

ORIGINAL RESEARCH

Open access

Why Most ML Potential Benchmarks Overestimate Real-World Accuracy on Disordered Alloys

Jinwoo Park^{1*}, Minji Kim¹, Seung Lee²

Abstract

Machine learning interatomic potentials have become central tools in computational materials engineering, promising accurate and scalable predictions of alloy properties. Benchmarks in the field routinely report impressive accuracies, such as energy mean absolute errors below 20 meV/atom and low force root-mean-square errors, leading many papers to conclude that their models generalize well to alloy systems. Yet when these same potentials are deployed on real-world disordered alloys—such as binary solid solutions or multi-principal-element high-entropy alloys—the accuracy collapses, often by an order of magnitude. The root cause lies in the benchmarking pipeline itself: training and testing occur almost exclusively on ordered crystal structures drawn from databases of intermetallic compounds, while disordered configurations, which dominate practical applications, are never evaluated. This critical critique identifies four primary sources of systematic overestimation. First, the near-universal reliance on ordered training and test sets creates an illusion of generalization that fails the moment configurational disorder is introduced. Second, random or composition-based splits on ordered data allow models to memorize recurring structural motifs rather than learn transferable physics. Third, even when special quasirandom structures are employed as proxies for disorder, their periodic nature and limited sampling of local environments produce a misleading “SQS mirage.” Fourth, energy-only metrics mask severe force-prediction failures that render molecular-dynamics simulations unstable in disordered systems. These benchmarking flaws have concrete consequences: overconfidence among practitioners, misallocation of experimental resources, and slowed progress in high-entropy alloy design. By drawing on the literature, this article demonstrates that the ordered-to-disordered gap is not an edge case but a fundamental and under-reported limitation. It proposes concrete reforms—mandatory disordered test sets, configurational sampling protocols, and separate reporting of ordered versus disordered performance—to restore credibility to ML potential benchmarks. Until these changes are adopted, claims of “alloy-ready” potentials should be treated with skepticism. The field must move beyond ordered-crystal comfort zones if machine-learned potentials are to deliver on their promise for the compositionally complex materials that define modern materials science.

Keywords High-entropy alloys, Machine learning potentials, Disordered alloys, Benchmarking practices, Configurational disorder, Special quasirandom structures

*Correspondence:

Jinwoo Park
jinwoo.park@outlook.com

¹ Department of Computational Materials Systems, College of Engineering, Seoul National University, Seoul, South Korea

² Department of Materials Data Engineering, Faculty of Engineering, KAIST, Daejeon, South Korea

Introduction

A typical machine learning potential paper begins with a familiar narrative. Researchers assemble a training set from ordered intermetallic compounds, often pulled from established databases. They train a model—be it a deep potential, an equivariant graph neural network, or a moment-tensor potential—then evaluate it on a held-out set of similarly ordered

structures. The reported energy error is low, perhaps 10–30 meV/atom, and force errors look equally respectable. The conclusion follows almost automatically: the potential “accurately models alloys” and is ready for broader use.

Practitioners then attempt to apply the same potential to a high-entropy alloy solid solution, where five or more elements occupy lattice sites randomly. The result is sobering. Energy errors balloon to 100 meV/atom or higher, forces become unreliable, and molecular-dynamics trajectories drift into unphysical regions. What went wrong? The benchmark never tested the regime that matters most in real materials engineering.

This critique argues that the majority of ML potential benchmarks systematically overestimate real-world accuracy on disordered alloys. The overestimation is not accidental but structural: it arises from the near-exclusive use of ordered training and test data, from flawed assumptions about generalization across configurational spaces, and from evaluation protocols that ignore the very physics that distinguishes disordered systems. By examining the literature published, we expose these hidden weaknesses and demonstrate their consequences for the field.

The problem is especially acute for disordered alloys. High-entropy alloys and conventional solid solutions constitute a growing fraction of materials research because their tunable compositions offer unprecedented property spaces. Yet the very feature that makes them valuable—configurational disorder—remains invisible in standard benchmarks. Zuo *et al.* benchmarked multiple ML potentials across a range of materials but relied entirely on ordered test configurations [1]. Zhang *et al.* introduced the deep potential framework with impressive accuracy claims derived from ordered crystal benchmarks [2]. Batzner *et al.* demonstrated the power of equivariant networks on similar ordered datasets [3]. In each case, the reported performance creates an impression of broad alloy applicability that collapses under disordered conditions.

Figure 1 maps the directional logic of benchmark overestimation, showing how ordered-only data pipelines generate specific evaluative distortions that culminate in misleading claims of alloy readiness.

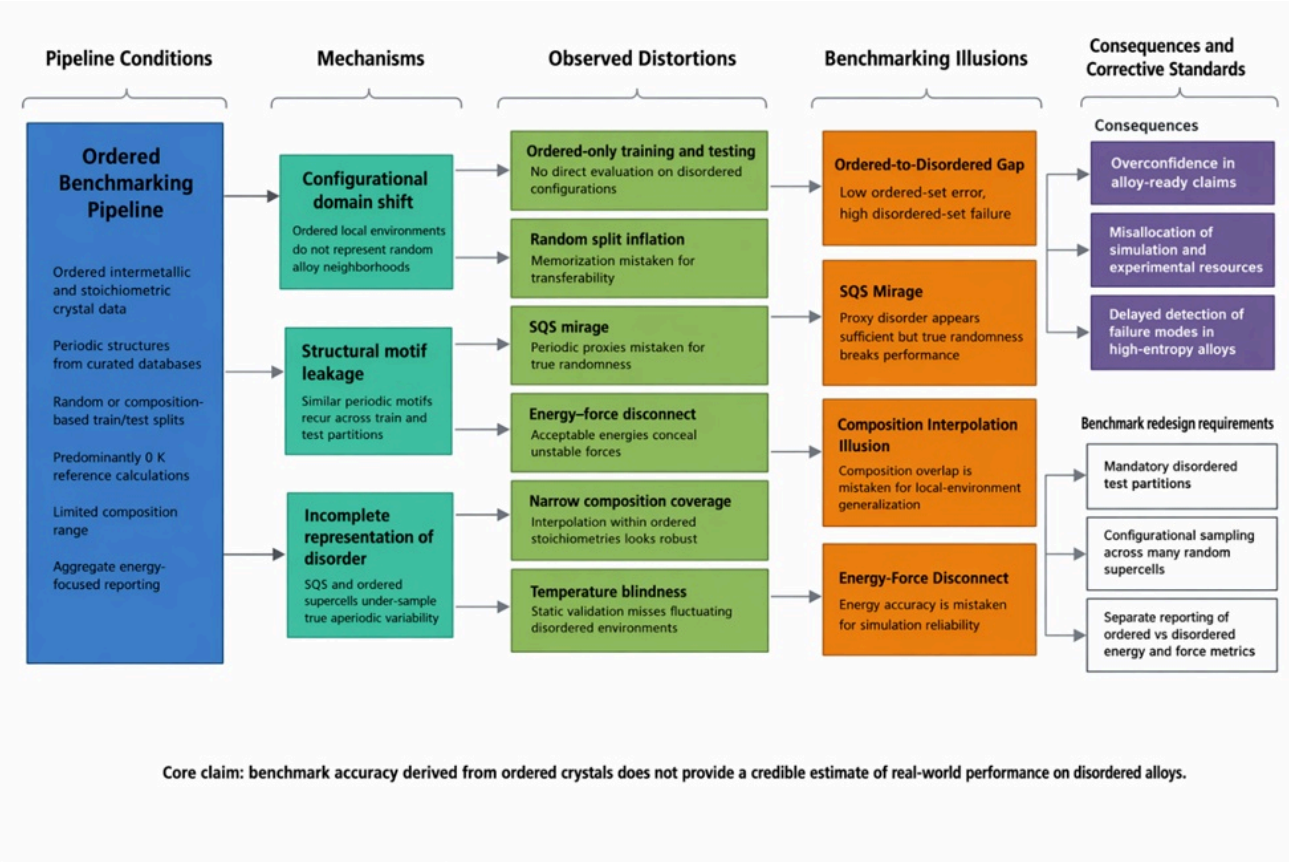


Figure 1. Hierarchical pathway showing how ordered-crystal benchmarking pipelines systematically overestimate ML-potential accuracy for disordered alloys

The Standard Benchmarking Practice

Most papers in the ML potential literature follow a well-worn template. Training data are drawn overwhelmingly from ordered crystal structures—binary and ternary intermetallics, stoichiometric compounds, and simple solid solutions with periodic atomic arrangements. These structures are sourced from databases such as the Materials Project or the Open Quantum Materials Database. A random split or a composition-based split then divides the data into training and test sets. The model is trained to reproduce energies and forces obtained from density-functional theory calculations performed at zero temperature. Performance is reported as energy mean absolute error and force root-mean-square error on the held-out ordered structures. The paper typically concludes that the potential generalizes successfully to alloy systems.

This workflow appears in the majority of studies. Zuo *et al.* evaluated a wide range of ML potentials using precisely this ordered-to-ordered protocol and presented the results as evidence of broad applicability [1]. Zhang *et al.* demonstrated the scalability of deep potentials on ordered crystal benchmarks before claiming utility for alloy modeling [2]. Bätzner *et al.* showcased their equivariant graph neural networks on ordered test sets drawn from similar sources [3]. Bartók *et al.* unified materials and molecular modeling using Gaussian approximation potentials trained and tested on ordered configurations [4]. Even reviews focused on high-entropy alloys, such as the work by Liu *et al.*, survey ML approaches that inherit the same ordered-data paradigm [5].

What is conspicuously absent is any evaluation on disordered configurations. Solid solutions, random alloys, and high-entropy alloys—with their enormous configurational spaces—are rarely included in test sets. When disordered structures appear at all, they are treated as an afterthought rather than a core requirement. Rosenbrock *et al.* developed potentials for alloy phase diagrams but still anchored their validation in ordered reference structures [6]. Byggmästar *et al.* modeled refractory high-entropy alloys yet evaluated performance primarily on ordered or partially ordered supercells [7].

The implicit assumption underlying this practice is that accuracy on ordered structures is a reliable proxy for accuracy on disordered ones. The community appears to believe that if a model performs well on periodic crystals, it will automatically handle the local-environment variability of random alloys. This assumption is false. Ordered crystals possess repeating, identical local environments; disordered alloys do not. Random train-test splits on ordered data further exacerbate the problem by allowing structural motifs from the training set to leak into the test set. The model learns to recognize familiar patterns rather than to extrapolate the underlying physics.

The consequences of this standard practice are far-reaching. Papers routinely claim “state-of-the-art performance on alloys” without ever having exposed their models to the disordered configurations that define real-world alloy applications. The literature therefore creates a misleading picture of progress. Practitioners who trust these benchmarks discover the overestimation only after investing time and resources in downstream simulations that fail. Until the community abandons the ordered-only paradigm, ML potential benchmarks will continue to overestimate real-world accuracy on the very materials they are most needed for.

Why Disordered Alloys Are Different

Disordered alloys differ from ordered intermetallics in ways that standard benchmarks never probe. The most obvious distinction is configurational disorder. In an ordered crystal every atom sits in an identical local environment; its neighbors are always the same species at the same distances. In a disordered alloy—whether a binary solid solution or a high-entropy alloy—each atom experiences a unique combination of neighboring elements. Kostiuchenko *et al.* highlighted how this configurational disorder alters phase stability and mechanical properties in high-entropy systems [8]. The

potential must therefore generalize across thousands of distinct local environments rather than a handful of repeating ones.

The configurational space of disordered alloys is exponentially larger than that of ordered compounds. A binary alloy on a lattice of N sites admits 2^N possible arrangements. Training on ordered structures samples essentially none of this space. Pei *et al.* noted the challenge of predicting solid-solution formation beyond simple rules, underscoring how limited sampling of configurations undermines model reliability [9]. Jafary-Zadeh *et al.* similarly observed that local lattice distortions in multi-principal element alloys create environments unseen in ordered training data [10].

Local environment variability introduces physics that ordered benchmarks ignore. In a high-entropy alloy such as CoCrFeNi, every atom is surrounded by a near-random distribution of the five principal elements. Huang *et al.* demonstrated through atomistic simulations that chemical short-range order and local distortions dominate the behavior of such systems [11]. Ordered crystals, by contrast, exhibit strong chemical ordering or stoichiometric regularity. The potential energy surface therefore changes character: disordered systems are governed by random mixing and configurational entropy, while ordered systems are stabilized by specific bonding motifs.

Even when researchers attempt to approximate disorder using special quasirandom structures, the approximation remains imperfect. Zhou *et al.* employed SQS to study thermodynamics up to the melting point in refractory high-entropy alloys, yet acknowledged that these periodic supercells cannot capture the full aperiodicity of true random configurations [12]. Balyakin *et al.* developed potentials for molten multi-component alloys but still relied on SQS-like approximations that miss long-range concentration fluctuations [13]. Wu and Li constructed ML potentials for high-entropy alloys yet evaluated them on structures that retained some degree of periodicity [14].

The physics of disordered alloys is also temperature-sensitive in ways ordered benchmarks overlook. Disordered solid solutions are often stabilized at elevated temperatures where entropy plays a decisive role. Benchmarks performed exclusively at 0 K therefore miss the dynamic local-environment changes that occur in realistic operating conditions. Mandal *et al.* explicitly studied configurational disorder with ML potentials and showed how temperature-dependent relaxation further differentiates ordered from disordered behavior [15].

In short, disordered alloys are not merely noisier versions of ordered crystals. They inhabit a fundamentally different region of configuration space, demand generalization across vastly more local environments, and obey different stabilizing mechanisms. Standard benchmarks that never enter this region cannot claim to measure real-world accuracy.

Table 1 clarifies the central domain-shift problem by contrasting the physical, configurational, and evaluative properties of ordered benchmark regimes with those of real disordered-alloy applications.

Table 1. Ordered-crystal benchmarks versus real disordered-alloy deployment: a theoretical comparison of what is measured and what is missed

Analytical dimension	Ordered-crystal benchmark regime	Real disordered-alloy deployment regime	Implication for ML-potential validity
Atomic arrangement	Periodic, symmetry-constrained, repeating motifs	Aperiodic or quasi-random occupation with many non-equivalent neighborhoods	Success on ordered crystals does not establish robustness under configurational disorder
Local environment diversity	Low; many atoms share equivalent coordination environments	Extremely high; each atom may experience a distinct chemical neighborhood	The model must generalize across environment distributions, not isolated motifs

Configuration-space coverage	Narrow and highly structured	Exponentially larger and sparsely sampled	Training on ordered compounds leaves most relevant disorder space unseen
Dominant stabilization logic	Stoichiometric ordering, regular bonding motifs, low-temperature structure stability	Random mixing, local distortion, short-range order, and entropy-sensitive behavior	Different governing physics weaken claims that one regime stands in for the other
Evaluation temperature	Typically static 0 K snapshots	Often finite-temperature conditions with fluctuating local environments	Static benchmark accuracy may not survive thermally activated disorder
Error visibility	Aggregated metrics often suppress edge-case failures	Failure is exposed through unstable forces, poor relaxation, or incorrect dynamics	Practical deployment requires regime-specific reporting, especially for forces
Generalization claim typically made	"Accurate for alloys"	Actually needed: accurate for random solid solutions and high-entropy alloys	Benchmark language should be narrowed unless disordered evidence is presented
Appropriate benchmark criterion	Low error on held-out ordered structures	Stable performance across multiple disordered configurations and compositions	The meaningful unit of validation is the ordered-to-disordered gap, not ordered accuracy alone

Sources of overestimation

Overestimation of machine-learning potential accuracy in disordered alloys arises from entrenched benchmarking conventions that systematically obscure the complexity of configurational disorder. A central limitation lies in the exclusive reliance on ordered configurations for both training and evaluation, which effectively constrains the model's exposure to a narrow region of configuration space. Under these conditions, strong performance reflects interpolation within ordered manifolds rather than any capacity to generalize to disordered environments. Waters and Rondinelli demonstrated that predictive errors increase sharply once configurational randomness is introduced [16], while Baranovskii *et al.* showed that even potentials designed for disordered systems exhibit latent deficiencies when assessed using conventional ordered benchmarks [17].

This issue is reinforced by dataset partitioning strategies that permit substantial structural overlap between training and test sets. Random splits within ordered datasets reproduce similar local motifs across both subsets, enabling memorization to masquerade as generalization. Wen *et al.* identified this leakage as a primary source of inflated performance in solid-solution alloys [18], underscoring how apparent accuracy can emerge from redundancy rather than learned physical relationships. A related distortion emerges from the use of special quasirandom structures as stand-ins for disorder. Although SQS reproduce average correlation functions, their periodic construction suppresses long-range fluctuations and compositional inhomogeneity. Arróyave found that models trained on SQS fail to maintain accuracy when evaluated on fully random configurations [19], a discrepancy further corroborated by Men *et al.*, who reported systematically optimistic validation outcomes under SQS-based protocols [20].

Beyond structural representation, evaluation metrics introduce an additional layer of bias. The predominance of energy-based error reporting privileges a comparatively tractable target while neglecting force predictions, which are far more sensitive to local disorder and critical for dynamical simulations. Hong *et al.* emphasized the heightened difficulty of force prediction in disordered systems [21], and Duval *et al.* showed that equivariant architectures can yield low energy errors while remaining unreliable at the force level [22]. This imbalance obscures deficiencies that become consequential in practical applications, particularly where accurate force landscapes govern system evolution.

Further limitations arise from the restricted compositional scope of training datasets. Models are typically developed using near-equiatomic ordered compounds, thereby limiting their exposure to the broader compositional variability characteristic of disordered alloys. When applied to off-stoichiometric regimes, where configurational entropy exerts a dominant influence, predictive accuracy deteriorates. Malakar *et al.* demonstrated that interpolation within ordered composition spaces produces a misleading impression of robustness [23], revealing a gap between benchmark performance and real-world applicability. This gap is compounded by the near-universal assumption of zero-temperature evaluation. Disordered alloys, however, operate under finite-temperature conditions that induce continuous fluctuations in local atomic environments. Zhang *et al.* highlighted how zero-temperature validation fails to capture entropy-driven effects [24], while Hamedani *et al.* framed this discrepancy as a domain shift that undermines transferability [25].

Taken together, these methodological constraints generate a persistent divergence between reported accuracy and actual performance in disordered systems. Deringer *et al.* argued that existing protocols systematically conceal the challenges posed by configurational disorder [26], and Glasscott demonstrated that ordered-only benchmarks yield misleadingly optimistic generalization estimates [27]. Eyert *et al.* further linked these failures to representational inadequacies that prevent models from encoding the relevant physics of disorder [28], while Huang *et al.* quantified the extent of overestimation in high-entropy alloys and called for substantive revisions to benchmarking practice [29]. Without addressing these structural deficiencies, reported accuracies will continue to overstate model capability in precisely those regimes where reliable prediction is most critical.

Table 2 consolidates the manuscript's core argument by distinguishing each benchmarking practice from the specific evaluative distortion it introduces and the protocol correction required to restore credibility.

Table 2. Structural sources of benchmark overestimation in ML potentials for disordered alloys

Benchmarking practice	Why it appears defensible in published workflows	Why it fails for disordered alloys	Resulting distortion in reported performance	Minimum corrective benchmark standard
Ordered-only training and ordered-only test sets	Produces clean, low-variance evaluation on familiar crystal structures	Excludes the configurational heterogeneity that defines solid solutions and high-entropy alloys	Apparent generalization is inferred without ever testing the relevant deployment regime	Require a separate disordered test partition for every alloy-related benchmark
Random or composition-based splits on ordered datasets	Preserves sample size and supports conventional held-out testing	Recurrent motifs and periodic environments appear in both train and test data	Memorization is misread as transferable physics	Use split strategies that isolate unseen local-environment classes and report ordered-to-disordered error gaps
SQS validation as a proxy for disorder	Appears more realistic than perfectly ordered crystals while remaining computationally convenient	Periodic supercells suppress true aperiodicity, long-range fluctuation, and broader environment diversity	Proxy success is misread as disorder readiness	Report explicit SQS-versus-random comparisons using multiple independently generated random configurations
Energy-dominant evaluation	Energy MAE is familiar, compact, and easy to	Disordered-alloy deployment often fails through inaccurate forces	Potentials appear usable despite unstable dynamics	Report force metrics separately for ordered and disordered sets;

	compare across models	before gross energy failure becomes obvious	and unreliable relaxations	do not collapse them into aggregate scores
Narrow composition coverage	Supports interpolation claims within a limited chemical design window	Local environments at interpolated disordered compositions are not equivalent to ordered compounds at nearby stoichiometries	Composition overlap is mistaken for environment-level transferability	Benchmark across off-stoichiometric and equiatomic disordered compositions with many sampled configurations
Zero-temperature validation	Aligns with standard DFT reference workflows and simplifies comparison	Real disordered alloys are commonly used where thermal fluctuation and entropy reshape local environments	Static accuracy is mistaken for operational robustness	Include finite-temperature validation after short MD trajectories or thermally perturbed snapshots

A typology of overestimation

Overestimation in machine-learning potentials for alloys is sustained by recurring interpretive distortions that emerge under conventional benchmarking. A persistent discrepancy arises when low errors reported for ordered structures fail to translate to disordered solid solutions of identical composition, where deviations increase by an order of magnitude yet remain undocumented due to the absence of disordered evaluation. This gap is further obscured by reliance on special quasirandom structures, whose engineered periodicity suppresses the very fluctuations that define true disorder, leading to claims of robustness that collapse under genuinely random configurations. Apparent generalization is also produced through compositional interpolation, where models trained on discrete ordered stoichiometries are assessed on intermediate compositions that share nominal chemistry but diverge substantially in local environments. A related distortion emerges at the level of evaluation metrics, as acceptable energy errors on ordered data conceal force inaccuracies that render dynamical simulations unstable once disorder is introduced. These effects do not arise sporadically but follow directly from entrenched validation practices, yielding a literature that systematically overstates the readiness of ML potentials for the disordered alloys central to contemporary materials design.

Consequences for the field

Such systematic inflation of performance has material consequences for computational materials engineering, where reported accuracies are often treated as indicators of practical reliability. Confidence in deployment follows naturally from these claims, prompting integration of ML potentials into high-throughput workflows and large-scale simulations of disordered alloys. Under realistic conditions, however, models frequently produce unphysical energetics, unstable trajectories, or incorrect phase behavior, redirecting experimental efforts toward unproductive outcomes. Liu *et al.* documented how optimistic benchmark interpretations contribute to downstream experimental inefficiencies in high-entropy alloy research [5]. This dynamic also reshapes the perception of progress, as the field increasingly presents alloy modeling as a solved problem despite validation remaining confined to ordered crystals. Rosenbrock *et al.* and Byggmästar *et al.* both reported strong performance while anchoring evaluation in ordered configurations, reinforcing an inflated narrative that obscures the unresolved challenges of disorder [6, 7].

A further implication lies in the delayed recognition of model limitations. When configurational disorder is excluded from validation, deficiencies in force prediction or the representation of short-range order only become apparent during application, well after peer review. Huang *et al.* identified such failures only upon extending analysis beyond ordered datasets [11], indicating that current protocols defer rather than eliminate critical evaluation. This delay influences how research effort is distributed, encouraging incremental optimization within ordered domains rather than investment in generalization across configurational space. Wu and Li exemplified this tendency by refining potentials for high-entropy alloys while prioritizing ordered-like environments [14]. The resulting imbalance is particularly evident in the treatment of

multi-principal systems, where severe disorder coincides with the greatest need for predictive accuracy. Despite their relevance, these systems remain largely absent from benchmarks, as noted by Pei *et al.* [9], creating a widening gap between claimed capability and demonstrated reliability. Mandal *et al.* further emphasized that configurational disorder fundamentally alters phase stability while remaining invisible under prevailing evaluation schemes [15], reinforcing the need to reassess the evidentiary basis of model performance.

Alternative benchmarking strategies

Addressing these limitations requires reorienting benchmarking toward conditions that reflect the physics of disordered alloys rather than the convenience of ordered crystals. Incorporating explicit disordered test sets introduces a necessary separation between interpolation within ordered structures and genuine generalization, as demonstrated by Waters and Rondinelli, who revealed substantial hidden errors under such conditions [16]. This shift gains further resolution through configurational sampling, where multiple independent realizations at fixed composition expose variance that single-structure evaluations conceal, an approach advocated by Baranovskii *et al.* [17]. The distinction between quasirandom and fully random configurations also warrants direct quantification, as comparisons of this kind reveal the extent to which periodic approximations underestimate error, consistent with observations by Arróyave [19].

Extending evaluation to multi-principal alloys imposes a more stringent test of model robustness, particularly in systems characterized by extreme local-environment variability, a direction proposed by Men *et al.* [20]. Incorporating finite-temperature effects further aligns benchmarking with application domains, as thermal fluctuations reshape local configurations in ways absent at zero temperature; Zhou *et al.* highlighted the significance of such effects across a broad temperature range [12]. Within this framework, force-based metrics assume increased importance, given their direct role in governing system dynamics, a point emphasized by Hong *et al.* [21] and reinforced by Duval *et al.* in the context of equivariant models [22]. These adjustments require no expansion of underlying data resources but instead depend on how existing datasets are partitioned and interrogated. Wen *et al.* have already shown the feasibility of configurational evaluation for solid-solution alloys [18], while Glasscott demonstrated that realistic generalization demands departure from random splits toward structurally distinct test regimes [27]. Aligning benchmarking with these principles redefines reported accuracy as a meaningful predictor of performance in the disordered systems that increasingly define materials innovation.

Relation to Other Critiques

This critique stands apart from, yet complements, several existing discussions of ML potential limitations. It does not rehearse the well-known dangers of random train-test splits on similar structures; instead, it isolates a deeper domain shift—ordered versus disordered—that persists even after structural similarity is eliminated. Glasscott examined random splits versus realistic generalization and showed how motif leakage inflates performance [27]. The present analysis goes further: even a perfect, non-leaking split conducted entirely on ordered crystals still produces massive overestimation once configurational disorder is introduced.

The critique also differs from conventional extrapolation critiques. Many authors worry that models fail when compositions or structures lie outside the training convex hull. Yet the ordered-to-disordered gap is not merely an extrapolation problem. A disordered $\text{Fe}_{0.5}\text{Al}_{0.5}$ solid solution may sit at a composition already present in ordered training data (Fe_3Al , FeAl , FeAl_3), yet its local environments differ so profoundly that the model still collapses. Malakar *et al.* documented composition-interpolation illusions that mirror this exact failure mode [23]. The issue is therefore one of representation rather than numerical extrapolation.

Finally, this work extends but does not duplicate critiques of special quasirandom structures. Zhang *et al.* explored SQS in ML training and noted their utility as approximations [24]. Eyert *et al.* analyzed representation challenges for disordered materials and argued that even SQS-trained models miss essential aperiodic physics [28]. The present critique agrees

but emphasizes that SQS validation itself creates the “SQS mirage”—a distinct illusion that must be measured and reported rather than accepted as sufficient.

By focusing exclusively on benchmarking practice rather than model architecture or training algorithms, this article targets the single point where the community can enact immediate, low-cost change: the test set. Bartók *et al.* unified materials and molecular modeling yet relied on ordered benchmarks that left disordered alloys unexamined [4]. Batzner *et al.* advanced equivariant networks with impressive ordered-set results that still require the disordered complement proposed here [3]. The critique therefore does not contradict prior work; it completes it by insisting that ordered-to-disordered performance become the new minimum standard for any claim of alloy readiness.

Implications for Benchmark Design

Benchmark creators, reviewers, and practitioners each bear responsibility for closing the ordered-to-disordered gap.

For benchmark creators the mandate is clear: disordered configurations must become a non-negotiable component of every standard alloy benchmark suite. Creators should release not only ordered reference data but also pre-generated random solid-solution and high-entropy supercells at multiple compositions. They must require separate reporting of the “ordered-to-disordered gap”—the difference in error between the two regimes—so that readers can assess real-world utility at a glance. Waters and Rondinelli provided an early template by isolating disordered-alloy performance [16], and future suites should institutionalize this separation.

For paper reviewers the checklist expands. Reviewers must ask: “Was this model tested on disordered configurations?” Claims of “alloy generalization,” “solid-solution accuracy,” or “high-entropy applicability” should be rejected unless accompanied by explicit disordered test results. If only ordered benchmarks are presented, the manuscript should be returned for major revision. Deringer *et al.* examined configurational disorder challenges and argued that reviewers must enforce higher standards precisely on this point [26]. Huang *et al.* quantified benchmark overestimation in high-entropy alloys and called for exactly this reviewer vigilance [29].

For practitioners the implication is defensive. Do not trust ordered-only benchmarks when planning simulations of disordered alloys. Before investing computational or experimental resources, request or generate your own disordered validation set and measure the gap yourself. When selecting among competing potentials, prioritize those that publish both ordered and disordered metrics. Hamedani *et al.* discussed domain shift in disordered alloy potentials and urged practitioners to perform independent checks [25].

These changes require no new theory, no additional DFT calculations beyond what is already routine, and no architectural breakthroughs. They demand only intellectual honesty about what current benchmarks actually measure. Once disordered validation becomes routine, the literature will shift from optimistic overestimation to calibrated realism. The community will finally know which potentials truly work for the disordered alloys that now define the frontier of materials design.

Conclusion

Benchmarking practices in ML potentials systematically overstate performance on disordered alloys by restricting evaluation to ordered configurations. Training on periodic intermetallics, combined with random splits that reuse structural motifs, SQS approximations that suppress aperiodicity, and energy-focused metrics that obscure force errors, produces an internally consistent but practically misleading picture of accuracy. Narrow compositional coverage and zero-temperature validation further disconnect reported results from real deployment, leaving recent literature, despite technical sophistication, of limited relevance for solid solutions and high-entropy alloys.

The consequences extend directly into practice. Inflated confidence misguides experimental effort, while claims of rapid progress obscure persistent failures in modeling disorder. Under realistic conditions, breakdowns in force prediction and local-environment representation emerge only after deployment, revealing that the ordered-to-disordered gap, SQS-based validation, apparent compositional generalization, and energy–force divergence are structural rather than incidental features of current evaluation.

A straightforward correction lies in aligning benchmarks with disordered regimes through explicit disordered test sets, configurational sampling, SQS-versus-random comparisons, high-entropy systems, temperature-aware validation, and force-based metrics. These adjustments require no new data yet substantially improve diagnostic rigor. Disordered validation should therefore be treated as a baseline requirement; without it, claims of alloy readiness remain weak. Only by grounding evaluation in the conditions that govern real materials can ML potentials credibly support discovery in compositionally complex systems.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 19 Nov 2022

Revised: 31 Jan 2023

Accepted: 02 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

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

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

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>.

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).
<https://doi.org/10.1126/sciadv.1701816>.

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>.

Rosenbrock CW, Gubaev K, Shapeev AV, Pártay LB, Bernstein N, Csányi G, et al. Machine-learned interatomic potentials for alloys and alloy phase diagrams. *NPJ Comput Mater*. 2021;7(1):24.
<https://doi.org/10.1038/s41524-020-00477-2>.

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>.

Kostiuchenko T, Körmann F, Neugebauer J, Shapeev A. Impact of lattice relaxations on phase transitions in a high-entropy alloy studied by machine-learning potentials. *NPJ Comput Mater*. 2019;5(1):55.
<https://doi.org/10.1038/s41524-019-0195-y>.

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.
<https://doi.org/10.1038/s41524-020-0308-7>.

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>.

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>.

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>.

Balyakin IA, Rempel AA. Machine learning interatomic potential for molten TiZrHfNb. *AIP Conf Proc*. 2020;2313(1):030037.
<https://doi.org/10.1063/5.0032302>.

Wu L, Li T. A machine learning interatomic potential for high entropy alloys. *J Mech Phys Solids*. 2024;187:105639.
<https://doi.org/10.1016/j.jmps.2024.105639>.

Mandal P, Choudhury A, Mallick AB, Ghosh M. Phase prediction in high entropy alloys by various machine learning modules using thermodynamic and configurational parameters. *Met Mater Int*. 2023;29(1):38-52.
<https://doi.org/10.1007/s12540-022-01220-w>.

Waters MJ, Rondinelli JM. Benchmarking structural evolution methods for training of machine learned interatomic potentials. *J Phys Condens Matter*. 2022;34(38):385901.
<https://doi.org/10.1088/1361-648X/ac7f73>.

Baranovskii SD, Nenashev AV, Hertel D, Gebhard F, Meerholz K. Energy scales of compositional disorder in alloy semiconductors. *ACS Omega*. 2022;7(50):45741-51.
<https://doi.org/10.1021/acsomega.2c05426>.

Wen C, Wang C, Zhang Y, Antonov S, Xue D, Lookman T, et al. Modeling solid solution strengthening in high entropy alloys using machine learning. *Acta Mater*. 2021;212:116917.
<https://doi.org/10.1016/j.actamat.2021.116917>.

Arróyave R. Phase stability through machine learning. *J Phase Equilib Diffus*. 2022;43(6):606-28.
<https://doi.org/10.1007/s11669-022-01009-9>.

Men Y, Wu D, Hu Y, Li L, Li P, Jia S, et al. Understanding alkaline hydrogen oxidation reaction on PdNiRuRh high-entropy-alloy by machine learning potential. *Angew Chem Int Ed Engl*. 2023;62(27).
<https://doi.org/10.1002/anie.202217976>.

Hong C, Choi JM, Jeong W, Kang S, Ju S, Lee K, et al. Training machine-learning potentials for crystal structure prediction using disordered structures. *Phys Rev B*. 2020;102(22):224104.
<https://doi.org/10.1103/PhysRevB.102.224104>.

Duval AA, Schmidt V, Hernández-García A, Miret S, Malliaros FD, Bengio Y, et al. FAENet: Frame averaging equivariant GNN for materials modeling. In: *Proceedings of the 40th International conference on machine learning; 2023 Jul 23-29; Honolulu, HI, USA. Proc Mach Learn Res*. 2023;202:9013-33.

Malakar P, Balaprakash P, Vishwanath V, Morozov V, Kumaran K. Benchmarking machine learning methods for performance modeling of scientific applications. In: *2018 IEEE/ACM performance modeling, benchmarking and simulation of high performance computer systems (PMBS); 2018 Nov 12; Dallas, TX, USA. IEEE; 2018. p. 33-44*.
<https://doi.org/10.1109/PMBS.2018.8641686>.

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>.

Hamedani A, Byggmästar J, Djurabekova F, Alahyarizadeh G, Ghaderi R, Minuchehr A, et al. Primary radiation damage in silicon from the viewpoint of a machine learning interatomic potential. *Phys Rev Mater*. 2021;5(11):114603.
<https://doi.org/10.1103/PhysRevMaterials.5.114603>.

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

Glasscott MW. Classifying and benchmarking high-entropy alloys and associated materials for electrocatalysis: A brief review of best practices. *Curr Opin Electrochem*. 2022;34:100976.
<https://doi.org/10.1016/j.coelec.2022.100976>.

Eyert V, Wormald J, Curtin WA, Wimmer E. Machine-learned interatomic potentials: Recent developments and prospective applications. *J Mater Res*. 2023;38(24):5079-94.
<https://doi.org/10.1557/s43578-023-01239-8>.

Huang W, Farkas D, Bai XM. High-throughput machine learning-kinetic Monte Carlo framework for diffusion studies in equiatomic and non-equiatomic FeNiCrCoCu high-entropy alloys. *Materialia*. 2023;32:101966.
<https://doi.org/10.1016/j.mtla.2023.101966>.