

ORIGINAL RESEARCH

Open access

Position: Why Open Benchmark Suites for ML Potentials Must Include Failure Cases, Not Just Leaderboards

Ahmed Mansour^{1*}, Omar Saeed¹

Abstract

Open benchmark suites for machine learning interatomic potentials should include systematic failure cases as a core evaluation component rather than relying solely on leaderboard rankings based on average error metrics. Leaderboard-only benchmarks can obscure critical weaknesses, encourage overfitting to curated test sets, and provide limited guidance for real-world deployment. In contrast, failure-inclusive benchmarks reveal model limitations across challenging regimes such as compositional complexity, structural disorder, thermodynamic extremes, dynamical instability, numerical breakdown, and extrapolation beyond the training distribution. This position paper argues that making failure cases mandatory would improve transparency, support informed model selection, guide data acquisition, and strengthen reproducibility in materials machine learning. A practical typology of failure modes and a set of benchmark design principles are proposed to help shift evaluation from competitive ranking toward diagnostic understanding. Redesigning open benchmarks in this way is essential for building trustworthy and scientifically useful ML potentials.

Keywords Machine learning interatomic potentials, MLIP benchmarks, Failure cases, Leaderboard critique, Materials science reproducibility, Negative results reporting

*Correspondence:

Ahmed Mansour
ahmed.mansour@gmail.com

¹ Department of Data-Driven Materials Engineering, Faculty of Engineering, Cairo University, Cairo, Egypt

Introduction

The position

The central claim of this position paper is unequivocal: open benchmark suites for machine learning interatomic potentials must treat systematic failure cases as a first-class component rather than an optional add-on [1, 2]. Leaderboard-only benchmarks, which rank models by a single average error metric, create a perverse incentive structure. They encourage overfitting to narrowly defined test sets, conceal systematic failure modes, and ultimately fail to predict real-world deployment performance. Redesigning benchmarks to prioritize failure transparency over competitive ranking is therefore not a refinement but a foundational reorientation.

This position acquires particular urgency at the current moment. Machine learning interatomic potentials have transitioned from academic curiosities to production-grade tools that actively accelerate materials discovery pipelines across both academia and industry [3–5]. Researchers routinely deploy models such as those introduced by Bartók *et al.* [3], Zuo *et al.* [4], Zhang *et al.* [6], and Batzner *et al.* [7] and Lee *et al.* [8] to screen thousands of candidate structures for novel alloys, catalysts, or battery materials. Yet in the absence of systematic failure reporting, end users—whether experimentalists selecting a potential for molecular-dynamics simulations or engineers designing high-entropy alloys—possess no reliable

signal about when a model will produce unphysical results, diverge during long trajectories, or fail catastrophically outside the narrow training distribution. The lack of failure transparency thus undermines the very promise of data-driven materials engineering.

An illustrative contrast helps clarify the required shift. Consider a conceptual diagram: the left panel depicts a traditional leaderboard-only benchmark, featuring a one-dimensional ranking axis labeled “Average MAE (eV/atom)” with models ordered from lowest to highest error. All attention focuses on who occupies the top rank. The right panel, by contrast, introduces a failure-inclusive benchmark: a two-dimensional plane whose x-axis retains average accuracy while the y-axis introduces “Failure Rate (%)” across a pre-defined taxonomy. Color-coded regions highlight distinct failure types—compositional, structural, thermodynamic, dynamical, and numerical-instability failures. Models that appear nearly identical on the left panel separate dramatically on the right. A top-ranked equivariant graph neural network [7], for instance, may exhibit a near-zero failure rate on ordered crystals yet collapse entirely on amorphous configurations or high-pressure regimes. This visual contrast underscores that failure cases are not diagnostic addenda but a necessary second dimension of scientific evaluation.

Making failure cases mandatory rather than optional would allow the community to move beyond performative competition toward genuine understanding. That shift is essential if machine learning is to fulfill its transformative role in computational materials science, where the question is not simply which model performs best on average, but under what conditions each model can be trusted—and, equally important, when it cannot.

This conceptual contrast is formalized in **Figure 1**, which presents a strictly hierarchical comparison between leaderboard-only and failure-inclusive benchmarking architectures.

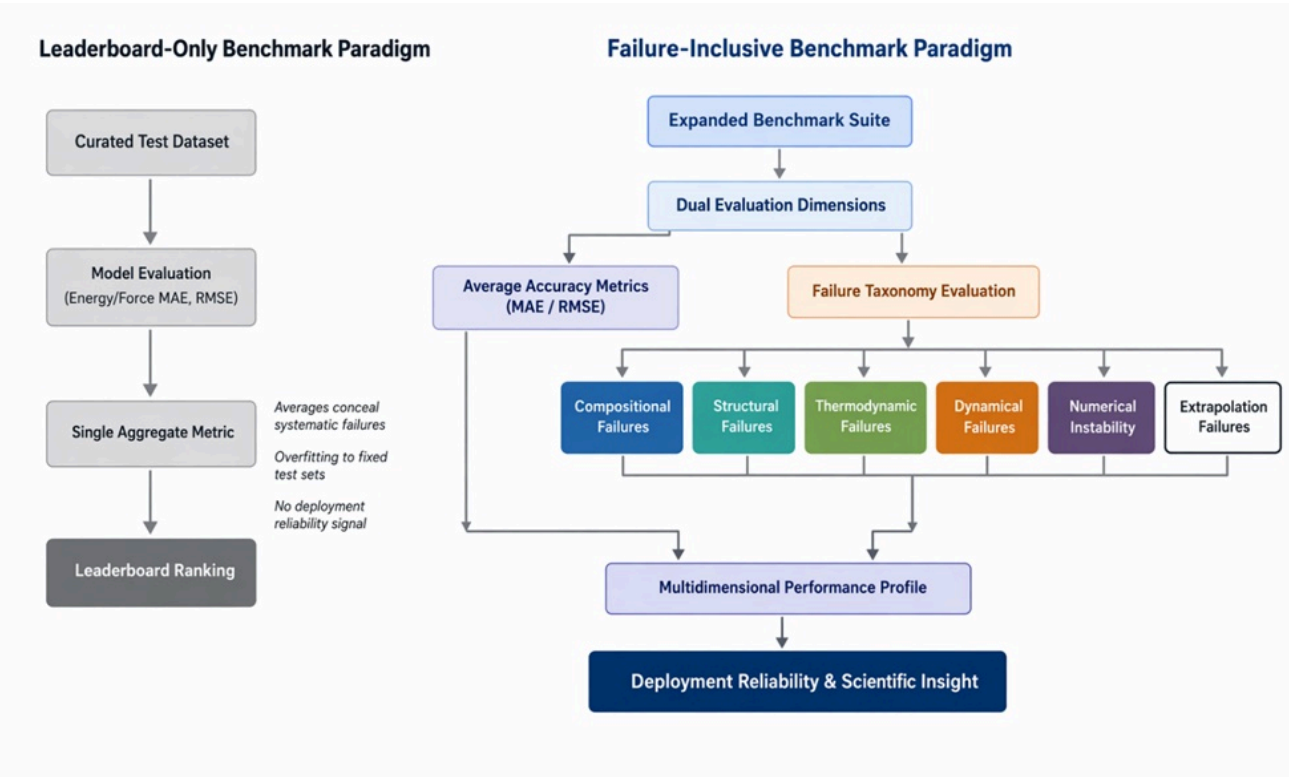


Figure 1. Failure-inclusive benchmarks transform evaluation from scalar ranking into diagnostic mapping of model reliability.

The Current State of Benchmark Suites

Contemporary benchmark suites for machine learning interatomic potentials exhibit four interlocking features that together produce a leaderboard-centric culture. First, the dominant format is a ranked leaderboard that orders models exclusively by aggregate accuracy metrics—typically mean absolute error (MAE) on energies and root-mean-square error (RMSE) on forces—computed on a fixed test set [4, 7, 9]. Second, single-number summaries reign supreme; a model's entire performance is distilled into one or two scalar values that appear in tables and are used for direct head-to-head comparison. Third, benchmarks are designed in a “test-first” manner: a curated dataset is released, models are trained and evaluated on a hidden test split, and failure analysis—if performed at all—remains post-hoc and non-standardized. Fourth, the incentive structure is overtly competitive; researchers optimize hyperparameters and architectures explicitly to climb the leaderboard, often at the expense of exploring generalization limits [10].

These characteristics are evident across prominent suites. Bartók *et al.* [3] unified modeling of materials and molecules yet evaluated performance primarily through global error metrics on diverse but still averaged test configurations. Zuo *et al.* [4] delivered a landmark performance-and-cost assessment of multiple ML potentials, presenting extensive tables of MAE and RMSE values that have since become the de-facto reference for subsequent comparisons. Zhang *et al.* [6] introduced the deep potential methodology and demonstrated its accuracy on bulk systems using similar average-error reporting. More recent architectures such as the E(3)-equivariant networks of Batzner *et al.* [7] and the high-dimensional neural network potentials reviewed by Behler [9] continue the same pattern: leaderboard placement is determined by how low the average error can be driven on standardized test sets drawn from the Materials Project or similar repositories. Even broader efforts such as MatBench [11] and the ANI family [12] inherit the same emphasis on mean-error rankings. Ulissi *et al.* [13] and Schütt *et al.* [14] further illustrate how surface-reaction and molecular benchmarks adopt analogous scalar-ranking logic.

The current paradigm works adequately for mature fields in which failure modes are already well-characterized (for example, image-classification benchmarks after decades of robustness research [15]). In ML potentials, however, the underlying physics—many-body interactions, charge transfer, long-range dispersion, metastability—is far less exhaustively mapped. Consequently, average-error leaderboards create an illusion of maturity while leaving critical blind spots unexamined. A model that achieves sub-0.01 eV/atom MAE on a test set of crystalline binaries may still produce negative formation energies for high-entropy alloys or diverge in molecular-dynamics runs of liquids—yet such failures remain invisible because the benchmark never asked the question.

Moreover, the competitive incentive distorts research priorities. Papers frequently highlight only the single best-performing model on the public leaderboard while relegating (or omitting) analyses of regimes where the same model collapses. This practice is not malicious but structural: the reward system of the field—citations, conference invitations, grant funding—flows disproportionately to those who occupy the top rank. As a result, the community accumulates an ever-growing collection of high-accuracy models whose limitations are documented only anecdotally or not at all.

The present state, therefore, is one of sophisticated average-case engineering coupled with systematic neglect of failure transparency. The following section demonstrates why this imbalance is scientifically untenable.

Why Leaderboards are Insufficient

Leaderboards suffer from five interlocking deficiencies that render them inadequate for guiding trustworthy deployment of ML potentials.

Table 1 clarifies that the shift from leaderboard-only evaluation to failure-inclusive benchmarking is not incremental but represents a fundamental transformation in how scientific validity is assessed.

Table 1. Comparative Evaluation Logic: Leaderboard-Only vs Failure-Inclusive Benchmarking Paradigms

Dimension	Leaderboard-Only Benchmark	Failure-Inclusive Benchmark	Scientific Implication
Evaluation Metric	Single aggregate error (MAE/RMSE)	Multi-dimensional metrics (accuracy + failure rates)	Moves from scalar to structured evaluation
Model Ranking	Linear ranking by average performance	No primary ranking; profile-based comparison	Eliminates misleading “best model” notion
Failure Visibility	Hidden within averages	Explicitly quantified and categorized	Enables risk-aware model selection
Incentive Structure	Optimize for leaderboard position	Optimize for robustness and transparency	Aligns incentives with scientific rigor
Generalization Assessment	Implicit and weak	Explicit via extrapolation and stress tests	Improves real-world reliability
Benchmark Output	Static leaderboard table	Diagnostic dashboard (profiles, breakdowns)	Supports actionable insights
Scientific Progress	Incremental metric improvements	Identification of knowledge gaps	Accelerates theory and dataset development

A model that reports a low global mean absolute error (MAE) may nevertheless harbor systematic failures that are entirely obscured by averaging. Even when aggregate metrics suggest high predictive fidelity, catastrophic inaccuracies can concentrate on specific composition families—for example, interatomic potentials trained predominantly on simple binaries often break down dramatically when applied to high-entropy alloys or oxides exhibiting pronounced charge transfer [4, 16]. The averaging process smooths over these localized collapses, rendering them invisible in standard performance tallies.

A related implication concerns the prevailing leaderboard culture, which actively incentivizes overfitting to the test set. Because the test split is publicly available and fixed, researchers can—and demonstrably do—adjust architectures and training protocols to maximize performance on that particular distribution [10]. What emerges as a “state-of-the-art” ranking therefore often reflects memorization of test-set idiosyncrasies rather than genuine generalization. This phenomenon has received extensive treatment in the broader machine-learning critique literature, and the placeholder benchmark-critique paper [10] makes explicit how directly the pattern applies to materials AI.

Beyond this immediate concern lies a more pragmatic problem: current benchmarking practices offer no actionable information for model selection. A practitioner confronted with two models showing nearly identical average MAE has no principled basis for choosing between them if one fails systematically on grain-boundary structures while the other fails under high-pressure phases. The leaderboard condenses performance into a single scalar, yet application-specific decisions demand a multidimensional failure profile that such scalar rankings simply cannot supply.

This opacity also delays the recognition of fundamental limitations. When every published model fails on the same unseen regime—for instance, liquid-phase dynamics or defect migration barriers—but that shared failure is never systematically reported, the community cannot identify the gap as a collective scientific frontier [3, 6, 13]. In the absence of such reporting, progress becomes trapped in a feedback loop of incremental average-case improvements that leave structural blind spots intact.

Compounding these issues is an emerging reproducibility illusion. A model that tops the leaderboard on the official test set may diverge or produce unphysical results when evaluated on a different draw from the same nominal distribution, let alone on deployment data drawn from real experimental conditions. This disconnect between benchmark performance

and real-world reliability has been observed anecdotally across multiple machine-learning interatomic potential (MLIP) families [7, 9, 12], yet it remains unquantified precisely because failure reporting is absent from standard practice.

These shortcomings are not unique to materials science. In computer vision, ImageNet leaderboards once celebrated top-1 accuracy while models remained brittle to distribution shifts; only the introduction of robustness challenge sets such as ImageNet-R and ImageNet-A exposed those weaknesses. In natural language processing, the GLUE leaderboard initially masked failures on adversarial or out-of-distribution examples until SuperGLUE and challenge sets were added. Materials machine learning is now repeating the same pattern: sophisticated architectures achieve impressive average metrics [7, 14] while their behavior on the edge cases that matter most for scientific discovery remains opaque.

The cumulative effect is a field that appears mature by leaderboard standards yet remains immature by scientific standards. Under these conditions, the next section therefore shifts from critique to a constructive examination of the positive epistemic value that failure cases can supply.

The Value of Failure Cases

Failure cases are not embarrassing outliers to be suppressed or hidden. On the contrary, they represent the most powerful instruments available for scientific advance in machine learning for materials, and five distinct lines of reasoning establish why their indispensable value follows directly from the logic of discovery rather than from mere engineering convenience.

A first argument concerns mechanistic insight. When a potential produces unphysical energies or forces on a particular class of structures, the precise location of that breakdown reveals missing physics—whether long-range electrostatics, many-body dispersion, or explicit electronic degrees of freedom [16, 17]. Historically, such diagnostic failures have accelerated theory development across the physical sciences; there is no reason to expect machine-learned potentials to differ in this respect. The breakdown pattern points toward specific inductive biases that require revision, turning what appears as an error into a signpost for model improvement.

A second argument addresses the practical challenge of model selection. A practitioner designing a catalyst surface can select a potential whose documented failure modes lie entirely outside the relevant pressure–temperature–composition window, while rejecting another potential whose failures overlap that window. This kind of informed decision is impossible to extract from average MAE alone [4, 11]; the failure profile, not the scalar benchmark, determines fitness for purpose under real-world conditions.

Beyond selection, failure cases guide targeted data acquisition. Identifying that all current models collapse on amorphous silicon–carbon compositions immediately signals the highest-leverage region for new ab-initio calculations or experiments. Blind data collection thereby converts into precision engineering of training distributions [12, 18, 19], transforming an undirected sampling problem into a hypothesis-driven campaign aimed at closing specific generalization gaps.

A related function concerns the prevention of overconfidence and the reduction of deployment risk. Explicit reporting of regimes where error exceeds five times the test-set baseline tempers hyperbolic claims of “quantum-mechanical accuracy” and supplies downstream users with uncertainty bounds that average metrics cannot provide [10]. Under these conditions, trust becomes calibrated rather than blind, and the gap between benchmark performance and real-world reliability becomes assessable.

Finally, failure cases create communal knowledge. Shared repositories of well-characterized failures allow the entire field to learn collectively rather than repeating the same mistakes in isolated laboratories. This acceleration of collective intelligence mirrors the role of negative-results publishing in other branches of science—yet in materials machine learning, such infrastructure has remained underdeveloped.

To make these arguments operational, a rigorous definition is required. A failure case for machine-learning interatomic potentials is any input structure, composition, thermodynamic condition, or dynamical regime for which a model's prediction error exceeds a pre-specified threshold—for example, >0.1 eV/atom energy error or $>5\times$ the average test-set MAE. This definition is more specific and actionable than the broader category of “negative results,” precisely because it ties the concept to concrete inputs and quantifiable error thresholds.

By elevating failure cases to first-class status, the community transforms benchmarking from a ranking exercise into a diagnostic engine whose output is not a winner but a map of the remaining scientific frontier. The following section supplies the taxonomy needed to construct that map.

Types of Failure Cases to Include

A failure-inclusive benchmark cannot rely on a single aggregate metric or a narrow set of held-out structures. Instead, it must systematically incorporate six distinct types of failure cases, each defined by a qualitatively different mode of breakdown. For each type, the following discussion provides a concrete definition, illustrative examples drawn from the literature, a justification of its scientific importance, and an actionable inclusion strategy.

Compositional failures occur when models break down on specific elements, concentration ranges, or element combinations—for instance, transition metals with unpaired d-electrons, high-entropy regions, or alloys exhibiting large electronegativity differences. Their importance stems from a central trend in the field: materials discovery increasingly targets chemically complex spaces where models trained on average compositions prove least reliable [4, 16]. A benchmark can include compositional failures by constructing dedicated test sets that span the periodic table in a stratified manner, thereby ensuring equal representation of under-sampled chemistries rather than allowing common elements to dominate.

A related but distinct category involves structural failures. Here models fail on low-symmetry crystals, defect configurations such as grain boundaries, vacancies, or dislocations, or disordered phases including amorphous, liquid, or glassy states. These failures are critical because real materials are rarely perfect crystals; defect energetics and amorphous structures govern mechanical, transport, and catalytic properties [6, 20]. Benchmarks should therefore embed standard defect supercells and melt-quench trajectories as mandatory test subsets, moving beyond equilibrium perfect crystals that represent only a narrow slice of relevant conditions.

Thermodynamic failures represent a third axis of breakdown. Predictions diverge at high temperature through force instability, under high pressure via unphysical compression, or in far-from-equilibrium metastable phases. Thermodynamic accuracy is foundational for phase-diagram construction and stability screening [9, 13, 21]; a model that performs well at ambient conditions but fails at extremes offers limited utility for discovery [22]. Inclusion requires temperature- and pressure-augmented test sets generated via ab-initio molecular dynamics at extreme conditions, deliberately probing regimes where interatomic potentials are most likely to leave their training distribution.

Beyond static and thermodynamic conditions lie dynamical failures [23, 24]. Long-time-scale drift, non-equilibrium quenching, and rare-event sampling (diffusion hops, nucleation events) expose failures that remain entirely invisible to static energy or force benchmarks. These matter because finite-temperature simulations constitute the majority of practical use cases [3, 7]; a model that conserves energy poorly or drifts unphysically over picosecond trajectories cannot be trusted for the microsecond simulations often required in practice. Benchmarks can incorporate short molecular-dynamics trajectories with ground-truth reference forces, requiring models to maintain energy conservation within tight bounds as a necessary condition for dynamical fidelity.

A fifth type concerns numerical instability failures. Models in this regime produce unphysical outputs such as negative formation energies or imaginary phonon frequencies, or outright numerical artifacts including NaN or overflow. Such

instabilities render a potential unusable regardless of its average accuracy, and they prove especially dangerous in automated high-throughput workflows where no human intervenes to flag suspicious outputs [14, 25]. A benchmark must therefore include stability filters that automatically flag and quantify these violations, treating them as first-class failure modes rather than as mere implementation details.

Finally, extrapolation failures occur for any structure lying outside the convex hull of the training distribution—whether compositionally or structurally. Extrapolation represents the regime where machine-learning potentials promise the greatest leverage, precisely because they offer predictions in regions without direct ab-initio reference data, yet it is also the regime where they currently deliver the highest risk [12, 18]. Stratified extrapolation test sets can be generated by systematically increasing the distance in descriptor space from the training manifold, allowing a benchmark to quantify how gracefully a model degrades as it moves beyond its comfort zone.

By embedding representative instances of each of these six types as non-optional components, benchmark suites cease to be mere accuracy contests. They become comprehensive stress tests that expose precisely where current architectures and datasets remain deficient, transforming what might otherwise appear as a ranking exercise into a diagnostic instrument for the field [26].

As systematized in Table 2, failure modes in ML interatomic potentials are not random anomalies but structured categories that expose distinct scientific and modeling limitations.

Table 2. Typology of Failure Modes in Machine Learning Interatomic Potentials and Their Diagnostic Roles

Failure Type	Definition	Typical Manifestation	Scientific Risk	Diagnostic Value	Benchmark Inclusion Strategy
Compositional	Failure across element combinations or concentration regimes	Large errors in high-entropy alloys or oxides	Misleading stability predictions	Reveals data imbalance and missing chemistry	Stratified datasets across periodic table
Structural	Failure on non-ideal or disordered structures	Errors in defects, amorphous phases, grain boundaries	Incorrect defect energetics	Exposes lack of structural generalization	Include defect supercells and disordered phases
Thermodynamic	Failure under extreme temperature/pressure	Instability at high T or unphysical compression	Invalid phase diagrams	Reveals missing physical constraints	High-T/P MD-derived test sets
Dynamical	Failure in time-evolution simulations	Energy drift, divergence in MD trajectories	Unreliable simulations	Tests temporal stability and conservation laws	Short MD trajectories with reference forces
Numerical Instability	Failure due to computational artifacts	NaN outputs, negative formation energies	Pipeline breakdown in automation	Identifies robustness limits	Stability filters and constraint checks

Extrapolation	Failure outside training distribution	Collapse on unseen compositions/structures	False discovery claims	Measures true generalization	Distance-based out-of-distribution test sets
---------------	---------------------------------------	--	------------------------	------------------------------	--

Principles for Failure-Inclusive Benchmarks

Seven interlocking design principles translate the above typology into operational reality.

Implementing a failure-inclusive benchmark requires more than a good-faith collection of challenging test cases; it demands a set of governing principles that structure evaluation, prevent strategic gaming, and ensure long-term scientific utility. Seven such principles emerge from the analysis above, each addressing a distinct mechanism by which current benchmarking practices distort incentives and obscure meaningful failure information.

The first principle concerns pre-identified failure taxonomies. Benchmarks must publish a canonical taxonomy—drawing on the six failure types outlined earlier—before any model evaluation begins. This prospective specification prevents post-hoc cherry-picking of convenient failure definitions after results are known, a practice that would otherwise allow researchers to retroactively define away inconvenient breakdowns. The taxonomy thereby becomes a binding contract between benchmark designers and model developers.

A second principle addresses the metrics used for failure detection. Standard error measures such as MAE and RMSE are insufficient on their own; suites must therefore report the failure rate (the fraction of inputs exceeding a pre-specified threshold), the catastrophic failure count (errors exceeding ten times the baseline), and the physical-constraint violation rate—for instance, the frequency of negative formation energies or imaginary phonon modes. These metrics capture qualitatively different aspects of model behavior that conventional averages systematically obscure, particularly tail performance and physical consistency.

Stratified reporting constitutes the third principle. Performance must be disaggregated by each failure type and by meaningful sub-regions within each type—reporting, for example, not merely an aggregate "structural failure" score but specifically the MAE on high-entropy alloys, the failure rate on triclinic crystals, and the catastrophic error count on grain-boundary structures. This stratified approach produces a multidimensional failure profile rather than a single scalar, enabling practitioners to match model capabilities to application requirements in a principled manner.

The fourth principle establishes a minimum failure-set requirement. No model may be listed on the official benchmark without evaluation on the complete standard failure-case collection; partial reporting, whether through selective omission of difficult cases or submission of results on only a subset of failure types, is disallowed. This requirement closes an obvious loophole in which models could claim benchmark compliance while avoiding precisely those evaluations most likely to expose weaknesses.

Because the failure landscape itself is not static, the fifth principle mandates failure-case versioning. As models improve, the failure-case corpus evolves: difficult cases that all contemporary models now pass are retired from the core suite, while newly discovered failure modes—identified through community reporting or systematic exploration—are added. This dynamic updating keeps the benchmark at the frontier of model capabilities rather than allowing it to become a saturated, uninformative test that no longer discriminates among state-of-the-art approaches.

A sixth principle acknowledges the heterogeneity of practical applications. Different use cases impose dramatically different tolerances for error: a qualitative phase classification may tolerate energy deviations that would completely invalidate a free-energy calculation. The suite therefore provides default thresholds but also exposes an application

programming interface allowing practitioners to specify application-specific error tolerances, ensuring relevance across diverse use cases from high-throughput screening to precise thermodynamics.

Finally, the seventh principle prioritizes transparency over ranking. The primary output of the benchmark is an interactive failure dashboard—visual error maps, spectrum plots showing the distribution of errors along the failure continuum, and stratified tables of performance by failure type. Any leaderboard, whether sorted by average MAE or by some composite score, is demoted to a secondary, optional view that users may consult but that does not define the benchmark's identity. This design choice signals a fundamental reorientation: the goal is not to declare a winner but to characterize capabilities and limitations in their full multidimensional complexity.

Taken together, these seven principles convert benchmarking from a zero-sum game—in which one model's gain is another's loss—into a collaborative diagnostic infrastructure. Under this alternative design, benchmark participation becomes an opportunity to map the boundaries of current methods rather than a competition to obscure those boundaries. The result is an infrastructure that accelerates the entire field by making failure visible, comparable, and actionable.

Objections and Responses

A common concern is that the space of potential failure modes is too expansive to meaningfully capture. Yet the analytical value of a taxonomy does not derive from exhaustive coverage but from its capacity to isolate recurrent, high-impact patterns that reveal systematic weaknesses across model classes. In practice, the absence of any structured account of failure obscures precisely those regularities that are most consequential for scientific deployment. An explicit, even if necessarily partial, taxonomy therefore serves as a critical epistemic scaffold, enabling cumulative understanding where ad hoc observations would otherwise remain fragmented.

This perspective also reframes the perceived tension between failure reporting and the preservation of state-of-the-art status. When models achieve superior aggregate performance while exhibiting instability under application-relevant conditions, their designation as state-of-the-art becomes analytically misleading rather than informative. The inclusion of failure cases does not unfairly penalize such systems; it recalibrates evaluation criteria to reflect robustness as an integral dimension of performance. Under these conditions, transparency operates not as a constraint but as a corrective mechanism, aligning reported capability with operational reliability.

The assumption that users prioritize only the highest-performing model further dissolves under closer scrutiny. Such a view presupposes the existence of a universally optimal system, an assumption that does not hold in materials science, where model suitability is inherently contingent on the target property space, data regime, and downstream application. What emerges instead is a landscape of conditional optimality, in which failure profiles provide indispensable information for navigating trade-offs. By exposing where and how models break down, these profiles directly inform model selection in ways that aggregate metrics cannot, thereby supporting more context-sensitive deployment decisions [4, 11].

Concerns regarding architectural specificity introduce an additional layer of complexity but ultimately reinforce the case for failure-inclusive design. While certain pathologies may indeed arise from architectural idiosyncrasies, a substantial proportion of failures originate from properties of the data distribution or domain constraints. Shifting the analytical focus toward input-driven failure modes allows benchmarks to transcend particular model instantiations, rendering them both architecture-agnostic and more resilient to future methodological shifts. This orientation ensures that benchmarks evolve alongside the field without becoming obsolete as new architectures emerge.

The anticipated increase in benchmarking complexity represents a legitimate operational consideration, yet it must be evaluated against its long-term epistemic consequences. Simplified evaluation protocols that systematically obscure critical weaknesses generate a form of methodological convenience that ultimately undermines scientific progress. By contrast, the additional effort required to incorporate structured failure analysis yields compounding returns, fostering

more reliable knowledge accumulation and reducing the risk of downstream misapplication. What appears as complexity in the short term thus functions as a safeguard against deeper forms of scientific error.

A related implication concerns the absence of established consensus on what constitutes a failure. Rather than constituting a barrier, this lack of agreement signals an open conceptual space in which the field has yet to formalize its evaluative standards. The introduction of a provisional taxonomy and associated principles should therefore be understood as an initiating move within an iterative process of collective refinement. Through sustained engagement—whether in workshops, shared benchmarks, or versioned evaluation frameworks—consensus emerges as a byproduct of structured discourse rather than a prerequisite for action.

Taken together, these considerations reveal that the inclusion of failure cases is not merely compatible with rigorous benchmarking but foundational to it. Each line of critique, when examined through the lens of scientific utility and methodological integrity, ultimately converges on the same conclusion: the systematic treatment of failure is indispensable for advancing trustworthy and context-aware AI in materials science.

Relation to Open Science Movements

The position advanced in this paper—that open benchmark suites for machine learning interatomic potentials must incorporate systematic failure cases—aligns directly with the broader open-science movement that has transformed computational research over the past decade. Failure reporting constitutes a concrete form of transparency: by making the limitations of models visible rather than hidden behind average metrics, the community fulfills the core open-science imperative of reproducible and verifiable knowledge [10]. Zuo *et al.* provided a comprehensive performance assessment [4], yet their benchmark—like most—does not include systematic failure cases; extending such suites with failure taxonomies would elevate reproducibility from a post-publication ideal to an intrinsic design feature.

This stance further resonates with the growing acceptance of negative-results publishing. Journals such as Digital Discovery have explicitly championed the value of well-documented failures [10], recognizing that negative outcomes accelerate progress by preventing redundant efforts. Failure cases for ML potentials are more actionable than generic negative results: they are tied to specific inputs, quantifiable thresholds, and diagnostic insights rather than vague underperformance. Bartók *et al.* unified materials and molecular modeling [3], yet without shared failure repositories the community cannot systematically build upon the breakdowns observed in that foundational work. By institutionalizing failure cases, benchmark suites transform negative results into structured community assets.

The proposal also parallels registered-reports frameworks now standard in psychology and social sciences. Pre-specifying failure taxonomies and evaluation protocols before model training or testing eliminates the bias inherent in post-hoc leaderboards. Batzner *et al.* demonstrated impressive accuracy on their test set [7], but a registered-report approach would have required upfront commitment to report failure rates on compositional and dynamical regimes, thereby preventing selective emphasis on successes. Similarly, the model-card and datasheet initiatives—originally developed for general machine learning—find natural extension here: every published ML potential should be accompanied by a standardized “failure card” that details performance across the six typology categories, echoing the documentation standards advocated in broader AI-for-science ecosystems [16].

Benchmarking standards in adjacent fields reinforce the argument [27, 28]. The GLUE and SuperGLUE suites in natural language processing succeeded precisely because they incorporated challenge sets that exposed robustness failures; likewise, robustness variants of ImageNet moved the field beyond average accuracy. In materials AI, Batra *et al.* described emerging intelligence ecosystems [16], yet these ecosystems will remain fragile without analogous challenge sets for interatomic potentials. Ulissi *et al.* addressed surface-reaction complexity through machine learning [13], but their insights remain under-leveraged because failure modes were not systematically catalogued. Schütt *et al.* introduced SchNet [14], an architecture whose limitations on disordered systems only became apparent anecdotally; an open-science-aligned benchmark would have made those limitations first-class knowledge from the outset.

Finally, the position advances the reproducibility and negative-results ethos already latent in the ML-potential literature. Behler's review of four generations of high-dimensional neural network potentials [9] acknowledges progressive improvements yet stops short of cataloguing persistent failure regimes; embedding failure cases would convert such reviews into living diagnostic maps. The open-science movement insists that scientific infrastructure must serve collective understanding rather than individual prestige. Leaderboard-only benchmarks serve the latter; failure-inclusive suites serve the former. By adopting this redesign, the materials AI community can align itself fully with the transparency, reproducibility, and cumulative-knowledge ideals that define modern open science.

Recommendations

To operationalize the position, concrete actions are required from four stakeholder groups.

For benchmark creators (including maintainers of suites such as MatBench [11], COMP6 extensions, and QM9-derived collections), four immediate steps are essential. First, every future release must include at least three dedicated failure-type test sets—one each for compositional, structural, and extrapolation failures—alongside the standard accuracy test set. Second, reporting must expand beyond MAE and RMSE to include failure rates and catastrophic failure counts, with stratified tables published for each of the six typology categories. Third, benchmark platforms should provide interactive visualization tools such as error maps in composition space and failure-spectrum plots to make limitations immediately interpretable. Fourth, benchmarks must adopt versioning protocols that retire obsolete failure cases and introduce newly discovered regimes annually, ensuring the suite remains at the scientific frontier.

For journal editors and reviewers, three policy changes will enforce accountability. First, manuscripts introducing new ML potentials must include a mandatory failure-analysis section that evaluates the six failure types; papers omitting this section should be returned for revision. Second, claims of “state-of-the-art” performance must be disallowed unless supported by both average metrics and a transparent failure profile; editors should require explicit statements such as “Model X achieves 0.015 eV/atom MAE yet exhibits 23 % failure rate on high-entropy alloys.” Third, journals should establish a dedicated “benchmark failure case” track—modeled on negative-results sections in Digital Discovery [10]—to publish standalone failure reports that receive equal citation weight.

For research funders, two funding priorities will accelerate adoption. First, dedicated calls should support the development and maintenance of failure-inclusive benchmark infrastructure, treating it as critical open-science infrastructure rather than ancillary tooling. Second, grant proposals that focus solely on incremental accuracy improvements should be deprioritized in favor of projects that systematically characterize and mitigate model failures; funders can require applicants to demonstrate how their work will contribute to shared failure repositories.

For the broader research community, three collective actions will embed the new norm. First, a shared failure-case repository—mirroring the Materials Project but dedicated to documented breakdowns—should be established under open governance. Second, an international working group on ML-potential benchmarking standards should be convened under the auspices of organizations already active in the field, tasked with ratifying and iteratively refining the typology and principles proposed here. Third, annual “failure challenge” competitions should be launched in which participants submit inputs designed to expose previously unreported failure modes, with prizes awarded for the most diagnostically valuable cases rather than lowest average error.

These recommendations are not aspirational; they are immediately actionable and directly address the structural incentives that currently sustain leaderboard culture. Zhang *et al.* introduced deep potential molecular dynamics [6], yet its long-term impact would be magnified if subsequent benchmarks had required failure reporting from the outset. By implementing these stakeholder-specific measures, the community converts the abstract position into operational reality.

Conclusion

Leaderboard-only benchmarking is no longer sufficient for evaluating machine learning interatomic potentials. Average error metrics alone cannot capture the systematic failures that determine whether a model is reliable in real scientific applications. Failure cases must therefore be treated as a first-class benchmark component, not as optional supplementary analysis. By embedding structured failure reporting into open benchmark suites, the field can improve transparency, reduce overconfidence, and accelerate more meaningful progress in materials discovery. The future of ML potential benchmarking should be defined not by who ranks first on average, but by how clearly the community understands where models succeed, where they fail, and why.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 08 Nov 2021

Revised: 11 Feb 2022

Accepted: 29 Apr 2022

Published online: 18 July 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

- Weichert D, Link P, Stoll A, Rüping S, Ihlenfeldt S, Wrobel S. A review of machine learning for the optimization of production processes. *Int J Adv Manuf Technol.* 2019;104(5-8):1889-902.
<https://doi.org/10.1007/s00170-019-03988-5>.
- Mishin Y. Machine-learning interatomic potentials for materials science. *Acta Mater.* 2021;214:116980.
<https://doi.org/10.1016/j.actamat.2021.116980>.

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>.

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>.

Himanen L, Geurts A, Foster AS, Rinke P. Data-driven materials science: Status, challenges, and perspectives. *Adv Sci.* 2019;6(21):1900808.
<https://doi.org/10.1002/advs.201900808>.

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>.

Lee K, Yoo D, Jeong W, Han S. SIMPLE-NN: An efficient package for training and executing neural-network interatomic potentials. *Comput Phys Commun.* 2019;242:95-103.
<https://doi.org/10.1016/j.cpc.2019.04.014>.

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>.

Alpaydin E. Machine learning. Cambridge (MA): MIT Press; 2021.

Dunn A, Wang Q, Ganose A, Dopp D, Jain A. Benchmarking materials property prediction methods: The Matbench test set and Automatminer reference algorithm. *NPJ Comput Mater.* 2020;6(1):138.
<https://doi.org/10.1038/s41524-020-00406-3>.

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>.

Ulissi ZW, Medford AJ, Bligaard T, Nørskov JK. To address surface reaction network complexity using scaling relations machine learning and DFT calculations. *Nat Commun.* 2017;8(1):14621.
<https://doi.org/10.1038/ncomms14621>.

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>.

Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A. Machine learning for molecular and materials science. *Nature.* 2018;559(7715):547-55.
<https://doi.org/10.1038/s41586-018-0337-2>.

Batra R, Song L, Ramprasad R. Emerging materials intelligence ecosystems propelled by machine learning. *Nat Rev Mater.* 2021;6(8):655-78.
<https://doi.org/10.1038/s41578-020-00255-y>.

Unke OT, Muwly M. PhysNet: A neural network for predicting energies, forces, dipole moments, and partial charges. *J Chem Theory Comput.* 2019;15(6):3678-93.

<https://doi.org/10.1021/acs.jctc.9b00181>.

Bernstein N, Csányi G, Deringer VL. De novo exploration and self-guided learning of potential-energy surfaces. *NPJ Comput Mater.* 2019;5(1):99.

<https://doi.org/10.1038/s41524-019-0236-6>.

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>.

Deringer VL, Caro MA, Csányi G. Machine learning interatomic potentials as emerging tools for materials science. *Adv Mater.* 2019;31(46):e1902765.

<https://doi.org/10.1002/adma.201902765>.

Schleder GR, Acosta CM, Fazzio A. Exploring two-dimensional materials thermodynamic stability via machine learning. *ACS Appl Mater Interfaces.* 2020;12(18):20149-57.

<https://doi.org/10.1021/acsami.9b14530>.

Behler J. First principles neural network potentials for reactive simulations of large molecular and condensed systems. *Angew Chem Int Ed.* 2017;56(42):12828-40.

<https://doi.org/10.1002/anie.201703114>.

Chmiela S, Sauceda HE, Müller KR, Tkatchenko A. Towards exact molecular dynamics simulations with machine-learned force fields. *Nat Commun.* 2018;9(1):3887.

<https://doi.org/10.1038/s41467-018-06169-2>.

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>.

Thompson AP, Aktulga HM, Berger R, Bolintineanu DS, Brown WM, Crozier PS, et al. LAMMPS: A flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Comput Phys Commun.* 2022;271:108171.

<https://doi.org/10.1016/j.cpc.2021.108171>.

Gonthier JF, Radin MD, Buda C, Daskoćil EJ, Abuan CM, Romero J. Measurements as a roadblock to near-term practical quantum advantage in chemistry: Resource analysis. *Phys Rev Res.* 2022;4(3):033154.

<https://doi.org/10.1103/PhysRevResearch.4.033154>.

Chanussot L, Das A, Goyal S, Lavril T, Shuaibi M, Riviere M, et al. Open Catalyst 2020 (OC20) dataset and community challenges. *ACS Catal.* 2021;11(10):6059-72.

<https://doi.org/10.1021/acscatal.0c04525>.

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>.