ORIGINAL RESEARCH **Open access**

A Conceptual Distinction between Generalization and Transfer in Materials Machine Learning

Hiroshi Nakamura[1*], Yuta Kato[1]

Abstract

In the rapidly expanding domain of artificial intelligence applied to materials science, a persistent conceptual ambiguity undermines the reliability of reported model capabilities. The terms “generalization” and “transfer” are routinely conflated, with authors claiming that a model “generalizes” when it is in fact being evaluated on samples drawn from a distinctly different distribution. This boundary/definitional paper draws a sharp conceptual distinction between the two notions. Generalization is defined as the expected performance of a trained model on new samples drawn independently and identically from the same underlying distribution as the training data. In contrast, transfer is defined as performance on samples drawn from a different distribution, where the I.I.D. assumption is violated by construction. The distinction matters because a model that generalizes excellently within its training distribution can fail dramatically under transfer conditions, and conversely, a successful transfer mechanism may mask poor generalization; treating the two interchangeably, therefore, produces overclaims about model robustness that cannot be sustained when materials discovery moves beyond the convex hull of available training data. The paper articulates a two-dimensional boundary framework—distribution-shift magnitude and feature-space overlap—that locates any given evaluation setting along a continuum from pure generalization to pure transfer, thereby enabling authors, reviewers, and practitioners to specify precisely which capability is being claimed and tested. By clarifying these boundaries and exposing the epistemic costs of current usage, the work supplies a conceptual foundation for more disciplined reporting standards and evaluation protocols in materials machine learning.

Keywords Transfer learning, Distribution shift, Generalization, Domain adaptation, Materials machine learning, Out-of-distribution performance

*Correspondence:
Hiroshi Nakamura
hiroshi.nakamura@outlook.com
[1] Department of Intelligent Materials Engineering, Nagoya University, Nagoya, Japan

Introduction

Materials-science applications of machine learning routinely announce that a new model “generalizes” to unseen compounds, crystal structures, or processing conditions. At the same time, parallel lines of work describe the same or similar models as achieving “transfer” across chemical families, property spaces, or experimental modalities. Yet these two descriptors are not interchangeable. The former invokes a statistical guarantee grounded in the i.i.d. assumption [1], whereas the latter presupposes a deliberate violation of that assumption through domain shift.

The present paper, therefore, undertakes a boundary/definitional analysis whose sole purpose is to disentangle the concepts of generalization and transfer as they appear in the materials-AI literature.

The problem is not merely terminological. When a graph-network model trained on a large computational database is said to “generalize” to experimental measurements [2-6], or when a deep-transfer pipeline is praised for its “generalization” across sparse experimental sets [7, 8], the claim implicitly equates two distinct epistemic

commitments. One commitment concerns the model's ability to interpolate reliably within the statistical envelope of its training distribution; the other concerns its ability to extrapolate or adapt when that envelope is left behind. Conflating them leads authors to overstate the readiness of models for real-world deployment, reviewers to accept performance numbers without scrutinizing the nature of the test distribution, and the broader community to underestimate the data-acquisition costs required for genuine progress in inverse design [7].

This study, therefore, proceeds in clearly demarcated stages. It first surveys the dominant usages of "generalization" in the materials-AI corpus, demonstrating that the term is most often employed under an implicit i.i.d. assumption even when the underlying data violate it. It then surveys usages of "transfer," revealing that the term is reserved for settings in which distribution shift is acknowledged yet optimistically bridged. The subsequent analysis isolates three concrete problems that arise from the conflation—category error, overclaiming, and evaluation mismatch—before advancing four numbered definitions that restore precision. These definitions are then contrasted with five neighboring concepts (interpolation, extrapolation, domain adaptation, robustness, and out-of-distribution detection) through a systematic textual comparison. The result is a compact yet rigorous conceptual scaffold that future work can use to locate any materials-model evaluation along an explicit generalization–transfer spectrum. By insisting on this precision, the paper does not diminish the impressive empirical successes already reported; rather, it supplies the linguistic and epistemic tools necessary to interpret those successes accurately and to design the next generation of experiments with clearer expectations about what has actually been demonstrated.

Generalization in Existing Literature

Within the materials-AI literature, the term "generalization" is most frequently invoked to describe a model's performance on held-out test samples that are assumed to be drawn from the same underlying distribution as the training data. Butler and co-workers, for example, frame generalization as the capacity of supervised models to predict properties of molecular and solid-state systems not encountered during training, provided those systems lie within the statistical manifold spanned by the training set [4]. Similarly, Schmidt *et al.* [5] emphasize that reliable generalization in solid-state materials science requires the test distribution to respect the i.i.d. condition implicit in standard cross-validation protocols.

This statistical usage recurs across multiple sub-domains. Chen *et al.* report that graph-network architectures achieve strong generalization on crystal-property prediction tasks precisely because the test crystals are sampled from the same computational database distribution used for training [6]. Kauwe *et al.* probe the limits of such generalization by asking whether models can identify "extraordinary" materials while remaining within the convex hull of known training examples, thereby treating generalization as an interpolative rather than extrapolative capacity [9-13]. Meredig and colleagues explicitly test whether machine-learning models retain generalization when asked to predict high-temperature superconductors, again anchoring their evaluation in the assumption that the candidate materials share the same feature-space statistics as the training library [14, 15].

A parallel thread appears in discussions of robustness under small perturbations. Chang *et al.* argue that generalization performance must be assessed not only on clean i.i.d. splits but also under controlled noise levels that simulate minor experimental variability, still within the original distribution [16]. Peterson and Brgoch likewise treat generalization as the ability of formation-energy models to maintain accuracy on unseen compositions drawn from the same compositional space [17-25]. Schmidt's later work on crystal-graph attention networks reinforces this view by benchmarking generalization error strictly on randomly held-out subsets of the same Materials Project-derived distribution [26, 27]. Fung *et al.*, in their systematic benchmarking of graph neural networks, repeatedly return to i.i.d. test errors as the primary measure of generalization quality [28].

Even when authors acknowledge that material data are heterogeneous, they often preserve the generalization label by subsampling or reweighting so that the effective test distribution matches the training distribution [26]. The dominant pattern, therefore, is that "generalization" is invoked whenever the evaluation protocol preserves the i.i.d. assumption, whether the data originate from density-functional theory, experimental measurements, or hybrid sources. This usage is internally consistent but becomes problematic precisely when the same papers later present results on chemically or experimentally distinct sets without

changing the label. The literature thus reveals a stable but narrow conception of generalization: performance under distributional invariance [1].

Transfer in Existing Literature

In contrast to the i.i.d.-centric usage of generalization, the materials-AI literature employs "transfer" when the source and target distributions are acknowledged to differ in statistically meaningful ways. Pan and Yang's foundational survey, although predating the materials focus, supplies the conceptual backbone that later works adopt: transfer learning is required precisely when the target domain violates the training distribution [2]. Jha *et al.* [8] exemplify this usage by demonstrating deep transfer learning from large computational datasets to sparse experimental measurements of materials properties, where the source and target distributions differ in both scale and noise characteristics.

The same framing appears in cross-property settings. Gupta and colleagues develop a deep-transfer framework that moves knowledge across mechanical, electronic, and thermal properties, explicitly noting that each property constitutes a distinct target domain with its own distribution [9]. Chen *et al.* [10] introduce AtomSets as a hierarchical transfer mechanism that enables knowledge to flow from large generic datasets to small specialized materials subsets, again highlighting the distributional gap between source and target. Yamada *et al.* [17] coined the term "shotgun transfer learning" to describe rapid adaptation across disparate material classes where the feature distributions share only partial overlap.

Further examples abound. De Breuck *et al.* [18] rely on feature selection and joint learning to achieve transfer across limited experimental datasets, underscoring that the target domain is deliberately undersampled relative to the source. Wang *et al.* [19] deploy compositionally restricted attention networks that facilitate transfer between different chemical families, treating the change in elemental composition as an explicit domain shift. Choudhary *et al.* [26], in their review of deep-learning applications, catalog multiple instances in which transfer learning is invoked to bridge computational-to-experimental or inter-property gaps. Finally, Kolluru *et al.* [29] demonstrate attention-based transfer across atomic systems whose local environments differ markedly, thereby confirming that transfer presupposes a change in underlying distribution.

These works share a defining characteristic: the label "transfer" is applied exactly when authors acknowledge that the i.i.d. assumption no longer holds and that some form of domain adaptation or fine-tuning is required to recover performance [2]. Unlike generalization papers, transfer papers do not claim that test samples are statistically interchangeable with training samples; they instead celebrate the model's ability to bridge the resulting distributional gap. This usage is consistent across the corpus yet remains conceptually entangled with generalization when the same manuscript switches terminology mid-argument or when reviewers treat transfer success as evidence of generalization.

The Problem with Current Usage

The persistent conflation of generalization and transfer in materials-AI research introduces a conceptual instability that propagates through both interpretation and evaluation. What appears, at first glance, to be a terminological imprecision in fact reflects a deeper misalignment between the statistical conditions under which models are trained and the epistemic claims subsequently attached to their performance. Within the bounded geometry of a training distribution, a model may exhibit highly reliable predictive behavior, yet this apparent robustness is contingent upon the preservation of distributional structure [13]. Once this structure is perturbed—particularly when target samples extend beyond the convex support of the training data—the same model can fail in ways that are not merely quantitative but qualitatively discontinuous [15]. At the same time, pipelines explicitly designed for cross-domain adaptation often succeed precisely because they relax the requirement of distributional identity. However, this flexibility can introduce instability when predictions are required within any single, fixed domain [8]. Treating these distinct regimes as interchangeable collapses an essential distinction and obscures the conditions under which model outputs can be interpreted as reliable scientific knowledge.

This conceptual slippage carries immediate consequences for how empirical results are communicated. When performance achieved under a distributional shift is framed as evidence of "good generalization" [9, 10], the implication extends beyond what the data can sustain. The reader is tacitly encouraged to assume that the reported behavior will persist under future i.i.d. sampling, despite the absence of evidence supporting such invariance. In practice, this

rhetorical transition is subtle yet pervasive, appearing most clearly in abstracts and concluding sections where transfer success is recast as a broader guarantee of predictive reliability. The resulting inflation of claims is particularly consequential in inverse design contexts, where expectations of model competence directly shape experimental prioritization and resource allocation [7]. Under these conditions, the distinction between interpolation within known regimes and adaptation across unknown ones becomes blurred, and the interpretive boundary between demonstrated capability and anticipated performance is eroded.

Beyond issues of interpretation, the conflation also destabilizes the evaluative framework through which progress is assessed. Measures designed to quantify generalization—typically grounded in i.i.d. assumptions and operationalized through cross-validation error—are structurally incompatible with those required to assess transfer performance, where distributional mismatch is intrinsic to the task [5, 17]. When identical metrics are reported without clarifying the underlying statistical context, numerical results become ambiguous indicators, unable to distinguish between robustness to sampling variability and successful compensation for distributional shift. This ambiguity accumulates across the literature, producing a form of epistemic opacity in which comparisons between studies lose their meaning and reproducibility becomes contingent on unstated assumptions. The resulting condition is not merely one of inconsistency but of reduced cumulative reliability, as otherwise rigorous contributions become difficult to situate within a coherent evaluative landscape.

Proposed Definitions

Clarifying this ambiguity requires a conceptual realignment that restores precision without imposing additional methodological burdens. The distinction between generalization and transfer can be anchored directly in the relationship between the probability distributions governing training and evaluation. Generalization is most coherently understood as the expected performance of a trained model when applied to new samples drawn independently from the same distribution that generated the training data, such that the i.i.d. assumption remains intact and no distributional shift is present. Under these conditions, predictive success reflects the model's capacity to capture stable structure within a fixed statistical environment.

A different regime emerges when this assumption is deliberately violated. Transfer describes model performance when evaluation occurs on data drawn from a distribution that is detectably distinct from that of the training set, requiring the model—explicitly or implicitly—to accommodate distributional change [8, 17]. The distinction is therefore not one of degree but of underlying statistical condition, marking a transition from invariance to adaptation as the central requirement.

Within this framework, further refinement becomes possible by attending to the geometric relationship between training and evaluation samples. When test points lie strictly within the convex hull of the training inputs, the task reduces to interpolation, and performance can be interpreted as a constrained form of generalization that does not require extrapolative inference [13]. In contrast, when evaluation samples fall outside this region, the model is confronted with a genuine distributional gap, and successful prediction necessarily entails bridging between source and target domains in a manner characteristic of transfer [15].

These definitions are intentionally operational, grounding conceptual distinctions in observable properties of data distributions rather than in abstract typologies. By tying classification directly to the presence or absence of distributional identity, they align naturally with existing experimental protocols and provide a stable basis for interpreting empirical claims. The result is not the introduction of new categories, but the restoration of a boundary that allows generalization, transfer, and their intermediate forms to be distinguished with clarity, thereby strengthening both the interpretability and comparability of results across the field.

Table 1 provides a claim-classification matrix that specifies the evidentiary conditions under which a materials-AI result may be described as generalization, robustness-qualified generalization, or transfer.

Table 1. Claim-classification matrix for distinguishing generalization, near-generalization, near-transfer, and transfer in materials-AI evaluation

Claim category	Distribution relation between train and test	Feature-space relation	I.I.D. status

Pure generalization	Statistically identical distributions	Complete overlap; test samples remain within the occupied training region	Holds exactly
Near-generalization	Small but limited shift; perturbative rather than domain-changing	High overlap; samples remain near training support	Approximately holds
Near-transfer	Meaningful shift between related source and target domains	Partial overlap; some shared descriptors, but new regions entered	Violated
Pure transfer	Substantively different source and target distributions	Negligible overlap; target occupies a new region of feature space	Violated by construction
Interpolative generalization	Same distribution	Strictly inside the convex hull	Holds
Extrapolative transfer	Different distribution or effective support beyond the training envelope	Outside convex hull	Violated in practice

Distinctions from Nearby Terms

Generalization and transfer must also be distinguished from five neighboring concepts that frequently appear in the same literature: interpolation, extrapolation, domain adaptation, robustness, and out-of-distribution detection. The relationships can be clarified by considering six core terms—generalization, transfer, interpolation, extrapolation, domain adaptation, and robustness—along five distinguishing dimensions: (i) assumption on data distribution, (ii) nature of test samples relative to training distribution, (iii) typical evaluation metric, (iv) relationship to the training convex hull, and (v) primary methodological approach.

Table 2 sharpens the framework by separating performance claims from geometric conditions, methodological tools, and diagnostic functions, thereby preventing adjacent concepts from being collapsed into the generalization–transfer distinction.

Table 2. Analytical separation of adjacent concepts: ontological status, evidentiary target, and methodological function

Concept	Ontological status in the manuscript	What is being claimed?	Distributi conditi
Generalization	Performance claim	Reliability on new samples from the same distribution	I.I.D. requ
Transfer	Performance claim	Reliability on samples from a different distribution	Non-I.I. require
Interpolation	Geometric subcase of generalization	Accuracy in the interior regions of known space	I.I.D. compati

Extrapolation	Geometric conditions are usually aligned with transfer	Accuracy beyond observed support	Typica implies effective
Domain adaptation	Methodological toolkit, not a claim	Reduction of the mismatch between the source and target domains	Assumes exists
Robustness	Stress-test qualifier and not identical to transfer	Stability under limited perturbation	Small lo shift
Out-of-distribution detection	Diagnostic function	Identification of when generalization assumptions fail	Shift dete rather th success c

The distinction between generalization and transfer becomes analytically tractable once it is anchored in the statistical relationship between training and evaluation distributions. Generalization presupposes an i.i.d. regime in which the data-generating process remains unchanged between training and testing [1]. In contrast, transfer is defined precisely by the breakdown of this assumption through a detectable distributional shift [2]. Within the former regime, interpolation represents a further restriction in which test samples remain confined to the convex support of the training data, thereby requiring no extension beyond observed feature space. By contrast, extrapolation emerges when evaluation samples lie outside this support, placing the task squarely within the domain of transfer, as the model must operate under conditions for which no direct empirical precedent exists. Domain adaptation occupies a different conceptual layer, functioning not as a performance category but as a set of algorithmic strategies designed to enable transfer under such shifts. Robustness, meanwhile, probes a narrower regime in which deviations from the training distribution are deliberately constrained, often remaining sufficiently small to preserve approximate statistical continuity and thus retaining a conceptual proximity to generalization.

This statistical distinction is mirrored in the test samples themselves, which serve as the operational substrate through which these regimes are instantiated. When samples are statistically interchangeable with those observed during training, as is the case for generalization and its interpolative subset, predictive success reflects the model's capacity to internalize stable regularities within a fixed distribution. A different epistemic condition arises when samples are not interchangeable, as in transfer and extrapolation, where predictive performance depends on the model's ability to accommodate structural differences between source and target domains. Domain adaptation explicitly intervenes at this juncture by constructing mappings that reconcile these differences, effectively reconfiguring the feature space to restore predictive coherence. Robustness occupies a boundary position in which test samples are no longer identical to training data but remain sufficiently proximate in distributional space to allow controlled evaluation of model sensitivity. In this context, out-of-distribution detection—often embedded implicitly within transfer-oriented pipelines—functions as a diagnostic mechanism, signaling when the statistical conditions underlying generalization no longer hold and when transfer-oriented reasoning must be invoked [16].

The divergence between these regimes is further reflected in the metrics through which performance is quantified. Under i.i.d. conditions, generalization is naturally assessed באמצעות test-set error and cross-validation procedures that assume distributional stability [5, 6]. When this stability is disrupted, evaluation must shift accordingly, with transfer performance measured in terms of target-domain accuracy following adaptation procedures that explicitly address distributional mismatch [8, 17]. Interpolation manifests as low predictive error within the convex hull of the training data. At the same time, success under extrapolation—or, more precisely, under transfer—requires the maintenance of predictive fidelity beyond this boundary. Domain adaptation introduces an additional evaluative layer, where performance is often gauged by the extent to which domain discrepancy is reduced, using measures such as maximum mean discrepancy or adversarial alignment objectives. Robustness, in turn, is operationalized through controlled perturbations, with performance degradation serving as an indicator of sensitivity to deviations that remain near the original distribution.

A geometric perspective provides a unifying lens through which these distinctions can be visualized and operationalized. The convex hull of the training data

delineates a boundary that separates regimes of statistical continuity from those of genuine distributional departure. Generalization and interpolation remain confined within or on this boundary, where predictive inference relies on variation already encoded in the training set. Transfer and extrapolation, by contrast, unfold beyond it, where the absence of direct empirical coverage necessitates either inductive extension or explicit adaptation. Domain adaptation may be interpreted as an attempt to map between such regions, effectively aligning distinct hulls to enable knowledge transfer. Robustness testing probes the immediate vicinity of the boundary, examining how models respond to small excursions that challenge, but do not fully violate, the assumptions of distributional identity.

These differences extend into the methodological choices that underpin model development. Generalization is typically pursued through conventional supervised learning frameworks augmented by regularization strategies that stabilize performance under i.i.d. conditions [4]. Transfer-oriented approaches, in contrast, rely on mechanisms such as pre-training followed by fine-tuning, feature alignment across domains, or meta-learning strategies that explicitly encode adaptability [9, 10]. Interpolation requires no additional machinery beyond the base learner, as it operates entirely within observed data regimes. In contrast, extrapolation demands the incorporation of domain-specific inductive biases or augmented data representations capable of supporting inference beyond the training manifold. Domain adaptation introduces its own class of algorithms, often leveraging adversarial or discrepancy-minimization techniques to bridge distributional gaps. Robustness-focused methods emphasize resilience, typically through regularization schemes or ensemble constructions that mitigate sensitivity to perturbations.

What emerges from this synthesis is not merely a refined taxonomy but a clarification of the conditions under which different forms of predictive success can be meaningfully interpreted. The ability of a model to interpolate does not imply competence in extrapolation, just as the successful application of domain adaptation does not constitute an intrinsic property of the model but rather reflects the efficacy of an external methodological intervention. Similarly, claims of robustness acquire significance only when the magnitude and nature of the permitted shift are explicitly specified. By aligning each concept with a distinct configuration of distributional assumptions, sample characteristics, evaluative criteria, geometric constraints, and methodological strategies, the framework removes the ambiguity that has allowed generalization and transfer to be conflated, thereby restoring interpretive precision to empirical claims in materials-AI research.

When Generalization Ends, and Transfer Begins

The transition from generalization to transfer is not abrupt. Still, it can be located precisely along two orthogonal conceptual dimensions that together define a practical boundary framework for any materials-AI evaluation setting. The first dimension is the magnitude of distribution shift between the training and test samples, ranging continuously from none (identical statistical properties) through small (minor perturbations in noise, sampling density, or measurement modality) to large (fundamental changes in chemical families, experimental conditions, or property spaces). The second dimension is the degree of feature-space overlap, ranging from complete (test samples lie entirely inside the convex hull of training inputs) through partial (some overlap exists but new regions are populated) to none (test samples occupy entirely disjoint regions of feature space). These two axes generate a four-zone map that makes the boundary between generalization and transfer explicit rather than intuitive.

Figure 1 locates any materials-AI evaluation within a left-to-right decision framework defined by distribution-shift magnitude and feature-space overlap, thereby showing where generalization ends and transfer begins.

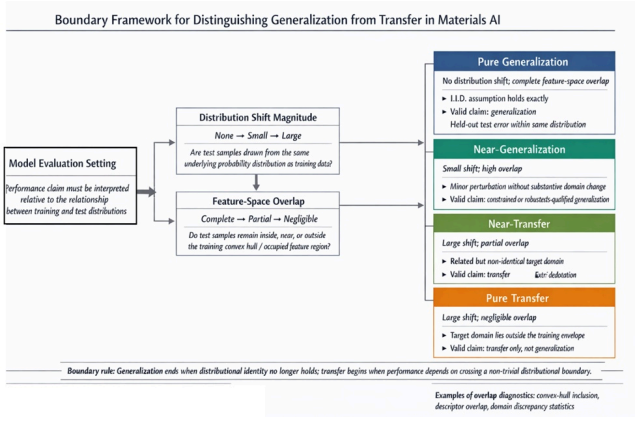

Figure 1. Any materials-AI evaluation within a left-to-right decision framework defined by distribution-shift magnitude and feature-space overlap

Zone 1—pure generalization—occupies the region of zero distribution shift and complete feature-space overlap. Here, the i.i.d. assumption holds without qualification, and performance claims rest solely on the model's ability to interpolate reliably within the statistical envelope of the training data, as illustrated by graph-network evaluations on held-out subsets of the same computational database [6]. Zone 2—near-generalization—permits small distribution shifts provided feature-space overlap remains high; this zone captures robustness to modest experimental variability while still qualifying as generalization because the underlying distribution identity is largely preserved [16]. Most contemporary materials-AI benchmarks that claim "generalization" actually operate in zone 2, yet authors rarely acknowledge the small shift, thereby blurring the boundary.

Zone 3—near-transfer—arises when distribution shift becomes large but feature-space overlap is still partial; adaptation mechanisms are required, yet some shared structure (for example, common atomic descriptors) can be leveraged. This zone is typical of cross-property or computational-to-experimental pipelines in which the source and target domains are related yet statistically distinct [9, 10]. Zone 4—pure transfer—occupies the far corner of a large shift and negligible overlap; here, the model must bridge entirely new regions of chemical or physical space with no statistical guarantee from the training distribution, as occurs in zero-shot adaptation across disparate materials classes [8, 29].

A conceptual diagram of this framework can be visualized as a two-dimensional grid: the horizontal axis labeled "Distribution Shift Magnitude" with arrows pointing rightward from "none" to "large," and the vertical axis labeled "Feature Space Overlap" with arrows pointing upward from "none" to "complete." Four labeled rectangles occupy the quadrants: zone 1 in the upper-left (no shift, complete overlap), zone 2 immediately to its right (small shift, high overlap), zone 3 below zone 2 (large shift, partial overlap), and zone 4 in the lower-right (large shift, negligible overlap). Arrows between zones indicate that real-world materials evaluations rarely sit at the extremes but migrate along the continuum as new data modalities or chemical spaces are introduced.

Crucially, the framework reveals that the majority of materials-AI claims labeled as "generalization" actually fall in zones 2 or 3. For instance, when a formation-energy model is tested on unseen compositions drawn from the same broad chemical space yet with different sampling densities [25], the evaluation is near-generalization rather than pure generalization; similarly, when transfer-learning pipelines move from large DFT libraries to sparse experimental sets while retaining partial elemental-feature overlap [17], the task is near-transfer yet is frequently reported under the generalization label [26]. By locating every evaluation protocol on this map, authors can state unambiguously whether their claim concerns performance under distributional invariance (zones 1–2) or under distributional adaptation (zones 3–4). The boundary is therefore not a philosophical nicety but a practical diagnostic that prevents the epistemic slippage documented in earlier sections. Once the framework is adopted, the question "Does this model generalize or transfer?" acquires a clear, two-dimensional answer rather than an ambiguous binary one.

Objections and Replies

The introduction of a sharper distinction between generalization and transfer often encounters resistance. Yet, these concerns tend to reveal more about prevailing interpretive habits than about any fundamental limitation of the framework itself. A frequent concern is that the relationship between the two should be understood as a continuum rather than a categorical separation, implying that any attempt to demarcate them risks imposing an artificial boundary. This intuition is not misplaced, but it mischaracterizes the intent of the proposed formulation. The framework does not deny continuity; rather, it renders that continuity intelligible by embedding tasks within a two-dimensional space defined by the magnitude of distributional shift and the degree of overlap between training and evaluation regimes. In doing so, it transforms what would otherwise remain an unstructured gradient into a measurable landscape in which claims can be situated with precision. Without such coordinates, the notion of a spectrum remains descriptively appealing but analytically inert, offering no basis for verifying whether two reported results occupy comparable positions within that space [1, 2].

A related concern arises from the empirical reality that materials datasets rarely satisfy strict i.i.d. conditions. Experimental variability, batch effects, and incomplete sampling introduce deviations that appear to undermine the very possibility of pure generalization. Under this view, the distinction collapses because all practical scenarios involve some degree of shift. Yet this observation, while empirically

valid, does not invalidate the framework; it instead clarifies how its categories should be applied. The presence of small or controlled deviations does not eliminate the conceptual boundary but calls for greater precision in labeling. Evaluations conducted under near-identical conditions should be interpreted as approximations to generalization rather than as its full realization. At the same time, stronger claims should be reserved for cases in which the i.i.d. assumption is demonstrably satisfied [5]. If, in practice, such conditions are rarely met, then the implication is not that the distinction is unnecessary, but that the field has been overstating generalization and should instead adopt transfer-oriented descriptions accompanied by appropriate adaptation metrics. What appears as a critique thus reinforces the need for terminological discipline.

Another line of criticism appeals to the maturity of machine-learning theory, noting that the distinction between generalization and transfer is already well established and therefore offers little conceptual novelty. While this is formally correct—these ideas are foundational within statistical learning theory and extensively treated in transfer-learning literature [1, 2]—their status within materials-AI practice tells a different story. Empirical studies conducted across the period routinely blur the boundary, with cross-domain performance described as "generalization," and such characterizations often pass unchallenged in peer review [8-10]. The issue, therefore, is not the absence of theoretical knowledge but its incomplete integration into a domain-specific epistemic context. Materials science introduces constraints—data scarcity, high-dimensional and sparsely populated feature spaces, and the strategic demands of inverse design—that amplify the consequences of conceptual imprecision [4, 7]. In this setting, the contribution of the framework lies not in redefining established theory but in translating it into a form that aligns with the practical realities and interpretive norms of the field.

Considered together, these objections do not erode the validity of the proposed distinction; instead, they underscore the importance of making its underlying structure explicit. By clarifying how continuous variation can be mapped, how real-world deviations should be interpreted, and how established concepts must be adapted to domain-specific conditions, the framework functions less as a rigid taxonomy than as an interpretive instrument. Its value lies precisely in its capacity to render implicit assumptions visible, thereby enabling more disciplined claims and more reliable accumulation of knowledge within materials-AI research.

Implications for Materials AI Practice

Adopting the proposed distinction and boundary framework carries direct consequences for how authors, reviewers, and the broader community conduct, evaluate, and communicate research.

For authors, the implications are threefold. First, every performance claim must specify whether it concerns generalization (zones 1–2) or transfer (zones 3–4) and must state the measured distribution shift and feature-space overlap explicitly. Second, the training–test split description must include quantitative diagnostics—such as maximum mean discrepancy or hull-overlap statistics—so readers can locate the evaluation on the framework map. Third, abstracts and conclusions must avoid sliding from transfer success to implied generalization guarantees; instead, they should report each capability separately, as Butler *et al.* [4] and Schmidt *et al.* [5] already recommend for related statistical claims.

For reviewers, the checklist expands. Reviewers should verify that the label "generalization" is used only when i.i.d. conditions are demonstrably satisfied and that "transfer" is reserved for acknowledged domain shifts [2]. When a manuscript reports strong performance on a cross-domain task yet labels it generalization, reviewers should request either relabeling or additional i.i.d. benchmarks within each domain. The framework supplies a shared vocabulary that makes such requests precise rather than subjective.

For the community at large, the implications concern infrastructure and standards. Shared benchmarks should be partitioned into separate generalization suites (i.i.d. splits within fixed distributions) and transfer suites (explicit source–target domain pairs), mirroring the separation already emerging in some cross-property efforts [9]. Reporting templates should mandate a one-sentence "claim locator" that places the evaluation in one of the four zones. Funding calls and editorial policies can require this locator, thereby accelerating the shift from ambiguous claims to reproducible epistemic commitments. Collectively, these changes transform materials AI from a field in which success is reported in interchangeable terms to one in which each capability—generalization inside the distribution

and transfer outside it—is measured, discussed, and improved on its own terms.

Conclusion

The central contribution of this boundary/definitional paper has been to demonstrate that generalization and transfer, though frequently treated as synonyms in materials machine learning, designate two fundamentally distinct capabilities grounded in the presence or absence of distribution shift. Generalization, as defined here, concerns performance under the i.i.d. assumption within the training distribution; transfer concerns performance when that assumption is violated and adaptation across domains is required. The conflation of these notions has produced category errors, overclaims, and evaluation mismatches that undermine the cumulative reliability of the field. By advancing four numbered definitions, a systematic comparison with neighboring terms, and a two-dimensional boundary framework that maps any evaluation into one of four zones, the paper supplies the conceptual precision necessary to locate claims accurately along the generalization–transfer spectrum.

Most existing materials-AI studies operate in the intermediate zones—near-generalization or near-transfer—yet are reported under the stronger label of generalization. The framework makes this slippage visible and therefore correctable. Its adoption will not diminish the genuine empirical advances already achieved; rather, it will allow those advances to be interpreted with greater epistemic clarity. Future work can now state with precision whether a model has been shown to generalize within a known distribution, to transfer across distributions, or to occupy some intermediate position. Reviewers can enforce this precision, and the community can develop separate benchmarks and reporting standards for each capability.

The ultimate implication is modest yet foundational: precise language about generalization and transfer is a prerequisite for trustworthy progress toward inverse design and materials discovery [7]. When authors, reviewers, and readers consistently distinguish the two concepts, claims about model performance become falsifiable, comparable, and cumulative. The present analysis, therefore, calls for a disciplined shift in vocabulary and evaluation practice across the materials-AI literature, ensuring that the impressive empirical successes of the past five years are placed on a conceptually sound footing for the decade ahead.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 10 May 2022 Revised: 30 Jun 2022 Accepted: 26 Jul 2022
Published online: 18 January 2023

References

De Mello RF, Ponti MA. Statistical learning theory. Mach Learn. 2018;3.

Panigrahi S, Nanda A, Swarnkar T. A survey on transfer learning. In: Intelligent and cloud computing; AHFE 2019. Singapore: Springer; 2021. p. 781-9.

Feng S, Fu H, Zhou H, Wu Y, Lu Z, Dong H. A general and transferable deep learning framework for predicting phase formation in materials. npj Comput Mater. 2021;7(1):10.

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

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

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.

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

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.

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.

Chen C, Ong SP. AtomSets as a hierarchical transfer learning framework for small and large materials datasets. npj Comput Mater. 2021;7(1):173.

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

Oviedo F, Ferres JL, Buonassisi T, Butler KT. Interpretable and explainable machine learning for materials science and chemistry. Acc Mater Res. 2022;3(6):597-607.

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

Ojih J, Al-Fahdi M, Rodriguez AD, Choudhary K, Hu M. Efficiently searching extreme mechanical properties via boundless objective-free exploration and minimal first-principles calculations. npj Comput Mater. 2022;8(1):143.

Meredig B, Antono E, Church C, Hutchinson M, Ling J, Paradiso S, et al. Can machine learning identify the next high-temperature superconductor? Examining extrapolation performance for materials discovery. Mol Syst Des Eng. 2018;3(5):819-25.

Chang R, Wang YX, Ertekin E. Towards overcoming data scarcity in materials science: Unifying models and datasets with a mixture of experts framework. npj Comput Mater. 2022;8(1):242.

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.

De Breuck PP, Hautier G, Rignanese GM. Materials property prediction for limited datasets enabled by feature selection and joint learning with MODNet. npj Comput Mater. 2021;7(1):83.

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

Gupta T, Zaki M, Krishnan NA, Mausam. MatSciBERT: A materials domain language model for text mining and information extraction. npj Comput Mater. 2022;8(1):102.

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

Banchi L, Pereira J, Pirandola S. Generalization in quantum machine learning: A quantum information standpoint. PRX Quantum. 2021;2(4):040321.

Stoll A, Benner P. Machine learning for material characterization with an application for predicting mechanical properties. GAMM-Mitt. 2021;44(1):e202100003.

Caro MC, Huang HY, Cerezo M, Sharma K, Sornborger A, Cincio L, et al. Generalization in quantum machine learning from few training data. Nat Commun. 2022;13(1):4919.

Peterson GG, Brgoch J. Materials discovery through machine learning formation energy. J Phys Energy. 2021;3(2):022002.

Choudhary K, DeCost B, Chen C, Jain A, Tavazza F, Cohn R, et al. Recent advances and applications of deep learning methods in materials science. npj Comput Mater. 2022;8(1):59.

Schmidt J, Pettersson L, Verdozzi C, Botti S, Marques MA. Crystal graph attention networks for the prediction of stable materials. Sci Adv. 2021;7(49):eabi7948.

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

Kolluru A, Shoghi N, Shuaibi M, Goyal S, Das A, Zitnick CL, et al. Transfer learning using attentions across atomic systems with graph neural networks (TAAG). J Chem Phys. 2022;156(18).