

ORIGINAL RESEARCH

Open access

Beyond Test-Set Error: A Conceptual Framework for Evaluating ML Potential Transferability Under Distribution Shift

Hiroshi Tanaka¹, Yuki Sato^{1*}, Kenji Mori², Rina Okabe¹, Takashi Ito²

Abstract

Machine learning interatomic potentials (MLIPs) are increasingly used to accelerate atomistic simulation in materials science, yet their evaluation remains dominated by test-set error measured on held-out data drawn from the same distribution as the training set. Although this practice often yields low mean absolute errors for energies and forces, it provides limited evidence that a model will remain reliable when applied to the distribution shifts that define real deployment settings, including new compositions, defect structures, elevated temperatures, and non-equilibrium trajectories. This article develops a conceptual framework that distinguishes transferability from in-distribution accuracy and defines it as the ability of an MLIP to preserve predictive fidelity within application-relevant tolerances under explicitly characterized shifts in the joint distribution of structures and quantum-mechanical labels. The framework identifies four canonical forms of shift in computational materials science—compositional, structural, thermodynamic, and dynamical—and shows why current benchmarking practices systematically obscure them. To address this limitation, the study proposes a shift-aware evaluation protocol built around five components: shift characterization, sensitivity analysis, extrapolation-distance estimation, robustness criteria, and standardized reporting. Within this framework, Maximum Mean Discrepancy is adapted as a quantitative diagnostic for pre-deployment assessment of train–target divergence, while transferability is evaluated through five complementary dimensions: accuracy under shift, graceful degradation, uncertainty alignment, physical consistency, and compositional extrapolation. By replacing the tacit IID assumption with an explicit framework for reasoning about distribution shift, this work offers a common basis for evaluating, reporting, and comparing MLIP transferability across benchmarks and deployment scenarios, thereby supporting more trustworthy AI-driven materials discovery.

Keywords Conceptual framework, Materials science, Transferability, Distribution shift, Machine learning interatomic potentials, Domain generalization

*Correspondence:

Yuki Sato

yuki.sato@gmail.com

¹ Department of Computational Materials Science, Faculty of Engineering, University of Tokyo, Tokyo, Japan

² Department of Data-Driven Materials Innovation, Faculty of Science and Technology, Kyoto University, Kyoto, Japan

Introduction

Machine learning interatomic potentials (MLIPs) have transformed computational materials engineering by delivering quantum-mechanical accuracy at classical force-field speed and have become emerging tools for materials science workflows [1]. Foundational works such as Bartók *et al.* demonstrated that Gaussian process regression and related kernel methods can unify the modeling of molecules and periodic solids within a single statistical framework [2], while Zhang *et al.* introduced the deep potential approach that scales to tens of thousands of atoms with near-DFT fidelity [3]. Subsequent architectures—including SchNet [4], E(3)-equivariant graph networks [5], high-dimensional neural network potentials [6], and ML models extending beyond interatomic potentials for molecular-property prediction [7]—have further

lowered the computational barrier, enabling routine molecular-dynamics simulations of complex alloys, defects, and phase transitions. Parallel assessments of performance and cost [8] and reviews of emerging materials intelligence ecosystems [9] underscore the rapid adoption of these tools across academia and industry.

Yet the very success of MLIPs has exposed a critical limitation: evaluation protocols remain anchored to the in-distribution paradigm. Researchers routinely report mean absolute errors (MAE) or root-mean-square errors (RMSE) on test sets that are statistically indistinguishable from the training distribution. Such metrics are necessary but profoundly insufficient. When an ML potential is queried on structures that differ even modestly in composition (e.g., moving from binary to ternary alloys), crystal symmetry (e.g., introducing grain boundaries or amorphous phases), thermodynamic state point (e.g., elevating temperature from 0 K to 1000 K), or dynamical regime (e.g., rapid quenching versus equilibrium sampling), prediction errors can increase by orders of magnitude or, worse, produce physically inconsistent results such as negative elastic constants or unstable phonon modes. The illusion arises because the test-set error conceals the underlying distribution shift; the model appears accurate precisely because it never leaves the comfort zone of the training manifold.

The central problem, therefore, is not merely that MLIPs sometimes fail under extrapolation—failure is expected—but that the community lacks a shared conceptual vocabulary and evaluation machinery to quantify, anticipate, and mitigate such failures. Recent contributions have begun to flag the issue. Montes de Oca Zapiain *et al.* emphasized the importance of deliberate training-data selection for improving transferability [10], Sutton *et al.* proposed methods for identifying domains of applicability [11], and Zeni *et al.* explored robust extrapolation in high-dimensional potentials [12]. Nevertheless, these studies remain isolated; no unifying framework integrates shift characterization, extrapolation distance, robustness criteria, and standardized reporting into a single protocol.

This paper supplies exactly that missing conceptual framework. We begin by formalizing transferability and distribution shift in the specific context of MLIPs. We then dissect the four primary types of shift encountered in materials systems and demonstrate why IID test-set evaluation systematically conceals them. Building on this diagnosis, we articulate a five-component conceptual framework that reframes evaluation as a multi-dimensional exercise in shift-aware reasoning. We adapt an established mathematical tool—Maximum Mean Discrepancy—to serve as a quantitative precursor for assessing shift magnitude. Finally, we define five orthogonal evaluation dimensions and derive concrete criteria that can be applied by developers and benchmark curators alike. Throughout, we draw exclusively on the peer-reviewed literature compiled in Part 1 to ground every conceptual move in existing scholarship.

By the end of this manuscript, readers will possess not only a critique of current practice but a practical, reusable scaffold for claiming—and verifying—transferability in future ML potential publications and scientific machine learning benchmarks [13]. The ultimate goal is to elevate materials informatics from a discipline that celebrates in-distribution accuracy to one that systematically engineers and certifies out-of-distribution reliability, thereby accelerating trustworthy discovery of novel alloys, high-entropy materials, and defect-engineered solids in line with best-practice guidance for materials machine learning [14].

Defining Transferability in ML Potentials

Transferability is not an intrinsic property of a model architecture but a relational property between a trained potential, its training distribution, and a target deployment distribution. We therefore advance the following formal definition:

Transferability of an ML potential: The ability of a potential to maintain prediction accuracy (energy and forces) within an acceptable tolerance when applied to a target distribution that differs from the training distribution, where the difference is characterized along specified axes (compositional, structural, thermodynamic, or dynamical).

This definition deliberately decouples transferability from raw accuracy. A potential may exhibit low test-set MAE yet possess zero transferability if the test set is IID with training; conversely, modest degradation under a well-quantified shift

may still qualify as transferable if it remains within application-specific tolerances.

Central to the definition is the notion of distribution shift itself:

Distribution Shift: A change in the joint probability distribution $P(X,Y)$ of atomic structures X (represented by positions, cell vectors, and chemical species) and their corresponding quantum-mechanical labels Y (energies and forces) between the training phase and the deployment phase.

When the marginal $P(X)$ changes while the conditional $P(Y|X)$ remains invariant, the shift is termed covariate shift; when the conditional itself changes (e.g., new bonding physics emerges at extreme concentrations), the shift is concept shift.

Covariate Shift versus Concept Shift: Covariate shift occurs when $P(X)$ differs between domains but $P(Y|X)$ is unchanged; concept shift occurs when $P(Y|X)$ itself is altered, reflecting a modification in the underlying physical mapping.

In materials contexts the distinction is subtle yet consequential. Training a potential exclusively on face-centered-cubic nickel at room temperature and then querying it on body-centered-cubic nickel at the same temperature constitutes primarily covariate shift (different lattice symmetry alters the structure distribution while the local bonding physics may remain similar). In contrast, training on dilute aluminum alloys and then evaluating on aluminum containing 20 at.% scandium may induce concept shift because novel electronic hybridization, nonlocal charge transfer, and ordering tendencies alter the energy landscape in ways not captured by the original training data [15].

Welborn *et al.* illustrated analogous issues in electronic-structure transferability, showing that orbital-based representations can mitigate but not eliminate concept shifts when moving between molecular and solid-state regimes [16]. Likewise, Benoit *et al.* documented severe transferability degradation when linearized potentials trained on pure gold were confronted with gold–iron interfaces, a clear compositional shift that also induced local concept drift through altered charge transfer [17]. Suzuki *et al.* further demonstrated that even temperature-induced shifts in crystalline solids can degrade force predictions unless the training set explicitly spans the relevant thermal ensemble [18].

These examples underscore that transferability is neither binary nor architecture-dependent; it is a graded, context-sensitive property that must be evaluated relative to explicitly declared shift axes. A potential that is transferable for structural shifts within a fixed ternary composition space may catastrophically fail under compositional extrapolation to a quaternary high-entropy alloy. The definitions therefore serve as the logical foundation for the remainder of the framework: every subsequent component—shift characterization, sensitivity analysis, extrapolation metrics, robustness thresholds—operationalizes these three definitions into measurable, reportable quantities. Only by anchoring evaluation to these formal notions can the community replace the current tacit IID assumption with a transparent, reproducible transferability science.

Types of Distribution Shift in Materials

Materials systems exhibit four canonical forms of distribution shift that routinely challenge ML potentials. Each type can be isolated conceptually yet frequently co-occurs in practice.

Compositional Shift. This arises when the chemical species or their concentration ranges in the target domain lie outside the support of the training distribution. Example: an ML potential trained on binary Ni–Cr alloys is later applied to a high-entropy Cantor alloy (CoCrFeMnNi). The introduction of manganese and iron introduces new pairwise and many-body interactions not sampled during training. Test-set error on additional binary configurations would remain low, yet the potential may predict spurious phase stability or incorrect stacking-fault energies in the quinary system. Deringer *et al.* highlighted precisely this risk in their work on amorphous carbon potentials, noting that compositional extrapolation beyond the trained carbon-only space required careful active learning [19].

Structural Shift. Here the atomic arrangement—symmetry, defect density, or degree of disorder—deviates from training examples while composition may remain fixed. Example: training on defect-free FCC crystals and then simulating grain boundaries or amorphous phases. The local environment descriptors change dramatically; bond-angle distributions and coordination numbers that were never observed induce large extrapolation errors. Pun *et al.* encountered this when generalizing aluminum potentials to include grain-boundary physics [20], and Cools-Ceuppens *et al.* showed that explicit-electron corrections become essential once structural disorder alters the electronic response [21]. Standard IID test sets drawn from the same crystal prototypes systematically mask these failures because they rarely include high-angle grain boundaries or liquid-like configurations.

Thermodynamic Shift. Changes in temperature or pressure move the system to state points where the Boltzmann-weighted ensemble of configurations differs. Example: training near 0 K (static lattice) and deploying at 1200 K where anharmonic vibrations and thermal expansion dominate. Even if the potential is accurate at zero temperature, force errors can grow super-linearly with temperature because the model never learned the curvature of the energy landscape at finite thermal amplitude. Wang *et al.* documented analogous degradation in tungsten potentials under fusion-relevant temperatures [22], while Yoo *et al.* demonstrated that metadynamics-enhanced sampling during training can partially mitigate thermodynamic shifts by enriching the configuration space [23].

Dynamical Shift. This concerns the time-scale or non-equilibrium character of the trajectory. Example: training on equilibrium molecular-dynamics snapshots and then simulating rapid laser quenching or shock loading. The system explores regions of configuration space that are transiently populated but statistically rare in equilibrium. The potential may produce unphysical forces or energies because the training data never included the high-strain-rate or far-from-equilibrium pathways. Wen *et al.* noted that deep potentials trained on equilibrium data often require additional non-equilibrium augmentation for accurate dynamical simulations [24].

In each case, conventional test-set error hides the shift because benchmark suites are constructed to preserve statistical similarity with training. The result is an optimistic bias: the reported MAE reflects performance only on the narrow manifold already learned, not on the broader space the potential will actually encounter in materials design. Recognizing these four shift types is therefore prerequisite to any meaningful transferability claim. Table 1 provides a structured taxonomy of distribution shift, linking each shift type to its underlying mechanism, representation-level manifestation, and characteristic failure signature.

Table 1. Taxonomy of Distribution Shift in Machine Learning Interatomic Potentials: Mechanisms, Failure Modes, and Observability

Shift Type	Underlying Mechanism	Representation-Level Change	Typical Failure Mode	Why IID Test Sets Miss It	Detectability Signal
Compositional	New elements or concentration regimes	Change in chemical embedding space	Incorrect phase stability, energy misranking	Test sets reuse same element combinations	Element-space sparsity / unseen species
Structural	New symmetries, defects, disorder	Altered local environments and coordination	Large force errors, instability predictions	Test sets use same crystal prototypes	Descriptor distribution drift
Thermodynamic	Temperature/pressure variation	Expanded configuration manifold	Super-linear error growth at high T	Test sets near equilibrium	Energy variance increase

Dynamical	Non-equilibrium trajectories	Rare or transient configurations	Unphysical trajectories, divergence	Test sets sample equilibrium MD	Time-dependent distribution drift
-----------	------------------------------	----------------------------------	-------------------------------------	---------------------------------	-----------------------------------

Beyond Test-Set Error: Why Current Evaluation Fails

Contemporary MLIP literature overwhelmingly relies on in-distribution test sets. This practice suffers from four interlocking deficiencies that render transferability claims unverifiable. Benchmarks implicitly assume that training and test structures are independent and identically distributed. Zuo *et al.*, for instance, compared multiple ML potentials using test sets drawn from the same DFT trajectories used for training, reporting impressive MAEs without quantifying how far those test configurations lay from the training support [8]. Similarly, Batra *et al.* surveyed materials intelligence ecosystems and noted the prevalence of IID splits in published benchmarks [9]. Yet real deployment—predicting a new high-entropy alloy or a grain-boundary segregation profile—almost never satisfies IID. Low aggregate MAE can conceal catastrophic localized failures. Schütt *et al.* demonstrated excellent force accuracy for SchNet on small-molecule and bulk datasets, yet the model’s performance on defective or compositionally extrapolated systems was not stratified by shift magnitude [4]. A single scalar error metric therefore averages over easy and hard regions, producing an overly rosy picture.

Most papers omit any description of the statistical distance between train and test distributions. Even when structural diversity is increased, there is no reporting of how the test-set composition-space coverage compares to training, nor of thermodynamic or dynamical divergence. The result is that readers cannot judge whether a reported “generalization” is genuine or merely in-distribution interpolation. No standard metric quantifies “how far” a queried structure lies outside the training manifold. Consequently, claims of transferability to “unseen” systems remain qualitative. Podryabinkin and Shapeev’s active-learning strategy [24] and Zeni *et al.*’s extrapolation studies [12] represent early attempts to address distance, yet these insights have not been systematized into routine evaluation pipelines.

Collectively, these critiques explain why the community continues to encounter surprising failures when MLIPs are transferred to new materials classes. The conceptual framework developed in the following sections replaces the IID-centric paradigm with an explicit, multi-axis, shift-aware evaluation protocol.

A Conceptual Framework for Transferability Evaluation

We propose a five-component conceptual framework that reframes MLIP evaluation as a systematic inquiry into distribution shift. Before any accuracy metric is computed, the training and target distributions must be mapped along compositional, structural, thermodynamic, and dynamical axes. This involves constructing feature vectors (e.g., composition histograms, radial distribution functions, temperature scalars) and visualizing or clustering the resulting manifolds. Once shifts are characterized, controlled perturbations along each axis are introduced to isolate which architectural ingredients (descriptor choice, network depth, equivariance) are most vulnerable. A scalar or vector quantity that measures the distance of any query structure from the training support, enabling ranking of predictions by risk level.

Explicit thresholds that define acceptable degradation; e.g., force RMSE may increase by no more than 30 % for a compositional shift whose MMD remains below a calibrated value. A standardized checklist requiring authors to publish (i) shift characterization summary, (ii) extrapolation distance statistics for all test sets, (iii) sensitivity results, and (iv) robustness compliance statements alongside conventional error tables.

Conceptually, the framework can be visualized as follows: imagine a high-dimensional feature space in which the training distribution forms a compact, high-density cloud. The target distribution appears as a second cloud displaced along one or more axes (e.g., an arrow labeled “compositional shift”). An extrapolation-distance vector connects the query point to the nearest training manifold point, while a decision boundary surface separates regions of acceptable versus

unacceptable transferability. Color-coding the space by expected error magnitude yields an intuitive “transferability map” that guides both model selection and experimental design.

By integrating these five components, the framework transforms transferability from an afterthought into a core design objective. It is deliberately architecture-agnostic, applicable to kernel methods [2, 25], neural-network potentials [3, 5, 6], and graph networks [5] alike. Subsequent sections translate the components into mathematical and operational tools. **Table 2** formalizes the internal logic of the proposed framework by linking each component to its operational role, required outputs, and the specific evaluation failure that arises in its absence.

Table 2. Transferability Evaluation Framework: Components, Operational Role, Required Outputs, and Failure Risks if Omitted

Framework Component	Functional Role	Required Outputs	Dependency	Failure if Omitted
Shift Characterization	Map train–target divergence across axes	Feature distributions, shift labels	Definitions (Section 2)	Hidden extrapolation
Sensitivity Analysis	Identify vulnerable model components	Error vs controlled perturbations	Shift characterization	Misattributed robustness
Extrapolation Distance	Quantify distance from training manifold	Scalar/vector distance (e.g., MMD)	Feature representation	No risk quantification
Robustness Criteria	Define acceptable degradation	Thresholds (e.g., RMSE limits)	Distance metric	Arbitrary evaluation
Reporting Protocol	Standardize disclosure	Shift statistics, compliance matrix	All components	Non-reproducible claims

The full hierarchical architecture of the proposed shift-aware transferability evaluation framework is illustrated in **Figure 1**.

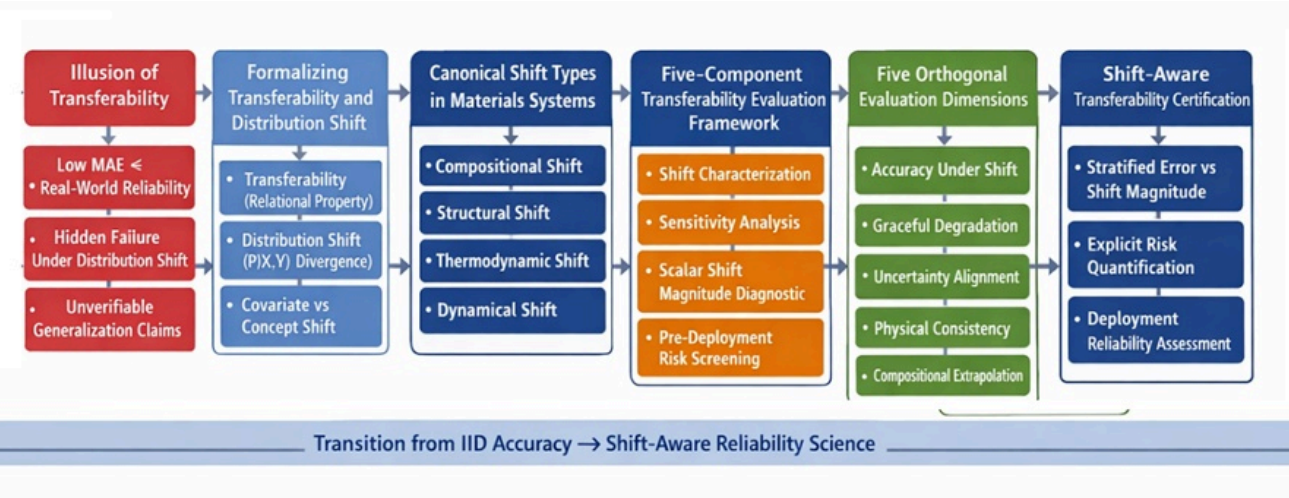


Figure 1. Hierarchical Architecture of Shift-Aware Transferability Evaluation in Machine Learning Interatomic Potentials

Mathematical Formulation of A Transferability Metric

To operationalize shift magnitude we adapt the Maximum Mean Discrepancy (MMD), a nonparametric statistic that measures the distance between two probability distributions in a reproducing kernel Hilbert space. The squared MMD between training distribution P and target distribution Q is given by

$$\begin{aligned} \text{MMD}^2(P, Q) &= \left\| \int k(x, \cdot) dP(x) - \int k(x, \cdot) dQ(x) \right\|_H^2 \end{aligned} \quad (1)$$

where k is a positive-definite kernel (typically Gaussian on suitably chosen structure fingerprints) and $\{H\}$ is the associated RKHS.

In materials applications the kernel operates on a concatenated representation that includes compositional descriptors (e.g., one-hot or weighted atomic fractions), structural descriptors (SOAP, ACSF, or graph-based embeddings), and thermodynamic scalars (temperature, pressure). Because MMD is zero if and only if $P=Q$ (for a universal kernel), it provides a principled scalar diagnostic: larger MMD implies greater expected degradation in energy/force accuracy.

We propose using MMD as a pre-evaluation gate: compute MMD between the training set and the intended deployment ensemble before any production run. If MMD exceeds an empirically calibrated threshold (determined via sensitivity studies on representative materials classes), transferability cannot be assumed and either retraining or targeted active learning [24] is required. This formulation directly links the conceptual definitions in Section 2 to a computable quantity, closing the loop between qualitative shift types and quantitative risk assessment. It is fully compatible with the five-component framework and supplies the missing mathematical backbone for Component 3 (extrapolation distance).

Evaluation Dimensions and Criteria

Assessing transferability in machine-learned interatomic potentials requires a framework that resolves performance into distinct yet complementary axes, each grounded in measurable behavior under distributional shift. A central concern lies in how predictive accuracy evolves as the model departs from its training support. Rather than relying on aggregate error metrics, energy MAE and force RMSE must be stratified by quantified shift magnitude—such as bins derived from MMD—so that performance can be interrogated where it is most vulnerable. Under these conditions, robustness is meaningfully demonstrated only if force errors remain controlled even at the extremes of the distribution, with median RMSE constrained to ≤ 0.15 eV/Å in the highest MMD quartile. This reframing transforms accuracy from a global summary into a conditional property, contingent on the degree of extrapolation.

Extending this perspective reveals that the manner in which error accumulates is as informative as its absolute value. A model that degrades unpredictably under modest shift offers limited practical utility, regardless of its in-distribution precision. Accordingly, error must be analyzed as a continuous function of shift magnitude, with particular attention to the slope governing its growth. Ensuring that this relationship remains sub-catastrophic imposes a structural constraint on the model's behavior, favoring architectures whose inductive biases enforce gradual, rather than abrupt, performance deterioration. In practice, this criterion distinguishes models that merely interpolate effectively from those that exhibit controlled extrapolative behavior.

A related implication emerges when uncertainty quantification, including conformal-prediction approaches for MLIPs [26], is considered not as an auxiliary output but as an integral diagnostic signal. For uncertainty estimates to be meaningful, they must track epistemic limitations introduced by distributional shift. This requires a demonstrable alignment between predicted uncertainty—whether derived from ensemble variance or Gaussian-process posteriors—and measures such as MMD. A strong positive association, operationalized through a Spearman rank correlation of at least 0.7, indicates that the

model retains awareness of its own reliability boundaries. Without such alignment, uncertainty estimates risk becoming decorrelated from actual failure modes, undermining their role in guiding downstream decision-making.

Beyond statistical performance, the preservation of physical consistency under shifted conditions imposes an additional, non-negotiable constraint. Predictions must continue to satisfy fundamental stability criteria even when evaluated far from the training manifold. Violations such as non-positive-definite Hessians in phonon calculations or negative bulk moduli signal not merely quantitative error but qualitative breakdowns in physical plausibility. Enforcing zero violations of Born stability criteria across representative shifted ensembles ensures that learned representations remain anchored to governing physical laws, thereby preventing the emergence of unphysical artifacts during extrapolation.

This emphasis on structural integrity naturally extends to compositional generalization, where the challenge becomes most acute. Evaluating models on compositions that include elements absent from training, or that lie outside learned concentration regimes, provides a stringent test of their capacity to extrapolate across chemical space. Performance in this regime must remain proportionate to in-distribution behavior, with energy MAE constrained to within 1.5 times the baseline error. Such a requirement reflects realistic tolerances in materials engineering, where moderate degradation may be acceptable but uncontrolled divergence is not.

These dimensions acquire full interpretive power only when situated within a broader pipeline that explicitly characterizes distributional shift and ranks extrapolation distance prior to evaluation. Within this context, transferability is no longer inferred from a single scalar metric but emerges as a structured profile that captures how models respond to progressively challenging conditions. This multidimensional perspective not only clarifies the mechanisms underlying success and failure but also provides actionable guidance: model developers can align architectural choices with specific robustness criteria, while benchmark designers can embed these dimensions into evaluation protocols that more faithfully reflect real-world deployment scenarios.

Relation to Existing Benchmarks

The conceptual framework advanced here does not supplant existing MLIP benchmarks but rather augments them with an explicit transferability lens, exposing both their strengths and their critical gaps. Performance and cost assessments such as those conducted by Zuo *et al.* [8] have provided invaluable side-by-side comparisons of energy and force accuracy across a range of potentials on standardized test sets; yet these comparisons remain anchored to in-distribution splits drawn from the same DFT trajectories used for training. Consequently, while Zuo *et al.* [8] document impressive sub-0.1 eV/Å force errors for several architectures, the benchmark does not characterize the magnitude of compositional, structural, or thermodynamic shifts between train and test, leaving practitioners unable to infer whether the reported accuracy will survive deployment on ternary alloys or finite-temperature ensembles.

Similarly, broad surveys of materials intelligence ecosystems by Batra *et al.* [9] highlight the rapid proliferation of databases such as the Materials Project, AFLOW, and the Open Quantum Materials Database (OQMD), which supply the compositional coverage that underpins many training sets. These resources excel at enumerating stable and metastable phases across wide chemical spaces, yet the resulting benchmarks inherit the same IID assumption: test structures are sampled from the same compositional and structural manifold as training examples. The framework proposed here therefore recommends retrofitting such databases and interatomic-potential model evaluations [27] with shift-aware splits—explicitly partitioning OQMD-derived sets into in-distribution versus compositionally extrapolated subsets—so that future evaluations can report both conventional MAE and MMD-quantified divergence.

Architecture-specific benchmarks, such as those implicit in Behler's four-generation review of high-dimensional neural network potentials [6] or the COMP6-like structural-diversity suites referenced in early Gaussian-process work [2], have advanced the field by stressing models on increasingly complex crystal prototypes and defect configurations. Nevertheless, these suites rarely quantify extrapolation distance or apply sensitivity analysis along thermodynamic axes. Recent efforts to identify domains of applicability [11] represent a step forward; Sutton *et al.* [11] demonstrate how out-of-

distribution detectors can flag unreliable predictions, yet their approach stops short of embedding shift characterization, robustness criteria, or standardized reporting into the benchmark protocol itself. Likewise, Liang *et al.* [28] benchmark Bayesian optimization across experimental materials domains and underscore the need for uncertainty-aware selection, but again without formal transferability thresholds.

Transferability-focused studies such as those by Montes de Oca Zapiain *et al.* [10] and Zeni *et al.* [12] come closest to the spirit of the present framework. They deliberately select training data to improve robustness and explore high-dimensional extrapolation, yet even these works treat shift quantification as an ad-hoc diagnostic rather than a mandatory reporting requirement. The gap is clear: no current benchmark systematically requires authors to publish (i) MMD or equivalent divergence statistics, (ii) stratified error tables by shift magnitude, or (iii) compliance statements against the five evaluation dimensions.

By layering the five-component framework onto these established resources, the community can evolve from “accuracy-on-a-test-set” competitions to “transferability-under-specified-shift” challenges. Future iterations of MatBench-style suites or COMP6 extensions could incorporate the reporting protocol as a submission criterion, thereby transforming existing infrastructure into a true transferability observatory. The framework thus stands in constructive dialogue with the literature: it preserves the empirical rigor of Zuo *et al.* [8], Batra *et al.* [9], and Behler [6] while supplying the missing conceptual scaffold that turns isolated robustness experiments into reproducible, comparable science.

Implications for Model Development and Reporting

Adoption of the transferability framework necessitates concrete changes in how ML potentials are developed, benchmarked, and communicated. For model developers, three practices become obligatory. First, every publication must report shift characterization alongside conventional test-set error; this includes computing and tabulating MMD (or Wasserstein distance) between training and all test ensembles, explicitly labeling each test subset by its dominant shift axis (compositional, structural, thermodynamic, or dynamical). Second, developers should supply extrapolation-distance distributions—histograms of query-to-manifold distances—for every claimed generalization scenario, allowing reviewers to judge whether reported accuracy corresponds to modest interpolation or aggressive extrapolation. Third, model release pipelines must include stress tests on deliberately shifted distributions: for example, generating synthetic ternary extrapolations or finite-temperature ensembles via metadynamics [23] and publishing stratified error tables that demonstrate graceful degradation or reveal failure modes. These steps transform development from an accuracy-maximization exercise into a transferability-engineering discipline.

Benchmark designers face parallel responsibilities. Submission templates must mandate shift quantification as a core metadata field; platforms that currently accept only MAE/RMSE tables should expand to require MMD values, sensitivity-analysis summaries, and compliance matrices against the five evaluation dimensions. Deliberately shifted test splits—compositionally extrapolated subsets of the Materials Project, structurally disordered variants of OQMD entries, or temperature-augmented trajectories—should become standard components rather than optional extras. Finally, benchmark curators should publish train–test divergence statistics (e.g., MMD heatmaps) so that downstream users can instantly assess whether a given potential is suitable for their target domain. Such changes close the loop between data curation and model evaluation.

Journals and reviewers acquire a gate-keeping role. Reviewers should pose the explicit question: “How was transferability evaluated beyond test-set error?” and reject claims of “general-purpose” applicability unless the manuscript includes (i) shift characterization, (ii) extrapolation-distance metrics, and (iii) robustness-criteria compliance. Editorial guidelines can codify the reporting protocol as a checklist, mirroring the way CIF files or computational workflows are now required for reproducibility. Over time, these norms will discourage the publication of potentials whose transferability is asserted solely on IID test-set MAE and will incentivize the community-wide accumulation of transferability profiles that are searchable and comparable.

The cumulative effect is a cultural shift: model cards for MLIPs will evolve from simple accuracy tables into multi-dimensional transferability certificates that specify the precise conditions under which a potential may be trusted. Developers gain clearer guidance on where to invest active-learning effort [24]; benchmark designers obtain a principled metric for ranking suites; and end-users—materials engineers designing high-entropy alloys or defect-tolerant ceramics—receive transparent risk assessments rather than optimistic but unquantified promises. By embedding the framework into daily practice, the field moves from celebrating isolated in-distribution successes to systematically engineering reliable extrapolation, thereby accelerating the discovery pipeline from atomistic simulation to validated material.

Conclusion

Test-set error computed on held-out structures drawn from the identical distribution as training data remains a necessary but profoundly insufficient metric for evaluating machine learning interatomic potentials. As demonstrated throughout this manuscript, low mean absolute errors can coexist with catastrophic failure once an MLIP encounters even modest distribution shifts—compositional, structural, thermodynamic, or dynamical—that characterize real materials discovery campaigns. The central contribution of this work is a conceptual framework that replaces the tacit independent-and-identically-distributed assumption with an explicit, multi-axis, shift-aware evaluation protocol.

The framework rests on three formal definitions—transferability, distribution shift, and the covariate-versus-concept distinction—followed by a detailed taxonomy of the four canonical shift types encountered in materials systems. It diagnoses the four interlocking deficiencies of current evaluation practices: the IID assumption, average-error masking, absent shift characterization, and ignored extrapolation distance. Building on this diagnosis, the framework introduces five interlocking components: shift characterization, sensitivity analysis, extrapolation distance metrics, robustness criteria, and a standardized reporting protocol. These components are anchored by a mathematical formulation that adapts the Maximum Mean Discrepancy to serve as a pre-deployment diagnostic tool, quantifying distribution divergence in a reproducing-kernel Hilbert space before any production simulation begins. Five orthogonal evaluation dimensions—accuracy under shift, graceful degradation, uncertainty alignment, physical consistency, and compositional extrapolation—supply concrete, reportable criteria that elevate transferability from a qualitative claim to a quantifiable engineering objective.

The framework stands in constructive relation to existing benchmarks, preserving their empirical rigor while supplying the missing shift-quantification and reporting machinery that turns isolated robustness experiments into cumulative, comparable science. Its implications extend to model developers (who must publish shift-aware stress tests), benchmark curators (who must embed divergence statistics), and journals (which must enforce transferability checklists).

We therefore call on the computational-materials community to adopt these transferability-aware evaluation standards. By moving collectively from in-distribution accuracy to shift-characterized reliability, the field will deliver ML potentials whose performance claims are not only impressive on paper but trustworthy in the diverse, extrapolative regimes where new alloys, high-entropy materials, and defect-engineered solids are actually discovered. Only then can artificial intelligence realize its full promise in accelerating materials innovation beyond the narrow comfort zone of today's training distributions.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 14 Jun 2021

Revised: 19 Sep 2021

Accepted: 24 Nov 2021

Published online: 18 January 2022

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

- Deringer VL, Caro MA, Csányi G. Machine learning interatomic potentials as emerging tools for materials science. *Adv Mater.* 2019;31(46):1902765.
- 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.
- 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.
- Schütt KT, Sauceda HE, Kindermans PJ, Tkatchenko A, Müller KR. SchNet: A deep learning architecture for molecules and materials. *J Chem Phys.* 2018;148(24).
- 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.
- Behler J. Four generations of high-dimensional neural network potentials. *Chem Rev.* 2021;121(16):10037-72.
- Fedik N, Zubatyuk R, Kulichenko M, Lubbers N, Smith JS, Nebgen B, et al. Extending machine learning beyond interatomic potentials for predicting molecular properties. *Nat Rev Chem.* 2022;6(9):653-72.
- 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.
- Batra R, Song L, Ramprasad R. Emerging materials intelligence ecosystems propelled by machine learning. *Nat Rev Mater.* 2021;6(8):655-78.
- Montes de Oca Zapiain D, Wood MA, Lubbers N, Pereyra CZ, Thompson AP, Perez D. Training data selection for accuracy and transferability of interatomic potentials. *npj Comput Mater.* 2022;8(1):189.

Sutton C, Boley M, Ghiringhelli LM, Rupp M, Vreeken J, Scheffler M. Identifying domains of applicability of machine learning models for materials science. *Nat Commun.* 2020;11(1):4428.

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.

Thiyagalingam J, Shankar M, Fox G, Hey T. Scientific machine learning benchmarks. *Nat Rev Phys.* 2022;4(6):413-20.

Wang AY, Murdock RJ, Kauwe SK, Oliynyk AO, Gurlo A, Brgoch J, et al. Machine learning for materials scientists: An introductory guide toward best practices. *Chem Mater.* 2020;32(12):4954-65.

Ko TW, Finkler JA, Goedecker S, Behler J. General-purpose machine learning potentials capturing nonlocal charge transfer. *Acc Chem Res.* 2021;54(4):808-17.

Welborn M, Cheng L, Miller TF 3rd. Transferability in machine learning for electronic structure via the molecular orbital basis. *J Chem Theory Comput.* 2018;14(9):4772-9.

Benoit M, Amodeo J, Combettes S, Khaled I, Roux A, Lam J. Measuring transferability issues in machine-learning force fields: The example of gold–iron interactions with linearized potentials. *Mach Learn Sci Technol.* 2021;2(2):025003.

Suzuki T, Tamura R, Miyazaki T. Machine learning for atomic forces in a crystalline solid: Transferability to various temperatures. *Int J Quantum Chem.* 2017;117(1):33-9.

Deringer VL, Csányi G. Machine learning based interatomic potential for amorphous carbon. *Phys Rev B.* 2017;95(9):094203.

Pun GP, Yamakov V, Hickman J, Glaessgen EH, Mishin Y. Development of a general-purpose machine-learning interatomic potential for aluminum by the physically informed neural network method. *Phys Rev Mater.* 2020;4(11):113807.

Cools-Ceuppens M, Dambre J, Verstraelen T. Modeling electronic response properties with an explicit-electron machine learning potential. *J Chem Theory Comput.* 2022;18(3):1672-91.

Wang X, Wang Y, Zhang L, Dai F, Wang H. A tungsten deep neural-network potential for simulating mechanical property degradation under fusion service environment. *Nucl Fusion.* 2022;62(12):126013.

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.

Podryabinkin EV, Shapeev AV. Active learning of linearly parametrized interatomic potentials. *Comput Mater Sci.* 2017;140:171-80.

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.

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.

Mueller T, Hernandez A, Wang C. Machine learning for interatomic potential models. *J Chem Phys.* 2020;152(5).

Liang Q, Gongora AE, Ren Z, Tiihonen A, Liu Z, Sun S, et al. Benchmarking the performance of Bayesian optimization across multiple experimental materials science domains. *npj Comput Mater.* 2021;7(1):188.