forms of domain shift without specifying which properties of the training support have actually been violated, rendering many OOD claims difficult to verify or compare. This article addresses that conceptual gap through a boundary-focused analysis of crystalline graph learning. It shows that prevailing usage conflates multiple non-equivalent shifts and argues that conventional vector-space definitions of OOD are inadequate for periodic, graph-structured materials data. In response, the paper identifies four primary dimensions along which crystalline graph distributions depart from training support: composition, structure, scale, and condition. It then proposes a dimension-explicit redefinition of OOD, together with measurable boundary criteria, operational detection rules, and a per-dimension domain-shift score that can be computed from characterized training distributions. Boundary cases and gray zones are examined to clarify how formally defined thresholds should be interpreted in practice. By distinguishing OOD from anomaly detection, novelty detection, extrapolation, and domain adaptation, the framework establishes a more precise conceptual foundation for evaluating generalization in crystalline materials machine learning. The central contribution is not a new predictive model, but a falsifiable vocabulary for reporting domain shift. Adopting dimension-specific OOD reporting would make claims of robustness more reproducible, benchmark design more informative, and model evaluation more scientifically defensible in AI-driven materials discovery.

Keywords Materials informatics, Graph neural networks, Domain shift, Out-of-distribution generalization, Crystalline materials graphs, Distribution shift dimensions

*Correspondence:

Bruno Martins

bruno.martins@gmail.com

¹ Department of Materials Data Analytics, Faculty of Engineering, University of Minho, Braga, Portugal

² Department of Computational Materials Innovation, Faculty of Science and Technology, University of Porto, Porto, Portugal

Introduction

The term “out-of-distribution” (OOD) appears frequently in materials ML papers. Authors claim their models generalize OOD — but what does this actually mean? A crystal tested from a different chemical family? A different crystal prototype? A larger unit cell? A higher temperature? The term is used ambiguously. This paper provides a boundary/definitional analysis of OOD for crystalline materials graphs and proposes a precise redefinition with operational criteria for domain shift detection.

Crystalline materials are routinely represented as graphs in contemporary machine-learning workflows. Foundational databases such as the Materials Project have supplied the large, structured datasets required for training graph-based

models [1]. Architectures including SchNet [2], crystal graph convolutional neural networks (CGCNN) [3], and universal graph networks [4] have demonstrated remarkable accuracy in property prediction when test and training distributions remain aligned [5]. Yet the field has rapidly moved toward more ambitious generalization settings. Transfer-learning strategies [6-10], self-supervised pretraining [11], and cross-property frameworks [9] are now routinely presented as solutions to "OOD" challenges. At the same time, reviews of OOD generalization [12] and domain shift in materials databases [13] have begun to document the practical difficulties that arise when models encounter unseen chemistries, structures, or conditions.

Despite this activity, the community lacks a shared operational definition of OOD tailored to crystalline graphs. In vector-based ML (images, tabular data), OOD is often framed via distance in feature space or via simple class novelty. Crystalline graphs, however, encode periodic atomic connectivity, variable cell sizes, space-group symmetry, and thermodynamic metadata. These attributes generate multiple, partially orthogonal axes of variation. A test crystal may be in-distribution with respect to elemental composition yet out-of-distribution with respect to structural prototype; current binary OOD labels cannot capture such nuance. Consequently, statements such as "our model achieves strong OOD performance" are unverifiable without explicit specification of which distribution(s) shifted and by how much.

The present work addresses this gap through a strictly conceptual, boundary-focused lens. It does not introduce new experiments, datasets, or performance metrics. Instead, it (i) catalogs the inconsistent usages of OOD currently found in the literature, (ii) explains why standard definitions collapse for graph-structured crystals, (iii) isolates the four dominant dimensions of shift that matter for crystals, (iv) supplies an explicit redefinition together with boundary criteria, (v) operationalizes detection along each dimension, and (vi) examines ambiguous boundary cases. The analysis draws exclusively on the cited body of work to ground each conceptual move. By replacing vague terminology with dimension-specific, measurable boundaries, the paper aims to make future claims of OOD generalization precise, falsifiable, and comparable across studies.

How "OOD" Is Currently Used in Materials ML

A survey of recent literature reveals at least six qualitatively different ways the term "OOD" (or closely related phrases such as "domain shift" or "generalization to unseen data") is invoked when crystalline materials are modeled as graphs.

Table 1 disentangles the major ways "OOD" is currently used in crystalline materials ML and shows that these usages refer to qualitatively different kinds of shift that should not be treated as analytically interchangeable.

Table 1. Typology of OOD usages in crystalline materials machine learning and their corresponding hidden sources of ambiguity

OOD usage in literature	What is implicitly being shifted	Typical example in crystalline ML	Why this is not equivalent to other OOD claims	Hidden ambiguity if reported only as "OOD"	Recommended reporting language
Elemental OOD	Atomic vocabulary / element identity support	Train on oxides, test on sulfides or nitrides	Changes node-type support and chemical priors rather than topology alone	A model may fail because of unseen species, not because of broader domain novelty	"Compositional OOD with respect to unseen element support"
Structural OOD	Prototype, space group,	Train on perovskites, test	Changes graph topology and	Can be mislabeled as	"Structural OOD with respect to unseen

	coordination topology	on spinels or layered phases	symmetry even when composition overlaps	chemistry shift when the true issue is structural support	prototype / space-group support"
Scale OOD	Graph size, unit-cell atom count, long-range periodic extent	Train on small primitive cells, test on larger supercells	Challenges message-passing depth and graph-size generalization rather than chemistry	Often conflated with structural novelty despite being a size-support problem	"Scale OOD with respect to unit-cell size / volume support"
Defect-concentration OOD	Disorder regime, occupancy state, defect density	Train on pristine crystals, test on vacancy-containing lattices	Alters physical state and local graph regularity, not merely composition	May be wrongly classified as structural OOD although the shift is regime-based	"Condition OOD with explicit defect-state shift"
Condition OOD	Temperature, pressure, thermodynamic regime	Train on 0 K DFT structures, test on finite-temperature MD snapshots	Same nominal material may occupy a different physical regime	Easy to overlook because composition and topology can appear unchanged	"Condition OOD with explicit thermodynamic shift"
Dataset shift	Repository composition, curation logic, sampling frame	Train on Materials Project, test on OQMD	Database change does not itself specify what material attributes shifted	Gives the illusion of OOD without identifying the actual shifted dimension	"Cross-dataset evaluation; specify compositional, structural, scale, and/or condition differences explicitly"

Within the emerging literature on machine learning for materials discovery, the notion of out-of-distribution generalization has been mobilized in ways that reflect markedly different underlying assumptions about what constitutes a meaningful shift. One prominent interpretation equates distributional departure with compositional novelty, such that a model trained exclusively on oxides is evaluated on sulfides or nitrides, and any introduction of an unseen element is treated as a boundary crossing [7, 12]. Here, the emphasis falls on the expansion of the atomic vocabulary encoded within the graph representation, implicitly framing generalization as the capacity to extrapolate across chemical identities. A related interpretation relocates this boundary from composition to structure, where the transition between crystal prototypes—such as from rocksalt or perovskite to spinel or layered configurations—defines the distributional shift, even when elemental constituents remain unchanged [11, 13]. In this case, the challenge is recast in terms of symmetry, coordination environments, and connectivity patterns, foregrounding structural diversity as the primary axis of variation.

This conceptual landscape becomes further differentiated when attention turns to scale and disorder. Variations in unit-cell size introduce a distinct form of shift, in which models trained on primitive cells comprising relatively few atoms are confronted with supercells or large-unit-cell compounds that extend graph diameter and amplify long-range periodic interactions [4, 14]. Under these conditions, the difficulty lies not in new chemical species or structural motifs per se, but in the altered spatial extent over which interactions must be coherently represented. Closely related, yet mechanistically distinct, is the transition from idealized to defective systems, where pristine lattices give way to vacancy- or substitution-containing configurations [15–17]. This shift introduces local perturbations and disorder, challenging the model's ability to reconcile deviations from periodicity with learned representations of crystalline order. Beyond these structural and spatial considerations, thermodynamic context introduces another layer of complexity. Training regimes grounded in zero-kelvin

density functional theory are frequently juxtaposed with test scenarios involving finite-temperature molecular dynamics or high-pressure phases, thereby extending the notion of distributional shift into the domain of physical conditions and state variables [18–20]. In parallel, some studies operationalize distributional change at the level of data provenance itself, treating transitions between repositories—such as from the Materials Project to OQMD—as indicative of out-of-distribution behavior, despite substantial overlap in the underlying physics and computational methodologies [1, 13].

Such heterogeneity in usage is not merely terminological; it shapes how empirical claims are constructed and interpreted. Work on transfer learning and data-efficient modeling often invokes one or more of these notions without articulating their distinctions, thereby collapsing multiple axes of variation into a single evaluative category [6–10]. Similarly, architectures ranging from self-supervised crystal-twin models to graph transformers are frequently described as exhibiting robustness under distributional shift when evaluated on datasets that differ along any of these dimensions [11, 21]. Even surveys explicitly concerned with generalization tend to enumerate these scenarios in parallel, without establishing criteria for differentiating their epistemic or methodological implications [12].

What emerges from this pattern is not a lack of recognition that these shifts matter, but rather an absence of conceptual separation between fundamentally different forms of generalization. A change in elemental composition modifies the set of permissible node types within a graph, whereas a shift in crystal prototype reorganizes the relational structure through which those nodes interact. Alterations in scale reshape the effective range of dependencies, while variations in thermodynamic conditions redefine the physical regime that the representation is intended to capture. When these distinct transformations are subsumed under a single label, the resulting category becomes analytically diffuse. A model may exhibit sensitivity to compositional novelty while remaining robust to structural variation under a fixed element set, yet both outcomes are often reported as equivalent instances of failure or success. Under these circumstances, claims regarding out-of-distribution generalization lose their discriminative power, rendering comparisons across studies increasingly ambiguous and, in practice, difficult to falsify.

Why Current Definitions Fail for Crystalline Graphs

Conventional formulations of out-of-distribution generalization, largely derived from vector-based learning paradigms, rest on the assumption that samples are independently drawn from a distribution embedded in a well-defined metric space. This premise becomes increasingly untenable in the context of crystalline materials, where graph-based representations encode interdependent chemical and structural information that resists reduction to Euclidean geometry. Nodes and edges simultaneously capture discrete elemental identities, continuous spatial coordinates, and periodic boundary conditions, producing representations for which no canonical embedding—and thus no universally meaningful distance metric—can be established [3, 4]. Although graph edit distances and kernel-based similarities offer partial remedies, their dependence on arbitrary design choices and their computational intractability at scale limit their utility as foundations for defining distributional proximity.

This representational tension is further complicated by the fact that distributional shift in materials systems is inherently multi-axial. A crystal may remain well within the training distribution in terms of elemental composition while diverging significantly in structural prototype or unit-cell scale, creating hybrid regimes that are simultaneously familiar and novel. Under such conditions, the imposition of a binary in-/out-of-distribution label collapses distinct modes of variation into a single category, obscuring the specific mechanisms through which generalization succeeds or fails [12, 13]. The notion of distributional membership thus becomes intrinsically vector-valued, demanding a more granular conceptualization than current practice typically affords.

Ambiguity intensifies in the absence of principled criteria for delimiting distributional boundaries. In vector spaces, thresholds based on statistical dispersion or geometric enclosure provide operational definitions of what lies beyond the training support. Crystalline systems, by contrast, lack any broadly accepted standard for determining when variation becomes qualitatively distinct. Incremental changes in composition, such as low-level impurities, or modest expansions in unit-cell size may be interpreted as either negligible perturbations or substantive departures, depending on context. The

literature offers no consistent guidance on where such thresholds should be drawn, leaving key distinctions effectively underdetermined [22].

The challenge is compounded by the intrinsic coupling between composition and structure. Variations in elemental makeup often induce corresponding transformations in coordination environments or even entirely different space groups, reflecting the underlying physicochemical constraints that govern crystal formation. As a result, attempts to isolate compositional shifts from structural ones frequently collapse under closer scrutiny, since changes along one dimension tend to propagate across others [14, 17]. What is nominally framed as a controlled perturbation may, in practice, entail a cascade of correlated alterations that undermine the coherence of single-axis interpretations.

A further obstacle lies in the limited characterization of training distributions themselves. Empirical studies typically report aggregate statistics—such as dataset size or elemental coverage—while omitting finer-grained descriptors, including the distribution of space groups, unit-cell sizes, or thermodynamic conditions represented in the data [1, 4]. Without an explicit account of this underlying support, claims regarding distributional novelty cannot be independently assessed, and the distinction between interpolation and extrapolation remains opaque.

Taken together, these considerations reveal that existing definitions of out-of-distribution generalization are ill-suited to the structural and physical complexity of crystalline graph learning. Apparent robustness under one interpretive framework may dissolve under another, exposing the extent to which current usage lacks falsifiability and comparability. Progress in this area depends on a reformulation that is commensurate with the representational realities of materials data—one that explicitly articulates the dimensions along which variation occurs, incorporates the relational structure of graphs, and establishes measurable criteria for identifying genuine departures from the training domain.

Dimensions of Distribution Shift for Crystals

Distribution shift in crystalline graph learning is more precisely understood as arising along four interdependent dimensions that reshape both representation and inference. Compositional variation manifests when test crystals introduce previously unseen elements, extend elemental concentrations beyond the empirical range, or form higher-order combinations such as the transition from binaries to ternaries; because node identities are element-specific, such variation perturbs the effective vocabulary and feature manifold available to the model [7–9]. This reconfiguration of chemical space is closely paralleled by structural divergence, in which crystals instantiate space groups or prototypes absent from training, exhibit altered coordination environments, or incorporate disorder through partial occupancy or amorphous character. Under these conditions, architectures grounded in message passing or attention over local neighborhoods encounter a genuine shift in graph topology rather than a simple perturbation of known patterns [3, 4, 11, 21, 23].

A further layer of complexity emerges through scale variation, where discrepancies in unit-cell size—reflected in atom counts, lattice parameters, or volumetric deviations—extend beyond the learned distribution. Such changes amplify the depth and range over which information must propagate, thereby modulating the balance between local and long-range interactions within the learned representation [14, 24, 25]. Beyond these geometric considerations, condition-dependent variation introduces shifts in thermodynamic context or defect states, as when models trained on static 0 K configurations are applied to finite-temperature structures, or when pressure regimes and defect densities diverge between training and evaluation. In practice, these factors reframe the physical meaning encoded in otherwise similar graph topologies [15, 18–20]. Although additional factors such as magnetic ordering or surface termination may intervene, they are typically absorbed within these broader categories. The significance of this framework lies in clarifying that generalization is not uniformly degraded; rather, it is differentially sensitive to how learned representations intertwine chemistry, structure, and scale, such that robustness along one dimension does not preclude failure along another.

Proposed Redefinition

Out-of-distribution (OOD) behavior in crystalline graph learning is more rigorously defined by departures from the training distribution along distinct but interacting dimensions. A test crystal is OOD when compositional variation introduces elements absent from training or drives elemental concentrations beyond the observed range by more than 2 standard deviations (or 20% absolute), thereby perturbing the learned chemical feature space. This disruption extends to structural divergence when the underlying space group or prototype is unrepresented in training, or when coordination environments deviate by over 30%, constituting a topological shift relative to known configurations. Such discrepancies are further amplified by scale variation, where unit-cell atom counts or volumes exceed the training distribution by more than 50%, altering the effective depth and range of graph-based interactions. Beyond these geometric and chemical factors, condition-dependent variation arises when temperature or pressure diverges from training regimes by more than 300 K or 10 GPa, unless such variability is explicitly encoded, thereby reframing the physical interpretation of identical graph structures.

This formulation imposes a necessary constraint on claims of OOD generalization, requiring explicit specification of the dimension and magnitude of shift; absent such detail, such claims remain underspecified. The framework is intentionally conservative, replacing diffuse notions of novelty with verifiable criteria grounded in observable graph attributes, including elemental composition, structural labels, scale, and thermodynamic metadata. Its compatibility with existing graph-based architectures follows directly from this attribute-level grounding [2-4, 11, 14, 17, 21, 23-26]. While the numerical thresholds serve as pragmatic initializations subject to refinement, the central contribution lies in isolating orthogonal axes of variation, thereby enabling precise attribution of model failure and a more interpretable account of generalization.

Operational Boundaries for Domain Shift Detection

Detection of distributional shift in crystalline graphs can be operationalized through boundary conditions defined over observable attributes of the training distribution. Compositional deviation is captured by the element-frequency vector, where the presence of any element with zero frequency constitutes OOD, while elements occurring in fewer than 1% of training instances indicate a near-OOD regime; the extent of deviation is naturally quantified by the count of such novel elements. This criterion is complemented by structural assessment, in which the empirical space-group distribution is coupled with normalized graph-edit distance to the closest training prototype, rendering a crystal OOD when its space group is unobserved or when topological divergence exceeds 0.3. A related constraint emerges at the level of scale, where the number of atoms per unit cell provides a direct proxy: values surpassing the maximum observed in training, or exceeding the median by more than twice the interquartile range, signal a breakdown of learned size-dependent regularities. Beyond these structural and geometric considerations, condition-dependent variation is evaluated through thermodynamic parameters, such that deviations in temperature beyond two standard deviations from the training mean indicate OOD when variability is present, whereas in fixed-temperature regimes—commonly centered at 0 K—any temperature above 100 K suffices; an analogous criterion applies to pressure. Under multi-dimensional shift, each axis is reported independently, preserving the structure of deviation rather than collapsing it into a single binary label.

Figure 1 visualizes the proposed decision architecture for defining OOD in crystalline materials graphs, showing how a characterized training distribution is translated into four explicit boundary checks—compositional, structural, scale, and condition—whose combined outcomes yield a dimension-specific OOD profile rather than a single binary label.

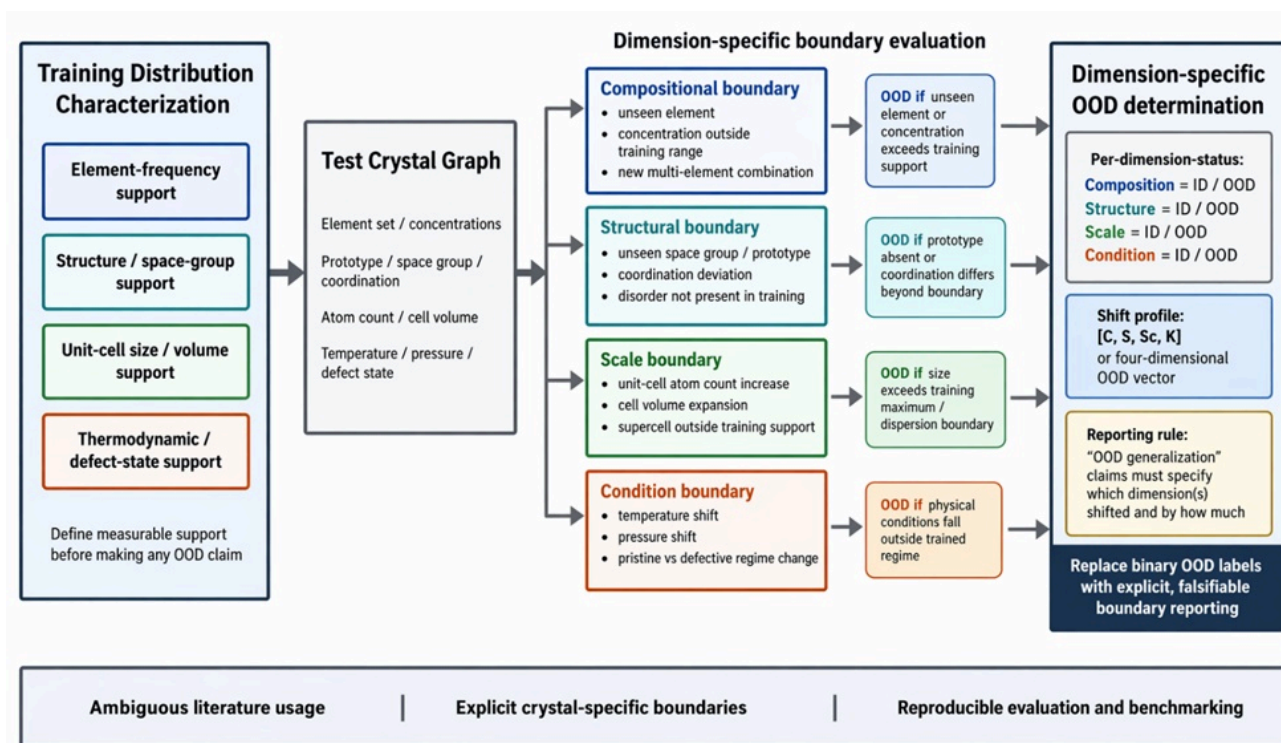


Figure 1. Hierarchical decision architecture for defining out-of-distribution (OOD) status in crystalline materials graphs. A characterized training distribution is evaluated against a test crystal along four explicit dimensions—composition, structure, scale, and condition. OOD status is assigned per dimension using boundary criteria, producing a reportable multi-dimensional shift profile rather than a single binary label.

A conceptual domain shift score along one dimension can be defined as $S_d = \frac{|\mu_{train,d} - \mu_{test,d}|}{\sigma_{train,d}}$ where $\mu_{train,d}$ and $\sigma_{train,d}$ are the mean and standard deviation of the training distribution along dimension d , and $\mu_{test,d}$ is the corresponding test value. Interpretation: $S_d > 2$ means the test distribution is significantly shifted (outside ~95 % of the training range). Scores are reported per dimension and can be inspected jointly to produce a multi-dimensional shift profile.

These operational boundaries are computationally cheap, require only summary statistics of the training set, and are independent of any particular model architecture. They therefore enable consistent, reproducible OOD labeling across studies.

Boundary Cases and Gray Zones

Although the proposed redefinition imposes sharp boundaries, crystalline graphs encountered in practice often reside in intermediate regimes that require careful interpretation. A compositional edge case arises when an element appears only sparsely in training—on the order of 0.1% and at trace concentrations—yet is introduced at substantially higher abundance in testing; once concentration exceeds the empirical range by more than 2σ or 20% absolute, the sample is unambiguously compositional OOD, regardless of prior rarity, precluding unwarranted claims of robustness to infrequent species [7, 8]. A related ambiguity emerges in structurally proximate systems, such as the transition from cubic to tetragonal perovskite, where the absence of the exact space group formally induces structural OOD, even as coordination environments remain nearly invariant; here, auxiliary similarity metrics become necessary to distinguish minor symmetry perturbations from substantive topological change [11, 23].

This tension between formal criteria and practical interpretation extends to scale, where modest increases in unit-cell size—such as a 10% expansion beyond the training maximum—do not activate the OOD threshold yet warrant explicit reporting as near-boundary cases to avoid overstating generalization [14, 24]. An analogous boundary condition appears under fixed thermodynamic training, particularly at 0 K, where any finite-temperature evaluation constitutes condition OOD by construction, rendering tolerance thresholds inapplicable and emphasizing the restrictive nature of such training regimes [18–20]. The complexity intensifies when multiple axes shift simultaneously, as in high-entropy compositions adopting previously unseen structures; in these cases, compositional and structural deviations must be reported independently, since their effects on model performance are neither interchangeable nor reducible to a single label [15, 17].

In each gray zone the operational rule is identical: declare the violated dimension(s), quantify the deviation using the domain-shift score $S_d = \frac{|\mu_{train,d} - \mu_{test,d}|}{\sigma_{train,d}}$, and report the margin to the boundary. Binary OOD/ID labels are replaced by a four-dimensional vector of shift magnitudes. This practice transforms ambiguous boundary cases into reproducible, diagnosable test instances rather than sources of irreproducible claims.

Table 2 converts the proposed framework into a practical decision matrix by specifying what must be characterized, what constitutes a boundary violation, how gray zones should be interpreted, and what authors must report for each shift dimension.

Table 2. Dimension-specific decision matrix for OOD classification, near-boundary interpretation, and reporting obligations in crystalline graph learning

Primary dimension	Core training support to characterize	Boundary trigger for OOD classification	Near-boundary / gray-zone interpretation	Minimum evidence authors should report	Main evaluation implication
Compositional	Element set, element frequencies, concentration ranges, combination order (binary/ternary/etc.)	Unseen element; concentration outside observed range beyond stated threshold; unseen multi-element combination	Rare but previously seen elements at modest concentration should be flagged as near-OOD, not automatically OOD	Element histogram, concentration range, exact violated rule, magnitude of deviation	Separates vocabulary failure from broader claims of generalization
Structural	Space-group counts, prototype inventory, coordination-number distribution, disorder representation	Unseen space group or prototype; coordination deviation beyond stated boundary; disorder regime absent from training	Closely related symmetry variants may be OOD but should be reported with similarity margin	Space-group/prototype support, nearest known prototype, coordination deviation, similarity metric if used	Reveals whether the model fails on topology rather than chemistry
Scale	Atom-count distribution, cell-volume distribution,	Unit-cell size or volume beyond training maximum or	Slightly larger cells should be treated as near-OOD when	Min/max, median, spread statistics, exact size difference, formal threshold applied	Tests graph-size robustness and limits of

	lattice-parameter range	robust spread boundary	marginally beyond common support but below formal threshold		message passing
Condition	Temperature range, pressure range, defect-state support, pristine/defective regime coverage	Test condition lies outside trained thermodynamic or defect-state regime	Fixed-condition training makes even modest departures analytically important	Temperature/pressure range, defect-state regime, whether training is fixed or variable, condition gap	Prevents false claims of robustness across physical regimes
Multi-dimensional coupled shift	Joint profile across all four primary dimensions	OOD on more than one axis simultaneously	Must not be collapsed into a single undifferentiated label; report the full shift profile	Per-dimension status, per-dimension margin, full OOD vector, any scalar summary used	Enables diagnosis of compound failure modes and cleaner benchmark design

Relation to Other Concepts

The proposed dimension-explicit OOD definition sits at the intersection of several neighboring concepts yet remains distinct.

OOD versus anomaly detection. Anomaly detection (e.g., unsupervised outlier methods applied to X-ray diffraction or atomic-resolution images) flags statistically rare samples within an assumed single distribution [27–29]. OOD, by contrast, is broader: a perfectly common perovskite at 1000 K is condition OOD even if its diffraction pattern is statistically typical. Anomaly detection asks “how unusual?”; OOD asks “does the sample lie inside the explicitly bounded training support along any of the four crystal-specific axes?”

OOD versus novelty detection. Novelty detection treats the appearance of an entirely new class absent from training [16]. The present framework subsumes novelty (e.g., entirely new element or new space group) but also encompasses continuous shifts such as concentration ranges or cell-size scaling that are not discrete “new classes.” A crystal with the same elements and space group but extreme concentration is OOD yet not novel.

OOD versus extrapolation. Extrapolation concerns prediction outside the convex hull of training feature vectors [22]. Crystal graphs lack a natural Euclidean hull; moreover, OOD can occur inside the convex hull of one dimension while outside another. Thus OOD properly includes—but is not limited to—extrapolation.

OOD versus domain adaptation. Domain adaptation assumes a known target distribution different from the source and seeks to align representations or reweight samples [9, 10]. OOD detection is the prerequisite step: it identifies which test crystals belong to a shifted domain and along which axis. Only after dimension-specific OOD labeling can targeted adaptation (e.g., compositional fine-tuning) be applied productively.

By separating these concepts, the framework prevents conflation. A model that succeeds on anomaly detection inside the training hyper-rectangle cannot claim OOD generalization. Conversely, a model that fails on a coupled compositional-structural shift cannot be rescued by generic anomaly detectors. The four-dimensional boundary criteria therefore serve as a common language that sharpens the distinction between related but non-identical generalization regimes.

Implications for Model Evaluation

Adoption of the proposed redefinition carries immediate, concrete consequences for three stakeholder groups.

For model developers: cease using the unqualified phrase “OOD generalization.” Every claim must be replaced by a four-tuple specifying the shifted dimension(s) and the corresponding Sd values. Training distributions must be fully characterized—element-frequency histograms, space-group counts, unit-cell-size statistics, and thermodynamic metadata—before any OOD statement is issued. Performance must be reported separately on each isolated dimension (e.g., compositional OOD only, structural OOD only) rather than on an undifferentiated “OOD test set.” This practice reveals which architectural choices (message-passing depth, attention mechanisms, invariant representations) are robust to specific shifts [3, 4, 11, 21].

For benchmark designers: construct test splits that isolate single dimensions while holding the others fixed. Materials Project-derived benchmarks [1] can be augmented with controlled compositional, structural, scale, and condition partitions. Each benchmark release must include the training hyper-rectangle boundaries and per-dimension Sd statistics. The “OOD gap” (in-distribution accuracy minus dimension-specific OOD accuracy) becomes the primary reporting metric. Cross-property or transfer-learning benchmarks [8, 9] should explicitly label which dimensions are being transferred.

For reviewers and editors: reject any manuscript that invokes “OOD generalization” without dimension specification and training-distribution characterization. Standard review questions should now include: “What does OOD mean in this paper? Which of the four dimensions shifted, and by how much according to the proposed criteria?” Supplementary materials must contain the training marginal statistics and the exact boundary thresholds applied. These requirements raise the evidentiary bar without demanding new experiments or simulations.

Collectively, these changes transform OOD claims from rhetorical flourishes into falsifiable scientific statements. Model cards and benchmark leaderboards will list performance vectors rather than single scalar “OOD accuracy” numbers. The field thereby moves from unverifiable assertions of robustness toward reproducible, dimension-resolved understanding of generalization limits in crystalline graph models.

Conclusion

Out-of-distribution generalization in crystalline materials machine learning cannot be reliably assessed when treated as a single binary category. Existing practice collapses distinct forms of distribution shift, despite the fact that compositional, structural, scale, and condition variation affect graph-based models through different mechanisms. Because crystalline graphs encode coupled chemical, topological, and thermodynamic information, conventional definitions of OOD do not provide a sufficient basis for evaluating generalization.

This work has introduced a dimension-explicit redefinition of OOD grounded in four primary axes of shift, each paired with measurable boundary criteria and operational detection procedures. The resulting framework enables OOD to be specified, quantified, and reported in a reproducible manner using only training distribution statistics. By preserving the structure of distributional variation rather than reducing it to a binary label, it supports clearer attribution of model performance and more interpretable evaluation outcomes.

The implications are methodological. Claims of OOD generalization should identify the shifted dimension, quantify the magnitude of deviation, and report results accordingly. Benchmark design and evaluation protocols should reflect the same structure. Adopting this standard provides a consistent and falsifiable basis for assessing generalization, improving comparability across studies and strengthening the reliability of machine learning approaches in materials discovery.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 16 Apr 2022

Revised: 28 Jul 2022

Accepted: 25 Nov 2022

Published online: 18 January 2023

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

- Wang Y, Liu Y, Song S, Yang Z, Qi X, Wang K, et al. Accelerating the discovery of insensitive high-energy-density materials by a materials genome approach. *Nat Commun.* 2018;9(1):2444.
<https://doi.org/10.1038/s41467-018-04897-z>.
- Schütt KT, Sauceda HE, Kindermans PJ, Tkatchenko A, Müller KR. SchNet: A deep learning architecture for molecules and materials. *J Chem Phys.* 2018;148(24):241722.
<https://doi.org/10.1063/1.5019779>.
- Xie T, Grossman JC. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys Rev Lett.* 2018;120(14):145301.
<https://doi.org/10.1103/PhysRevLett.120.145301>.
- Chen C, Ye W, Zuo Y, Zheng C, Ong SP. Graph networks as a universal machine learning framework for molecules and crystals. *Chem Mater.* 2019;31(9):3564-72.
<https://doi.org/10.1021/acs.chemmater.9b01294>.
- Zhou Q, Lu S, Wu Y, Wang J. Property-oriented material design based on a data-driven machine learning technique. *J Phys Chem Lett.* 2020;11(10):3920-7.
<https://doi.org/10.1021/acs.jpclett.0c00665>.

Zhang Y, Ling C. A strategy to apply machine learning to small datasets in materials science. *NPJ Comput Mater.* 2018;4(1):25.
<https://doi.org/10.1038/s41524-018-0081-z>.

Yamada H, Liu C, Wu S, Koyama Y, Ju S, Shiomi J, et al. Predicting materials properties with little data using shotgun transfer learning. *ACS Cent Sci.* 2019;5(10):1717-30.
<https://doi.org/10.1021/acscentsci.9b00804>.

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.
<https://doi.org/10.1038/s41467-019-13297-w>.

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.
<https://doi.org/10.1038/s41467-021-26921-5>.

Lee J, Asahi R. Transfer learning for materials informatics using crystal graph convolutional neural network. *Comput Mater Sci.* 2021;190:110314.
<https://doi.org/10.1016/j.commatsci.2021.110314>.

Magar R, Wang Y, Barati Farimani A. Crystal twins: Self-supervised learning for crystalline material property prediction. *NPJ Comput Mater.* 2022;8(1):231.
<https://doi.org/10.1038/s41524-022-00921-5>.

Liu J, Shen Z, He Y, Zhang X, Xu R, Yu H, et al. Towards out-of-distribution generalization: A survey. *arXiv:2108.13624 [Preprint]*. 2021.
<https://doi.org/10.48550/arXiv.2108.13624>.

Sankaranarayanan S, Balaji Y, Jain A, Lim SN, Chellappa R. Learning from synthetic data: Addressing domain shift for semantic segmentation. In: 2018 IEEE/CVF conference on computer vision and pattern recognition; 2018 Jun 18-22; Salt Lake City, UT. Los Alamitos (CA): IEEE Computer Society; 2018. p. 3752-61.
<https://doi.org/10.1109/CVPR.2018.00395>.

Dai M, Demirel MF, Liang Y, Hu JM. Graph neural networks for an accurate and interpretable prediction of the properties of polycrystalline materials. *NPJ Comput Mater.* 2021;7(1):103.
<https://doi.org/10.1038/s41524-021-00574-w>.

Tian L, Fan Y, Li L, Mousseau N. Identifying flow defects in amorphous alloys using machine learning outlier detection methods. *Scr Mater.* 2020;186:185-9.
<https://doi.org/10.1016/j.scriptamat.2020.05.038>.

Guo Y, Kalinin SV, Cai H, Xiao K, Krylyuk S, Davydov AV, et al. Defect detection in atomic-resolution images via unsupervised learning with translational invariance. *NPJ Comput Mater.* 2021;7(1):180.
<https://doi.org/10.1038/s41524-021-00642-1>.

Yang Z, Buehler MJ. Linking atomic structural defects to mesoscale properties in crystalline solids using graph neural networks. *NPJ Comput Mater.* 2022;8(1):198.
<https://doi.org/10.1038/s41524-022-00879-4>.

Zuo Y, Chen C, Li X, Deng Z, Chen Y, Behler J, et al. Performance and cost assessment of machine learning interatomic potentials. *J Phys Chem A.* 2020;124(4):731-45.
<https://doi.org/10.1021/acs.jpca.9b08723>.

Zhang L, Han J, Wang H, Car R, E W. Deep potential molecular dynamics: A scalable model with the accuracy of quantum mechanics. *Phys Rev Lett.* 2018;120(14):143001.
<https://doi.org/10.1103/PhysRevLett.120.143001>.

Axelrod S, Schwalbe-Koda D, Mohapatra S, Damewood J, Greenman KP, Gómez-Bombarelli R. Learning matter: Materials design with machine learning and atomistic simulations. *Acc Mater Res.* 2022;3(3):343-57.
<https://doi.org/10.1021/accountsmr.1c00238>.

Yan K, Liu Y, Lin Y, Ji S. Periodic graph transformers for crystal material property prediction. In: *Advances in neural information processing systems 35*; 2022 Nov 28-Dec 9; New Orleans, LA. Neural Information Processing Systems Foundation, Inc.; 2022. p. 15066-80.
<https://doi.org/10.52202/068431-1096>.

Sutton C, Boley M, Ghiringhelli LM, Rupp M, Vreeken J, Scheffler M. Identifying domains of applicability of machine learning models for materials science. *Nat Commun.* 2020;11(1):4428.
<https://doi.org/10.1038/s41467-020-17112-9>.

Das K, Samanta B, Goyal P, Lee SC, Bhattacharjee S, Ganguly N. CrysXPP: An explainable property predictor for crystalline materials. *NPJ Comput Mater.* 2022;8(1):43.
<https://doi.org/10.1038/s41524-022-00716-8>.

Cheng J, Zhang C, Dong L. A geometric-information-enhanced crystal graph network for predicting properties of materials. *Commun Mater.* 2021;2(1):92.
<https://doi.org/10.1038/s43246-021-00194-3>.

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.
<https://doi.org/10.1109/ACCESS.2022.3181750>.

Louis SY, Zhao Y, Nasiri A, Wang X, Song Y, Liu F, et al. Graph convolutional neural networks with global attention for improved materials property prediction. *Phys Chem Chem Phys.* 2020;22(32):18141-8.
<https://doi.org/10.1039/D0CP01474E>.

Chung HW, Freitas R, Cheon G, Reed EJ. Data-centric framework for crystal structure identification in atomistic simulations using machine learning. *Phys Rev Mater.* 2022;6(4):043801.
<https://doi.org/10.1103/PhysRevMaterials.6.043801>.

Banko L, Maffettone PM, Naujoks D, Olds D, Ludwig A. Deep learning for visualization and novelty detection in large X-ray diffraction datasets. *NPJ Comput Mater.* 2021;7(1):104.
<https://doi.org/10.1038/s41524-021-00575-9>.

Cohn R, Holm E. Unsupervised machine learning via transfer learning and k-means clustering to classify materials image data. *Integr Mater Manuf Innov.* 2021;10(2):231-44.
<https://doi.org/10.1007/s40192-021-00205-8>.