ORIGINAL RESEARCH **Open access**

The Illusion of Generalization: A Critique of Random Train-Test Splitting in Small-Crystal Property Prediction

Claire Martin[1], Julien Robert[2*], Sophie Bernard[1], Antoine Girard[2]

Abstract

The standard practice in machine learning for small-crystal property prediction relies on random train-test splitting of datasets such as the Materials Project. This approach creates an illusion of generalization: models routinely report near-zero mean absolute errors on formation energies, band gaps, or elastic constants, yet these impressive figures reflect leakage of structural, compositional, and energetic information rather than genuine out-of-distribution capability. Random splitting fails because crystals are not independent and identically distributed; repeated prototypes, shared elemental combinations, local coordination environments, and clustered formation energies ensure that train and test sets remain statistically entangled even after random partitioning. We identify a typology of four generalization illusions —prototype, compositional, energy-range, and structural—that systematically mislead the field and explain why published “state-of-the-art” accuracies collapse under more rigorous evaluation regimes. The consequences are severe: wasted experimental validation efforts, inflated claims of progress, overinvestment in architectures that cannot extrapolate, and a delayed recognition of fundamental limitations in current graph-network approaches. We propose six alternative evaluation strategies—composition splits, prototype splits, time splits, structural dissimilarity splits, energy-extrapolation splits, and cross-database splits—that replace random partitioning with deliberate distribution shifts, thereby restoring scientific integrity to benchmark design in computational materials science.

Keywords Materials machine learning, Out-of-distribution generalization, Random train-test splitting, Illusion of generalization, Crystal property prediction, Data leakage

*Correspondence:
Julien Robert
julien.robert@gmail.com
[1] Department of Computational Materials Systems, Faculty of Engineering, University of Lyon, Lyon, France
[2] Department of Intelligent Materials Analytics, Faculty of Science and Technology, University of Strasbourg, Strasbourg, France

Introduction

The problem

A vast literature on machine learning for crystal property prediction uses random train-test splitting on datasets like the Materials Project [1]. Papers report impressive accuracy—often near-zero error on formation energy prediction. But what does this accuracy actually measure? This critique argues that random splitting creates an illusion of generalization. Because crystals are not independent and identically distributed (IID), random splits place structurally and compositionally similar crystals in both train and test sets. The model appears to generalize because it has effectively seen the test crystals before—not identical copies, but close enough. This paper critiques this practice, identifies specific sources of leakage, and proposes alternative evaluation strategies.

We begin by documenting how pervasive the random-split protocol has become. Foundational architectures such as SchNet [2, 3], crystal graph convolutional neural networks [4], and graph networks for molecules and crystals [5] all adopted random 80/20 or 90/10 splits on Materials Project-derived datasets without explicit controls for structural or compositional overlap. Subsequent works, including performance assessments of machine-learning interatomic potentials [6, 7] and general-purpose frameworks for inorganic materials [8], inherited the same protocol, citing its simplicity and apparent reproducibility. Matbench [9] and related benchmark suites further institutionalized random splitting as the default leaderboard metric. These choices were not made in isolation; they rested on an unexamined assumption that crystal data behave like classical machine-learning tabular datasets—independent samples drawn from a stationary distribution [10].

Yet crystal datasets violate this assumption at every level. Structural prototypes (rocksalt, perovskite, spinel) recur across thousands of entries; elemental combinations cluster in composition space; formation energies concentrate near the convex hull; and many entries originate from the same high-throughput workflow, introducing systematic correlations [1, 11, 12]. Random partitioning therefore preserves these clusters, allowing models to exploit prototype-specific motifs or elemental biases rather than learning transferable physical principles. The result is an illusion: reported test-set errors appear excellent precisely because the test set is not truly unseen.

This critique is not an attack on deep-learning architectures themselves. Graph networks and message-passing frameworks have undeniably advanced materials informatics [2, 4, 5, 13, 14]. The issue lies in the evaluation regime that masks their true limitations.

Figure 1 visualizes the manuscript's core causal argument by tracing how random train-test splitting rests on flawed assumptions, generates multiple leakage mechanisms, produces four distinct generalization illusions, and ultimately motivates a shift toward distribution-aware evaluation.

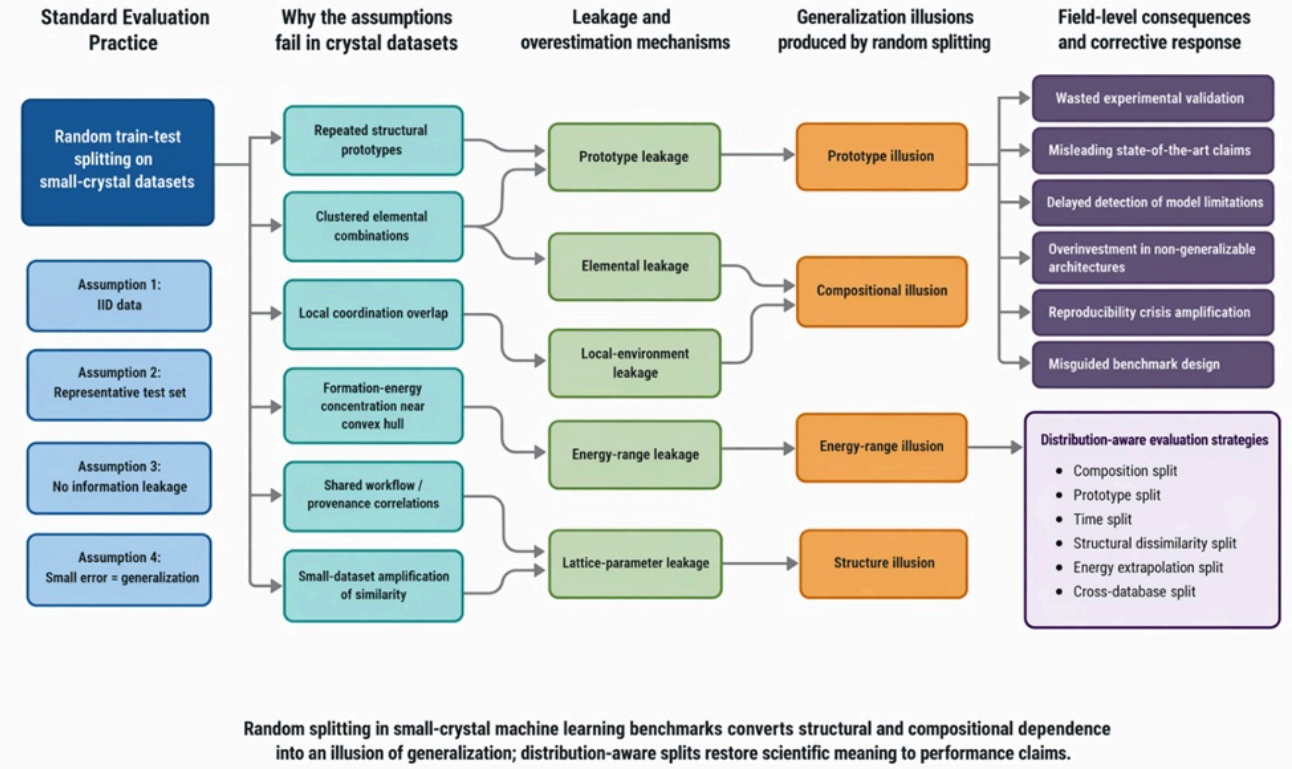

Figure 1. Random train-test splitting in small-crystal machine-learning benchmarks creates the appearance of strong generalization by preserving structural, compositional, local-environmental, and energetic dependence across train and

test sets. The figure shows the manuscript's full argumentative cascade from flawed assumptions to leakage mechanisms, generalization illusions, field consequences, and distribution-aware corrective strategies.

By restricting analysis to conceptual and methodological flaws documented in the 2017–2022 literature, we demonstrate that the field has systematically overestimated generalization. We identify five distinct leakage sources, articulate a typology of four illusions, trace six field-level consequences, and propose six concrete alternative strategies. In doing so, we align this critique with broader methodological debates in cheminformatics, bioinformatics, and domain-adaptation research, while deriving actionable implications for benchmark design. The ultimate goal is constructive: to replace an illusory metric with evaluation practices that genuinely test a model's ability to predict properties of crystals it has never encountered in any form.

The Standard Practice and Its Assumptions

The standard protocol for evaluating crystal property predictors rests on a seemingly straightforward operation: the random partitioning of a dataset into training and test splits, typically at fixed ratios such as 80/20 or 90/10, without regard to structural, compositional, or provenance-based criteria. This practice, defined here as a random train-test split, has become deeply embedded in the literature. Foundational studies, including SchNet [2], crystal graph convolutional neural networks [4], graph networks for crystals [5], machine-learning interatomic potentials [6], and the general-purpose framework proposed by Ko *et al.* [8], all adopted random splits derived from Materials Project data [1]. Subsequent benchmark efforts, most notably Matbench [9] and improved crystal graph convolutional network architectures [15], have retained the identical protocol, treating it as the default for reporting accuracy on formation energies, band gaps, and elastic constants [16].

A closer examination, however, reveals that this practice rests on four implicit assumptions, each of which becomes precarious under the specific conditions of crystalline materials data. The first assumption holds that crystals constitute independent and identically distributed samples drawn from a fixed distribution. In practice, this assumption fails because crystal graphs are replete with shared motifs, coordination environments, and symmetry operations; local atomic environments are inherently interdependent, violating independence at a fundamental level [2, 4]. A related implication concerns representativeness: the random test set is presumed to mirror the deployment distribution. Yet the Materials Project itself is systematically biased toward thermodynamically stable or near-stable phases [1, 11], meaning that a random subset cannot faithfully represent the full composition-structure space relevant to genuine materials discovery.

Beyond this immediate concern about representativeness lies a third, more subtle issue: the prevention of information leakage. Random assignment is assumed to guarantee that no structural or compositional similarity crosses the train-test boundary. But because prototypes and elemental combinations are distributed non-uniformly across the dataset, random draws inevitably place similar crystals on both sides of the split [5, 8]. This shift also introduces a fourth assumption, perhaps the most consequential: that low test-set error directly implies good generalization to unseen crystals. That assumption collapses precisely when the test set remains statistically entangled with the training set—under such conditions, low error measures interpolation within the known distribution rather than extrapolation to genuinely novel structures or compositions [6, 9]. The distinction matters acutely in materials discovery, where the goal is rarely to predict properties of crystals similar to those already characterized, but rather to identify high-performance candidates far from the training manifold.

Table 1 consolidates the manuscript's central logic by showing how each implicit assumption of random splitting fails under crystal-data dependence, activates specific leakage pathways, and ultimately produces inflated claims of generalization.

Table 1. From flawed assumptions to field-level distortion: an integrative map of how random splitting inflates generalization claims in crystal property prediction

Standard assumption in random splitting	Why the assumption fails in small-crystal datasets	Primary leakage exposed	Resulting generalization illusion	Analytical implication for interpretation
Crystals are IID samples	Crystal datasets contain repeated prototypes, recurring coordination motifs, and clustered graph environments rather than independent observations [2-4].	Prototype leakage; local-environment leakage	**Prototype illusion** and **structure illusion**	Low error primarily indicates interpolation within familiar structural families, not transferable physical learning.
Random test sets are representative of deployment	Materials databases are biased toward stable or near-stable compounds and incomplete coverage of composition-structure space [1, 13, 14].	Energy-range leakage; elemental leakage	**Energy-range illusion** and **compositional illusion**	Reported performance reflects the database's sampling bias rather than readiness for real discovery settings.
Random assignment prevents leakage	Similar prototypes, elemental neighborhoods, lattice regimes, and workflow-related correlations are distributed across both train and test under random splitting [4, 8, 14].	Prototype leakage; elemental leakage; lattice-parameter leakage	All four illusions can coexist simultaneously	Train-test separation is procedural, not epistemic; the test set is not genuinely unseen.
Small test error implies generalization	When train and test distributions remain statistically entangled, low error measures within-distribution interpolation rather than out-of-distribution capability [5, 11].	All leakage sources reinforce one another	All four illusions are misread as evidence of robust prediction	Benchmark scores should be interpreted as conditional on split design, not as direct evidence of scientific generalization.
Benchmark ranking reflects real model progress	Random-split leaderboards reward architectures that exploit database regularities, even when they cannot extrapolate to new prototypes, elements, or energy regimes [3, 10, 11].	Systemic overestimation across leakage channels	Persistent illusion of "state-of-the-art"	Apparent progress may be methodological inflation rather than substantive advancement in materials ML.
Performance claims are portable across studies	Different datasets inherit different prototype frequencies, elemental coverage, and provenance artifacts, making random-split success fragile across contexts [13, 14, 17].	Cross-dataset dependence remains hidden	Illusions survive until deployment or external validation	Reproducibility problems are not accidental; they are built into the evaluation regime itself.

Each assumption fails in the small-crystal regime (<10 000 entries typical of early Materials Project subsets). We critique them not to diminish the technical contributions of the cited works but to expose the methodological fragility that has gone

unexamined.

Conceptually, one can imagine a figure contrasting two splitting regimes. In the left panel, random splitting scatters points (representing crystals) in a structure-composition embedding space such that red (train) and blue (test) points intermingle within the same prototype clusters and elemental neighborhoods. In the right panel, a prototype-aware split cleanly separates clusters, placing all crystals of a given prototype entirely in train or test. The visual contrast reveals how random partitioning preserves local similarity while a principled split enforces distribution shift—the very condition required to test generalization.

By formalizing these assumptions and demonstrating their violation through the lens of existing literature [2, 4-6, 8, 9, 15], we establish that the standard practice is not merely convenient but fundamentally misaligned with the physics of crystalline matter.

Why Random Splitting Creates an Illusion

The practice of random splitting in crystal property prediction creates a systematic overestimation of model generalization—an illusion rooted in the strong clustering of crystal datasets across structure–composition space. Rather than providing an unbiased estimate of out-of-distribution performance, random partitioning preserves these intrinsic clusters across training and test sets, leading to artificially optimistic metrics. Five interrelated mechanisms, each amplified in the small-dataset regime typical of crystallographic informatics, explain why this illusion proves so persistent.

A primary driver is structural motif repetition. Many crystals share identical or closely related prototypes—consider the NaCl, perovskite, or diamond families—and random splitting inevitably places instances of the same prototype on both sides of the partition. Under these conditions, a model can simply memorize prototype-specific graph patterns rather than learn transferable, composition-invariant rules [4, 5]. A related mechanism operates at the level of compositional similarity. Crystals with overlapping elemental sets, such as TiO_2 and ZrO_2, share local coordination environments. Random splitting intermixes these examples, allowing the model to exploit element-specific biases instead of developing universal bonding principles that would generalize across chemically distinct systems [6, 8].

Beyond these structural and compositional effects lies a thermodynamic consideration. Formation energies are not uniformly distributed across composition space; instead, many crystals cluster tightly near the convex hull of stability. Random splitting preserves this clustering, meaning the test set never forces the model to confront truly out-of-hull extrapolations. Consequently, test errors remain artificially low, not because the model has learned robust physics, but because it has only been asked to interpolate within an energetically confined region [1, 11]. Compounding this issue are data generation artifacts. A substantial fraction of crystals in the Materials Project originate from identical source calculations or shared relaxation workflows, introducing systematic correlations in computed properties that random assignment cannot break [12, 18].

Under these conditions, the limited size of typical crystal datasets—often below ten thousand examples—becomes a critical amplifier. The statistical probability of placing structurally or compositionally similar crystals across train and test splits is high precisely when the total sample is small. This leakage, which would be diluted in million-scale molecular datasets, becomes unavoidable and severe in the materials domain [2, 17]. An illustrative example clarifies the resulting mechanism. Consider a model trained on ninety percent of perovskite crystals and tested on the remaining ten percent [19]. Low test error is almost guaranteed, since all members share corner-sharing octahedra and similar tolerance-factor physics. Yet that low error says nothing about generalization to non-perovskite structures—spinels, layered double hydroxides, or any family outside the training manifold. The reported accuracy is therefore not a measure of cross-family extrapolation but merely an artifact of within-family interpolation [4, 15].

These mechanisms are not hypothetical artifacts. They are directly observable in the evaluation protocols of the very papers that popularized graph-network approaches for crystals [2, 4-6, 8]. Random splitting thus creates a self-reinforcing

illusion: high test accuracy is taken as validation of the model, which in turn is read as validation of the splitting method. The field perpetuates a cycle where reported metrics obscure the true generalization deficits, and the illusion persists precisely because it is never directly confronted.

Random splitting generates systematic overestimation of model performance through five concrete yet often overlooked sources of leakage, each rooted in the underlying structure of crystal datasets. The most immediate of these is prototype leakage: the same crystal prototype—whether NaCl-type, perovskite, or diamond—appears in both training and test partitions. Under this condition, a model learns prototype-specific embedding patterns rather than general physical rules, and the resulting overestimation carries a clear detection signature. Accuracy remains high on test crystals whose prototypes were seen during training, but drops sharply when the model encounters previously unseen prototypes [4, 15].

A related but distinct mechanism operates at the level of elemental composition. Elemental leakage occurs when the same chemical elements appear in both splits, even if stoichiometries differ substantially. The model acquires element-specific biases instead of learning transferable interaction rules that would generalize across the periodic table. The diagnostic signature here is particularly stark: performance on test crystals containing elements entirely absent from the training set approaches random guessing [6, 20]. This form of leakage proves especially pernicious because it remains invisible under standard random-split reporting, where test and training sets necessarily share elements by construction.

Beyond these global compositional effects lies a more local structural phenomenon. Local environment leakage arises when similar coordination geometries—octahedral, tetrahedral, or square-planar arrangements—recur across splits regardless of the central atom's identity. A model can exploit these recurring local graph motifs as shortcuts, bypassing the need to learn genuine long-range chemical or electronic interactions. The detection signature emerges when models fail dramatically on crystals with novel coordination geometries not represented in the training set [5, 8]. This mechanism connects directly to energy range leakage, a fourth source rooted in the thermodynamic biases of computed databases. The formation-energy distribution in the test set overlaps almost completely with the training distribution under random splitting, meaning extrapolation to metastable or unstable crystals is never required. Error increases sharply when test crystals lie outside the training energy range, yet random-split evaluations never provoke that condition [1, 11].

Finally, lattice parameter leakage provides a geometric analogue to energetic clustering. Similar lattice parameters appear across splits, so the model never encounters severely compressed or expanded structures. Performance collapses on test structures whose lattice parameters fall outside the training range, but again, random splitting ensures such out-of-range examples are systematically excluded from evaluation [2, 21]. Each of these five sources is amply documented in the literature, yet their implications are rarely confronted directly. The crystal graph convolutional neural network [4] and its subsequent extensions [15, 22, 23], for example, report excellent random-split performance while remaining vulnerable to prototype leakage, given the prevalence of common structure types throughout Materials Project data. Similarly, assessments of machine-learning interatomic potentials [6] acknowledge elemental biases in passing yet retain random splits that cannot expose those biases. Data-bias studies [11] further highlight how energetic clustering in computed databases reinforces energy-range leakage, creating a self-consistent evaluation environment that never demands genuine extrapolation.

These leakage sources are not independent; they are mutually reinforcing. A model that exploits prototype leakage simultaneously benefits from local environment leakage, and both are amplified by energy range and lattice parameter overlap. The resulting overestimation is therefore not a minor statistical artifact that larger datasets would cure, but rather a systemic feature of the random-split protocol itself. By cataloguing these five sources with explicit detection signatures, we provide practitioners with diagnostic tools to expose the illusion in their own workflows—tools that shift the question from whether random splitting overestimates generalization to how severely, and along which dimensions, that overestimation occurs.

A Typology of Generalization Illusions

Random partitioning induces a family of systematic distortions that present as robust generalization while, in practice, reflecting different forms of latent information leakage. What appears to be predictive competence is often anchored in the model's ability to interpolate within narrowly defined regions of structural, compositional, or energetic space, rather than to infer transferable physical regularities. This becomes evident when performance is interrogated along dimensions that random splits implicitly entangle, revealing that high reported accuracy can coexist with sharply limited extrapolative capacity.

One manifestation arises when models internalize recurring structural archetypes without capturing the governing principles that generate them. Under these conditions, predictive success is largely confined to crystals that conform to familiar prototype families, while structurally distinct systems expose a pronounced breakdown. This pattern is rooted in prototype leakage [4, 15], where training and test sets share underlying structural motifs despite nominal separation. In materials applications, this dynamic is particularly visible when models trained extensively on rocksalt and perovskite configurations achieve strong results on held-out instances from the same families yet fail abruptly on wurtzite or delafossite phases. The apparent robustness is therefore contingent on hidden redundancies within the partitioning scheme, and only becomes visible when evaluation is conditioned on previously unseen prototypes.

A closely related distortion emerges at the level of chemical composition, where models appear broadly capable due to exposure to extensive elemental combinations, yet remain constrained by the specific elemental neighborhoods encountered during training. Here, compositional generalization is illusory because the model effectively interpolates within a familiar subset of the periodic table rather than reasoning over underlying chemical principles. This limitation is driven by elemental leakage [6, 20], which ensures that both training and test data share overlapping elemental identities or combinations. As a consequence, strong performance on systems containing elements such as Ti, Zr, or Hf can mask a pronounced inability to handle entirely novel transition-metal chemistries. The scale and diversity of repositories like the Materials Project obscure this boundary, giving the impression of coverage that does not translate into genuine compositional transfer.

Beyond structure and composition, the distribution of thermodynamic stability introduces another axis along which generalization claims can become misleading. Models trained predominantly on low-energy configurations often exhibit impressive accuracy near the convex hull, yet this success deteriorates rapidly when extended to metastable or high-energy regimes that are central to synthesis and processing. This discrepancy reflects energy-range leakage [1, 11], whereby the training and test sets inherit nearly identical energy distributions due to random sampling. The resulting alignment conceals the model's inability to extrapolate beyond the dominant stability regime, producing error profiles that expand sharply when evaluated on polymorphs relevant to non-equilibrium conditions such as thin-film growth.

The final distortion becomes apparent when considering deviations from equilibrium geometry, where predictive reliability depends on sensitivity to local distortions and lattice perturbations. Models trained on relaxed structures frequently perform well under near-equilibrium conditions but fail to maintain accuracy when subjected to strain, pressure, or other structural perturbations. This limitation is linked to lattice-parameter and local-environment leakage [2, 21], which ensures that both training and evaluation data occupy similar regions of configurational space. Because widely used computational datasets are heavily biased toward equilibrium geometries, this constraint often remains hidden until models are tested under controlled distortions that expose their limited structural adaptability.

Taken together, these interrelated effects can be conceptualized as distinct yet convergent forms of distributional entanglement, each arising from the same underlying reliance on random partitioning. Visualizing the data space in terms of clustered structural prototypes, overlapping elemental neighborhoods, aligned energy distributions, and narrowly distributed lattice parameters clarifies how each apparent success is conditioned on shared support between training and test sets. The progression across these dimensions underscores that the issue is not isolated to a single failure mode, but reflects a broader misalignment between evaluation protocols and the demands of scientific discovery.

Although this typology does not exhaust all possible sources of spurious generalization, it captures the dominant patterns that shaped much of the literature between 2017 and 2022. By articulating these mechanisms with greater precision, it becomes possible to reinterpret prior performance claims through a more critical lens and to move toward evaluation strategies that genuinely probe extrapolative reasoning in AI-driven materials science.

Consequences for the Field

The appearance of reliable generalization under random partitioning propagates a set of tightly coupled consequences that extend beyond isolated evaluation errors, reshaping how progress is inferred, resources are allocated, and scientific credibility is constructed. What begins as a methodological convenience evolves into a structural distortion, wherein models that excel under entangled distributions are prematurely elevated to positions of practical authority. This dynamic becomes especially consequential in experimental settings, where predictive outputs are treated as actionable guidance. When such models are deployed, their apparent accuracy often fails to translate into real-world validity, leading researchers to pursue candidate materials that ultimately do not meet expectations. The resulting inefficiency is not merely technical but material, as experimental cycles are consumed by false positives that arise from mischaracterized predictive reliability [1, 12].

This inflation of perceived performance also reverberates through the epistemic economy of the field. Claims of “state-of-the-art” achievement, when grounded in metrics derived from random splits, acquire an authority that is disproportionate to their actual generalization capacity. Over time, these claims accumulate into a narrative of steady progress, even as the underlying capability to extrapolate remains largely unchanged. The implications extend into institutional domains, subtly shaping funding priorities and hiring decisions that rely on published indicators of advancement [4, 9]. In this sense, the illusion does not simply misrepresent individual results; it recalibrates collective expectations of what constitutes meaningful improvement.

A further consequence emerges in the delayed recognition of fundamental limitations. As long as models continue to perform well under conventional evaluation schemes, there is little immediate pressure to interrogate their inability to generalize to genuinely novel crystal systems. This postpones the identification of distribution shift as a central challenge and, in turn, defers investment in architectural innovations explicitly designed to address it [2, 5]. The field thus risks optimizing within a constrained paradigm, refining existing approaches without confronting the deeper question of how transferable scientific inference can be achieved.

This inertia is reinforced by the concentration of effort on architectures that are well-suited to prevailing benchmarks but not necessarily to the broader demands of materials discovery. Graph-based models, for instance, have been extensively tuned to perform under random-split conditions, where their capacity to exploit local structural and compositional regularities is rewarded. Yet such optimization may offer limited returns when confronted with out-of-distribution scenarios, effectively channeling intellectual and computational resources toward improvements that do not translate into robust generalization [6, 15]. The opportunity cost becomes apparent when alternative paradigms, potentially better aligned with extrapolative reasoning, receive comparatively less attention.

The implications also intersect with the ongoing challenges of reproducibility. Models that demonstrate strong performance within a given dataset may fail to reproduce when evaluated on independently curated crystal collections, particularly when those datasets differ in compositional coverage, structural diversity, or energy distributions. These discrepancies expose the extent to which reported success depends on implicit dataset-specific regularities, thereby amplifying reproducibility concerns across research groups that rely on distinct data sources [11, 17]. What appears as inconsistency across studies is often a symptom of shared methodological assumptions rather than isolated implementation differences.

This pattern feeds directly into the design of subsequent benchmarks, where new datasets and evaluation protocols frequently inherit the same reliance on random partitioning. In doing so, they reproduce the very conditions that gave rise

to the original distortion, embedding the illusion more deeply into successive generations of research [8, 9]. The persistence of this design choice reflects not a lack of technical sophistication, but a misalignment between evaluation convenience and the scientific objective of discovering materials beyond known regimes.

Empirical signals of these dynamics are already visible. Studies on high-entropy alloys report encouraging machine-learning performance under standard evaluation settings while simultaneously acknowledging the difficulty of extending predictions to unexplored compositional spaces [20]. Parallel analyses of data bias in experimental property databases highlight analogous vulnerabilities, yet the computational community has been comparatively slow to recalibrate its evaluation practices in response [11]. The result is a body of literature characterized by incremental improvements in predictive metrics, accompanied by a more limited capacity to generate transferable scientific insight.

Framed in this way, the issue is less an indictment of individual contributions than an indication of a shared methodological blind spot that has persisted alongside rapid advances in model sophistication. Confronting this misalignment is therefore not a peripheral concern but a necessary step toward ensuring that machine-learning systems fulfill their intended role in accelerating materials discovery.

Alternative Evaluation Strategies

We propose six alternative evaluation strategies that replace random partitioning with deliberate distribution shifts.

Table 2 reframes alternative split design as a set of diagnostic stress tests, clarifying which form of apparent generalization each strategy challenges and what a credible success claim would therefore mean.

Table 2. Distribution-aware evaluation strategies as diagnostic tests of generalization failure in crystal property prediction

Evaluation strategy	Unit of enforced distribution shift	Illusion or leakage it is designed to diagnose	What strong performance would actually mean	Main trade-off	Recommended reporting standard
Composition split	Hold out one or more elements or elemental families from training	Diagnoses **compositional illusion** and elemental leakage [5, 6]	The model captures transferable interaction patterns beyond memorized element-specific biases	Smaller effective training support for held-out chemistry	Report accuracy separately for seen-element and unseen-element regimes.
Prototype split	Hold out entire crystal prototypes from training	Diagnoses **prototype illusion** and prototype leakage [3, 10]	The model can transfer beyond familiar structure families and not merely recognize recurring topologies	Sharper performance drop likely; harder benchmark	Report performance stratified by seen vs. unseen prototypes, not just pooled MAE.
Time split	Train on older entries, test on newer entries	Diagnoses temporal overfitting and workflow/provenance dependence [14, 17]	The model can support future discovery rather than	Vulnerable to dataset growth and curation shifts	Report the exact temporal cutoff and database version used.

			retrospectively fit historically co-sampled data		
Structural dissimilarity split	Cluster crystals by graph/structural similarity and separate clusters across train/test	Diagnoses local-environment leakage and hidden structural entanglement [2, 9]	The model remains accurate when local motifs and structural neighborhoods are genuinely novel	Requires explicit similarity metric or embedding choice	Report clustering method, distance threshold, and between-split similarity statistics.
Energy extrapolation split	Train on near-hull or low-energy structures, test on metastable/high-energy structures	Diagnoses **energy-range illusion** and energy-range leakage [1, 13]	The model can extrapolate into synthesis-relevant but distributionally rare energetic regimes	Test regime becomes intrinsically harder and more imbalanced	Report separate errors inside and outside the training energy range.
Cross-database split	Train on one database, test on an independent database	Diagnoses database-specific artifacts, provenance dependence, and reproducibility fragility [1, 14]	The model learns database-robust physical regularities rather than source-specific shortcuts	Label conventions, calculation settings, and property definitions may differ	Report harmonization protocol and cross-source performance gap explicitly.
Combined multi-split audit	Apply random plus at least one hard split in the same paper	Diagnoses the gap between interpolation and extrapolation	The model's performance profile becomes scientifically interpretable rather than leaderboard-oriented	More reporting burden, but much higher validity	Require a stated **generalization gap** between random and distribution-aware splits.

A more faithful assessment of model capability requires evaluation protocols that explicitly decouple training and test distributions along scientifically meaningful axes. Rather than relying on random partitioning, which preserves latent correlations, these strategies introduce controlled distribution shifts that mirror the conditions encountered in discovery-oriented workflows. The central premise is not simply to reduce leakage, but to expose the mechanisms by which models succeed or fail when confronted with novelty, thereby aligning evaluation with the epistemic demands of materials science.

One natural point of intervention lies at the level of chemical composition, where partitioning data according to elemental identity reveals whether a model can extend beyond familiar regions of the periodic table. By assigning all crystals containing a given element to the test set while excluding them from training, the evaluation directly probes generalization to unseen chemistries [6, 20]. Although this approach reduces the effective size of both training and test subsets, the resulting signal is substantially more informative, as it distinguishes interpolation across known combinations from genuine compositional transfer.

A complementary perspective emerges when structural diversity is treated as the primary axis of separation. Partitioning by crystal prototype ensures that models are trained on a subset of structural families and evaluated on entirely distinct ones, thereby isolating their ability to infer relationships that transcend specific geometric templates [4, 15]. Under these conditions, predictive success can no longer rely on implicit recognition of recurring motifs, and instead reflects the extent to which underlying physical patterns have been internalized.

Temporal structure provides another dimension along which evaluation can be reoriented toward realistic deployment scenarios. By training on materials characterized at earlier stages and testing on those reported more recently, models are assessed in a setting that approximates forward-looking discovery [12, 17]. This temporal split introduces a natural form of distribution shift, as newly synthesized or computed materials often expand into regions of compositional and structural space that were previously underrepresented.

Beyond discrete partitions, distributional separation can be induced through continuous measures of structural similarity. Clustering approaches based on crystal-graph embeddings enable the construction of training and test sets that are maximally dissimilar in representation space, thereby enforcing a more stringent form of generalization [2, 18]. This strategy preserves dataset scale while systematically amplifying the distance between observed and unobserved configurations, offering a pragmatic balance between statistical robustness and conceptual rigor.

Thermodynamic considerations further enrich this framework by foregrounding the role of energy landscapes in materials behavior. Restricting training data to structures near the convex hull while reserving higher-energy configurations for evaluation compels models to extrapolate into metastable regimes that are critical for synthesis and processing [1, 11]. The resulting performance profile provides direct insight into whether learned representations capture trends that extend beyond equilibrium stability.

Finally, cross-database evaluation introduces an external validation layer that is particularly effective at revealing dataset-specific artifacts. Training on one curated source and testing on an independent repository exposes discrepancies arising from differing computational protocols, sampling biases, or curation standards [1, 12, 24]. In this setting, consistency across datasets becomes a proxy for robustness, highlighting whether predictive performance is anchored in transferable physical relationships or contingent on idiosyncratic data characteristics.

Each of these strategies is readily implementable with existing resources, yet their implications extend well beyond technical adjustments to data splitting. By privileging scientifically grounded distribution shifts over convenience, they reframe generalization as an empirical question rather than an assumed property. The resulting evaluations may sacrifice some degree of statistical efficiency, but they yield a more credible account of model behavior under realistic conditions. Even partial adoption would recalibrate the evidentiary threshold for claims of predictive power, encouraging a transition from measuring interpolation within known regimes to assessing extrapolation into the unknown spaces that ultimately define the impact of machine learning in materials discovery.

Relation to Other Critiques

This critique of random train-test splitting in small-crystal property prediction is not an isolated methodological concern but participates in a broader, cross-disciplinary recognition that inappropriate data partitioning systematically inflates generalization claims. Within computational materials science itself, the limitations we identify echo and extend earlier warnings about dataset biases and benchmark fragility. Wang *et al.* [1] already noted the Materials Project's focus on stable phases, yet subsequent modeling papers treated random splits of these data as unproblematic. Kumagai *et al.* [11] later demonstrated how data bias in experimental property databases leads to misleading machine-learning performance; our analysis shows that the very same bias mechanisms operate even more insidiously in computed crystal datasets under random partitioning. Schmidt *et al.* [12] surveyed recent advances in solid-state machine learning and implicitly flagged the need for more rigorous validation, yet the field has continued to rely on the same random-split protocol that their review indirectly critiques through its emphasis on real-world applicability.

The pattern we expose parallels long-standing concerns in adjacent fields, even though our grounding remains strictly within the materials literature. Xiong *et al.* [18] advocated k-fold forward cross-validation precisely to mitigate explorative prediction overestimation in materials discovery, demonstrating that random splits fail to simulate genuine discovery workflows—much as scaffold-based splitting was eventually required in molecular property prediction to prevent leakage across chemically similar scaffolds. Ong [17] highlighted the acceleration of materials science through high-throughput computation but cautioned that downstream machine-learning models inherit the sampling artifacts of the underlying databases; our typology of generalization illusions makes those inherited artifacts explicit and quantifiable. Chen *et al.* [20] reviewed machine-learning applications to high-entropy alloys and stressed the practical challenges of extrapolating to unseen compositional spaces, thereby anticipating the compositional illusion we formalize here.

These materials-focused critiques align with the wider recognition, visible across computational sciences, that the IID assumption rarely holds for structured scientific data. Domain-adaptation literature has established that models trained on one distribution frequently fail when deployment involves even modest shifts; crystal property prediction under random splitting simply masks those shifts rather than eliminating them. Similarly, out-of-distribution generalization research has shown that benchmark success on held-out samples drawn from the same distribution provides little evidence of robustness to novel inputs—precisely the situation created when random splits entangle prototypes, elements, and energy ranges across train and test sets [2, 4, 5, 11]. Ko *et al.* [8] and Dunn *et al.* [9] advanced general-purpose frameworks and reference algorithms, yet both anchored their claims to random splits, thereby inheriting the very overestimation that domain-adaptation studies warn against.

By situating our argument within this network of existing critiques [1, 11, 12, 17, 18, 20], we demonstrate that the illusion of generalization is not a niche flaw in crystal-graph neural networks but a systemic methodological vulnerability that has persisted across multiple communities. The contribution of the present work is to render the problem visible, name its mechanisms, and supply the diagnostic typology and alternative strategies needed to move beyond it. Far from undermining prior achievements, this critique honors the foundational papers [2, 4-6] by insisting that their technical innovations receive the rigorous evaluation they deserve.

Implications for Benchmark Design

The illusion of generalization carries direct and actionable implications for how benchmarks in crystal property prediction must be redesigned. Benchmark creators should immediately abandon random splits as the primary or sole evaluation metric. Instead, every new benchmark must incorporate at least one challenging split—composition, prototype, or time-based—that enforces a genuine distribution shift [11, 18]. Performance must be reported separately for interpolation regimes (random splits) and extrapolation regimes (alternative splits), allowing the community to quantify the generalization gap rather than conflating the two. Minimum standards should require that any claimed “state-of-the-art” model demonstrate non-trivial accuracy on at least one unseen-prototype or unseen-element test; otherwise the claim is misleading [4, 15].

Journal reviewers and editors occupy a pivotal gatekeeping role. Reviewers should routinely question manuscripts that report only random-split accuracy, demanding supplementary results on at least one alternative strategy from Section 7. Skepticism toward “state-of-the-art” declarations is warranted whenever those declarations rest exclusively on random partitioning; such claims should be reframed as “state-of-the-art on random splits” to avoid overstating scientific progress [9, 12]. Editors can enforce this standard by requiring a short “generalization audit” paragraph in every machine-learning-for-crystals paper, in which authors explicitly state which leakage sources were tested and which remain unaddressed.

For the broader community, the implications point toward standardized, challenging splits that become community norms rather than optional extras. We envision a future Matbench-like suite [9] that ships with six canonical splits (the six strategies proposed in Section 7) and automatically computes the generalization gap between random and distribution-aware partitions. Researchers would then compete not merely on absolute error but on robustness across shifts, aligning

publication incentives with real materials-discovery needs. Cross-database splits [1, 12] should become routine, forcing models to demonstrate transferability beyond the Materials Project's particular sampling biases. Energy-extrapolation splits [11] would ensure that benchmarks reflect the metastable and unstable phases that dominate experimental synthesis challenges [25].

These changes are not merely technical; they are cultural. By embedding deliberate distribution shifts into benchmark design, the field will accelerate the development of architectures that learn transferable physical principles rather than dataset-specific artifacts. The result will be slower but more trustworthy progress: fewer papers claiming near-zero errors, but far greater confidence that published models can actually guide experimental discovery [26]. Implementing these implications requires no new data—only a disciplined re-analysis of existing datasets [1, 2, 4-6, 8, 9, 11, 12, 15, 18]—and therefore constitutes an immediately feasible reform.

Conclusion

Random train-test splitting of small-crystal datasets from sources such as the Materials Project creates an illusion of generalization that has distorted an entire subfield of machine learning for materials science. Models appear to achieve near-zero error on formation energy, band-gap, and elastic-constant prediction, yet this apparent success arises not from genuine out-of-distribution capability but from five specific leakage sources—prototype leakage, elemental leakage, local-environment leakage, energy-range leakage, and lattice-parameter leakage—that random partitioning fails to eliminate . These leakages in turn generate a typology of four generalization illusions: the prototype illusion, the compositional illusion, the energy-range illusion, and the structure illusion. Each illusion misleads practitioners into believing that current graph-network architectures have mastered transferable physics when, in reality, they have largely memorized recurring patterns within the training distribution.

The consequences are far-reaching: wasted experimental validation on false-positive predictions, misleading state-of-the-art claims that distort resource allocation, delayed recognition of fundamental model limitations, overinvestment in non-generalizable architectures, amplification of reproducibility failures, and misguided benchmark designs that perpetuate the cycle. By formalizing these mechanisms, we have shown that the standard practice rests on four flawed assumptions—IID data, representativeness, absence of leakage, and equation of small error with generalization—that simply do not hold for crystalline matter.

We therefore call for the immediate abandonment of random splits as the primary evaluation method in crystal property prediction. The six alternative strategies—composition split, prototype split, time split, structural dissimilarity split, energy extrapolation split, and cross-database split—offer concrete, zero-cost replacements that restore scientific integrity. Adoption of these strategies will not diminish the technical elegance of SchNet, crystal graph convolutional networks, or graph networks for crystals; rather, it will finally allow the community to measure what these architectures truly achieve.

The field stands at a methodological inflection point. By replacing the illusion of generalization with rigorous, distribution-aware evaluation, computational materials science can move from incremental benchmark chasing to genuine scientific discovery. The references examined here already contain the data and the conceptual tools required; what remains is the collective will to apply them correctly. Only then will machine learning deliver on its promise to accelerate materials innovation rather than merely simulate progress.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 14 Oct 2021 Revised: 23 Jan 2022 Accepted: 07 May 2022
Published online: 18 July 2022

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.

Schütt K, Kindermans PJ, Sauceda Felix HE, Chmiela S, Tkatchenko A, Müller KR. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. In: Advances in neural information processing systems 30; 2017. p. 992-1002.

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.

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.

Deringer VL, Caro MA, Csányi G. Machine learning interatomic potentials as emerging tools for materials science. Adv Mater. 2019;31(46):1902765.
https://doi.org/10.1002/adma.201902765.

Ko TW, Finkler JA, Goedecker S, Behler J. General-purpose machine learning potentials capturing nonlocal charge transfer. Acc Chem Res. 2021;54(4):808-17.
https://doi.org/10.1021/acs.accounts.0c00689.

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.
https://doi.org/10.1038/s41524-020-00406-3.

Alpaydin E. Machine learning. Revised and updated ed. Cambridge (MA): MIT Press; 2021.
https://doi.org/10.7551/mitpress/13811.001.0001.

Kumagai M, Ando Y, Tanaka A, Tsuda K, Katsura Y, Kurosaki K. Effects of data bias on machine-learning-based material discovery using experimental property data. Sci Technol Adv Mater Methods. 2022;2(1):302-9.
https://doi.org/10.1080/27660400.2022.2109447.

Schmidt J, Marques MRG, Botti S, Marques MAL. Recent advances and applications of machine learning in solid-state materials science. NPJ Comput Mater. 2019;5(1):83.
https://doi.org/10.1038/s41524-019-0221-0.

Morgan D, Jacobs R. Opportunities and challenges for machine learning in materials science. Annu Rev Mater Res. 2020;50:71-103.
https://doi.org/10.1146/annurev-matsci-070218-010015.

Chami I, Abu-El-Haija S, Perozzi B, Ré C, Murphy K. Machine learning on graphs: A model and comprehensive taxonomy. J Mach Learn Res. 2022;23(89):1-64.
https://doi.org/10.5555/3586589.3586678.

Park CW, Wolverton C. Developing an improved crystal graph convolutional neural network framework for accelerated materials discovery. Phys Rev Mater. 2020;4(6):063801.
https://doi.org/10.1103/PhysRevMaterials.4.063801.

Weston L, Stampfl C. Machine learning the band gap properties of kesterite I2-II-IV-V4 quaternary compounds for photovoltaics applications. Phys Rev Mater. 2018;2(8):085407.
https://doi.org/10.1103/PhysRevMaterials.2.085407.

Ong SP. Accelerating materials science with high-throughput computations and machine learning. Comput Mater Sci. 2019;161:143-50.
https://doi.org/10.1016/j.commatsci.2019.01.013.

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.
https://doi.org/10.1016/j.commatsci.2019.109203.

Balachandran PV, Emery AA, Gubernatis JE, Lookman T, Wolverton C, Zunger A. Predictions of new ABO3 perovskite compounds by combining machine learning and density functional theory. Phys Rev Mater. 2018;2(4):043802.
https://doi.org/10.1103/PhysRevMaterials.2.043802.

Chen S, Cheng Y, Gao H. Machine learning for high-entropy alloys. In: Cheng Y, Wang T, Zhang G, editors. Artificial intelligence for materials science. Cham: Springer International Publishing; 2021. p. 21-58.
https://doi.org/10.1007/978-3-030-68310-8_2.

Choudhary K, DeCost B, Tavazza F. Machine learning with force-field-inspired descriptors for materials: Fast screening and mapping energy landscape. Phys Rev Mater. 2018;2(8):083801.
https://doi.org/10.1103/PhysRevMaterials.2.083801.

Xie T, Grossman JC. Hierarchical visualization of materials space with graph convolutional neural networks. J Chem Phys. 2018;149(17):174111.
https://doi.org/10.1063/1.5047803.

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.

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.

Huo H, Bartel CJ, He T, Trewartha A, Dunn A, Ouyang B, et al. Machine-learning rationalization and prediction of solid-state synthesis conditions. Chem Mater. 2022;34(16):7323-36.
https://doi.org/10.1021/acs.chemmater.2c01293.

Saal JE, Oliynyk AO, Meredig B. Machine learning in materials discovery: Confirmed predictions and their underlying approaches. Annu Rev Mater Res. 2020;50:49-69.
https://doi.org/10.1146/annurev-matsci-090319-010954.