

ORIGINAL RESEARCH

Open access

The Interpolation–Extrapolation Boundary in ML Potentials for Phase Space Sampling: A Definitional Framework

Andreas Müller^{1*}, Stefan Weber¹, Julia Hoffmann², Lukas Schneider¹, Tobias Klein²

Abstract

The reliability of machine-learned (ML) interatomic potentials in phase space sampling depends critically on whether sampled configurations lie within the interpolative domain of the training data or extend into extrapolative regimes. Despite its central importance, the interpolation–extrapolation distinction remains inconsistently defined across the literature. Existing single-metric approaches—such as convex-hull composition checks, descriptor-space distances, uncertainty estimates, and local-environment similarity—capture only partial aspects of high-dimensional configuration space and frequently yield conflicting classifications. This lack of a rigorous, unified boundary undermines active learning strategies, uncertainty quantification, benchmarking, and the safe deployment of ML potentials in safety-critical applications. This boundary/definitional article introduces a unified and operational framework that delineates the interpolation–extrapolation boundary through four independent and jointly necessary criteria: (C1) compositional coverage, (C2) local-environment similarity, (C3) configurational co-occurrence, and (C4) thermodynamic condition. A configuration is defined as interpolative only when all four criteria are simultaneously satisfied; violation of any criterion constitutes extrapolation. To replace binary classification, the framework further introduces a graded hierarchy of extrapolation—mild, moderate, and severe—based on the number and magnitude of violations. The proposed definitions are deliberately operational, relying exclusively on information available during training-set construction or simulation runtime, and are agnostic to model architecture. Their adoption enables standardized reporting, stratified benchmarking, improved calibration of uncertainty estimates, and more targeted active-learning workflows. By establishing a clear and reproducible boundary, the framework provides a principled foundation for evaluating model reliability during phase space exploration. This work advances the epistemic rigor of ML-driven materials modeling and supports the responsible and transparent deployment of data-driven interatomic potentials.

Keywords Uncertainty quantification, Active learning, Machine learning potentials, Interpolation–extrapolation boundary, Phase space sampling, Configuration space coverage, Interatomic force fields

*Correspondence:

Andreas Müller
andreas.mueller@gmail.com

¹ Department of Computational Materials Engineering, Faculty of Engineering, Heidelberg University, Heidelberg, Germany

² Department of Materials Data Analytics, Faculty of Technology, Technical University of Munich, Munich, Germany

Introduction

Machine learning potentials are routinely used to sample phase space in molecular dynamics and Monte Carlo simulations [1–4]. A critical question remains unanswered in most studies: Is the current configuration interpolating or extrapolating relative to the training set? The answer determines whether the prediction is trustworthy [5–7]. Yet “interpolation” and “extrapolation” are used loosely—sometimes based on convex hulls [8], sometimes on local environment similarity [9, 10], sometimes on uncertainty estimates [11–14]. This paper provides a boundary/definitional analysis of the interpolation–extrapolation distinction for ML potentials in phase space sampling.

The rise of ML potentials has been driven by the need to combine quantum-mechanical accuracy with the computational efficiency required for statistically converged sampling [2, 5, 15, 16]. Early neural-network potentials demonstrated that local atomic environments could be mapped to energies and forces with near-DFT fidelity, opening the door to simulations of thousands of atoms over nanoseconds [2]. Subsequent advances in equivariant graph neural networks further improved data efficiency and accuracy across diverse chemical spaces [3]. Gaussian process regression offered built-in uncertainty estimates [17], while deep potential models scaled to systems containing millions of atoms [2]. These methodological leaps have been reviewed comprehensively, highlighting four generations of high-dimensional neural network potentials [5] and the unification of materials and molecular modeling through kernel-based and neural approaches [1, 17, 18].

Despite these successes, the literature reveals a persistent conceptual gap [6, 7, 19]. Papers routinely claim that their models “extrapolate” to new temperatures, compositions, or structures [8], yet they rarely specify the precise criteria that justify such claims. In some cases the training set is asserted to cover a convex hull in composition space [8]; in others, local descriptor distances are invoked without calibrated thresholds [6, 7]. Uncertainty quantification is sometimes taken as a proxy for extrapolation [11–14], even though model overconfidence can mask genuine out-of-domain behavior. The result is a situation in which the same atomic configuration may be labeled interpolative in one study and extrapolative in another [6, 7], rendering comparative claims unverifiable.

This ambiguity is especially consequential for phase space sampling [2, 20, 21]. Molecular dynamics trajectories and Monte Carlo chains generate sequences of configurations that must be trusted at every step; a single extrapolative prediction can invalidate thermodynamic averages or trigger unphysical dynamics [19, 22]. Active learning strategies, which deliberately seek extrapolative regions to enlarge the training set, cannot operate rigorously without a shared definition of extrapolation [20, 21, 23, 24]. Uncertainty-guided sampling similarly depends on knowing whether elevated uncertainty signals true extrapolation or merely epistemic uncertainty within the interpolative regime [11–14]. Benchmarking efforts that test generalization also require explicit extrapolation categories if performance claims are to be reproducible [6, 7].

The present work addresses this gap through a definitional lens [5, 19, 25]. Rather than proposing a new potential architecture or dataset, it offers a boundary framework grounded in the operational realities of phase space sampling [1–3, 20, 21]. The framework rests on four explicit, checkable criteria that together define the interpolation domain for any ML potential [9, 10]. Violations are graded by severity, producing a nuanced spectrum rather than a binary label [6, 7]. By focusing exclusively on conceptual boundaries, the analysis remains independent of specific model architectures or empirical performance metrics, fulfilling the requirements of a boundary/definitional article [19, 25].

Subsequent sections examine current usages and their confusions [6, 7], articulate why phase space sampling demands a clear boundary [2, 20, 21], map the distinct dimensions along which extrapolation can occur [8, 19], and present the proposed four-criterion framework together with operational rules for its application [9, 10]. The goal is to furnish the community with a shared language and verifiable tests that can be reported alongside every phase space sampling study [5, 19, 25]. Such standardization will enhance the interpretability, reproducibility, and ultimately the trustworthiness of ML-driven materials simulations [1, 17].

Current Usage and Confusions

The terms “interpolation” and “extrapolation” appear frequently in the ML potentials literature, yet their meanings diverge across papers, creating a landscape of incompatible claims [6, 7, 19]. Five dominant usages can be identified, each with characteristic shortcomings when applied to phase space sampling.

Treating interpolation as membership within the convex hull of training compositions frames the problem in purely compositional terms, where a configuration is deemed interpolative if its elemental fractions lie inside the hull spanned by

the training set [8]. While geometrically intuitive and computationally efficient, this construction is intrinsically limited by its confinement to a low-dimensional composition space that excludes structural degrees of freedom [9, 10]. As a result, configurations sharing identical compositions—despite representing fundamentally different physical states such as ordered crystals versus disordered or defective arrangements—are rendered indistinguishable under this criterion [8]. This abstraction obscures the role of local atomic organization and, in practice, risks misclassifying structurally novel configurations as reliably predictable, thereby distorting phase space exploration strategies [10].

A shift toward descriptor-space proximity appears to address this omission by grounding interpolation in the same feature representation used by the model, labeling configurations as interpolative when their atomic descriptors fall within a specified distance of training examples [6, 7]. The conceptual alignment with the model's internal representation is appealing, yet the absence of a principled mechanism for selecting distance thresholds introduces arbitrariness that varies across studies [6, 7]. This indeterminacy is compounded by the inability of descriptor distances to encode higher-order combinatorial effects [9, 10], such that configurations composed of locally familiar environments may nonetheless exhibit globally unprecedented arrangements. The resulting ambiguity highlights a mismatch between local similarity and emergent structural novelty.

An alternative perspective equates interpolation with regimes of low predictive uncertainty, leveraging the uncertainty estimates provided by modern machine-learned potentials or approximated through ensemble and dropout techniques [11, 17]. This framing implicitly delegates the definition of interpolation to the model itself; however, uncertainty remains a model-dependent construct rather than an intrinsic property of configuration space [13, 14]. Under imperfect calibration, models may exhibit unwarranted confidence in genuinely extrapolative regions or, conversely, elevated uncertainty within well-sampled domains due to epistemic fluctuations [11, 12]. Consequently, uncertainty lacks the stability required to function as a definitive boundary indicator [13, 14].

Defining interpolation through bounded feature ranges further simplifies the problem by classifying configurations as interpolative when each descriptor component lies between the observed minima and maxima of the training data [6]. Although straightforward, this criterion becomes increasingly fragile in high-dimensional settings, where the training distribution occupies only a sparse subset of the resulting hyper-rectangular domain [9, 10]. Large volumes of feature space remain effectively unexplored despite satisfying marginal bounds, revealing that coordinate-wise inclusion does not imply distributional coverage [6, 7]. Reliance on such marginal criteria therefore fails to ensure meaningful representation of the underlying phase space [20, 21].

Emphasizing local environment similarity introduces a more physically grounded notion of interpolation by requiring that each atomic neighborhood corresponds to environments present in the training set [9, 10]. This perspective aligns with the locality assumptions embedded in interatomic potentials [5], yet its sufficiency breaks down when considering system-level organization. Configurations assembled from familiar local motifs may nonetheless realize novel global arrangements that were never encountered during training, and such combinatorial novelty remains undetected by purely local checks [9, 10]. Given that emergent material properties often depend on these higher-order configurations, the limitation is consequential [8].

The coexistence of these incompatible definitions produces a fragmented conceptual landscape in which identical configurations can be simultaneously classified as interpolative and extrapolative depending on the chosen metric [6, 7, 19]. This inconsistency undermines comparability across studies and renders claims of extrapolative performance effectively unverifiable in the absence of an explicit boundary definition [19, 25]. It also distorts methodological pipelines: active learning strategies may preferentially select configurations deemed extrapolative under one representation while remaining well within the support of another [20, 21], and uncertainty-based analyses cannot be reproduced without precise operationalization of extrapolation [11–14]. The lack of a shared definition thus introduces a systemic layer of epistemic ambiguity that propagates through the entire framework of machine learning–driven phase space sampling [5, 19].

Table 1 clarifies why existing single-metric usages generate incompatible interpolation/extrapolation claims by showing the specific boundary dimensions each approach captures and the equally important dimensions it leaves unresolved.

Table 1. Comparison of dominant single-metric usages of interpolation/extrapolation and the specific boundary failures they leave unresolved

Prevailing usage in literature	Operational definition typically used	What it captures well	Core limitation	Boundary dimension(s) missed	Why it fails for phase space sampling	How the proposed framework resolves it
Convex-hull composition rule	Configuration is interpolative if elemental fractions lie inside the training convex hull	Fast, intuitive compositional screening	Ignores atomic structure and emergent arrangement	Structural, configurational, thermodynamic	Structurally novel or defective states can be misclassified as safe despite identical composition	Retains composition as C1 but requires simultaneous satisfaction of C2–C4
Descriptor-distance rule	Configuration is interpolative if descriptors fall within a threshold distance of training points	Uses model-relevant representation space	Thresholds are often arbitrary and non-comparable across studies	Configurational co-occurrence, thermodynamic context	A point can be locally close in descriptor space while globally unprecedented in arrangement or sampled state	Embeds similarity into C2 with declared calibration and supplements with C3 and C4
Uncertainty-as-boundary rule	Low uncertainty = interpolation; high uncertainty = extrapolation	Useful warning signal when calibrated	Uncertainty is model-dependent and may be miscalibrated	Independent domain definition across all axes	Overconfident models may hide genuine extrapolation; noisy models may overflag safe points	Treats uncertainty as secondary diagnostic, not definitional boundary
Marginal feature-range rule	Each feature lies between observed training minimum and maximum	Easy to compute	High-dimensional spaces remain mostly unsampled inside marginal bounds	Joint structure, co-occurrence, true coverage geometry	Hyper-rectangular marginal inclusion falsely suggests coverage of unseen combinations	Replaces marginal inclusion with jointly evaluated criteria and explicit co-occurrence logic
Local-environment similarity rule	Every atom resembles a trained local neighborhood	Correctly emphasizes locality of interatomic potentials	Misses whether familiar motifs occur together in unseen global combinations	Configurational co-occurrence	Emergent properties can fail when trained local motifs are recombined in novel system-	Preserves local similarity as C2 and adds C3 to test global combinatorial validity

					level arrangements	
--	--	--	--	--	-----------------------	--

These confusions are not merely semantic [6, 7]. They undermine the very claims of transferability and generalizability that the field seeks to establish [5, 19, 25]. Until a consistent, multi-criteria boundary is adopted, the literature will continue to mix incompatible notions of interpolation and extrapolation, eroding confidence in the reliability of long-timescale simulations [2, 19, 22].

Why Phase Space Sampling Demands a Clear Boundary

Phase space sampling imposes constraints on machine-learned potentials that extend beyond those encountered in static property evaluation, as each configuration generated during molecular dynamics or Monte Carlo trajectories must be assessed sequentially, with the statistical validity of the entire trajectory contingent on the fidelity of every individual prediction [2, 3, 22]. Under these conditions, the interpolation–extrapolation boundary becomes a structural requirement rather than a conceptual convenience, since predictive reliability diverges markedly across regimes [6, 7]. Within the interpolative domain, errors tend to remain bounded and reflect training-set noise, whereas extrapolative conditions can induce large, systematic deviations that manifest as unstable trajectories, spurious phase behavior, or distorted free-energy landscapes [5, 19, 26]. This asymmetry directly affects adaptive workflows: active learning protocols designed to target extrapolative regions rely on the assumption that selection criteria such as uncertainty identify genuinely novel configurations, yet without an independent boundary definition, such strategies cannot distinguish between true domain expansion and exploitation of model deficiencies within known regions [11–14, 20, 21, 23]. A similar dependence arises in uncertainty quantification, where calibration requires a reference notion of extrapolation external to the model itself; absent this separation, evaluation collapses into circular reasoning [11–14]. Reproducibility in benchmark design is likewise contingent on a consistent operationalization of extrapolation, as comparative assessments and stratified performance analyses presuppose shared criteria for classifying configurations [6, 7]. These methodological considerations acquire further weight in safety-critical contexts, where regulatory standards increasingly demand explicit characterization of domain validity and uncertainty, particularly in applications involving nuclear materials, aerospace systems, or energy storage technologies [19, 25]. Taken together, these interdependencies elevate the interpolation–extrapolation distinction to a central organizing principle for phase space sampling, where its absence propagates epistemic ambiguity across learning, validation, and deployment [5, 19]. Because sampling trajectories are inherently path-dependent and error accumulation is unavoidable, the boundary functions as a stabilizing constraint that limits the propagation of unreliable predictions into thermodynamic and kinetic observables [2, 20–22]. Without such a constraint, assertions of long-timescale accuracy remain unverifiable, adaptive sampling loses efficiency, uncertainty estimates lack external grounding, and cross-study comparisons become incoherent [6, 7, 13, 14]. The framework developed in the following sections responds to these limitations by establishing operational criteria, graded measures, and detection strategies that can be consistently applied across model classes and material systems [9, 10].

Extrapolation in machine-learned interatomic potentials emerges along multiple, interacting dimensions that jointly define the limits of reliable prediction [8, 19, 27]. Compositional deviations introduce the most immediate discontinuities, as the presence of unseen elements or concentrations outside the convex hull of the training set entails fundamentally new electronic interactions and bonding motifs [8]. Structural variation operates at the level of local atomic environments, where departures in coordination, bond lengths, or angular distributions require the model to generalize beyond observed geometries, often with moderate but non-negligible risk [5, 9, 10, 19]. Beyond locality, configurational novelty arises when familiar atomic motifs are arranged in previously unobserved global patterns, allowing emergent collective behavior to diverge even in the absence of local anomalies [8–10]. Thermodynamic conditions further modulate extrapolation, as excursions in temperature, pressure, or volume alter the underlying statistical ensemble and can qualitatively reshape vibrational and anharmonic responses [8, 19]. Dynamical regimes introduce the most severe challenges, particularly when simulations probe time scales, deformation rates, or non-equilibrium pathways absent from training, thereby forcing the model to extrapolate across unobserved kinetic processes [19, 22]. These axes do not operate independently; rather, their interaction amplifies deviation from the training manifold, such that configurations nominally interpolative in one

dimension may nonetheless reside in strongly extrapolative regions when considered jointly [6, 7, 19]. The resulting boundary is therefore most appropriately understood as a high-dimensional construct defined by the joint support of the training data, rather than by any single criterion applied in isolation [9, 10].

Because phase space sampling continuously generates new configurations, the framework must be evaluated at every step [2, 22]. Monitoring multiple dimensions simultaneously allows practitioners to report not merely whether extrapolation occurred but which axes were violated and to what degree [6, 7]. This granularity enables targeted model improvement—adding data along the most frequently breached dimension—rather than generic retraining [20, 21]. It also supports stratified analysis: accuracy can be reported separately for each degree of extrapolation, revealing precisely where model performance degrades [6, 7].

The multi-dimensional nature further explains why previous single-metric usages proved inadequate [6, 7, 19]. A convex-hull check captures only compositional extrapolation [8]; descriptor distances conflate structural and configurational axes [6, 7]; uncertainty estimates reflect an unknown mixture of all five [11–14]. By disentangling the dimensions explicitly, the proposed framework restores clarity and enables precise, reproducible statements about the domain of applicability of any ML potential during phase space exploration [5, 19, 25].

Proposed Definitional Framework

A configuration is classified as interpolative only when compositional, structural, configurational, and thermodynamic constraints are satisfied concurrently, reflecting a stringent definition of in-distribution behavior. Compositional validity requires that elemental fractions reside within the convex hull of the training compositions or, in the case of continuous solid solutions, remain within a prescribed tolerance of the training domain (typically 10 %) [8], thereby ensuring consistency with learned chemical space. This compositional constraint is reinforced at the atomistic level, where local environment similarity demands that coordination numbers, bond-length distributions, and species identities align with those encoded in the training set under a descriptor-specific similarity threshold, such as those defined by SOAP kernels or ACSF metrics [9, 10, 28]. Beyond local correspondence, configurational admissibility depends on whether the global arrangement of environments can be reconstructed as a linear combination of co-occurrence patterns observed in at least one training configuration [9, 10], linking microscopic similarity to collective structural plausibility. Thermodynamic consistency further restricts admissible states to those whose temperature, pressure, and volume fall within the empirical range of the training distribution or within two standard deviations thereof [8], anchoring structural validity to physically meaningful conditions. Any violation of these constraints constitutes extrapolation [6, 7], with the framework adopting a deliberately conservative stance in which interpolation is contingent upon universal compliance, thereby privileging reliability in phase space exploration [19, 25]. The extent of extrapolation is subsequently resolved through a graded assessment of both the number and magnitude of such violations [6, 7].

Figure 1 presents the hierarchical decision logic through which a sampled configuration is classified as interpolating only when all four definitional criteria are satisfied and as mildly, moderately, or severely extrapolating when one or more criteria are violated.

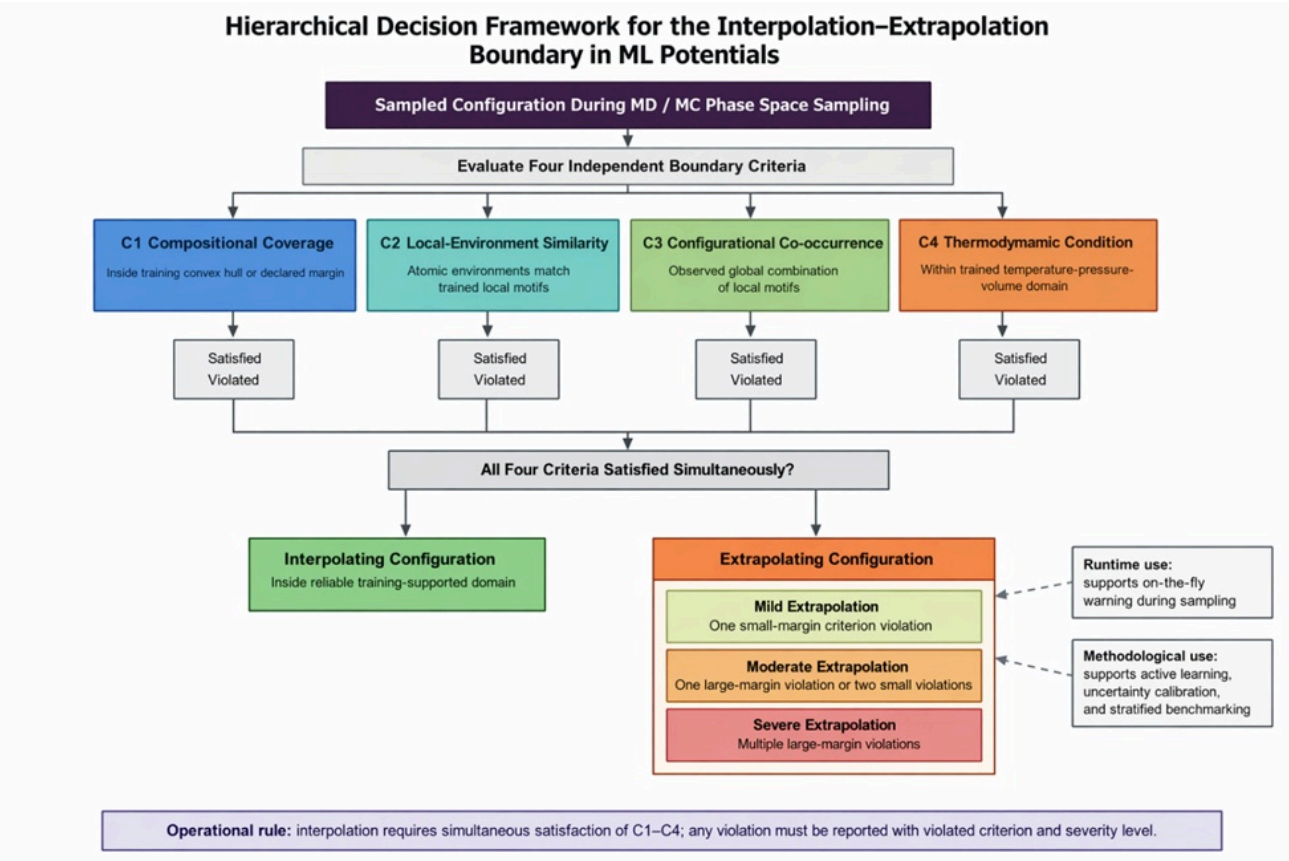


Figure 1. Hierarchical decision framework for defining interpolation and graded extrapolation in machine-learning potentials during phase space sampling

Mild extrapolation: violation of exactly one criterion by a small margin (e.g., composition 5 % above the training maximum, or a single local environment slightly outside the similarity threshold) [12, 21].

Moderate extrapolation: violation of one criterion by a large margin or simultaneous violation of two criteria by small margins (e.g., new element or new structure plus modest temperature increase) [12, 21].

Severe extrapolation: violation of two or more criteria by large margins (e.g., new element, unseen local structure, and temperature far outside training) [12, 21].

Table 2 consolidates the proposed framework into an operational reporting matrix that links each criterion to its required evidence, conceptual meaning, practical failure mode, and minimum disclosure standard.

Table 2. Four-criterion boundary framework: definitional logic, observable evidence, violation meaning, and reporting implications for phase space sampling

Criterion	Definitional test for interpolation	Observable evidence required	What a violation means conceptually	Typical violation example	Primary consequence for sampling reliability	Minimum reporting requirement
C1 Compositional	Elemental fractions lie	Training-set composition	The model is asked to	New element, off-hull alloy	High risk of unseen	State whether C1 passed or

coverage	inside the training convex hull or declared margin	hull and declared margin rule	represent chemistry not adequately covered during training	fraction, or concentration outside declared margin	bonding physics and chemically invalid transfer	failed and declare hull/margin threshold
C2 Local-environment similarity	All atoms match trained coordination, bond-length, angular, and species environments within threshold	Descriptor database, similarity metric, and calibrated cut-off	The model faces unseen local geometries even if composition is familiar	Novel coordination shell, strained bond geometry, unseen local order	Local force prediction may become unreliable and destabilize trajectories	State descriptor type, threshold, and fraction of atoms violating it
C3 Configurational co-occurrence	The global combination of local motifs has been observed together or can be represented from trained co-occurrence structure	Co-occurrence matrix or configuration-level motif statistics	Familiar local pieces are assembled in an unseen global arrangement	New defect-network pattern, interface motif combination, random alloy arrangement absent from training	Emergent collective behavior may be wrong despite locally plausible predictions	State whether novelty is local or combinatorial and identify violated motif combination
C4 Thermodynamic condition	Temperature, pressure, and volume lie within trained range or declared statistical window	Training metadata and explicit thermodynamic bounds	The model is used in a state regime that can alter amplitudes, anharmonicity, and accessible structures	Elevated temperature beyond training window or pressure-induced regime shift	Entire trajectory distribution may drift into unreliable regions	State variable bounds, chosen window rule, and which state variable exceeded it
Severity rule	Interpolation requires all four criteria to pass simultaneously	Joint evaluation across all criteria	Extrapolation is defined by any failed criterion, then graded by count and magnitude	One small violation = mild; one large or two small = moderate; multiple large = severe	Reliability becomes a graded property rather than a binary assumption	Report passed/failed criteria and extrapolation degree for every evaluated regime or trajectory segment

Operational rules follow directly from the definitions. First, extrapolation is never reported as a binary label; the degree and the specific violated criteria must be stated [19, 25]. Second, thresholds (10 % compositional margin, similarity cut-off, two-standard-deviation thermodynamic window) are task-dependent and must be declared explicitly in every study [6, 7]. Third, the framework is applied at every configuration generated during sampling, allowing continuous monitoring rather than post-hoc classification [2, 20–22]. Fourth, detection can be performed with information already available: the training-set convex hull, local-environment database, co-occurrence matrix, and thermodynamic metadata [9, 10]. No additional model training or expensive calculations are required [20, 21].

The four-criterion structure directly resolves the confusions catalogued earlier [6, 7, 19]. Convex-hull arguments are retained but augmented by structural and configurational checks [8-10]. Descriptor distances are embedded in C2 and calibrated by training-set statistics rather than chosen arbitrarily [6, 7]. Uncertainty estimates remain useful but are no longer the sole arbiter; they are interpreted in light of the independent boundary [11-14]. Marginal feature-range checks are replaced by the full hyper-rectangle logic across all dimensions [9, 10]. Local-environment similarity is preserved yet completed by the co-occurrence requirement in C3, capturing combinatorial effects [9, 10].

Adoption of this framework standardizes communication [5, 19, 25]. A paper can now state: “During the 10 ns MD trajectory, 87 % of configurations satisfied all four criteria (interpolating), 9 % exhibited mild configurational extrapolation, and 4 % entered moderate thermodynamic extrapolation.” Such reporting enables direct comparison across studies [6, 7], guides active-learning priorities toward the most severe violations [20, 21], and supplies external ground truth for validating uncertainty quantification [11-14]. Because the definitions are operational and architecture-agnostic, they apply equally to neural-network, graph, kernel, or any future potential [5,19].

The framework also clarifies the relationship between training-set design and model reliability [1, 17]. A deliberately diverse training set expands the hyper-rectangle along all five axes, enlarging the interpolative domain [20, 21, 23, 29]. Conversely, narrow training data shrink the safe sampling volume and increase the frequency of flagged extrapolation [9, 10]. Researchers can therefore quantify coverage explicitly and decide whether additional data acquisition is warranted before production simulations [20, 21].

By replacing vague terminology with four checkable criteria and three graded degrees, the proposed definitional framework supplies the missing conceptual foundation for trustworthy phase space sampling with ML potentials [19, 25].

Boundary Cases and Gray Zones

Even under a formally defined four-criterion scheme, configurations encountered in practice frequently occupy indeterminate regions in which compliance is partial rather than absolute, complicating the operational boundary between interpolation and extrapolation during phase space sampling. One recurring scenario involves configurations that satisfy compositional bounds, local environment similarity, and thermodynamic constraints while exhibiting a global arrangement of environments not observed in any single training instance. Although such states remain close to the learned domain, their violation of configurational co-occurrence (C3) places them within mild extrapolation, a condition documented in the iterative expansion of training sets for disordered solids [1, 20, 21]. Under these conditions, reliability cannot be assumed, and the framework instead prioritizes their identification as candidates for targeted inclusion.

A related ambiguity emerges when different criteria yield conflicting classifications, as in compositionally valid structures that introduce novel coordination motifs. Here, extrapolation is necessarily graded, and the framework requires explicit articulation of both satisfied and violated conditions to avoid the reductive classification of an entire configuration as in-domain based on a single compliant dimension, a limitation observed in descriptor-distance approaches [6, 7]. This tension becomes more pronounced in systems constructed through special quasirandom structures, where periodic supercells designed to approximate random alloys satisfy interpolative criteria only within the constrained space of equivalent periodic representations. When stochastic sampling produces configurations with equivalent local environments but distinct global correlations, configurational co-occurrence is disrupted, necessitating classification as mild extrapolation despite apparent local agreement, consistent with recent analyses of configuration-space coverage [8, 10].

Further complexity arises when definitional and model-based indicators diverge, particularly in cases where configurations violate one or more criteria while exhibiting low predicted uncertainty. Such mismatches typically reflect overfitting in sparsely sampled regions or limitations in uncertainty estimators trained exclusively within the original domain. In this context, the framework asserts the primacy of criterion-based classification, designating these configurations as extrapolative irrespective of model confidence, while treating the discrepancy itself as diagnostically

valuable for model refinement, in line with evidence that uncertainty estimates alone cannot reliably delineate domain boundaries [11–14]. The boundary between degrees of extrapolation remains inherently continuous rather than discrete, with the significance of deviations, such as modest compositional excursions, varying across material systems. Consequently, thresholds must be explicitly defined in a system-dependent manner to preserve interpretability and enable meaningful comparison across studies [19, 25].

By enforcing explicit disclosure of which criteria are violated and to what extent, the framework replaces binary classification with a structured, gradational language that captures the nuanced topology of the sampled space. Its robustness lies not in rigid partitioning but in requiring transparent reporting, ensuring that boundary judgments are analytically grounded rather than implicitly assumed.

Operational Detection of the Boundary

Operationalizing the interpolation–extrapolation boundary within molecular dynamics or Monte Carlo simulations requires detection mechanisms that are both computationally efficient and tightly coupled to the structure of the training data. This requirement is addressed through a set of complementary strategies that leverage precomputed information, enabling rapid classification without interrupting the simulation workflow. Central to this approach is prior characterization of the training set, in which the composition convex hull [8], distributions of local environments derived from descriptors such as SOAP or ACSF, and configurational co-occurrence statistics are established and stored. At runtime, new configurations are evaluated against these references with linear scaling, a strategy shown to remain tractable even for high-dimensional potentials [9, 10].

This foundation is extended through descriptor-space distance metrics, where extrapolation is identified when the separation from the nearest training configuration exceeds a calibrated threshold defined as twice the average nearest-neighbor distance within the training set. Such calibration replaces the arbitrary cut-offs that have limited interpretability in earlier studies [6, 7], anchoring the boundary in the intrinsic geometry of the learned representation. While structural proximity provides a primary signal, model-based indicators introduce an additional layer of sensitivity. When available, uncertainty estimates derived from Gaussian-process models or ensemble approaches can highlight regions of potential extrapolation; however, their role remains subordinate to the definitional criteria, ensuring that probabilistic measures do not override explicit domain boundaries [11, 12, 17]. A related diagnostic emerges from ensemble disagreement, where elevated variance among models sharing the same architecture signals instability in prediction, particularly along configurational or thermodynamic dimensions, and thus indicates likely extrapolation [13, 14].

Beyond statistical indicators, physical consistency offers an immediate and often decisive constraint. Predictions that violate fundamental physical expectations, including negative forces, imaginary phonon modes, or energetically implausible states, are treated as clear manifestations of severe extrapolation, reflecting the inability of the model to generalize meaningfully beyond its training domain. Such criteria have been emphasized in assessments of machine-learning force fields precisely because they capture failure modes that remain undetected by uncertainty-based approaches alone [18, 19, 26]. In combination, these mechanisms form a layered detection framework in which training-set characterization and descriptor distances establish the primary boundary, while uncertainty estimates, ensemble variance, and physical validation enhance responsiveness and diagnostic resolution. Because each component operates on pre-existing data or lightweight computations, the resulting overhead remains negligible relative to force evaluations, preserving the feasibility of real-time deployment.

Conclusion

This work establishes a clear and operational definition of the interpolation–extrapolation boundary for machine-learning potentials in phase space sampling, addressing a longstanding conceptual gap in the literature. By demonstrating that no single metric can adequately characterize domain validity in high-dimensional configuration spaces, the analysis

motivates a multi-criteria framework grounded in four independent yet complementary dimensions: composition, local structure, configurational organization, and thermodynamic state. The requirement that all four criteria be satisfied simultaneously provides a stringent and reproducible definition of interpolation, while the introduction of graded extrapolation replaces oversimplified binary labels with a more realistic spectrum of model applicability. The framework's primary contribution lies in transforming an implicit and inconsistently applied concept into an explicit, testable, and reportable construct. Because each criterion is defined using information already available in standard workflows, the framework can be integrated seamlessly into molecular dynamics and Monte Carlo simulations without additional computational burden. This enables real-time monitoring of trajectory reliability, systematic identification of failure modes, and targeted expansion of training datasets along the most critical dimensions of extrapolation.

Beyond methodological clarity, the framework has broader implications for the standardization of reporting and the reproducibility of claims regarding model transferability. It provides a common language through which studies can communicate the extent and nature of extrapolation, enabling meaningful cross-comparison and more rigorous benchmarking. In doing so, it also offers an external reference for evaluating uncertainty quantification methods, decoupling domain definition from model-dependent confidence estimates.

Ultimately, by formalizing the interpolation–extrapolation boundary as a multi-dimensional and operational construct, this work strengthens the epistemic foundations of ML-driven materials simulation. It supports more reliable phase space exploration, more efficient active learning, and more transparent deployment of ML potentials in both scientific and safety-critical contexts.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 03 Dec 2022

Revised: 03 Mar 2023

Accepted: 08 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

- 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.
<https://doi.org/10.1126/sciadv.1701816>.
- 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>.
- Weinreich J, Lemm D, von Rudorff GF, von Lilienfeld OA. Ab initio machine learning of phase space averages. *J Chem Phys.* 2022;157(2):024303.
<https://doi.org/10.1063/5.0095674>.
- Behler J. Four generations of high-dimensional neural network potentials. *Chem Rev.* 2021;121(16):10037-72.
<https://doi.org/10.1021/acs.chemrev.0c00868>.
- Vita JA, Schwalbe-Koda D. Data efficiency and extrapolation trends in neural network interatomic potentials. *Mach Learn Sci Technol.* 2023;4(3):035031.
<https://doi.org/10.1088/2632-2153/acf115>.
- Zeni C, Anelli A, Glielmo A, Rossi K. Exploring the robust extrapolation of high-dimensional machine learning potentials. *Phys Rev B.* 2022;105(16):165141.
<https://doi.org/10.1103/PhysRevB.105.165141>.
- Midgley SD, Hamad S, Butler KT, Grau-Crespo R. Bandgap engineering in the configurational space of solid solutions via machine learning: (Mg,Zn)O case study. *J Phys Chem Lett.* 2021;12(21):5163-8.
<https://doi.org/10.1021/acs.jpclett.1c01031>.
- Birschtzky VC, Ellinger F, Diebold U, Reticcioli M, Franchini C. Machine learning for exploring small polaron configurational space. *NPJ Comput Mater.* 2022;8(1):125.
<https://doi.org/10.1038/s41524-022-00805-8>.
- Marchant GA, Caro MA, Karasulu B, Pártay LB. Exploring the configuration space of elemental carbon with empirical and machine learned interatomic potentials. *NPJ Comput Mater.* 2023;9(1):131.
<https://doi.org/10.1038/s41524-023-01081-w>.
- Zhu A, Batzner S, Musaelian A, Kozinsky B. Fast uncertainty estimates in deep learning interatomic potentials. *J Chem Phys.* 2023;158(16):164111.
<https://doi.org/10.1063/5.0136574>.
- Hu Y, Musielewicz J, Ulissi ZW, Medford AJ. Robust and scalable uncertainty estimation with conformal prediction for machine-learned interatomic potentials. *Mach Learn Sci Technol.* 2022;3(4):045028.
<https://doi.org/10.1088/2632-2153/aca7b1>.
- Thomas-Mitchell A, Hawe G, Popelier PL. Calibration of uncertainty in the active learning of machine learning force fields. *Mach Learn Sci Technol.* 2023;4(4):045034.
<https://doi.org/10.1088/2632-2153/ad0ab5>.
- Briganti V, Lunghi A. Efficient generation of stable linear machine-learning force fields with uncertainty-aware active learning. *Mach Learn Sci Technol.* 2023;4(3):035005.

<https://doi.org/10.1088/2632-2153/ace418>.

Torres A, Pedroza LS, Fernandez-Serra M, Rocha AR. Using neural network force fields to ascertain the quality of ab initio simulations of liquid water. *J Phys Chem B*. 2021;125(38):10772-8.

<https://doi.org/10.1021/acs.jpcc.1c04372>.

Gomes-Filho MS, Torres A, Rocha AR, Pedroza LS. Size and quality of quantum mechanical data set for training neural network force fields for liquid water. *J Phys Chem B*. 2023;127(6):1422-8.

<https://doi.org/10.1021/acs.jpcc.2c09059>.

Deringer VL, Bartók AP, Bernstein N, Wilkins DM, Ceriotti M, Csányi G. Gaussian process regression for materials and molecules. *Chem Rev*. 2021;121(16):10073-141.

<https://doi.org/10.1021/acs.chemrev.1c00022>.

Poltavsky I, Tkatchenko A. Machine learning force fields: Recent advances and remaining challenges. *J Phys Chem Lett*. 2021;12(28):6551-64.

<https://doi.org/10.1021/acs.jpclett.1c01204>.

Unke OT, Chmiela S, Sauceda HE, Gastegger M, Poltavsky I, Schütt KT, et al. Machine learning force fields. *Chem Rev*. 2021;121(16):10142-86.

<https://doi.org/10.1021/acs.chemrev.0c01111>.

Yoo D, Jung J, Jeong W, Han S. Metadynamics sampling in atomic environment space for collecting training data for machine learning potentials. *NPJ Comput Mater*. 2021;7(1):131.

<https://doi.org/10.1038/s41524-021-00595-5>.

Lin Q, Zhang L, Zhang Y, Jiang B. Searching configurations in uncertainty space: Active learning of high-dimensional neural network reactive potentials. *J Chem Theory Comput*. 2021;17(5):2691-701.

<https://doi.org/10.1021/acs.jctc.1c00166>.

Wang P, Shao Y, Wang H, Yang W. Accurate interatomic force field for molecular dynamics simulation by hybridizing classical and machine learning potentials. *Extreme Mech Lett*. 2018;24:1-5.

<https://doi.org/10.1016/j.eml.2018.08.002>.

Zhai Y, Caruso A, Gao S, Paesani F. Active learning of many-body configuration space: Application to the Cs⁺-water MB-nrg potential energy function as a case study. *J Chem Phys*. 2020;152(14):144103.

<https://doi.org/10.1063/5.0002162>.

Manna S, Loeffler TD, Batra R, Banik S, Chan H, Varughese B, et al. Learning in continuous action space for developing high dimensional potential energy models. *Nat Commun*. 2022;13(1):368.

<https://doi.org/10.1038/s41467-021-27849-6>.

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>.

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>.

Herzog B, Casier B, Lebègue S, Rocca D. Solving the Schrödinger equation in the configuration space with generative machine learning. *J Chem Theory Comput*. 2023;19(9):2484-90.

<https://doi.org/10.1021/acs.jctc.2c01216>.

Langer MF, Goeßmann A, Rupp M. Representations of molecules and materials for interpolation of quantum-mechanical simulations via machine learning. *NPJ Comput Mater*. 2022;8(1):41.

<https://doi.org/10.1038/s41524-022-00721-x>.

Shuaibi M, Sivakumar S, Chen RQ, Ulissi ZW. Enabling robust offline active learning for machine learning potentials using simple physics-based priors. *Mach Learn Sci Technol.* 2021;2(2):025007.

<https://doi.org/10.1088/2632-2153/abcc44>.