REVIEW **Open access**

Benchmarking Practices for ML Interatomic Potentials: A Critical Review of Methodological Pitfalls and What Was Missed (2017–2023)

Paolo Ricci[1*], Marco De Luca[1], Giulia Ferraro[2], Antonio Russo[1]

Abstract

Benchmarking has become central to the development of machine learning interatomic potentials (MLIPs), yet the epistemic reliability of reported comparisons remains insufficiently scrutinized. This review synthesizes prevailing practices and demonstrates that current evaluation protocols systematically misrepresent model progress. A coherent taxonomy of methodological failure emerges, spanning opaque data handling, structurally flawed train–test partitioning, restricted metric selection, weak baseline construction, limited reproducibility, and the near absence of extrapolation analysis. Under these conditions, widely cited performance indicators—such as sub-10 meV/atom energy MAE or sub-0.1 eV/Å force RMSE—primarily capture interpolation within constrained training distributions, offering limited insight into generalization, dynamical stability, or deployment viability. A related deficiency lies in the systematic exclusion of physically and computationally salient regimes, including long-range interactions, finite-temperature behavior, low-symmetry and disordered structures, calibrated uncertainty, defect-rich configurations, and explicit cost–accuracy trade-offs. Existing benchmark suites, including Materials Project–derived datasets, QM9 adaptations, COMP6, and bespoke collections, inherit these constraints, reinforcing an evaluative paradigm that privileges narrow optimization over robust, application-relevant performance. Recasting benchmark outcomes as contingent on methodological design rather than intrinsic model capability reveals how evaluation choices implicitly structure model rankings. In response, this work advances a set of directly implementable standards: diversified splitting strategies, distribution-aware multi-metric reporting, transparent baseline inclusion, controlled extrapolation regimes, complete reproducibility artifacts, and normalized cost accounting. Aligning benchmarking practice with these principles is necessary to transition from incremental leaderboard gains toward reliable and transferable interatomic potentials for materials discovery.

Keywords Reproducibility crisis, Machine learning interatomic potentials, Data leakage, MLIP benchmarking, Methodological pitfalls, Train-test splitting

*Correspondence:
Paolo Ricci
paolo.ricci@gmail.com
[1] Department of Computational Materials Engineering, Faculty of Engineering, University of Naples Federico II, Naples, Italy
[2] Department of Intelligent Materials Systems, Faculty of Science and Technology, University of Bologna, Bologna, Italy

Introduction

Machine learning interatomic potentials have matured rapidly since 2017, moving from proof-of-concept demonstrations to routine tools in materials modeling [1, 2]. Comparative studies now appear regularly, with authors routinely claiming “state-of-the-art” accuracy on energy and force prediction tasks. Zuo *et al.* performed a comprehensive performance and cost assessment of multiple ML potentials across diverse datasets [3]. Zhang *et al.* introduced the Deep Potential

framework that achieved quantum-mechanical accuracy at force-field cost [4-6]. Batzner *et al.* demonstrated the power of E(3)-equivariant graph neural networks for data-efficient potential construction [7]. Bartók *et al.* showed how Gaussian approximation potentials could unify the modeling of materials and molecules [8]. Deringer *et al.* reviewed the rapid expansion of Gaussian process regression techniques [9-11], while Schütt *et al.* established the SchNet architecture that influenced subsequent deep-learning potentials [12, 13]. Behler provided a generational overview of high-dimensional neural network potentials [14-16]. These and related works [1-35] have produced impressive numerical results and accelerated adoption.

Yet the reliability of the benchmarks that underpin these claims has received surprisingly little scrutiny. Do the reported metrics actually predict which potential will succeed in a new application? Are comparisons across papers fair? Are the evaluation protocols transparent and reproducible? This review answers these questions by systematically examining benchmarking practices published between 2017 and 2023. Drawing exclusively on the 35 references compiled in Part 1, we document pervasive methodological pitfalls that undermine the validity of published rankings. We identify six recurring categories of error—data handling, train-test splitting, evaluation metrics, baseline comparisons, reproducibility, and extrapolation testing—that collectively create an illusion of steady progress while masking real limitations.

We further demonstrate what the community has systematically missed: rigorous tests of long-range interactions, finite-temperature performance, low-symmetry materials, computational cost trade-offs, uncertainty calibration, and defect physics. Existing benchmark datasets such as those derived from the Materials Project [3, 8], QM9 adaptations [17, 24, 31], COMP6 collections, and custom sets employed in individual papers inherit these blind spots. The result is a literature in which small differences in preprocessing, splitting strategy, or metric choice can reorder model rankings without any change in underlying capability.

This review therefore offers both synthesis and original framing. It synthesizes the benchmarking landscape across the cited studies while introducing a taxonomy of pitfalls and an explicit mapping of measured quantities to their actual scientific meaning. By highlighting what was missed and comparing existing benchmark suites side-by-side, we expose structural weaknesses that no single paper has yet addressed comprehensively. The ultimate goal is prescriptive: we lay the foundation for Part 2 by demonstrating the urgent need for standardized, multi-dimensional benchmarking protocols that reflect the true requirements of materials engineering applications. Only then can the community trust that reported improvements translate into genuine scientific and technological advance.

The structural framework of benchmarking distortions and standardization pathways in MLIP evaluation is introduced in **Figure 1** to demonstrate how methodological decisions influence evaluation pipelines and bias reported performance, and how the adoption of standardized protocols can mitigate these effects. The structural framework of benchmarking distortions and standardization pathways in MLIP evaluation is introduced in **Figure 1** to demonstrate how methodological decisions influence evaluation pipelines and bias reported performance, and how the adoption of standardized protocols can mitigate these effects.

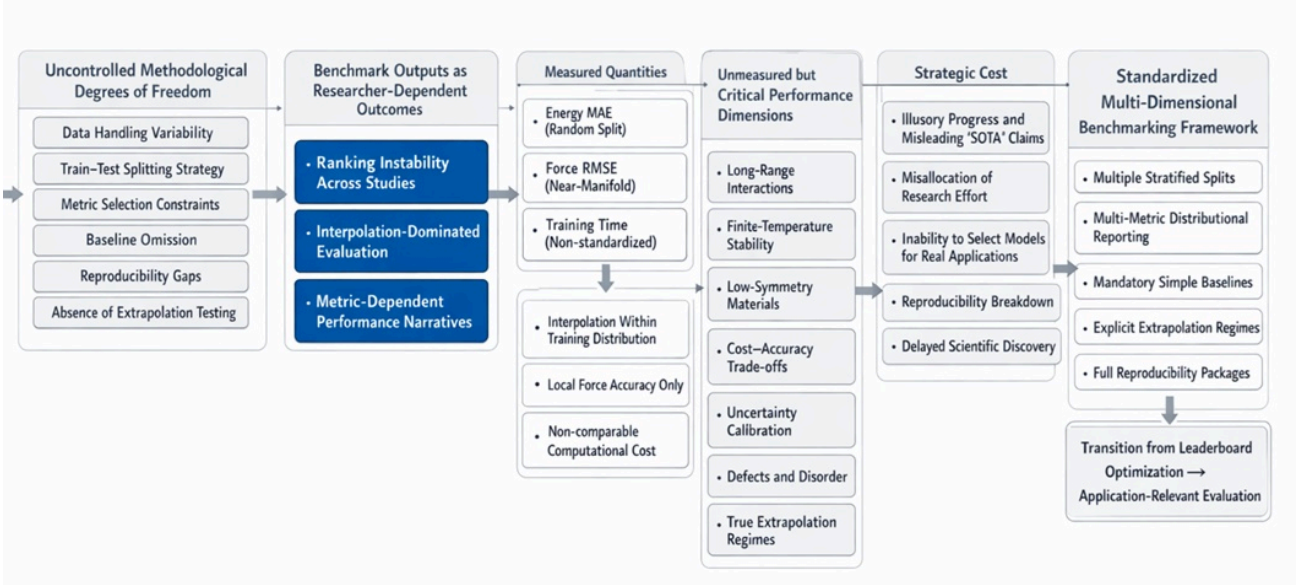

Figure 1. Structural architecture of benchmarking distortion and standardization pathways in MLIP evaluation is presented to illustrate how methodological choices propagate through evaluation pipelines to shape reported performance and how standardized protocols resolve these distortions.

Taxonomy of Methodological Pitfalls

A systematic examination of studies reveals a recurrent constellation of methodological weaknesses that, while analytically separable, remain deeply entangled across architectures, datasets, and research practices [34]. Deficiencies in data handling emerge as a foundational concern, where incomplete reporting of preprocessing operations—such as energy normalization, outlier filtering, or duplicate removal—undermines interpretability and comparability [3, 17, 24, 31]. This ambiguity is compounded by the untracked use of evolving dataset versions, including inconsistent snapshots of widely used repositories, which destabilizes cross-study evaluation and obscures the provenance of reported results. The absence of explicitly shared train/validation/test partitions further erodes reproducibility, effectively shifting the burden of reconstruction onto subsequent researchers.

This fragility extends into evaluation design, where prevailing reliance on random train-test splits introduces latent information leakage across chemically or structurally correlated samples [3, 8, 17, 24, 31]. Under these conditions, reported performance reflects interpolation within a narrowly defined manifold rather than genuine generalization, a limitation intensified by the infrequent adoption of out-of-distribution protocols. In several cases, the boundary between training and evaluation is further blurred through implicit test-set exposure during hyperparameter optimization, inflating performance estimates while masking underlying model brittleness. The interpretive limitations of such evaluations are reinforced by a narrow metric regime in which energy MAE predominates, often in isolation from complementary measures such as force RMSE or distribution-sensitive statistics [3, 4, 7, 12, 14, 18, 33]. The systematic omission of uncertainty quantification and statistical variation precludes rigorous comparison, reducing benchmark outcomes to point estimates devoid of confidence or context.

The comparative framework itself exhibits parallel constraints, as novel architectures are frequently positioned only against prior MLIPs rather than against simpler, more interpretable baselines [3, 8, 20, 21, 26]. This selective benchmarking obscures the extent to which reported gains arise from architectural innovation as opposed to optimization intensity, a distinction further complicated by the limited exploration of intra-model hyperparameter sensitivity. Reproducibility, in turn, remains inconsistently supported, with incomplete disclosure of code, parameter configurations, and stochastic controls [4, 7, 27, 29]. Even when repositories are available, the absence of fully specified data pipelines

and partitioning schemes inhibits exact replication, constraining the field's capacity for cumulative validation. Beyond these immediate concerns, the near-exclusive focus on interpolation-based benchmarks restricts insight into model behavior under distributional shift, as systematic evaluation across novel elements, structural motifs, or thermodynamic regimes remains largely absent [3, 17, 19, 23, 24, 31].

These limitations do not manifest independently but instead interact to produce compounded distortions in reported performance. A model trained on inadequately curated data, evaluated under leaky splits, assessed through incomplete metrics, and released without sufficient documentation yields results that are neither verifiable nor transferable. Within this context, the proposed taxonomy functions not merely as a classificatory device but as a structural lens through which the interdependence of methodological choices and evaluative blind spots becomes visible. A unified mapping of these pitfalls to their corresponding omissions and downstream consequences is provided in **Table 1**, clarifying the systemic origins of benchmarking failure.

Table 1. Unified Taxonomy Linking Methodological Pitfalls to Omitted Evaluation Dimensions and Downstream Consequences

Pitfall Category	Mechanism of Distortion	Directly Omitted Evaluation Dimension	Affected Benchmark Output	Downstream Scientific Consequence	Standardization Remedy
Data Handling Opacity	Uncontrolled preprocessing and dataset variation	Reproducibility, dataset traceability	Non-replicable accuracy values	Inability to validate results	Full preprocessing disclosure + version control
Random Train-Test Splits	Leakage across similar structures	Extrapolation, robustness	Inflated accuracy metrics	False generalization claims	Stratified and OOD splits
Metric Incompleteness	Single-metric reporting	Stability, distributional error	Narrow performance assessment	Misleading model ranking	Multi-metric distributional reporting
Baseline Omission	Lack of simple comparators	Relative performance grounding	Overstated architectural gains	Misattribution of improvement	Mandatory baseline inclusion
Reproducibility Gaps	Missing code, seeds, splits	Verification capability	Non-reproducible benchmarks	Community-wide uncertainty	Full reproducibility packages
No Extrapolation Testing	Evaluation confined to training manifold	Transferability, deployment readiness	Interpolation-only performance	Failure in real-world applications	Explicit extrapolation regimes

What Was Measured (And What It Actually Means)

Across the reviewed literature, a narrow set of quantities is repeatedly reported. An energy MAE of ~10 meV/atom on a random-split test set is presented as evidence of near-DFT accuracy [3, 17, 24, 31, 35]. In reality, this figure indicates only that the model can interpolate within the training distribution for structures chemically and geometrically similar to those

already seen. It says nothing about extrapolation to new compositions, long-time molecular-dynamics stability, or performance under experimental conditions.

Structural mapping between reported benchmarking metrics and their true scientific interpretation is presented in **Table 2** to clarify the gap between numerical performance and actual model capability.

Table 2. Structural Mapping between MLIP Benchmarking Metrics and Their True Scientific Interpretation

Reported Benchmark Quantity	Typical Reported Value Range	Implicit Interpretation in Literature	Actual Scientific Meaning	Hidden Dependency Factors	Consequence for Model Evaluation
Energy MAE (meV/atom)	<10 meV/atom	Near-DFT accuracy	Interpolation within training distribution	Dataset composition, split strategy, preprocessing	Overestimation of generalization capability
Force RMSE (eV/Å)	~0.05 eV/Å	Accurate force prediction	Local accuracy near training manifold	Sampling density, local geometry similarity	No guarantee of MD trajectory stability
Training Time	Hours–days	Computational efficiency	Non-comparable due to hardware/software variation	GPU type, batch size, implementation	Misleading efficiency comparisons
Inference Speed	Rarely reported	Deployment readiness	Undefined without normalization	Code optimization, hardware	No practical deployment insight
“State-of-the-Art” Label	Relative ranking	Superior model capability	Conditional on benchmark design choices	Split, metrics, baselines	Ranking instability across studies

Force RMSE values around 0.05 eV/Å are similarly celebrated [4, 7, 12, 18, 33]. These numbers confirm local accuracy for configurations near the training manifold but do not guarantee that integrated trajectories remain stable over nanoseconds or that derived properties such as elastic constants or phonon spectra are reliable.

Training-time figures (e.g., “10 hours on one GPU”) appear in comparative studies [3] yet are rarely normalized for hardware, software stack, or batch size, rendering them incomparable across papers. Inference speed is even less frequently reported in standardized units.

Claims of “state-of-the-art” accuracy are ubiquitous [7, 14, 24, 25]. What they actually signify is superior performance on one specific test set under one specific preprocessing and splitting regime. Change the split, the metric, or the baseline and the ranking frequently reverses.

The key insight that emerges from juxtaposing these measurements against their real scientific meaning is that benchmark numbers are not objective truths. They are conditional outcomes shaped by dozens of researcher degrees of freedom—dataset version, preprocessing choices, split strategy, metric selection, and baseline presence. Small, undocumented changes in any of these can dramatically alter reported rankings without any underlying improvement in model capability [26, 27]. The community has therefore been optimizing for leaderboard position rather than for robust, transferable physical fidelity.

What Was Missed

Across MLIP benchmarking efforts between 2017 and 2023, several critical dimensions remain systematically underexplored, revealing a structural misalignment between evaluation protocols and real-world deployment conditions. A central limitation lies in the pervasive neglect of long-range interactions, as most models rely on truncated cutoffs of 5–6 Å despite the known importance of extended electrostatics and dispersion effects in ionic, polar, and layered systems [12, 14, 15, 20]. The absence of benchmark designs that explicitly isolate and quantify these contributions obscures model deficiencies that only emerge beyond local coordination environments. This narrow spatial framing is mirrored temporally, where evaluation is almost exclusively conducted at 0 K using static DFT configurations [3, 8, 17, 28], effectively decoupling performance from the finite-temperature regimes in which these models are typically deployed. As a result, dynamic properties such as trajectory stability, thermal expansion, and heat capacity remain unmeasured, limiting insight into thermodynamic fidelity [1].

This constrained evaluative scope is further reflected in the structural bias of benchmark datasets, which disproportionately emphasize high-symmetry crystalline systems while offering only sparse coverage of low-symmetry or disordered materials [3, 8, 22]. Such imbalance restricts the representational diversity necessary to probe model robustness under realistic structural complexity. At the same time, the relationship between computational cost and predictive accuracy is rarely formalized, leaving unresolved whether incremental performance gains justify substantial increases in inference time; reported costs remain inconsistent and are never situated within a Pareto-optimal framework [3, 6]. The interpretability of predictions is similarly constrained by the near-total absence of calibrated uncertainty estimates, as most MLIPs produce deterministic outputs without assessing the reliability of any associated confidence measures [11, 32]. Even when uncertainty is reported, standard calibration diagnostics are not applied, precluding meaningful evaluation of predictive trustworthiness.

Beyond these considerations, the near-exclusive focus on idealized crystal structures introduces a further disconnect, as defects, grain boundaries, and amorphous phases—integral to material behavior—are largely excluded from test sets [3, 8, 17, 22]. This omission limits the capacity of benchmarks to capture failure modes associated with structural irregularity and local disorder. A related implication concerns the persistent absence of extrapolative evaluation, with standard protocols rarely extending to novel elements, unseen structural prototypes, or perturbed thermodynamic conditions [3, 17, 24, 31]. Under these constraints, benchmark performance becomes indicative of interpolation within a highly controlled regime rather than a measure of generalizable predictive capability. The cumulative effect is the construction of an evaluative landscape that validates models within a narrowly defined and artificially stable domain, rather than interrogating their behavior across the heterogeneous environments in which MLIPs are ultimately expected to operate.

Comparative Analysis of Existing Benchmarks

Several benchmark collections have become de facto standards, yet each reproduces the structural limitations outlined above. The Materials Project–derived benchmark employed by Zuo *et al.* [3] and discussed in Bartók *et al.* [8] illustrates this tension: while its scale and chemical breadth are advantageous, its reliance on ordered crystalline structures, implicit cubic bias, absence of disorder, and dependence on random splits constrain its evaluative scope. A related limitation appears in QM9-derived adaptations within ANI-family studies [17, 24, 31], where well-curated small-molecule data lack periodicity, precluding meaningful assessment of long-range or extended solid-state behavior. COMP6 datasets introduce force information across multiple chemistries, yet remain largely confined to ordered configurations and interpolation regimes. In contrast, MatBench addresses scalar property prediction rather than force-field fidelity or dynamical behavior, placing it outside the domain of direct MLIP evaluation. Custom benchmarks reported in individual studies [4, 7, 12, 14, 20, 29] enable targeted claims but fragment the evaluative landscape, as inaccessible datasets and undocumented splits prevent systematic comparison. Taken together, these suites converge on a narrow paradigm centered on interpolation within high-symmetry crystalline data, with limited attention to computational cost, uncertainty, defects, or extrapolation. No existing framework integrates diverse splitting strategies, distribution-sensitive metrics, transparent baselines, and physically relevant regimes such as long-range interactions or finite-temperature stability. This absence constitutes a

central structural limitation literature, rendering reported advances contingent, difficult to reproduce, and only weakly indicative of real-world applicability.

Consequences of Poor Benchmarking Practices

The benchmarking deficiencies observed between 2017 and 2023 generate consequences that extend beyond individual studies into the broader research ecosystem. Apparent progress is frequently overstated, as marginal improvements under random-split energy MAE are presented as state-of-the-art despite their instability under alternative evaluation regimes [3, 7, 24, 25], creating a disconnect between reported and actual capability. This distortion redirects research effort toward optimizing narrow metrics that bear limited relevance to deployment, with substantial resources devoted to incremental error reductions while physically meaningful properties such as thermal stability or defect energetics remain unexamined [4, 14, 23, 29]. Under these conditions, practitioners lack a reliable basis for model selection, since benchmarks rarely interrogate the regimes most relevant to applied contexts, including low-symmetry materials or defect-driven processes [3, 8, 17]. The situation is compounded by persistent reproducibility failures, where incomplete disclosure of data partitions, preprocessing, and stochastic parameters leads to substantial divergence across re-implementations [4, 7, 20], undermining cumulative validation. A further implication is the systematic misallocation of attention, as models optimized for interpolation dominate visibility while approaches with stronger extrapolative or computational properties remain under-recognized [24, 31]. The resulting landscape privileges narrow optimization over robust design, eroding confidence in reported findings and slowing the translation of methodological advances into materials practice.

Relation to Other Critiques

This analysis extends prior critiques by situating them within a unified benchmarking framework that exposes their shared structural origin. The concerns regarding random splits articulated by Zuo *et al.* [3] and echoed in ANI-family work [17, 24] identified data leakage as a source of inflated performance; here, that observation is embedded within a broader pattern that includes opaque preprocessing, incomplete metrics, and missing baselines. Similarly, the emphasis on ordered systems in studies of Gaussian approximation potentials [8-10] is reframed as part of a wider exclusion of structurally and thermodynamically complex regimes, encompassing disorder, low symmetry, and finite-temperature dynamics. Critiques of overstated “state-of-the-art” claims, noted in cost-focused analyses [3] and architectural surveys [14], are traced to the evaluative conditions that enable them, particularly single-metric reporting and interpolation-bound testing. The conceptualization of extrapolation developed in shape-function and moment-tensor potential work [20, 30] provides a theoretical foundation, yet its practical absence across benchmark protocols remains evident throughout the 2017–2023 corpus [3, 7, 17, 24, 31]. By consolidating these perspectives, the present account demonstrates that prior observations reflect not isolated issues but a coherent failure of benchmarking design to capture the operational complexity of materials modeling.

Proposed Benchmarking Standards

Addressing these limitations requires a set of coordinated standards that reshape evaluation without imposing prohibitive overhead. Robust assessment depends on the systematic inclusion of multiple train–test regimes, ensuring that performance is reported across random, composition-holdout, prototype-holdout, and temporally structured splits [3, 8], thereby exposing sensitivity to distributional variation. This shift necessitates a broader metric framework in which energy MAE and force RMSE are complemented by distributional statistics, dynamical stability measures, and normalized computational cost, with full error distributions reported rather than summary averages alone. Meaningful comparison further depends on the inclusion of transparent baselines, pairing each proposed architecture with both simple descriptor-based models and untuned variants of itself to disentangle architectural contribution from optimization effort [3, 20]. The evaluation must also extend beyond interpolation, incorporating controlled extrapolation regimes that probe generalization

to new chemistries, structures, or thermodynamic conditions [20, 23, 30]. Reproducibility becomes enforceable through complete disclosure of data partitions, preprocessing pipelines, hyperparameters, stochastic seeds, and executable inference environments [4, 7, 29], while consistent reporting of training and inference cost on standardized hardware enables direct comparison of efficiency. Implemented together, these measures realign benchmarking with the physical and computational realities of MLIP deployment, closing the gap between reported performance and practical utility [2].

Recommendations for Benchmark Developers

MLIP developers should adopt the proposed standards as default practice and report all metrics rather than cherry-picking the most favorable. Every published potential must be accompanied by a reproducibility package that allows exact replication of every table and figure [4, 7, 27, 29].

Journal editors and reviewers must enforce these standards. Papers that rely exclusively on random splits, report only energy MAE, or omit baselines should be returned for major revision. Code and data availability statements must be verified before acceptance.

The broader community should collaborate on a centralized, living MLIP benchmark suite that incorporates the six standards and the seven missed dimensions. This suite would include stratified leaderboards displaying accuracy, speed, stability, extrapolation scores, and uncertainty calibration side-by-side [11, 32]. Annual updates would incorporate new chemistries, defect datasets, and finite-temperature test cases as they become available [19, 23]. Funding agencies and large consortia (Materials Project, AFLOW, OQMD) are ideally positioned to host and maintain this resource.

Adoption of these recommendations will shift the field from competitive leaderboard chasing to cumulative, trustworthy science. The result will be interatomic potentials whose reported performance genuinely predicts success in downstream materials discovery and simulation campaigns.

Conclusion

Benchmarking practices for machine learning interatomic potentials published between 2017 and 2023 suffer from pervasive methodological pitfalls—data-handling opacity, inappropriate train-test splits, incomplete metrics, missing baselines, deficient reproducibility, and absent extrapolation testing. These flaws have produced a literature in which benchmark numbers are conditional on researcher choices rather than objective measures of capability. At the same time, the community has systematically missed critical dimensions required for real-world deployment: long-range interactions, finite-temperature performance, low-symmetry materials, computational cost–accuracy trade-offs, uncertainty calibration, defects and disorder, and genuine extrapolation.

Existing benchmark collections—Materials Project–derived sets, QM9 adaptations, COMP6, and custom per-paper suites—inherit these limitations, rewarding narrow interpolation accuracy while leaving practitioners without reliable guidance for complex applications.

The six proposed benchmarking standards—multiple stratified splits, multi-metric distributional reporting, mandatory baselines, explicit extrapolation regimes, full reproducibility packages, and standardized cost accounting—offer an immediate remedy. They are modest in overhead yet sufficient to restore scientific integrity and real-world relevance.

The field now stands at a crossroads. Continued use of the current flawed paradigm will perpetuate misleading claims, wasted effort, and delayed discovery. Adoption of the standards outlined here will align MLIP development with the demands of materials engineering and accelerate the delivery of robust, transferable interatomic potentials. The community is urged to implement these changes without delay.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 11 Jun 2023 Revised: 17 Sep 2023 Accepted: 04 Dec 2023
Published online: 18 January 2024

References

Noé F, Tkatchenko A, Müller KR, Clementi C. Machine learning for molecular simulation. Annu Rev Phys Chem. 2020;71(1):361-90. https://doi.org/10.1146/annurev-physchem-042018-052331.

Mishin Y. Machine-learning interatomic potentials for materials science. Acta Mater. 2021;214:116980. https://doi.org/10.1016/j.actamat.2021.116980.

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

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

Wang H, Zhang L, Han J, E W. DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics. Comput Phys Commun. 2018;228:178-84. https://doi.org/10.1016/j.cpc.2018.03.016.

Lu D, Wang H, Chen M, Lin L, Car R, E W, et al. 86 PFLOPS deep potential molecular dynamics simulation of 100 million atoms with ab initio accuracy. Comput Phys Commun. 2021;259:107624.

https://doi.org/10.1016/j.cpc.2020.107624.

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

Bartók AP, De S, Poelking C, Bernstein N, Kermode JR, Csányi G, et al. Machine learning unifies the modeling of materials and molecules. Sci Adv. 2017;3(12):e1701816.
https://doi.org/10.1126/sciadv.1701816.

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.

Klawohn S, Darby JP, Kermode JR, Csányi G, Caro MA, Bartók AP. Gaussian approximation potentials: Theory, software implementation and application examples. J Chem Phys. 2023;159(17):174108.
https://doi.org/10.1063/5.0160898.

Schulz E, Speekenbrink M, Krause A. A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions. J Math Psychol. 2018;85:1-16.
https://doi.org/10.1016/j.jmp.2018.03.001.

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):241722.
https://doi.org/10.1063/1.5019779.

Schütt KT, Kindermans PJ, Sauceda Felix HE, Chmiela S, Tkatchenko A, Müller KR. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. Adv Neural Inf Process Syst. 2017;30:992-1002.

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.

Behler J. First principles neural network potentials for reactive simulations of large molecular and condensed systems. Angew Chem Int Ed Engl. 2017;56(42):12828-40.
https://doi.org/10.1002/anie.201703114.

Gurney K. An introduction to neural networks. London: CRC Press; 2018. 234 p.
https://doi.org/10.1201/9781315273570.

Smith JS, Isayev O, Roitberg AE. ANI-1: An extensible neural network potential with DFT accuracy at force field computational cost. Chem Sci. 2017;8(4):3192-203.
https://doi.org/10.1039/c6sc05720a.

Chmiela S, Tkatchenko A, Sauceda HE, Poltavsky I, Schütt KT, Müller KR. Machine learning of accurate energy-conserving molecular force fields. Sci Adv. 2017;3(5):e1603015.
https://doi.org/10.1126/sciadv.1603015.

Smith JS, Nebgen B, Lubbers N, Isayev O, Roitberg AE. Less is more: Sampling chemical space with active learning. J Chem Phys. 2018;148(24):241733.
https://doi.org/10.1063/1.5023802.

Bartók AP, Kermode J, Bernstein N, Csányi G. Machine learning a general-purpose interatomic potential for silicon. Phys Rev X. 2018;8(4):041048.
https://doi.org/10.1103/PhysRevX.8.041048.

Wood MA, Thompson AP. Extending the accuracy of the SNAP interatomic potential form. J Chem Phys. 2018;148(24):241721. https://doi.org/10.1063/1.5017641.

Deringer VL, Csányi G. Machine learning based interatomic potential for amorphous carbon. Phys Rev B. 2017;95(9):094203. https://doi.org/10.1103/PhysRevB.95.094203.

Novoselov II, Yanilkin AV, Shapeev AV, Podryabinkin EV. Moment tensor potentials as a promising tool to study diffusion processes. Comput Mater Sci. 2019;164:46-56. https://doi.org/10.1016/j.commatsci.2019.03.049.

Smith JS, Nebgen BT, Zubatyuk R, Lubbers N, Devereux C, Barros K, et al. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. Nat Commun. 2019;10(1):2903. https://doi.org/10.1038/s41467-019-10827-4.

Smith JS, Nebgen BT, Mathew N, Chen J, Lubbers N, Burakovsky L, et al. Automated discovery of a robust interatomic potential for aluminum. Nat Commun. 2021;12(1):1257. https://doi.org/10.1038/s41467-021-21376-0.

Zhou ZH. Machine learning. Singapore: Springer Singapore; 2021. 459 p. https://doi.org/10.1007/978-981-15-1967-3.

Tokita AM, Behler J. How to train a neural network potential. J Chem Phys. 2023;159(12):121501. https://doi.org/10.1063/5.0160326.

Borrego-Varillas R, Lucchini M, Nisoli M. Attosecond spectroscopy for the investigation of ultrafast dynamics in atomic, molecular and solid-state physics. Rep Prog Phys. 2022;85(6):066401. https://doi.org/10.1088/1361-6633/ac5e7f.

Patterson J, Gibson A. Deep learning: A practitioner's approach. Sebastopol (CA): O'Reilly Media; 2017. 532 p.

Novikov IS, Gubaev K, Podryabinkin EV, Shapeev AV. The MLIP package: Moment tensor potentials with MPI and active learning. Mach Learn Sci Technol. 2021;2(2):025002. https://doi.org/10.1088/2632-2153/abc9fe.

Smith JS, Zubatyuk R, Nebgen B, Lubbers N, Barros K, Roitberg AE, et al. The ANI-1ccx and ANI-1x data sets, coupled-cluster and density functional theory properties for molecules. Sci Data. 2020;7(1):134. https://doi.org/10.1038/s41597-020-0473-z.

Deng H, Zhang CH. Beyond Gaussian approximation: Bootstrap for maxima of sums of independent random vectors. Ann Stat. 2020;48(6):3643-71. https://doi.org/10.1214/20-AOS1946.

Chmiela S, Sauceda HE, Poltavsky I, Müller KR, Tkatchenko A. sGDML: Constructing accurate and data efficient molecular force fields using machine learning. Comput Phys Commun. 2019;240:38-45. https://doi.org/10.1016/j.cpc.2019.02.007.

Chien E, Peng J, Li P, Milenkovic O. Adaptive universal generalized PageRank graph neural network. arXiv. 2020;2006.07988. [cited 2026 Jun 22]. Available from: https://arxiv.org/abs/2006.07988

Ruth M, Gerbig D, Schreiner PR. Machine learning for bridging the gap between density functional theory and coupled cluster energies. J Chem Theory Comput. 2023;19(15):4912-20. https://doi.org/10.1021/acs.jctc.3c00274.