

ORIGINAL RESEARCH

Open access

A Framework for Assessing Compositional Generalizability in High-Entropy Alloy ML Potentials

Ali Hassan^{1*}, Noor Siddiqui¹, Bilal Khan², Sana Malik¹

Abstract

High-entropy alloys (HEAs) and multi-principal element alloys (MPEAs) occupy an enormous compositional space that conventional computational approaches cannot fully explore. A five-element system with 10 % concentration steps already contains millions of distinct nominal compositions, each further multiplied by exponentially many atomic configurations arising from configurational disorder. Machine learning (ML) interatomic potentials have been proposed as a scalable solution to accelerate property prediction and materials design in this vast space. Yet the compositional generalizability of these potentials—their capacity to make reliable predictions for compositions and local environments lying outside the training distribution—remains largely unexamined in a systematic way. Most existing benchmarks focus on interpolation within narrow ranges of ordered or equiatomic compounds, leaving critical extrapolation scenarios untested. This conceptual framework article identifies four distinct but interrelated dimensions of compositional generalizability in HEA ML potentials: (i) element extrapolation, (ii) concentration interpolation, (iii) multi-element recombination, and (iv) local environment diversity. It proposes a four-component assessment framework—training-set characterization, test-set design, dimension-specific generalization metrics, and a structured validation protocol—that enables researchers to quantify generalization gaps without relying on performance numbers or simulation results. For each dimension, explicit validation strategies and success criteria are defined, grounded in the literature on special quasirandom structures, graph-network representations, and equivariant architectures. The framework is deliberately conceptual, emphasizing definitions, relationships among components, assessment criteria, and diagnostic reasoning rather than empirical data. By adopting this framework, the community can move beyond ad-hoc testing and develop ML potentials that truly generalize across the high-dimensional alloy landscape. The implications extend to more reliable high-throughput screening, accelerated discovery of novel HEAs, and clearer guidance for training-set design and model architecture choices. Ultimately, systematic assessment of compositional generalizability will help ensure that ML potentials fulfill their promise for complex concentrated alloys.

Keywords High-entropy alloys, Multi-principal element alloys, Machine learning potentials, Configurational disorder, Special quasirandom structures, Compositional generalizability

*Correspondence:

Ali Hassan

ali.hassan@gmail.com

¹ Department of Materials Data Science and Engineering, Faculty of Engineering, Aga Khan University, Karachi, Pakistan

² Department of Computational Materials Analytics, Faculty of Technology, Qatar University, Doha, Qatar

Introduction

High-entropy alloys (HEAs) and multi-principal element alloys (MPEAs) have emerged as a new paradigm in materials design, offering exceptional mechanical, thermal, and functional properties that arise precisely from their compositional complexity [1–5]. Unlike conventional alloys built around one or two principal elements, HEAs contain three to six or more elements in near-equiatomic or non-equiatomic proportions. This multi-principal nature generates an enormous

compositional space: even a modest five-element system sampled at 10 % concentration increments yields millions of possible nominal compositions. Each composition, moreover, corresponds to an astronomical number of distinct atomic arrangements because of configurational disorder.

Machine learning interatomic potentials have been widely recognized as a powerful route to navigate this space efficiently [1, 6–10]. By learning directly from density-functional-theory data, these potentials can predict energies, forces, and derived properties orders of magnitude faster than first-principles calculations while retaining near-ab-initio accuracy within their domain of applicability. Recent advances in graph-network architectures [8] and equivariant neural networks [7] have further improved data efficiency and physical fidelity, making it feasible to train potentials on the disordered structures that characterize HEAs.

Nevertheless, a fundamental limitation persists. The compositional generalizability of HEA ML potentials—the ability of a model to produce accurate predictions for compositions and local atomic environments that differ substantially from those seen during training—is rarely assessed in a structured or reproducible manner. Existing literature reviews highlight progress in model development and applications to specific HEAs yet acknowledge that generalization across composition remains an open challenge [1]. Performance assessments of ML potentials typically focus on ordered compounds or narrow interpolation regimes, which do not capture the extrapolation demands inherent to HEA design [6, 11, 12]. Benchmarks that do include HEAs often rely on special quasirandom structures (SQS) for a single nominal composition and do not systematically probe how well the model handles new elements, new concentrations, novel element-pair combinations, or truly random (non-SQS) configurations [13, 14].

This conceptual framework article therefore introduces a systematic approach for assessing compositional generalizability in HEA ML potentials. The framework rests on four core dimensions of generalization that together span the practical needs of alloy discovery. It defines four assessment components that any research group can apply without additional experimentation or simulation, only conceptual reorganization of existing training and test data. Validation strategies tailored to each dimension provide clear success criteria expressed in terms of relative error ratios and diagnostic patterns rather than absolute metrics. By separating generalization performance along these four axes, the framework reveals specific weaknesses—whether in element coverage, concentration sampling, interaction learning, or configurational representation—and guides targeted improvements in training-set design and model architecture.

What Makes HEA Compositional Generalization Hard

Compositional generalization in HEA ML potentials is intrinsically more demanding than in conventional alloy systems for six interrelated reasons that stem from the fundamental physics and data realities of complex concentrated alloys.

First, the composition space is high-dimensional. With three to six principal elements, each concentration can vary continuously between zero and one (subject to the sum-to-one constraint). Even coarse 10 % sampling steps produce combinatorial explosion. No training set, however large, can densely cover all combinations, forcing every practical model to extrapolate in composition space [1, 11].

Second, configurational disorder multiplies the difficulty. For any fixed nominal composition, the number of possible atomic arrangements grows exponentially with system size. A model must therefore learn not only how properties depend on average composition but also how they depend on the specific local atomic environments that realize that composition. Training data typically capture only a tiny fraction of these environments, so generalization requires the model to infer unseen local neighborhoods from limited examples [13–15].

Third, local-environment coupling is pervasive. Changing the concentration of one element alters the probability of every possible nearest-neighbor and second-nearest-neighbor shell for all other elements. Consequently, a model trained on a particular set of element-pair statistics cannot simply be queried at a new concentration; the entire local-environment distribution shifts in a coupled, non-independent manner [12, 16, 17].

Fourth, property landscapes are strongly non-linear. Mechanical, thermal, and electronic properties of HEAs are not simple linear superpositions of elemental contributions. Lattice distortion, short-range ordering, and electronic hybridization create intricate, non-monotonic dependencies on both composition and local chemistry. Linear interpolation in composition space therefore provides no guarantee of accurate property prediction, and small concentration changes can produce disproportionately large property shifts [3, 18, 19].

Fifth, data scarcity is acute. Generating high-quality density-functional-theory reference data for disordered HEAs is computationally expensive, especially when many distinct configurations must be sampled to represent the same nominal composition. Most published training sets remain modest in size and are heavily biased toward equiatomic or near-equiatomic compositions and ordered or SQS supercells [6, 20, 21].

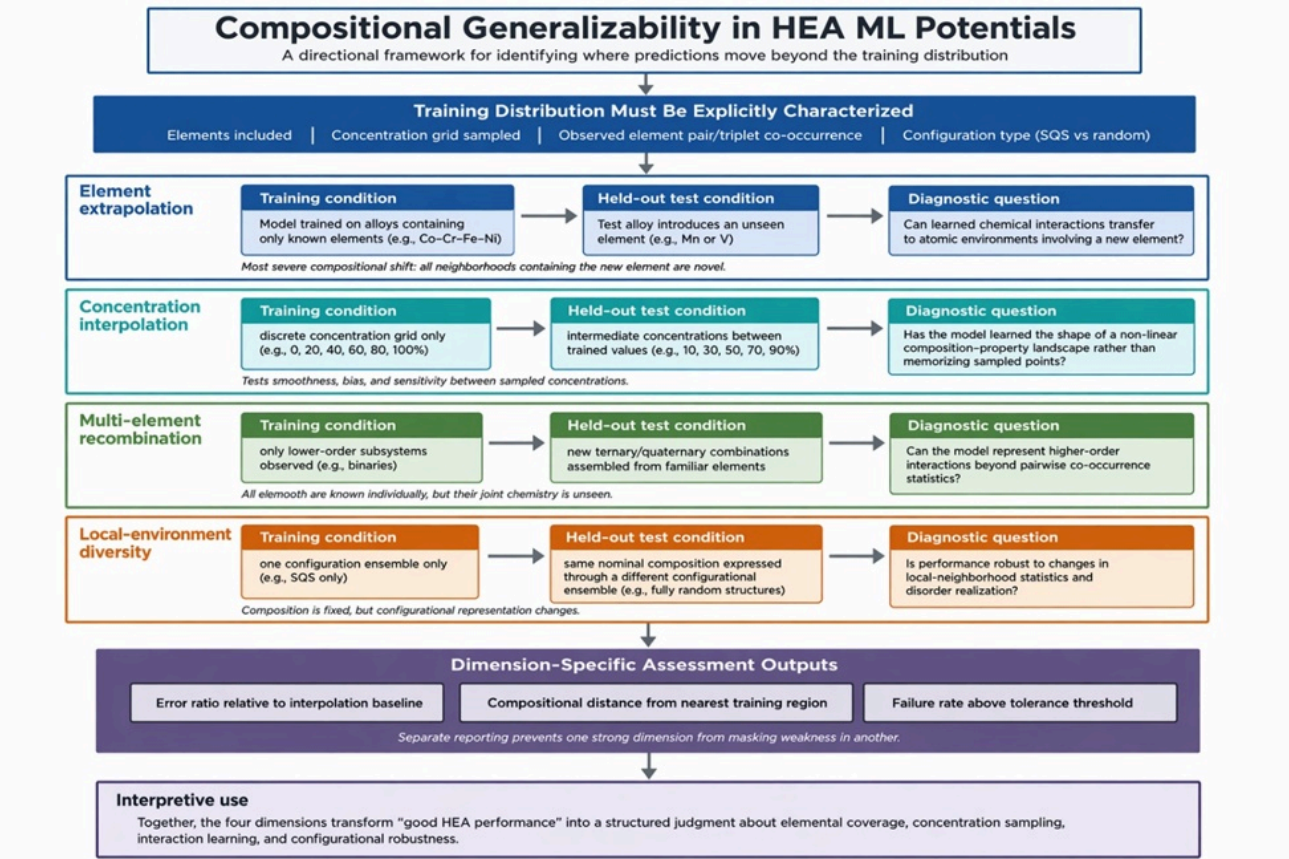
Sixth, the very tools used to generate training structures have built-in limitations. Special quasirandom structures [13] are excellent for approximating the average disordered state of a given nominal composition within a periodic supercell, yet they cannot capture the full spectrum of concentration fluctuations, long-range disorder, or rare local environments that exist in truly random solid solutions. Consequently, a potential trained exclusively on SQS data may fail when confronted with configurations drawn from a genuinely random ensemble [14, 22, 23].

These six factors interact. High dimensionality and data scarcity force sparse sampling; configurational disorder and local-environment coupling demand that the model learn subtle, many-body interactions from that sparse data; non-linear landscapes amplify any extrapolation error; and SQS approximations further restrict the diversity of local environments presented during training. The result is that standard interpolation tests—however successful—provide little insight into whether a potential will remain reliable when applied to the new compositions that alloy designers actually wish to explore. A dedicated conceptual framework for assessing generalization is therefore not merely useful but essential if ML potentials are to realize their transformative potential for HEA discovery.

Dimensions of Compositional Generalizability

Compositional generalizability in HEA ML potentials can be decomposed into four orthogonal yet interacting dimensions. Each dimension isolates a distinct type of extrapolation that arises in practical alloy-design workflows.

Figure 1 maps compositional generalizability into four diagnostically distinct dimensions—element extrapolation, concentration interpolation, multi-element recombination, and local-environment diversity—arranged as a directional assessment structure from training-domain definition to out-of-distribution evaluation.



Element extrapolation probes whether a model can infer properties for an HEA incorporating an element absent from the training distribution, as when a potential calibrated on Co–Cr–Fe–Ni is deployed on Co–Cr–Fe–Ni–Mn or Co–Cr–Fe–Ni–V. Under such conditions, all atomic environments involving the added species are both statistically unseen and chemically distinct, leaving the model without direct empirical grounding for those neighborhoods. A related challenge emerges when composition varies continuously rather than at discretely sampled points: concentration interpolation evaluates behavior at intermediate concentrations not represented during training, where non-monotonic response surfaces or abrupt transitions render simple compositional interpolation inadequate. This issue extends further when previously unobserved combinations of elements must be resolved, since training confined to binary subsystems does not guarantee that higher-order interactions in ternary or quaternary alloys are correctly encoded, thereby testing whether the learned representation captures genuine many-body effects rather than pairwise regularities. Beyond compositional considerations, robustness also depends on sensitivity to configurational disorder, as models trained on SQS supercells may not generalize to fully random arrangements of identical nominal composition, exposing limitations in representing long-range fluctuations that SQS constructions intentionally minimize [13]. Although these facets are analytically separable, their coupling in realistic settings implies that deficiencies rarely occur in isolation, complicating attribution while underscoring the need for diagnostic decomposition.

This decomposition allows researchers to pinpoint whether a model's weakness lies in elemental coverage, sampling density, interaction learning, or configurational representation.

Table 1 consolidates the four dimensions into an analytical typology by distinguishing the boundary crossed in each case, the form of novelty introduced at test time, and the specific failure logic each dimension is designed to expose.

Table 1. Analytical typology of compositional generalizability dimensions in HEA machine-learning potentials

Dimension	What is held constant	What changes at test time	Boundary being crossed	Dominant source of novelty	Main failure mode revealed	Why the dimension must be reported separately
Element extrapolation	Broad alloy family logic and nominal assessment objective	Introduction of at least one element absent from training	Elemental identity boundary	New chemical neighborhoods involving the unseen element	Element-specific memorization instead of transferable chemical representation	Failure here can be hidden if a model performs well only on familiar chemistries
Concentration interpolation	Element set remains the same	Intermediate compositions not present on the training grid	Sampling-density boundary	Shift in concentration-dependent neighborhood probabilities	Piecewise memorization of sampled grid points; systematic bias across concentration range	Strong accuracy at trained concentrations does not establish continuity between them
Multi-element recombination	Individual elements are familiar	New pair/triplet/higher-order co-occurrences appear jointly	Interaction-order boundary	Emergent many-body chemistry from unseen combinations	Inability to transfer from isolated subsystem statistics to coupled multi-element behavior	Binary success can falsely imply higher-order transferability
Local-environment diversity	Nominal composition remains fixed	Disorder realization or configuration ensemble changes	Configurational-representation boundary	Different local-neighborhood statistics despite same average composition	Overfitting to one structural generator such as SQS	Composition-level validation alone cannot detect sensitivity to disorder realization

A Framework for Assessment

The proposed framework comprises four interlocking components that together transform the abstract concept of compositional generalizability into an actionable evaluation protocol.

Rigorous evaluation begins with precise delineation of the training distribution, as meaningful claims of generalization depend on knowing the compositional and configurational scope the model has encountered. This entails explicit reporting of the elemental set, concentration ranges and their discretization, the actual co-occurrence of element pairs or higher-order combinations within training structures, and the number and type of configurations sampled per composition, including the distinction between SQS and fully random realizations. Such characterization establishes the effective support of the learned representation and, in turn, defines where extrapolative demands arise; absent this information, the boundary between interpolation and genuine out-of-distribution prediction remains indeterminate.

Building on this foundation, evaluation must isolate distinct sources of distributional shift through deliberately constructed hold-out sets that probe unseen elements, intermediate concentrations, novel multi-element combinations, and alternative configurational statistics. The key requirement is independence from the training workflow, since any procedural overlap risks subtle leakage that would inflate apparent robustness. Interpreting performance across these regimes further necessitates dimension-specific metrics rather than aggregate scores: an error ratio benchmarking degradation relative to in-distribution interpolation, an assessment of compositional distance in element–concentration space, and a failure rate defined by exceedance of a property-dependent tolerance threshold [24, 25]. Disaggregating these quantities prevents the obscuring of systematic weaknesses that would otherwise be averaged out.

A consistent validation protocol operationalizes these principles by coupling controlled training on a well-defined base set with targeted evaluation across the constructed test domains, followed by comparison against an interpolation baseline drawn from within the training support. The resulting generalization gap, expressed as the degradation in accuracy under these shifts, provides a model-agnostic measure of robustness. Crucially, this procedure abstracts from specific architectures or target properties, enabling reproducible assessment and facilitating cumulative progress across studies.

When applied consistently, the four components convert anecdotal observations about “good performance on HEAs” into precise, diagnosable statements about where and how generalization succeeds or fails. The framework therefore supplies the conceptual scaffolding needed to move HEA ML potential development from case-by-case validation to systematic, reproducible assessment.

Validation Strategies for Each Dimension

Each dimension is accompanied by a concrete validation strategy that operationalizes the framework while remaining purely conceptual.

Strategy for element extrapolation begins with a base quaternary system such as Co–Cr–Fe–Ni. The potential is trained on structures that exclude one target element (for example, no Mn). It is then tested on structures of the quinary Co–Cr–Fe–Ni–Mn at the same nominal concentrations used in training. The success criterion is that the error ratio remains below 2× relative to the interpolation baseline. This threshold reflects the expectation that introducing an entirely new element is the most severe test; any larger degradation signals insufficient transferability of learned chemical interactions.

Strategy for concentration interpolation trains the model exclusively at discrete concentration steps (0 %, 20 %, 40 %, 60 %, 80 %, 100 % for each element, with other elements adjusted to maintain the total). Testing occurs only at the exact midpoints (10 %, 30 %, 50 %, 70 %, 90 %). Success is defined by an error ratio below 1.5× the interpolation baseline and the absence of systematic bias (errors not consistently positive or negative). This criterion ensures the model has captured the non-linear shape of the property landscape rather than merely memorizing the trained points.

Strategy for multi-element recombination trains the potential only on binary subsystems (all possible pairs among the constituent elements). Testing proceeds on ternary and quaternary compositions formed by recombining those binaries. The success criterion is that the error on the recombined higher-order systems remains correlated with the binary error; no abrupt jump in error ratio should occur. This confirms that the model has learned transferable many-body interactions rather than isolated pairwise statistics.

Strategy for local environment diversity trains exclusively on SQS supercells and tests on random configurations of identical nominal compositions (or the reverse). Because true configurational disorder is expected to be the hardest aspect, the tolerance is relaxed to an error ratio below 3× the baseline. The criterion acknowledges that SQS and random ensembles differ in subtle long-range statistics; a modest degradation is acceptable provided the model still captures average properties reliably.

These four strategies are modular. A research group can apply any subset or all of them depending on the intended use case of the potential. Because each strategy maps directly onto one dimension and one component of the framework, results are immediately interpretable: a failure in element extrapolation points to insufficient elemental diversity in training, a failure in concentration interpolation signals inadequate sampling of the property landscape, and so on. Collectively the strategies transform the abstract framework into a practical checklist that any HEA ML potential developer can follow.

Relation to Existing Benchmarks

Existing materials-property benchmarks such as the Materials Project, OQMD, and AFLOW have been instrumental in advancing ML potentials for crystalline solids [6, 8, 11]. These databases predominantly contain ordered compounds with well-defined stoichiometries and minimal configurational disorder. While they provide excellent testbeds for interpolation within narrow composition ranges, they do not represent the high-dimensional, disordered nature of HEAs and MPEAs. Consequently, a potential that performs well on these ordered benchmarks may still fail when confronted with the combinatorial complexity of multi-principal element systems [1, 12].

Benchmarking studies of ML interatomic potentials have similarly focused on performance metrics for elemental metals, binary alloys, or simple ternary compounds [6, 20, 26, 27]. For instance, large-scale assessments compare energy and force errors across thousands of structures drawn from equilibrium lattices or defect-containing supercells, yet they rarely include the vast compositional spaces characteristic of HEAs [6, 11]. Even when HEAs appear in such evaluations, testing is typically limited to a handful of equiatomic compositions generated with SQS and evaluated only for interpolation within that narrow slice of composition space [14, 15]. No systematic probe of element extrapolation, concentration interpolation, multi-element recombination, or local-environment diversity is performed.

The gap is particularly evident in reviews of ML for HEAs. Progress is documented in model architectures and applications to specific alloy families, but compositional generalization is acknowledged as an unresolved challenge rather than a quantified property [1, 28]. Transfer-learning studies across alloy families exist [7, 8], yet they do not formalize the four dimensions introduced here. Similarly, work on SQS for HEAs emphasizes their utility for approximating average disorder within a fixed composition but does not address how well potentials generalize when the underlying configurational ensemble changes [13, 14, 22].

This framework directly fills the missing structure. It provides HEA-specific test-set designs that can be layered onto any existing benchmark suite without requiring new data generation. By insisting on separate reporting for each of the four dimensions, the framework prevents the common practice of declaring “good performance on HEAs” after testing only equiatomic SQS structures. It therefore complements rather than replaces current benchmarks, supplying the additional layer of scrutiny needed for complex concentrated alloys. Adoption of the framework will enable future benchmark curators to include standardized HEA generalization suites, transforming ad-hoc validation into community-wide, reproducible assessment [1, 6, 11].

Interpreting Assessment Results

Interpreting results from the framework requires attention to the pattern of generalization gaps across the four dimensions rather than any single scalar value. Good compositional generalization is indicated when error ratios remain close to 1.0 for all dimensions, failure rates stay below 10 %, and no systematic bias (consistent over- or under-prediction) appears. Such outcomes demonstrate that the model has learned transferable representations of local chemistry, composition-property relationships, and configurational statistics.

Poor generalization manifests differently in each dimension. An error ratio exceeding 5× for element extrapolation signals that the model cannot transfer learned interactions to unseen elements; the potential is effectively memorizing element-specific features rather than general chemical principles [7, 12]. Systematic over- or under-prediction in concentration

interpolation reveals that the property landscape is strongly non-linear and undersampled; the model is essentially performing piecewise interpolation instead of capturing the true functional form [3, 18]. Abrupt jumps in error for multi-element recombination indicate insufficient learning of higher-order interactions; the potential treats elements as independent rather than coupled [15, 16]. Finally, a large gap in local-environment diversity (error ratio > 3×) shows that SQS-based training has not equipped the model to handle realistic concentration fluctuations and long-range disorder [13, 14].

Table 2 converts the framework from a reporting device into a diagnostic tool by mapping distinct error patterns to likely structural causes, model-development responses, and scope-of-use decisions.

Table 2. Diagnostic interpretation matrix linking observed validation patterns to probable causes and corrective design actions

Observed validation pattern	Most plausible structural interpretation	What it suggests about the training set	What it suggests about the model representation	Highest-priority corrective action	Deployment implication
High error ratio mainly in element extrapolation; other dimensions relatively stable	Learned representation is chemistry-specific rather than transferable across elemental identities	Elemental coverage is too narrow for intended discovery space	Latent features do not encode enough chemically transferable structure	Add chemically diverse held-out/near-held-out elements and retrain with broader elemental span	Safe only for alloys composed of already represented elements
High error ratio mainly in concentration interpolation with directional bias	Model has not captured the non-linear composition–property landscape	Concentration grid is too sparse or unevenly distributed	Model behaves like a local interpolator between sampled points	Densify intermediate concentrations, especially near steep property regions	Use cautiously between sampled compositions; avoid continuous screening claims
Abrupt degradation in multi-element recombination after good binary performance	Pairwise familiarity does not translate into higher-order interaction learning	Training data lack sufficient ternary/quaternary co-occurrence structure	Representation undercaptures many-body coupling	Introduce explicit higher-order combinations in training and test suites	Do not deploy on novel multi-element blends assembled from validated binaries alone
Large gap only in local-environment diversity	Model is sensitive to the structural generator rather than the nominal chemistry	Configuration ensemble is too narrow, often over-reliant on SQS	Learned neighborhood statistics are brittle to disorder realization	Mix SQS and random configurations within training and validation	Suitable only for configuration classes similar to those used in training
Simultaneous failure in element extrapolation and multi-element recombination	New elements are entering through unseen coupled chemistries, creating compounded novelty	Both elemental diversity and co-occurrence diversity are insufficient	Architecture may not separate elemental transfer from interaction transfer	Expand chemistry coverage first, then add structured higher-order combinations	Out-of-domain use is high risk even when single-dimension tests look acceptable

Uniformly low generalization gaps across all four dimensions	Representation is robust across elemental, compositional, interactional, and configurational shifts	Training coverage is balanced rather than concentrated in one region	Model likely captures transferable local chemistry and disorder structure	Preserve reporting discipline; extend only with property-specific validation	Appropriate candidate for broader HEA screening within the validated alloy family
--	---	--	---	--	---

These patterns carry immediate diagnostic value. Failure confined to element extrapolation points to the need for broader elemental diversity in the training set. A concentration-interpolation shortfall suggests denser sampling of the composition grid or inclusion of more intermediate structures. Multi-element recombination problems call for explicit training on higher-order combinations or architectures that better capture many-body effects [8, 28]. Local-environment diversity failures indicate that SQS alone is insufficient; random configurations must be added to the training workflow.

Because the framework reports metrics separately, researchers can prioritize fixes. A model that excels on three dimensions but collapses on the fourth can still be used confidently within its validated scope. Over time, consistent application of the framework will reveal which architectures—graph networks [8], equivariant models [7], or hybrid approaches—naturally exhibit superior generalization, guiding the community toward more robust HEA potentials.

Limitations of The Framework

While the framework supplies a clear conceptual scaffold, it is subject to five practical limitations that users should acknowledge.

Despite its conceptual economy, comprehensive assessment imposes a tangible computational burden, as constructing independent test domains often necessitates generating and labeling additional structures when such data are absent. In high-dimensional compositional spaces, this may translate into hundreds of DFT evaluations, placing practical constraints on resource-limited settings [6, 20]. A related consideration concerns how success is defined: the use of relative error ratios, rather than fixed thresholds, reflects the context-dependent nature of acceptable performance, which varies with material system, target property, and application-specific tolerances. Under these conditions, a model adequate for elastic constants may remain insufficient for vacancy formation energies within the same alloy family, requiring careful calibration of criteria [1, 3].

Beyond these practicalities, interpretability is complicated by intrinsic coupling among generalization regimes, since extrapolation to unseen elements simultaneously introduces novel local environments and interaction patterns. Although separate reporting attenuates this effect, interdependence persists and can obscure attribution of failure modes [12, 16]. This limitation is further compounded by the restriction to zero-temperature compositional and configurational variability, leaving thermally driven phenomena—such as vibrational contributions, phase stability, and temperature-dependent ordering—outside the current scope, and thus pointing to an unresolved extension of the framework [18, 29]. Moreover, generalization remains inherently property-specific: predictive fidelity may be retained for total energies yet degrade for forces or elastic responses, reflecting the distinct sensitivities of these observables to local structural features. Consequently, robust evaluation requires property-resolved analysis rather than implicit assumptions of transferability [6, 26].

Despite these limitations, the framework remains a substantial advance over the absence of any structured assessment. Its modular design allows incremental adoption—starting with one or two dimensions—and future extensions can address the gaps identified here.

Implications for Hea MI Potential Development

The framework carries direct implications for three stakeholder groups: model developers, benchmark designers, and end users.

For model developers the primary recommendation is to report compositional generalization metrics alongside standard accuracy benchmarks. Rather than stating that a potential “works well for HEAs,” developers should publish error ratios, failure rates, and generalization gaps for each of the four dimensions. This transparency will accelerate progress by highlighting which architectures—equivariant graph networks [7], universal graph frameworks [8], or hybrid descriptors—naturally generalize better. Developers should also redesign training sets to maximize coverage of element pairs and random configurations rather than relying solely on equiatomic SQS structures [13, 14].

Benchmark designers can use the framework to create standardized HEA generalization suites. By publishing pre-defined training and test splits that embody the four dimensions, the community can establish common reference sets analogous to those already used for ordered compounds. Inclusion of both SQS and random configurations will become a default requirement, elevating the quality of future ML potential evaluations [6, 11, 28].

For practitioners selecting a potential for alloy design campaigns, the framework supplies a simple decision checklist. Before deploying a model on novel compositions, practitioners should verify that the relevant dimensions have been tested and that generalization gaps remain within acceptable bounds. A potential that has not been subjected to this assessment should be treated with caution, especially when the target composition lies far from the training distribution [1, 15].

Collectively these implications shift the culture of HEA ML potential development from case-by-case validation to systematic, reproducible assessment. Over time the framework will reduce wasted effort on models that appear accurate in narrow tests yet fail in real design scenarios, accelerate discovery of truly generalizable potentials, and increase confidence in ML-driven HEA screening.

Conclusion

High-entropy alloys and multi-principal element alloys present an enormous compositional space whose exploration demands ML potentials capable of reliable extrapolation. Yet current benchmarks and validation practices rarely assess the very generalizability that is most critical for practical alloy design. This conceptual framework addresses that gap by defining four distinct dimensions of compositional generalizability—element extrapolation, concentration interpolation, multi-element recombination, and local environment diversity—and by organizing assessment around four interlocking components: training-set characterization, test-set design, dimension-specific metrics, and a structured validation protocol.

Explicit validation strategies for each dimension translate the framework into an actionable checklist that any research group can adopt without new experimentation. By reporting generalization gaps separately rather than as a single performance number, the framework reveals precise model weaknesses and guides targeted improvements in training data and architecture. Its relation to existing benchmarks is complementary: it adds the HEA-specific layer that ordered-compound suites lack. Interpretation guidelines clarify what constitutes success or failure, while acknowledged limitations ensure realistic application.

The implications are immediate and far-reaching. Model developers gain diagnostic clarity, benchmark curators obtain a blueprint for HEA-specific suites, and practitioners receive a defensible basis for model selection. Widespread adoption of this framework will elevate compositional generalization from an afterthought to a core reporting requirement, ensuring that ML potentials fulfill their promise for the discovery and design of next-generation complex concentrated alloys. The community is therefore urged to integrate these assessment practices into standard workflows so that future HEA ML potentials are not only accurate but demonstrably generalizable across the vast alloy landscape.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 06 Nov 2022

Revised: 19 Feb 2023

Accepted: 31 May 2023

Published online: 18 July 2023

Rights and permissions

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

References

Liu X, Zhang J, Pei Z. Machine learning for high-entropy alloys: Progress, challenges and opportunities. *Prog Mater Sci.* 2023;131:101018.

<https://doi.org/10.1016/j.pmatsci.2022.101018>.

Byggmästar J, Nordlund K, Djurabekova F. Modeling refractory high-entropy alloys with efficient machine-learned interatomic potentials: Defects and segregation. *Phys Rev B.* 2021;104(10):104101.

<https://doi.org/10.1103/PhysRevB.104.104101>.

Zhou Y, Srinivasan P, Körmann F, Grabowski B, Smith R, Goddard P, et al. Thermodynamics up to the melting point in a TaVCrW high entropy alloy: Systematic ab initio study aided by machine learning potentials. *Phys Rev B.* 2022;105(21):214302.

<https://doi.org/10.1103/PhysRevB.105.214302>.

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

<https://doi.org/10.1038/s41578-019-0121-4>.

Murty BS, Yeh JW, Ranganathan S, Bhattacharjee PP. High-entropy alloys. 2nd ed. Amsterdam: Elsevier; 2019. 388 p.

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

Batzner S, Musaelian A, Sun L, Geiger M, Mailoa JP, Kornbluth M, et al. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat Commun*. 2022;13(1):2453.
<https://doi.org/10.1038/s41467-022-29939-5>.

Chen C, Ye W, Zuo Y, Zheng C, Ong SP. Graph networks as a universal machine learning framework for molecules and crystals. *Chem Mater*. 2019;31(9):3564-72.
<https://doi.org/10.1021/acs.chemmater.9b01294>.

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

Friederich P, Häse F, Proppe J, Aspuru-Guzik A. Machine-learned potentials for next-generation matter simulations. *Nat Mater*. 2021;20(6):750-61.
<https://doi.org/10.1038/s41563-020-0777-6>.

Hodapp M, Shapeev A. Machine-learning potentials enable predictive and tractable high-throughput screening of random alloys. *Phys Rev Mater*. 2021;5(11):113802.
<https://doi.org/10.1103/PhysRevMaterials.5.113802>.

Gubaev K, Ikeda Y, Tasnádi F, Neugebauer J, Shapeev AV, Grabowski B, et al. Finite-temperature interplay of structural stability, chemical complexity, and elastic properties of bcc multicomponent alloys from ab initio trained machine-learning potentials. *Phys Rev Mater*. 2021;5(7):073801.
<https://doi.org/10.1103/PhysRevMaterials.5.073801>.

Zhang J, Zhang YP, Su CM. First-principles study of FeNi_{1-x}Crx (0 ≤ x ≤ 1) disordered alloys from special quasirandom structures. *Calphad*. 2020;71:102007.
<https://doi.org/10.1016/j.calphad.2020.102007>.

Wang S, Xiong J, Li D, Zeng Q, Xiong M, Chai X. Comparison of two calculation models for high entropy alloys: Virtual crystal approximation and special quasi-random structure. *Mater Lett*. 2021;282:128754.
<https://doi.org/10.1016/j.matlet.2020.128754>.

Santos-Florez PA, Dai SC, Yao Y, Yanxon H, Li L, Wang YJ, et al. Short-range order and its impacts on the BCC MoNbTaW multi-principal element alloy by the machine-learning potential. *Acta Mater*. 2023;255:119041.
<https://doi.org/10.1016/j.actamat.2023.119041>.

Jafary-Zadeh M, Khoo KH, Laskowski R, Branicio PS, Shapeev AV. Applying a machine learning interatomic potential to unravel the effects of local lattice distortion on the elastic properties of multi-principal element alloys. *J Alloys Compd*. 2019;803:1054-62.
<https://doi.org/10.1016/j.jallcom.2019.06.318>.

Xing B, Wang X, Bowman WJ, Cao P. Short-range order localizing diffusion in multi-principal element alloys. *Scr Mater*. 2022;210:114450.
<https://doi.org/10.1016/j.scriptamat.2021.114450>.

Vazquez G, Singh P, Saucedo D, Couperthwaite R, Britt N, Youssef K, et al. Efficient machine-learning model for fast assessment of elastic properties of high-entropy alloys. *Acta Mater*. 2022;232:117924.
<https://doi.org/10.1016/j.actamat.2022.117924>.

Kim G, Diao H, Lee C, Samaei AT, Phan T, de Jong M, et al. First-principles and machine learning predictions of elasticity in severely lattice-distorted high-entropy alloys with experimental validation. *Acta Mater*. 2019;181:124-38.

<https://doi.org/10.1016/j.actamat.2019.09.026>.

Sorkin V, Chen S, Tan TL, Yu ZG, Man M, Zhang YW. First-principles-based high-throughput computation for high entropy alloys with short range order. *J Alloys Compd.* 2021;882:160776.

<https://doi.org/10.1016/j.jallcom.2021.160776>.

Wang J, Kwon H, Kim HS, Lee BJ. A neural network model for high entropy alloy design. *NPJ Comput Mater.* 2023;9(1):60.

<https://doi.org/10.1038/s41524-023-01010-x>.

Wong ZM, Tan TL, Yang SW, Xu GQ. Optimizing special quasirandom structure (SQS) models for accurate functional property prediction in disordered 2D alloys. *J Phys Condens Matter.* 2018;30(48):485402.

<https://doi.org/10.1088/1361-648X/aae764>.

Coregliano LN, Razborov AA. Natural quasirandomness properties. *Random Struct Algorithms.* 2023;63(3):624-88.

<https://doi.org/10.1002/rsa.21153>.

Zhou ZH. Machine learning. Singapore: Springer Singapore; 2021.

<https://doi.org/10.1007/978-981-15-1967-3>.

Alpaydin E. Machine learning. Rev and updated ed. Cambridge (MA): MIT Press; 2021. 280 p.

<https://doi.org/10.7551/mitpress/13811.001.0001>.

Li XG, Chen C, Zheng H, Zuo Y, Ong SP. Complex strengthening mechanisms in the NbMoTaW multi-principal element alloy. *NPJ Comput Mater.* 2020;6(1):70.

<https://doi.org/10.1038/s41524-020-0339-0>.

Anstine DM, Isayev O. Machine learning interatomic potentials and long-range physics. *J Phys Chem A.* 2023;127(11):2417-31.

<https://doi.org/10.1021/acs.jpca.2c06778>.

Anand A, Liu SJ, Singh CV. Recent advances in computational design of structural multi-principal element alloys. *iScience.* 2023;26(10):107751.

<https://doi.org/10.1016/j.isci.2023.107751>.

Huang X, Liu L, Duan X, Liao W, Huang J, Sun H, et al. Atomistic simulation of chemical short-range order in HfNbTaZr high entropy alloy based on a newly-developed interatomic potential. *Mater Des.* 2021;202:109560.

<https://doi.org/10.1016/j.matdes.2021.109560>.