ORIGINAL RESEARCH **Open access**

What Does "Compositional Generalization" Mean for Multi-Principal Element Alloys? A Boundary Problem for GNNs

Maria Hernandez[1*], Carlos Vega[1]

Abstract

The term compositional generalization is increasingly invoked in machine learning for multi-principal element alloys (MPEAs), yet its meaning remains inconsistent and often conflated with interpolation or extrapolation. In standard machine learning, compositional generalization denotes the structured recombination of known components into novel configurations. Translating this concept to MPEAs is non-trivial due to their continuous composition space, permutation symmetries, and overlapping local atomic environments. Graph neural networks (GNNs), while effective for property prediction, are not inherently designed for such recombination beyond interpolation. This article identifies a resulting boundary problem and proposes an operational definition of compositional generalization based on four criteria: novel element–concentration co-occurrences, extrapolation beyond the convex hull of training compositions, controlled performance degradation, and permutation invariance. The framework clarifies widespread misuses of the term and provides a concrete basis for evaluating and designing models for MPEA discovery.

Keywords Materials informatics, Compositional generalization, High-entropy alloys, Graph neural networks, Out-of-distribution generalization, Multi-principal element alloys

*Correspondence:
Maria Hernandez
maria.hernandez@gmail.com
[1] Department of Computational and Smart Materials Systems, Faculty of Engineering, University of Valencia, Valencia, Spain

Introduction

"Compositional generalization" has become a buzzphrase in materials machine learning, particularly for multi-principal element alloys (MPEAs, including high-entropy alloys). Papers claim that their GNNs achieve compositional generalization—but what exactly does this mean? Does it mean predicting properties for unseen element combinations? For new concentrations of seen elements? For element permutations? For compositions outside the training range? The term is used ambiguously, and this ambiguity hinders scientific progress. This paper provides a boundary/definitional analysis of compositional generalization for MPEAs in the context of GNNs.

The ambiguity is not merely terminological. In the broader machine-learning literature, compositional generalization is understood as the capacity to recombine known primitives according to rules that were never explicitly observed during training [1, 2]. Lake *et al.* argued that human-like generalization requires compositional structure [2], yet the translation of this idea to materials science has been largely implicit and unexamined. Within MPEA research, the term appears in discussions of extrapolation to higher-order systems [3], unseen concentrations [4], and element-substitution scenarios [5], but without a shared operational definition. Bartók *et al.* showed that machine-learning interatomic potentials can unify modeling across molecules and materials [6], yet their framework treats composition as a fixed input rather than a

recombinable structure. Batra *et al.* surveyed emerging materials intelligence ecosystems and highlighted the need for robust generalization [7, 8], but again left "compositional" unspecified.

The stakes are high. MPEAs are prized precisely because their vast composition space—spanning quinary and higher systems—promises property landscapes unreachable by conventional alloys [9, 10]. If GNNs are to accelerate discovery in this space, they must move beyond memorizing known points or smoothly interpolating between them. Yet without a precise boundary between interpolation and compositional generalization, it is impossible to evaluate whether current models truly recombine elemental contributions in novel ways or merely exploit statistical regularities in the training distribution. Zuo *et al.* benchmarked multiple machine-learning interatomic potentials and demonstrated impressive accuracy on held-out compositions [11], yet nowhere defined what they meant by compositional generalization. Behler's review of high-dimensional neural network potentials similarly emphasizes transferability without addressing the compositional dimension explicitly [12]. Researchers surveyed machine learning for high-entropy alloys and noted progress in property prediction, but the absence of a formal definition of compositional generalization limits the interpretability of reported successes [3].

This article therefore undertakes a conceptual boundary analysis. Section 2 clarifies what generalization means in standard machine learning and isolates the specific demands of compositional generalization. Section 3 articulates why MPEAs constitute a special case that renders the concept both essential and ambiguous. Section 4 documents existing usages and confusions in the literature. Section 5 formulates the resulting boundary problem for GNNs. Section 6 proposes a definitional framework with formal criteria. Section 7 examines boundary cases and gray zones. Subsequent sections (to appear in Part 2) relate the framework to systematicity, compositionality, and inductive bias, then derive concrete implications for GNN design and evaluation. By the end, the field will possess an operational definition that replaces vague claims with testable, comparable assertions.

Generalization in machine learning resists reduction to a single, unified notion; instead, it comprises a set of related but non-equivalent capacities whose distinctions become analytically consequential when translated into materials science contexts. At its most conventional, generalization is understood through the lens of the independent and identically distributed assumption, where model performance is evaluated on data drawn from the same underlying distribution as the training set [13]. This formulation stabilizes the train–test paradigm but implicitly constrains the scope of inference to regimes where statistical continuity is preserved. Once this constraint is relaxed, a broader and more demanding interpretation emerges, in which models are expected to maintain predictive fidelity under distributional shifts. Such out-of-distribution behavior encompasses covariate shift, label shift, and concept drift, each introducing a distinct mechanism by which learned associations may degrade or require adaptation.

This expansion of scope foregrounds a further layer of complexity associated with how models manipulate internal representations. Compositional generalization, as discussed in machine learning and cognitive science, captures the capacity to recombine previously acquired components into configurations not encountered during training [1, 2]. The conceptual significance of this ability lies not merely in novelty, but in the structured reuse of learned primitives. A stricter interpretation is introduced by systematic generalization, where successful recombination must exhibit rule-governed consistency rather than context-specific adaptation. The distinction is consequential: whereas compositional behavior may arise through flexible pattern matching, systematicity implies an underlying algebraic or symbolic regularity that constrains how new combinations are processed. Empirical work underscores this divide, with Mankowitz *et al.* demonstrating that compositional behavior in neural systems can be induced through targeted inductive biases that privilege recombination [1], while Lake *et al.* argue that the systematic nature of human cognition remains only partially captured by current architectures [2].

Clarifying these conceptual layers requires disentangling compositionality from more familiar geometric notions of interpolation and extrapolation. Interpolation operates within the convex hull defined by observed data, effectively smoothing over sparsity by leveraging local continuity in feature space. Extrapolation, by contrast, extends learned relationships beyond this boundary, requiring models to project patterns into regions devoid of empirical support.

Compositional generalization cuts across this geometric distinction. It is not defined by spatial position relative to the training distribution, but by the novelty of how constituent elements are combined. In a materials context—consider a compositional phase space represented by a ternary diagram—this distinction becomes concrete. A candidate alloy may lie well within the convex hull of known compositions and yet embody a previously unobserved pairing of elements and concentrations. Conversely, a composition outside the hull may or may not introduce such novelty. The defining feature is therefore not location, but the absence of prior co-occurrence among constituent components.

This perspective exposes a subtle but critical limitation in how generalization is often inferred in materials modeling. Convex-hull coverage, while visually intuitive, does not guarantee that a model has internalized the combinatorial structure necessary for meaningful recombination. Hupkes *et al.* provide a useful analytical lens by demonstrating that neural architectures frequently achieve strong interpolation performance while failing to exhibit systematic compositional behavior [14]. A similar pattern is likely to arise in atomistic and crystal graph models. Approaches such as graph neural networks for crystalline systems [15] and continuous-filter architectures like SchNet [16, 17] have demonstrated impressive performance under IID conditions, yet their representational adequacy for recombining elemental interactions in unseen configurations remains underexplored. The implication is not merely methodological but conceptual: claims of generalization in complex alloy systems, including multi-principal element alloys, must be situated within a framework that distinguishes distributional robustness from true compositional reasoning [18]. Without this distinction, apparent predictive success may obscure a deeper fragility when models encounter genuinely novel chemical arrangements.

Multi-principal element alloys inhabit a compositional regime that both necessitates and destabilizes conventional interpretations of compositional generalization. The challenge begins with the nature of composition itself: rather than discrete symbols, alloy systems are defined over continuous concentration vectors constrained to the (n-1) simplex, where each coordinate is interdependent through the unity constraint. Under these conditions, recombination cannot be understood as the juxtaposition of separable units; instead, it corresponds to movement within a metric space in which even small perturbations in concentration can induce qualitatively different physicochemical behavior. This continuous embedding complicates any direct transfer of intuitions derived from symbolic domains, where compositionality is typically framed as rule-based concatenation.

A related complication emerges from the representational symmetry inherent to alloy systems. Physically equivalent compositions admit multiple permutations in vector form, yet a model that fails to recognize this invariance risks encoding spurious distinctions. At the same time, enforcing permutation invariance does not, in itself, resolve the deeper issue of whether the model can meaningfully recombine elemental contributions. The difficulty intensifies as the number of principal elements increases. Composition spaces in MPEAs expand combinatorially while remaining geometrically confined, producing a structure that is neither well-approximated by low-dimensional intuition nor trivially navigable through standard Euclidean reasoning. In practice, this means that proximity in compositional space does not straightforwardly correspond to similarity in underlying atomic configurations or emergent properties.

This tension becomes particularly evident when considering the role of local atomic environments. Distinct global compositions may generate nearly indistinguishable local coordination patterns, effectively collapsing differences that would otherwise signal novelty at the compositional level. As emphasized by Bartók *et al.*, local environments often dominate property prediction in atomistic models, suggesting that learned representations may privilege structural motifs over explicit compositional diversity [6]. The implication is subtle but important: a model may appear to generalize across compositions while, in effect, reusing familiar local patterns rather than engaging in genuine recombination of elemental identities and concentrations.

The situation is further complicated by the intrinsically non-linear character of materials property landscapes. Mechanical, thermodynamic, and chemical properties emerge from many-body interactions that resist decomposition into additive elemental contributions [19, 20]. Consequently, any notion of compositional generalization that relies on linear superposition is fundamentally misaligned with the governing physics. This non-linearity has been repeatedly highlighted in the development of machine-learned interatomic potentials, where capturing higher-order interactions is essential for

predictive accuracy [21, 22]. Behler's account of high-dimensional neural network potentials underscores that meaningful representations must encode complex many-body effects rather than simple pairwise relationships [12].

These structural and physical constraints are well documented across the materials informatics literature. Batra *et al.* draw attention to the data sparsity and scaling challenges associated with high-dimensional composition spaces, particularly in the context of integrated materials intelligence systems [7]. Zuo *et al.* further illustrate the difficulty of extending machine-learning potentials across compositionally diverse regimes, where extrapolative reliability remains limited despite strong in-distribution performance [11]. Complementing these observations, Researcher identify the continuous and multi-element character of high-entropy alloys as a persistent barrier to robust predictive modeling, noting that existing approaches often struggle to reconcile compositional complexity with limited training data [3].

Taken together, these considerations suggest that the notion of "recombining known components" cannot be directly imported from cognitive science without substantial reinterpretation. In the context of MPEAs, a model trained on lower-order subsystems must implicitly disentangle elemental identity, concentration-dependent effects, and interaction terms in a manner that supports transfer to higher-order compositions. Achieving this requires more than smooth interpolation within a densely sampled region; it demands a representational structure capable of extrapolating combinatorial relationships under shifting constraints. Without a precise conceptual framework to distinguish these mechanisms, empirical success risks being misattributed. What appears as compositional generalization may, under closer scrutiny, reflect only the exploitation of local regularities within a geometrically constrained and smoothly varying landscape.

Existing Usage and Confusions

Examination of the recent literature reveals that the term "compositional generalization" is invoked across multiple research contexts in ways that conflate qualitatively distinct phenomena. One common usage involves extrapolation to higher-order systems, such as training on binary or ternary alloys and testing on quaternary or quinary formulations. Several studies describe this as compositional generalization [3, 4], yet what is actually being evaluated is extrapolation within the simplex; the model may succeed merely by extending smooth compositional trends rather than by recombining learned primitives in a genuinely novel manner. A related implication is that such performance does not necessarily require symbolic-like compositionality, a point often overlooked.

Beyond this immediate concern, another usage emerges when models are trained on discrete concentration slices—for instance, 0%, 25%, 75%, or 100%—and tested on intermediate values such as 50%. This scenario constitutes textbook interpolation, not compositional generalization, despite occasional claims to the contrary precisely because the exact ratio is novel [5]. The conflation here is subtle but consequential: interpolative success says little about a model's ability to handle inputs that lie outside the convex hull of training distributions. In practice, distinguishing between interpolation and composition demands explicit geometric criteria, which the literature rarely provides.

A third usage, training on CoCrFe and CoCrNi then testing on CoCrFeNi, comes closest to the cognitive-science ideal of compositionality [1]. Even here, however, ambiguity persists because the convex hull of training points may still contain the test point via linear mixing, meaning that purely numerical interpolation could suffice. This ambiguity points to a deeper issue: without operational tests for whether a model actually recombines element–concentration pairs in a rule-like manner, claims of compositional generalization remain underdetermined.

Compounding the difficulty, a fourth usage centers on element permutation invariance—asserting generalization on the grounds that predictions remain unchanged under reordering of elements in the input vector. While such invariance is a necessary symmetry requirement for any physically consistent model [15, 16], it is not generalization at all but rather an architectural design choice. Treating it as evidence of compositionality confounds necessary conditions with sufficient evidence of novel recombination. A fifth and final usage involves transfer across similar alloy families, such as training on CoCrFeNi and testing on CuNi-based systems. This scenario is best understood as domain adaptation driven by shared local environments rather than compositional recombination [7, 23].

Across all five usages, the literature typically offers no operational criteria for what would count as genuine compositional generalization. Pei *et al.*, for example, predicted high-entropy solid-solution formation but provided no explicit test of whether the model recombined element–concentration pairs in a compositional manner [23]. Similarly, Wen *et al.* designed high-entropy alloys via machine learning yet evaluated performance only on independent and identically distributed splits [5]. The cumulative effect is a proliferation of incomparable claims: one paper's "compositional generalization" becomes another's interpolation, undermining cumulative progress in the field. Under these conditions, what is urgently needed is not more demonstrations of empirical fit but a shared framework of operational definitions and falsifiable tests tailored to the combinatorial structure of materials systems. **Table 1** distinguishes the principal usages currently labeled as "compositional generalization" and shows why most of them fall short of the stricter boundary advanced in this article.

Table 1. Boundary Demarcation of Competing Usages of "Compositional Generalization" in MPEA Modeling

Usage in the Literature	What Is Actually Being Tested	Why It Is Often Misclassified	Boundary Status Under the Proposed Framework	What Authors Should Call It Instead
Extrapolation to higher-order systems	Prediction on quaternary or quinary systems after training on lower-order alloys	Higher-order novelty is assumed to imply recombinational novelty, even when the model is only extending smooth compositional trends	Not sufficient on its own	Higher-order extrapolation
Unseen concentration points	Prediction at new concentration ratios between previously observed compositions	Exact concentration novelty is mistaken for compositional novelty, although the test may remain inside the convex hull	Usually not compositional generalization	Concentration-space interpolation
Unseen element combinations	Prediction for new combinations of previously seen elements	This comes closest to genuine compositional recombination, but may still occur inside the training hull	Necessary but not sufficient	Recombinational novelty test
Permutation invariance	Invariance of predictions under reordering of element indices	Architectural symmetry is confused with generalization	Baseline requirement only, not evidence of generalization	Physical invariance check
Transfer across related alloy families	Prediction across chemically similar systems with overlapping local environments	Success may reflect domain similarity rather than structured recombination	Distinct from compositional generalization	Domain transfer or family transfer
Distance-based OOD testing	Prediction on test points that are compositionally distant from training data	Distance from training points is treated as equivalent to recombinational novelty	Insufficient unless novel joint occurrences are also established	OOD extrapolation test
Random train-test split	Good performance on held-out samples from	Generic generalization performance is rhetorically	Outside the definition	IID generalization

success	the same data distribution	inflated into compositional generalization		

The diagnosis is clear. The field lacks a shared operational definition. This absence produces incomparable claims, hidden assumptions about what counts as "compositional," and systematic confusion between interpolation in a smooth landscape and genuine recombination of elemental contributions. Without boundary-drawing, progress toward models that truly navigate the MPEA simplex remains unmeasurable.

A Boundary Problem for GNNs

Graph neural networks occupy a peculiar position with respect to compositional generalization in multiprincipal element alloys. Their message-passing architecture encodes local atomic environments with high fidelity, yet this very locality gives rise to a set of interconnected boundary problems that the field has only begun to recognize. The first of these concerns the threshold between compositional and non-compositional behavior. Where exactly does the boundary lie between a GNN that merely interpolates within composition space and one that genuinely generalizes compositionally? Current evaluation practices rarely distinguish the two, treating smooth extrapolation as evidence of recombination when in fact the two capacities rest on fundamentally different mechanisms. This ambiguity is not merely semantic: it directly affects whether a model can be trusted to predict properties outside the convex hull of training compositions.

A related but distinct boundary question concerns representational requirements. For a GNN to support genuine compositional generalization, it must maintain disentangled embeddings for each element and its concentration, support compositional operators that combine these embeddings in a systematic manner, and remain aware of global simplex constraints that govern physically realizable alloys. Chen *et al.* introduced graph networks as a universal framework [15], but the paper does not specify how such disentanglement is enforced for continuous compositions, leaving a critical gap between architectural capability and compositional necessity. In practice, without explicit mechanisms for keeping element-specific and concentration-specific information separate, the network risks entangling them in ways that preclude systematic recombination.

Beyond representational issues lies the data regime boundary: how much compositional diversity in the training set is sufficient to claim generalization rather than memorization? Schütt *et al.* demonstrated that SchNet generalizes across molecular and materials domains when local environments overlap [16], yet the precise amount of required recombination diversity remains unquantified. This matters because a model trained on sufficiently dense sampling of composition space may appear to generalize compositionally when in fact it has simply memorized a fine-grained grid [24]. Conversely, a model trained on sparse but strategically chosen compositions might genuinely recombine primitives. The literature provides no guidance on where that threshold lies.

The evaluation boundary follows naturally from these concerns. What test protocols can reliably distinguish compositional generalization from interpolation or domain adaptation? Standard random splits or distance-based out-of-distribution splits fail to isolate novel joint occurrences of element-concentration pairs. A test point that lies far from any training point in Euclidean distance may still be interpolative in a compositional sense if it falls inside the convex hull of training compositions when represented appropriately. Conversely, a test point close in distance may require genuine recombination if it pairs elements and concentrations that never co-occurred during training. Developing protocols that operationalize this distinction remains an open methodological challenge.

Finally, the architectural boundary asks whether GNNs, by virtue of their message-passing inductive bias, possess inherent limitations compared with architectures that might embed explicit compositional operators—transformers with concentration-conditioned attention, for example, or neural-symbolic hybrids. Xie and Grossman’s crystal graph convolutional networks excel at property prediction [25] but inherit the same localism that may limit true recombination. Message passing aggregates information from local neighborhoods; whether this aggregation can ever amount to the

systematic, rule-like recombination of discrete primitives is not obvious. Each of these five boundary questions matters because the answers collectively determine whether GNNs can ever serve as reliable oracles for the unexplored regions of multiprincipal element alloy space.

The literature has not yet addressed them systematically. Schmidt *et al.* reviewed advances in machine learning for solid-state materials and called for better generalization benchmarks [26], yet offered no multiprincipal element alloy-specific compositional criteria. Ramprasad *et al.* surveyed materials informatics and noted the promise of GNNs [27], again without boundary definitions. The consequence is that claims of compositional generalization remain unverifiable, and architectural innovations proceed without a clear target. Until the field confronts these boundary problems explicitly—operationalizing each, proposing falsifiable tests, and distinguishing necessary from sufficient architectural conditions—progress will continue to be measured against ill-defined benchmarks, and the gap between claimed and actual compositional capacity will remain unclosed.

Proposed Definitional Framework

Figure 1 maps the article's central argument by showing how broad machine-learning notions of generalization become ambiguous in MPEA composition space and how that ambiguity culminates in the proposed operational boundary for GNN evaluation.

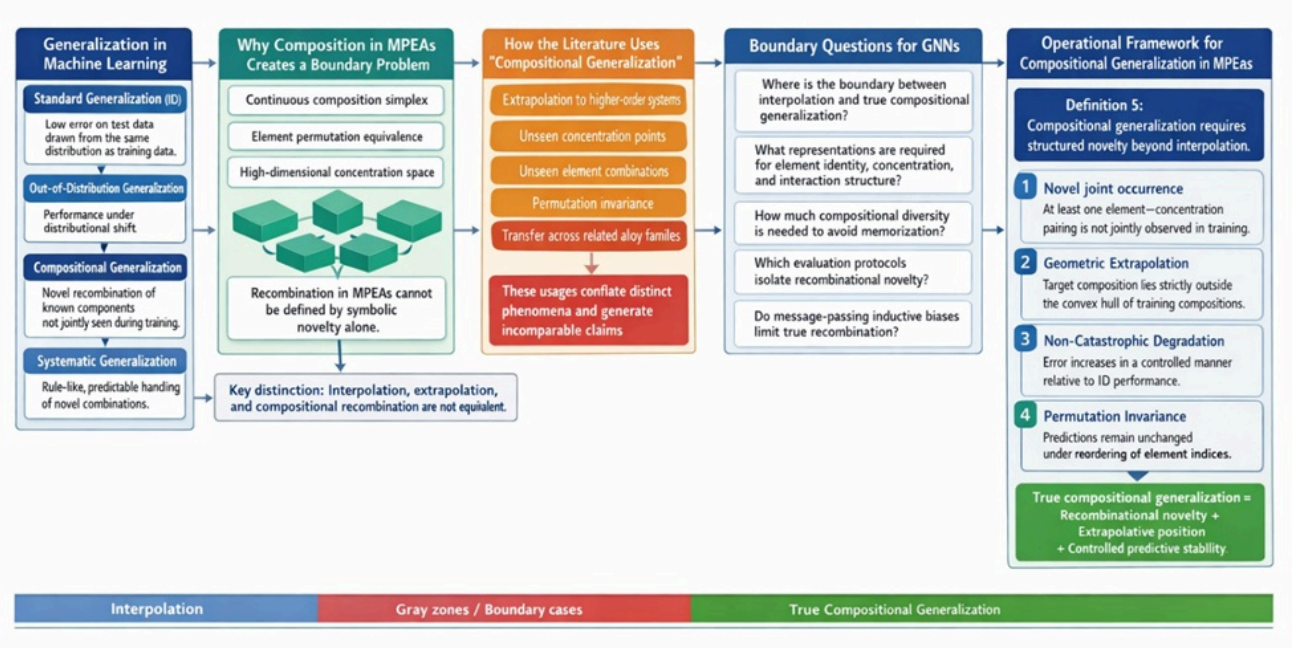

Figure 1. Generalization in machine learning within the MPEA composition space and the resulting proposed operational boundary for GNN evaluation.

We propose the following operational framework to resolve the boundary problem.

A model is said to exhibit *compositional generalization* in multi-principal element alloy (MPEA) property prediction if it satisfies the following conditions simultaneously:

1. **Novel compositional recombination**: The target composition includes at least one pair of elements with associated concentrations that were not jointly observed in any training instance.

2. **Geometric extrapolation**: The target composition lies strictly outside the convex hull of the training compositions in the composition simplex.
3. **Non-catastrophic performance degradation**: The prediction error on such target compositions does not exceed a constant-factor increase relative to the error observed on held-out IID test data.
4. **Permutation invariance**: Model predictions are invariant under permutation of element indices.

Together, these conditions distinguish compositional generalization from interpolation and unconstrained extrapolation by requiring simultaneous recombinatorial novelty, geometric extrapolation, and controlled predictive stability.

Boundary Cases and Gray Zones

Even with a formally specified framework in place, certain test cases resist clean classification. These ambiguous instances are not merely edge cases to be dismissed; they illuminate the subtle, graded nature of the boundary between interpolation and genuine compositional generalization in graph neural networks for multiprincipal element alloys.

Consider first the missing intermediate problem. Training on equiatomic CoCr and equiatomic FeNi, then testing on equiatomic CoCrFeNi, places the test point inside the convex hull of training compositions via linear mixing. Under a strict reading of standard criteria, this would count as interpolation. Yet the joint four-element occurrence is undeniably novel: no training sample contains all four elements simultaneously. The model must still discover quaternary interaction terms that are not reducible to pairwise or ternary combinations. One could argue for inclusion as a form of compositional generalization because the recombination of element sets is required. The counterargument is equally forceful: because the hull contains the point, a purely linear model could succeed without any nonlinear recombination. A reasonable boundary proposal treats this as a gray-zone interpolation case unless empirical evidence shows that nonlinear interaction terms beyond simple mixing are necessary for accurate prediction.

A related difficulty appears in the concentration continuum problem. Training on Co-rich (80% Co, 10% Cr, 10% Ni) and Ni-rich (10% Co, 10% Cr, 80% Ni) compositions, then testing on equiatomic (33-33-33), involves interpolation in concentration space. Nevertheless, the test point recombines element-concentration pairs that never appeared together during training: for instance, the combination of 33% Co with 33% Ni is jointly novel even though each concentration appears elsewhere with different partners. This case tests whether the model has learned smooth, continuous concentration dependence that extrapolates systematically to intermediate values, or whether it has merely memorized discrete regimes and performs piecewise interpolation. The ambiguity here is particularly instructive because it reveals that geometric position alone—inside or outside the convex hull—is insufficient; one must also examine the joint distribution of element-concentration pairs.

Beyond geometric considerations, local environment similarity creates another source of ambiguity. In the similar local environment problem, training on CoCrFeNi and CuNi separately, then testing on CoCrFeNiCu, places the test point outside the convex hull but only marginally so. More importantly, the local chemical environments around Cu atoms closely resemble those around Ni atoms in the training set. The model may therefore succeed via local pattern matching—exploiting the fact that Cu behaves analogously to Ni in similar coordination shells—rather than through true recombination of compositional primitives. Whether such success counts as compositional generalization depends on whether one emphasizes the novelty of the element combination (Cu appears with Co, Cr, Fe, and Ni for the first time) or the mechanism of transfer (the model essentially treats Cu as a substitute for Ni, which is domain adaptation under another name).

The element substitution problem sharpens this tension further. Training on CoCrFeMn and testing on CoCrFeNi involves introducing a completely novel element (Ni) that was absent from all training compositions. Yet Ni and Mn have similar atomic sizes and crystal structure preferences, so local atomic environments around Ni in the test alloy may be well approximated by environments around Mn in the training alloys. Success here could reflect domain adaptation driven by shared local structural motifs rather than any systematic recomposition of element-concentration pairs. The vector

substitution crossing an element-type boundary challenges the field to specify whether compositionality requires element-level novelty, environment-level similarity, or something in between.

These four cases are usefully visualized on a two-dimensional projection of the composition simplex. Training points form a sparse cloud with a shaded convex hull. The missing intermediate case lies inside the hull but marked with a dashed recombination circle, signaling its ambiguous status. The concentration continuum case sits on the hull edge, annotated with concentration-gradient arrows that indicate smooth variation along one direction but joint novelty across element-concentration pairs. The similar local environment case appears just outside the hull, with overlapping local-environment shading to indicate that geometric extrapolation is minimal and pattern matching may suffice. The element substitution case is shown as a vector substitution arrow crossing an element-type boundary, emphasizing the shift in elemental identity while preserving local structural analogies. Decision boundaries drawn according to the proposed criteria would separate clear successes—cases requiring both strict extrapolation and novel joint element-concentration pairs—from gray zones, visually clarifying where current claims about GNN generalization in multiprincipal element alloys actually fall.

What these cases demonstrate collectively is that the boundary between interpolation and compositional generalization is not binary but graded. A test point can be interpolative in one sense (geometric position inside the hull) yet compositionally novel in another (joint occurrence of elements or concentration pairs). Conversely, a point can be extrapolative geometrically yet rely on local environment similarity that undermines claims of true recomposition. Rigorous evaluation protocols must therefore interrogate both geometric position and combinatorial novelty, and they must do so jointly rather than sequentially. Without such dual interrogation, the field risks celebrating as compositional generalization what may simply be interpolation in disguise, domain adaptation, or local pattern matching—each of which has different implications for whether GNNs can serve as reliable oracles in unexplored regions of multiprincipal element alloy space.

Table 2 converts the proposed definition into a practical decision matrix that allows researchers to classify evaluation scenarios without conflating interpolation, extrapolation, and true compositional generalization.

Table 2. Decision Matrix for Classifying MPEA Prediction Tasks as Interpolation, Gray-Zone Generalization, or True Compositional Generalization

Evaluation Condition	Inside Convex Hull?	Novel Element–Concentration Joint Occurrence?	New Element Combination?	Controlled Error Degradation?	Permutation Invariance Satisfied?	Classification
New concentration point between observed training compositions	Yes	Sometimes no	No	Yes	Yes	Interpolation
New quaternary composition linearly reachable from observed lower-order mixtures	Yes	Possibly yes	Yes	Yes	Yes	Gray-zone case
New composition outside hull but assembled from	No	No	Sometimes yes	Yes	Yes	Extrapolation, not compositional generalization

already observed joint pairings						
New composition outside hull with at least one unobserved element–concentration pairing	No	Yes	Yes or partial	Yes	Yes	True compositional generalization
New composition outside hull with novel recombination but severe performance collapse	No	Yes	Yes	No	Yes	Failed compositional generalization
Prediction unchanged under element reordering only	Variable	No	No	Variable	Yes	Invariance only, not generalization
Transfer from one alloy family to another with overlapping local environments	Variable	Variable	Sometimes	Yes	Yes	Domain transfer / adaptation
Boundary case with quaternary novelty inside hull but emergent interaction demands	Yes	Partial	Yes	Variable	Yes	Gray zone requiring explicit reporting

Relation to Other Key Concepts

The proposed definitional framework for compositional generalization in MPEAs does not exist in isolation; it must be situated within a network of neighboring concepts that have shaped both machine-learning theory and materials informatics. Clarifying these relations prevents the framework from being misinterpreted as a narrow technical stipulation and instead positions it as a conceptual bridge.

Systematicity is perhaps the closest relative. Systematic generalization, as articulated by Lake *et al.*, requires that the ability to handle novel combinations follows a predictable, rule-governed pattern rather than isolated memorization [2]. In the cognitive-science literature, this often takes an algebraic form: once a model learns “add red” and “add blue,” it should systematically produce “red + blue” without further examples. Mankowitz *et al.* showed that neural networks can be engineered for such systematicity when provided with appropriate compositional inductive biases [1]. For MPEAs, however, full algebraic systematicity is neither necessary nor always desirable. The composition simplex is continuous, not discrete, and property landscapes are governed by non-linear many-body physics rather than symbolic rules. A GNN

may achieve the operational criteria of Definition 5—novel joint occurrences outside the convex hull with controlled error degradation—without exhibiting strict algebraic recombination. Thus, compositional generalization in MPEAs is a weaker but still demanding condition: it insists on structured recombination of elemental contributions yet tolerates the smooth, physics-constrained variations inherent to continuous concentration spaces.

Compositionality, in the linguistic or cognitive sense, supplies the philosophical underpinning. The principle states that the meaning of a whole is a function of the meanings of its parts and the mode of their combination. Applied to MPEAs, this would imply that the property of a quinary alloy is systematically derivable from the contributions of its constituent elements and their local interaction modes. Yet the non-linear property landscapes documented by Batra *et al.* [7] and Behler [12] demonstrate that simple functional composition rarely holds; emergent phenomena such as lattice distortion or short-range ordering break strict additivity. The proposed framework therefore adopts a pragmatic, operational compositionality: it requires only that the model recombine element-concentration pairs in a manner that respects the simplex geometry and local-environment similarity, without demanding that the underlying mapping be factorizable into independent elemental functions. This pragmatic stance distinguishes MPEA compositional generalization from the stronger linguistic ideal while preserving its spirit.

Out-of-distribution (OOD) generalization provides the broader umbrella. Compositional generalization is one specific species of OOD generalization—structured recombination rather than arbitrary distribution shift. Standard OOD benchmarks in materials science often test extrapolation along a single axis (temperature, pressure) or domain adaptation across chemically similar families [26]. The framework advanced here is stricter: it simultaneously demands geometric extrapolation in the simplex (Criterion 2) and combinatorial novelty (Criterion 1). Consequently, a model that passes generic OOD tests may still fail the compositional criteria if the test points lie inside the convex hull or reuse previously observed element-concentration pairs. This distinction is crucial because many GNN evaluations reported in the literature rely on random or distance-based OOD splits that conflate these phenomena [27].

Inductive bias enters the picture as the architectural enabler. Graph neural networks possess strong relational inductive biases through message passing and permutation invariance [15, 16], yet these biases alone do not guarantee compositional recombination. The framework implies that additional biases—disentangled element embeddings, concentration-conditioned attention, and explicit simplex constraints—are required if GNNs are to cross the boundary from interpolation to genuine compositional generalization. Schütt *et al.* demonstrated that continuous-filter convolutions improve generalization across molecular and materials domains by respecting local geometry [16], but the same architecture may still collapse elemental identities unless explicitly regularized for disentanglement.

Finally, extrapolation capacity is necessary but not sufficient. The framework elevates extrapolation (Criterion 2) to a core requirement while insisting that it be driven by compositional recombination rather than mere extension of smooth trends. In the MPEA simplex, a model can extrapolate linearly across the hull boundary without ever recombining element-concentration pairs in a novel way; such behavior satisfies geometric extrapolation yet fails Definition 5. By subordinating extrapolation to the stricter compositional criteria, the framework prevents over-optimistic claims based solely on hull-distance metrics.

Taken together, these relations demonstrate that compositional generalization for MPEAs is neither identical to systematicity, nor reducible to compositionality in the linguistic sense, nor synonymous with generic OOD or extrapolation. It occupies a distinct conceptual niche that respects the continuous, high-dimensional, and physics-governed character of MPEA composition spaces while still demanding structured recombination. Only by maintaining these distinctions can the community evaluate GNNs against a shared, non-vague standard.

Implications for GNN Design and Evaluation

Adoption of the proposed definitional framework directly reshapes both the design of graph neural networks and the protocols used to evaluate them in MPEA property prediction. The implications are concrete, actionable, and extend

beyond mere terminology.

For GNN design, four architectural priorities emerge. First, GNNs must represent elements and concentrations in a disentangled manner rather than collapsing them into a single node embedding [28]. Current message-passing schemes [15, 25] often fuse elemental identity with concentration early in the forward pass; the framework requires separate channels or conditional embeddings so that recombination can occur at higher layers. Second, message-passing layers must support explicit compositional operators—mechanisms that combine neighbor information in a manner sensitive to both elemental type and concentration magnitude. Without such operators, the network cannot satisfy Criterion 1 (novel joint occurrence). Third, positional encodings or attention mechanisms conditioned on continuous concentration variables become essential. The simplex geometry is not Euclidean; standard graph positional encodings designed for discrete graphs are insufficient. Fourth, permutation invariance over element ordering must be enforced as a baseline (Criterion 4), but it is no longer viewed as an endpoint; additional symmetry-breaking regularizers may be needed during training to encourage compositional rather than rote invariance. These design directives build directly on the foundations laid by Chen *et al.* [15] and Xie and Grossman [25] yet push beyond them toward architectures explicitly engineered for recombination.

For evaluation, the framework mandates a shift from conventional splits to compositionally structured test protocols. Test sets must include novel joint occurrences of element-concentration pairs that satisfy both Criterion 1 and Criterion 2; random or k-fold splits no longer suffice. Error must be reported separately for interpolation regions (inside the convex hull) and true compositional-extrapolation regions (outside the hull with novel pairs). Degradation curves should be plotted as a function of compositional distance—defined, for example, via earth-mover's distance on the simplex or pairwise concentration co-occurrence mismatch—so that non-catastrophic degradation (Criterion 3) can be quantitatively verified. Finally, researchers must explicitly test boundary cases such as the "missing intermediate" and "concentration continuum" problems described in Section 7; models that succeed on IID validation yet collapse on these gray zones cannot claim compositional generalization.

For reporting, transparency becomes obligatory. Every manuscript must state which definition of compositional generalization is being invoked and must supply the operational criteria used to construct test sets. Claims based solely on IID performance or simple hull-distance extrapolation must be rephrased as "interpolation success" rather than "compositional generalization." Boundary cases where the model fails should be documented rather than omitted, turning potential weaknesses into opportunities for targeted improvement.

These changes collectively raise the bar for GNN research in MPEAs. What was previously an ill-defined buzzphrase becomes a measurable engineering objective. Architectural innovations that once appeared incremental—such as concentration-conditioned attention or disentangled embeddings—now acquire clear justification and evaluation criteria. Evaluation protocols that previously rewarded smooth interpolation now penalize the absence of true recombination. The net result is a more rigorous, comparable, and ultimately more trustworthy literature on machine-learning potentials for compositionally complex alloys.

Conclusion

This article resolves a central ambiguity in machine learning for MPEAs by defining the boundary of compositional generalization. Current usage conflates interpolation, extrapolation, and domain transfer, making claims difficult to compare or validate. We propose a stricter operational definition requiring simultaneous novelty in element–concentration co-occurrences, extrapolation beyond the training convex hull, and non-catastrophic error, with permutation invariance as a baseline condition.

This boundary shows that many reported successes of GNNs reflect interpolation or local-environment transfer rather than true recombinational capability. It also establishes a clear standard for future work: models must be evaluated on structured test sets that enforce both geometric and combinatorial novelty. By replacing an ill-defined term with

measurable criteria, this framework enables more rigorous evaluation, clearer reporting, and targeted architectural development. Progress in GNN-based discovery of MPEAs depends on moving from claims of generalization to demonstrable compositional reasoning under these conditions.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 31 May 2021 Revised: 07 Oct 2021 Accepted: 18 Dec 2021
Published online: 18 January 2022

References

Hertz JA. Introduction to the theory of neural computation. Boca Raton: CRC Press; 2018.

Lake BM, Ullman TD, Tenenbaum JB, Gershman SJ. Building machines that learn and think like people. Behav Brain Sci. 2017;40:e253.

Chen S, Cheng Y, Gao H. Machine learning for high-entropy alloys. In: Artificial intelligence for materials science. Cham: Springer International Publishing; 2021. p. 21-58.

Rao Z, Tung PY, Xie R, Wei Y, Zhang H, Ferrari A, et al. Machine learning-enabled high-entropy alloy discovery. Science. 2022;378(6615):78-85.

Wen C, Zhang Y, Wang C, Xue D, Bai Y, Antonov S, et al. Machine learning assisted design of high entropy alloys with desired property. Acta Mater. 2019;170:109-17.

Bartók AP, De S, Poelking C, Bernstein N, Kermode JR, Csányi G, et al. Machine learning unifies the modeling of materials and molecules. Sci Adv. 2017;3(12):e1701816.

Batra R, Song L, Ramprasad R. Emerging materials intelligence ecosystems propelled by machine learning. Nat Rev Mater. 2021;6(8):655-78.

Nguyen G, Dlugolinsky S, Bobák M, Tran V, Lopez Garcia A, Heredia I, et al. Machine learning and deep learning frameworks and libraries for large-scale data mining: A survey. Artif Intell Rev. 2019;52(1):77-124.

Miracle DB, Senkov ON. A critical review of high entropy alloys and related concepts. Acta Mater. 2017;122:448-511.

George EP, Raabe D, Ritchie RO. High-entropy alloys. Nat Rev Mater. 2019;4(8):515-34.

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.

Behler J. Four generations of high-dimensional neural network potentials. Chem Rev. 2021;121(16):10037-72.

Zhou ZH. Machine learning. Cham: Springer Nature; 2021.

Hupkes D, Dankers V, Mul M, Bruni E. Compositionality decomposed: How do neural networks generalise? J Artif Intell Res. 2020;67:757-95.

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.

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).

Schütt K, Kindermans PJ, Sauceda HE, Chmiela S, Tkatchenko A, Müller KR. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. Adv Neural Inf Process Syst. 2017;30.

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

Szabó A, Bán K, Hlinka V, Pásztor J, Lovas A. The formation and stability of bulk amorphous and high entropy alloys. Acta Mater Transylv. 2021;4(1):51-7.

Ikeda Y, Grabowski B, Körmann F. Ab initio phase stabilities and mechanical properties of multicomponent alloys: A comprehensive review for high entropy alloys and compositionally complex alloys. Mater Charact. 2019;147:464-511.

Byggmästar J, Nordlund K, Djurabekova F. Simple machine-learned interatomic potentials for complex alloys. Phys Rev Mater. 2022;6(8):083801.

Behler J. First principles neural network potentials for reactive simulations of large molecular and condensed systems. Angew Chem Int Ed. 2017;56(42):12828-40.

Pei Z, Yin J, Hawk JA, Alman DE, Gao MC. Machine-learning informed prediction of high-entropy solid solution formation: Beyond the Hume-Rothery rules. npj Comput Mater. 2020;6(1):50.

Jeong W, Lee K, Yoo D, Lee D, Han S. Toward reliable and transferable machine learning potentials: uniform training by overcoming sampling bias. J Phys Chem C. 2018;122(39):22790-5.

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.

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.

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

Seko A, Togo A, Tanaka I. Descriptors for machine learning of materials data. In: Nanoinformatics. Singapore: Springer Singapore; 2018. p. 3-23.