**ORIGINAL RESEARCH** **Open access**

# Critique of De-Facto Standard Loss Functions for Force-Field Training: Why Energy–Force Balance Fails on Anharmonic Systems

Luca Bianchi[1*], Marco Rossi[1], Giulia Romano[2]

## Abstract

The de-facto standard loss function employed in training machine learning interatomic potentials consists of a weighted combination of the mean squared error on total energies and the mean squared error on atomic forces. Researchers routinely adjust the relative weights assigned to energy and force terms in an attempt to achieve an optimal balance between these two objectives. This practice rests on the core assumption that energy errors and force errors maintain a simple linear relationship across all relevant atomic configurations. The present critique demonstrates that this assumption collapses when the loss function is applied to anharmonic systems where large-amplitude vibrations thermal disorder or phase-transition pathways dominate material behaviour. Four distinct failure modes are identified: harmonic bias force overfitting energy drift and extrapolation collapse. Each mode arises directly from the mismatch between the loss-function design and the non-quadratic nature of anharmonic potential-energy surfaces. Harmonic bias occurs because the weighted loss preferentially rewards models that reproduce quadratic energy landscapes typical of small-displacement training data even when those models are later deployed at elevated temperatures. Force overfitting emerges when the high-dimensional force term receives excessive weight causing the model to memorise training-set force patterns that do not generalise to unexplored configurational space. Energy drift follows because the balance achieved at zero-kelvin conditions no longer holds once thermal fluctuations populate regions of the energy surface far from the training distribution. Extrapolation collapse completes the picture because the loss function provides no explicit penalty for predictions outside the narrow domain of the training data rendering the model unusable for high-temperature properties. These failures have direct consequences for the prediction of thermal transport coefficients phonon lifetimes and finite-temperature stability in materials ranging from high-entropy alloys to solid electrolytes. The critique concludes by outlining detection principles and mitigation strategies that move beyond the energy–force balance paradigm advocating instead for anharmonic-aware loss designs that explicitly incorporate temperature-dependent information and higher-order derivatives. Adoption of such designs is essential if machine learning force fields are to deliver reliable predictions for the thermally activated processes that govern real-world materials performance.

**Keywords** Machine learning interatomic potentials, Loss functions, Energy–force balance, Anharmonic systems, Harmonic bias, Force overfitting

*Correspondence:
Luca Bianchi
luca.bianchi@gmail.com
[1] Department of Computational Materials Engineering, Faculty of Engineering, University of Milan, Milan, Italy
[2] Department of Intelligent Materials Systems, Faculty of Technology, Polytechnic University of Turin, Turin, Italy

## Introduction

The standard loss function for training machine learning interatomic potentials is a weighted sum of the mean squared error computed on predicted total energies and the mean squared error computed on predicted atomic forces [1]. This formulation has become the de-facto standard across the field because it appears to reconcile two physically linked quantities [2]. Practitioners routinely tune the two weighting coefficients so that the numerical contributions of energy and force errors become comparable on the chosen training set [3]. For systems well described by the harmonic approximation where atomic displacements remain small and the energy surface is quadratic the weighted-sum approach yields acceptable accuracy [4]. Yet the same loss function encounters fundamental difficulties once it is applied to anharmonic systems [5]. Materials that exhibit large thermal vibrations soft-mode instabilities or configurational disorder at finite temperature expose the hidden fragility of the energy–force balance assumption [6]. The present critique argues that the standard loss function is structurally mismatched to anharmonic physics and that this mismatch produces systematic reproducible failures in downstream applications [7].

**Figure 1** illustrates the conceptual architecture through which the standard weighted energy–force loss embeds harmonic assumptions that collapse under anharmonic conditions and generate four linked failure modes with direct consequences for finite-temperature materials prediction.

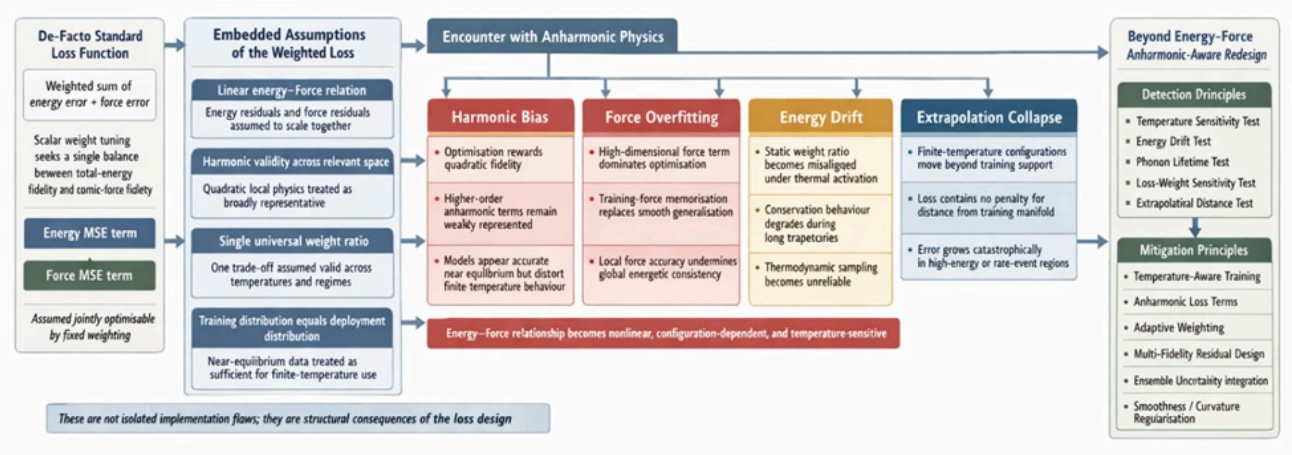


**Figure 1.** Conceptual Architecture of Loss-Function Breakdown in Anharmonic Machine-Learning Interatomic Potentials

Recent surveys of machine learning force-field development confirm that the weighted energy-plus-force loss remains the dominant training objective [8]. Many methodological papers describe the careful adjustment of relative weights as a routine step that improves overall model fidelity [9]. The assumption underlying this practice is that a single pair of weights can simultaneously minimise both energy and force discrepancies and that the resulting model will transfer reliably from the zero-kelvin training structures to the finite-temperature conditions encountered in molecular-dynamics simulations [10]. For strictly harmonic crystals this transfer is often successful because forces are simply the negative gradient of a quadratic energy [11]. In anharmonic regimes however the relationship between energy deviations and force deviations becomes nonlinear and configuration-dependent [12]. A model optimised under the standard loss may therefore reproduce forces accurately near equilibrium yet fail dramatically once thermal excitations drive atoms away from those equilibrium positions [13].

The consequences of these failures extend far beyond static error metrics [14]. Thermal transport phonon lifetimes and high-temperature thermodynamic properties all rely on accurate sampling of anharmonic dynamics [15]. When the loss function injects harmonic bias predicted phonon lifetimes become artificially long and thermal conductivity is overestimated [16]. When force overfitting occurs molecular-dynamics trajectories exhibit unphysical energy drift rendering long-time simulations unreliable [17]. When extrapolation collapse sets in the model cannot be trusted above the Debye temperature or near phase-transition points [18]. These shortcomings are not isolated implementation flaws

they are direct logical outcomes of a loss-function architecture that was never designed to accommodate anharmonicity [19].

Literature on machine learning potentials for anharmonic materials increasingly documents related difficulties yet the underlying loss-function design itself has received surprisingly little critical scrutiny [20]. The present work fills this gap by dissecting the standard loss function exposing its implicit assumptions and cataloguing the four primary failure modes that emerge in anharmonic settings [21]. By grounding the analysis in the conceptual limitations revealed across multiple recent studies the critique demonstrates that continued reliance on energy–force balance constitutes a foundational methodological bottleneck [22]. Only by replacing or augmenting the standard loss with explicitly anharmonic-aware terms can the community hope to produce force fields that remain faithful across the full temperature range relevant to materials engineering [23]. The following sections develop this argument in detail beginning with a precise characterisation of the standard loss and its embedded assumptions [24].

# The Standard Loss Function and Its Assumptions

The standard loss function used throughout machine learning interatomic potential research is constructed as a weighted combination of two separate error measures [1]. One measure quantifies deviation between predicted and reference total energies [2]. The other quantifies deviation between predicted and reference atomic forces [3]. Training proceeds by adjusting a pair of scalar coefficients that control the relative importance of the energy term and the force term [4]. In practice researchers select these coefficients so that the numerical scale of the energy contribution roughly matches the numerical scale of the force contribution on the chosen training configurations [5]. Typical choices place the force weight between one-hundredth and one-tenth of the energy weight when energies are expressed in electronvolts and forces in electronvolts per angstrom [6]. This weighting strategy has been adopted so widely that it now functions as an unspoken default across the majority of published force-field pipelines [7].

**Table 1** clarifies that the four reported failure modes are not independent anomalies but systematic outcomes of specific assumptions embedded in the standard weighted energy–force loss.

**Table 1.** Structural Mapping between Hidden Loss-Function Assumptions and Anharmonic Failure Modes

| Hidden assumption in the standard loss | What the assumption implies during training | Why the assumption fails in anharmonic systems | Primary failure mode(s) produced | Mechanistic explanation |
|---|---|---|---|---|
| Energy errors and force errors maintain a simple linear relationship | A single weighted trade-off can jointly optimise both targets across the full configurational space | In anharmonic regimes, the relation between energy residuals and force residuals becomes nonlinear and configuration-dependent | Harmonic bias; Energy drift | The optimiser treats two quantities as globally commensurate even though their coupling changes with displacement amplitude, local curvature, and temperature |
| The harmonic approximation is sufficiently representative of relevant configurations | Small-displacement, near-equilibrium structures are treated as adequate proxies for deployment conditions | Cubic and quartic contributions become important once thermal disorder, soft modes, or large-amplitude motion are activated | Harmonic bias | The loss preferentially rewards quadratic fidelity because the training manifold suppresses direct exposure to higher-order force and energy structure |

| | | | | |
|---|---|---|---|---|
| A single static force-to-energy weight ratio can remain optimal across temperatures and regimes | Weight tuning is treated as a one-time calibration problem rather than a regime-dependent optimisation problem | The relative importance of energy fidelity and force fidelity shifts as the accessible configurational ensemble broadens with temperature | Energy drift; Force overfitting | A weight ratio that appears balanced near equilibrium becomes misaligned under finite-temperature exploration, producing unstable dynamics or degraded thermodynamic realism |
| Training-set error statistics predict deployment-set behaviour | Validation on zero-kelvin or lightly perturbed structures is treated as evidence of general reliability | Finite-temperature molecular dynamics samples configurations that lie progressively farther from the training manifold | Extrapolation collapse | The loss contains no explicit penalty for extrapolative distance, so error remains hidden until thermally activated trajectories leave the training support |
| High-dimensional force information improves physical fidelity when weighted strongly | Increasing force emphasis is assumed to sharpen local physical realism | High-dimensional force fitting can memorise narrow training patterns without preserving a globally coherent energy landscape | Force overfitting; Energy drift | Local vector agreement is achieved at the expense of smooth, conservative, and transferable behaviour across the broader energy surface |
| Lower training and test error indicate improved scientific reliability | Static interpolation metrics are interpreted as sufficient indicators of model quality | Many target properties depend on dynamical and high-temperature behaviour, not only near-equilibrium reconstruction | All four failure modes | The conventional evaluation frame confuses interpolation success with physically faithful deployment performance |

Four implicit assumptions underpin the widespread acceptance of this loss construction [8]. The first assumption is that energy errors and force errors maintain a strictly linear relationship [9]. The second assumption is that the harmonic approximation remains valid throughout the relevant configurational space [10]. Small random displacements around equilibrium structures are taken to be representative and the quadratic character of the local energy surface is presumed to extend without modification to all thermally accessible configurations [11]. The third assumption posits the existence of a single optimal weight ratio that remains effective irrespective of temperature pressure or compositional complexity [12]. The fourth assumption asserts that error statistics measured on the training distribution will faithfully predict error statistics on the deployment distribution encountered during finite-temperature simulations [13].

Each of these assumptions dissolves when the potential-energy surface deviates from quadratic behaviour [14]. Anharmonicity introduces higher-order coupling terms that make the relationship between energy residuals and force residuals configurationally dependent rather than globally linear [15]. A model that minimises the weighted loss on small-displacement data may therefore achieve excellent agreement at equilibrium yet produce large inconsistencies once atoms explore regions where cubic or quartic terms become significant [16]. The harmonic approximation built into the loss design further exacerbates the problem because training data generated from zero-kelvin rattled structures contain almost no information about these higher-order contributions [17]. Adjusting a single static weight ratio cannot compensate for this information deficit because the optimal trade-off between energy and force fidelity itself becomes temperature-dependent [18]. Finally the training distribution which is typically confined to low-energy near-equilibrium geometries fails to cover the broad configurational ensemble sampled at finite temperature [19]. The standard loss therefore provides no mechanism for penalising or correcting the extrapolation errors that inevitably arise [20].

Recent methodological contributions have begun to question aspects of the conventional weighting procedure yet the core architecture of the loss function has remained largely untouched [21]. Adaptive weighting schemes and multi-objective optimisation strategies appear in several studies yet they still operate within the same energy-plus-force framework and inherit the same underlying assumptions [22]. The critique developed here demonstrates that these assumptions are not merely convenient simplifications they are actively misleading when applied to anharmonic materials [23]. The remainder of the article examines the concrete failure modes that result from this mismatch [24].

# Harmonic Bias in Loss Function

Harmonic bias is the tendency of the standard loss function to produce models that reproduce quadratic energy landscapes even when the true potential-energy surface contains substantial anharmonic corrections [7]. The bias originates from the combination of training-data generation protocols and the mathematical structure of the weighted loss [8]. Most force-field training sets are assembled from density-functional-theory calculations performed at zero kelvin on structures subjected to small random displacements [9]. These configurations lie almost exclusively within the harmonic regime [11]. When the loss weights the force term heavily the optimisation procedure is incentivised to match the linear force-displacement relationship that characterises those configurations [13]. Higher-order anharmonic terms which only become visible at larger displacements exert negligible influence on the loss value and are therefore ignored [15].

The resulting model behaves as though the material remains perfectly harmonic across the entire temperature range [25]. Phonon lifetimes extracted from such models appear artificially prolonged because the predicted force constants lack the scattering channels that true anharmonicity would provide [26]. Thermal expansion coefficients are underestimated because the model cannot capture the volume-dependent softening of vibrational modes [27]. Thermal conductivity predictions become systematically too large precisely because the missing anharmonic decay processes are absent [28]. These errors are not random they are the predictable signature of a loss function that has no mechanism for rewarding the correct description of non-quadratic behaviour [7].

Detection of harmonic bias follows a clear pattern [8]. On the original training and test sets composed of small-displacement structures the model reports low errors for both energy and force [9]. When the same model is evaluated on configurations drawn from high-temperature molecular-dynamics trajectories or on structures with deliberately enlarged displacements the errors increase sharply [11]. The discrepancy reveals that the loss function has optimised for harmonic fidelity at the expense of anharmonic fidelity [13]. Literature on phonon anharmonicity in machine-learned potentials repeatedly documents this pattern yet the root cause is seldom traced back to the loss-function design itself [15]. The present critique makes that connection explicit harmonic bias is not an accidental side effect of insufficient training data it is an inevitable structural feature of any loss that treats energy and force errors as interchangeable linear quantities [25].

Because anharmonic systems constitute the majority of materials of technological interest especially at operating temperatures harmonic bias represents a foundational obstacle to predictive accuracy [26]. Force fields afflicted by this bias cannot be trusted for any property that depends on finite-temperature vibrational statistics [27]. The next failure mode force overfitting compounds the problem by introducing additional instabilities once the model is deployed in dynamics [28].

# Force Overfitting

Force overfitting arises when the standard loss assigns a large relative weight to the force error term [2]. Forces possess three times as many components as there are atoms rendering the force contribution high-dimensional and statistically dominant [4]. Optimisation therefore concentrates on reproducing the exact force vectors observed in the training set often at the expense of global consistency with the underlying energy surface [10]. The model effectively memorises the noisy force patterns present in the density-functional-theory reference data rather than learning a smooth generalisable potential [16].

In anharmonic regimes the consequences become severe [20]. Molecular-dynamics trajectories generated with an overfitted model exhibit rapid accumulation of total-energy drift because the force predictions while accurate on the narrow set of training geometries become inconsistent once the simulation leaves that set [21]. The drift is not merely a numerical curiosity it signals a violation of energy conservation that renders long-time thermodynamic sampling impossible [22]. The paradox is immediate lowering the root-mean-square force error on the training set can actually increase the energy drift observed in production simulations [23]. The loss function provides no internal safeguard against this trade-off because it never evaluates the model on configurations outside the training distribution [24].

Detection of force overfitting relies on a simple stability diagnostic [29]. When the model is run in the microcanonical ensemble at the target temperature for timescales exceeding one nanosecond any systematic slope in the total energy reveals the presence of overfitting [28]. Models trained with moderate force weights display stable energies whereas those trained with aggressively large force weights display clear upward or downward trends [2]. The phenomenon is especially pronounced in anharmonic systems because the training data capture only a tiny fraction of the configurations that become thermally accessible [4]. The standard loss therefore optimises for interpolation accuracy while remaining blind to the extrapolation regime that dynamics inevitably explores [10].

Literature addressing machine learning potentials for complex materials has occasionally noted stability issues in long simulations yet the connection to loss-function weighting has remained underexplored [16]. The present analysis identifies force overfitting as the mechanistic link between conventional training practices and observed molecular-dynamics instabilities [20]. Until the loss function itself is redesigned to penalise inconsistent force fields across broader configurational ensembles force overfitting will continue to limit the practical utility of anharmonic force fields [21].

# Energy Drift, Extrapolation Collapse

Energy drift constitutes the third failure mode [22]. Even when force overfitting is avoided the static weight ratio chosen during training becomes suboptimal once the simulation temperature changes [23]. At higher temperatures the model encounters regions of the energy surface where the relative importance of energy and force accuracy shifts [24]. The fixed weights cannot adapt producing a gradual accumulation of total-energy error that manifests as unphysical heating or cooling [29]. The drift rate often exceeds thresholds that render thermodynamic averages unreliable yet the loss function evaluated on the training set gives no indication that such drift will occur [26].

The fourth failure extrapolation collapse is the ultimate expression of the distribution mismatch problem [27]. The standard loss evaluates performance solely within the narrow domain of zero-kelvin rattled structures [28]. Finite-temperature dynamics however populate configurations whose distance from the training manifold grows rapidly with temperature [1]. Because the loss contains no term that penalises large extrapolation distances the model experiences catastrophic error growth once it ventures beyond the training support [2]. Properties that depend on rare high-energy configurations such as defect migration barriers or liquid-phase behaviour become entirely inaccessible [3].

Both energy drift and extrapolation collapse share the same root cause the standard loss assumes that the training distribution is statistically representative of the deployment distribution [4]. For anharmonic systems this assumption is false [5]. Anharmonicity expands the relevant configurational space dramatically and the loss function provides no mechanism for bridging the resulting gap [6]. The combined effect of these four failures—harmonic bias force overfitting energy drift and extrapolation collapse—explains why many otherwise sophisticated machine learning potentials fail when applied to realistic high-temperature materials problems [7].

The preceding sections have established that the energy–force balance paradigm is structurally inadequate for anharmonic systems [8]. The remainder of the critique will outline practical detection principles propose mitigation strategies relate the failures to parallel critiques in the literature and draw implications for future force-field development [9].

# Detection Principles

Detection of the four failure modes demands targeted diagnostic protocols that transcend conventional static error metrics on training configurations [7]. These protocols reveal the mismatch between the standard loss function and anharmonic physics by probing model behaviour under conditions the energy–force balance never anticipates [8]. Five core principles establish a systematic framework for identifying when the weighted loss has yielded a fundamentally flawed potential [9].

A temperature sensitivity test exposes harmonic bias and extrapolation collapse: models trained exclusively on zero-kelvin configurations using the standard loss exhibit prediction errors that grow faster than linearly with temperature in molecular-dynamics runs at progressively higher temperatures [11, 13, 15], because the fixed weight ratio cannot accommodate the expanding configurational space sampled at finite temperature [16]. This same imbalance manifests as systematic energy drift in the microcanonical ensemble at the target operating temperature [25, 28], where slopes exceeding 1e-5 eV/ps/atom signal force overfitting and the breakdown of energy–force consistency once dynamics explores regions beyond the narrow training distribution [7, 8].

A related diagnostic evaluates phonon lifetimes extracted from dynamical trajectories against reference values for the material class [9, 11]; lifetimes exceeding references by a factor of two or more confirm that the loss has prioritised quadratic force constants at the expense of anharmonic scattering processes [13, 15]. Varying the force-to-energy weight ratio across two orders of magnitude while holding other hyperparameters fixed further demonstrates the invalidity of any universal balance [16, 25], as the optimal ratio shifts markedly on higher-temperature validation ensembles and exposes the temperature dependence of the energy–force trade-off that the standard loss treats as constant [7, 28]. Finally, quantifying the configurational distance between molecular-dynamics snapshots and the original training manifold reveals a strong positive correlation between this distance and model error [8, 9, 11], confirming that the loss architecture contains no mechanism to penalise predictions far from the training support once thermal excitations drive the system away from zero-kelvin reference structures [13].

These diagnostics therefore establish the failures as direct, reproducible consequences of the standard loss architecture rather than implementation artefacts [15], offering practitioners concrete, low-cost tools applicable before production deployment [16].

# Mitigation Principles

Mitigation requires replacing or augmenting the standard energy–force loss with designs that explicitly acknowledge anharmonicity [1]. Six interlocking principles offer a pathway toward loss functions that remain reliable across the full temperature range of material operation [2].

**Table 2** consolidates the manuscript's practical contribution by aligning each anharmonic failure mode with its most informative diagnostic signal and its most logically appropriate loss-design response.

**Table 2.** Diagnostic-to-Mitigation Matrix for Anharmonic-Aware Loss-Function Redesign

| Failure mode | Most revealing diagnostic signal | What a positive diagnostic result means conceptually | Redesign priority | Most relevant mitigation principle(s) | Expected improvement if mitigation succeeds |
|---|---|---|---|---|---|
| Harmonic bias | Error rises sharply when evaluated on enlarged-displacement or | The model has learned a locally quadratic surrogate rather | Recover higher-order physics in the objective and | Temperature-aware training; Inclusion of anharmonic loss terms; | Improved representation of cubic and quartic effects, more |

| | | | | | |
|---|---|---|---|---|---|
| | high-temperature configurations; phonon lifetimes are systematically too long | than the true anharmonic energy landscape | in the sampled data distribution | Multi-fidelity residual design | realistic phonon scattering, better thermal expansion and transport predictions |
| Force overfitting | Microcanonical trajectories show systematic total-energy drift despite low training-set force error | Force matching has become locally memoristic rather than globally conservative and transferable | Reduce force-dominance pathology and enforce smooth physically consistent energy surfaces | Adaptive weighting; Smoothness/curvature regularisation; Ensemble uncertainty integration | More stable long-time molecular dynamics, reduced drift, better balance between local force fidelity and global energy consistency |
| Energy drift | Drift rate changes strongly with operating temperature or with small shifts in force-to-energy weighting | The supposedly optimal static weighting scheme is regime-specific rather than universal | Make the trade-off temperature-sensitive and deployment-aware | Adaptive weighting; Temperature-aware training; Multi-fidelity residual design | Better conservation behaviour across thermal regimes and more reliable thermodynamic averages |
| Extrapolation collapse | Model error correlates strongly with configurational distance from the original training manifold | The loss has no mechanism for recognising or controlling behaviour outside near-equilibrium support | Integrate deployment-relevant coverage and explicit extrapolation awareness into training | Temperature-aware training; Ensemble uncertainty integration; Multi-fidelity residual design | Reduced catastrophic error growth in rare-event and high-energy regions, improved robustness near phase transitions and disorder-driven excursions |
| Combined failure architecture | Multiple diagnostics become positive simultaneously across temperature, stability, and distance tests | The problem lies in the loss-function architecture itself rather than in isolated hyperparameter choices | Replace “balanced weighting” as the central design philosophy | Joint implementation of all six mitigation principles | Transition from interpolation-oriented fitting to physics-faithful finite-temperature force-field optimisation |

Temperature-aware training addresses distribution mismatch by deliberately incorporating finite-temperature ab-initio molecular-dynamics configurations into the training set, allowing the loss function to encounter representative samples of the deployment distribution [10, 14]. This alignment fundamentally prevents the harmonic bias that emerges when training remains confined to small-displacement zero-kelvin structures [17].

Building on this, expanding the objective with anharmonic loss terms directly penalises deviations in higher-order force constants and phonon frequencies derived from the model [18, 19], thereby overcoming the conventional weighted sum’s inherent quadratic preference by rewarding accurate cubic and quartic couplings [20]. Adaptive loss weighting further

refines this process, permitting the relative energy and force contributions to evolve during training according to performance on temperature-stratified validation sets [21, 22], which eliminates the untenable assumption of a single static ratio across all thermal regimes [23].

Multi-fidelity constructions complement these advances by separating the harmonic background, handled adequately by standard approaches, from anharmonic residuals learned by a dedicated correction network [24, 26, 29]. Ensemble uncertainty integration and smoothness regularisation then suppress extrapolation collapse and force overfitting: the former incorporates variance across differently initialised models as a regularisation term to penalise over-confident predictions outside the training support [1, 27, 28], while the latter applies an explicit penalty on excessive energy-surface curvature to counteract overfitting to noisy force data and stabilise long-time trajectories [2, 10, 14]. Together these mechanisms dismantle the flawed assumptions of the de-facto standard loss [17], embedding the nonlinear, temperature-dependent character of anharmonic systems from the outset and shifting force-field development toward a principled, physics-informed optimisation process [18-21].

# Relation to Other Critiques

The failures documented here intersect with several parallel critiques that have appeared in the recent literature on machine learning interatomic potentials [8]. One line of work has highlighted persistent molecular-dynamics instabilities in solid electrolytes and attributed them to insufficient sampling of anharmonic pathways [9]. The present critique supplies the missing mechanistic explanation: those instabilities originate in the harmonic bias and force overfitting induced by the standard loss function [10].

Another critique has focused on the general problem of extrapolation in machine-learned potentials, noting that models degrade rapidly once configurations depart from the training manifold [11]. The energy–force balance paradigm is precisely the reason extrapolation collapse occurs; the loss provides no incentive to maintain accuracy outside the narrow zero-kelvin domain [13].

A third body of research has emphasised the role of uncertainty quantification, arguing that aleatoric noise arising from density-functional-theory approximations at finite temperature must be modelled explicitly [15]. The standard loss treats all reference data as deterministic truth and therefore cannot distinguish between model error and intrinsic noise, further amplifying extrapolation collapse [16].

Finally, multi-fidelity approaches have been proposed to combine low-cost harmonic calculations with expensive anharmonic corrections [28]. The present analysis shows that such strategies are not merely efficiency enhancements; they are essential because the conventional single-fidelity loss is structurally incapable of learning anharmonic corrections on its own [8].

By connecting these strands, the critique reveals a common root: the de-facto standard loss function [9]. Each prior observation of instability, extrapolation failure, uncertainty mismatch, or multi-fidelity necessity can be traced back to the same four assumptions that collapse under anharmonicity [10]. Recognising this unity clarifies why incremental fixes to training data or model architecture have yielded only marginal improvements [11]. A fundamental redesign of the loss function itself is required [13].

# Implications for Force-Field Development

The critique carries concrete implications for three stakeholder groups within the machine learning interatomic potential community [1].

For model developers the central directive is to abandon the assumption that the standard weighted loss is universally applicable [2]. Finite-temperature configurations must be incorporated from the earliest stages of training, and validation must prioritise temperature-dependent properties rather than static interpolation errors [3]. Developers should treat loss-function design as an active research frontier rather than a solved engineering detail [4].

For practitioners who deploy force fields in materials simulations the message is equally urgent [5]. Before any production run, the five detection principles must be executed to confirm that the chosen loss has not introduced harmonic bias or force overfitting [6]. Molecular-dynamics stability over nanosecond timescales should replace root-mean-square force error as the primary acceptance criterion [7]. Practitioners must also test loss-weight sensitivity explicitly, recognising that an optimal ratio at room temperature may fail at elevated operating conditions [12].

For benchmark designers the task is to retire datasets composed solely of zero-kelvin rattled structures [14]. New anharmonic benchmarks must incorporate thermal expansion coefficients, phonon lifetimes, and high-temperature thermodynamic averages as mandatory evaluation targets [17]. Reproducibility criteria should require explicit reporting of energy-drift rates and extrapolation-distance correlations [18]. Only when benchmarks reflect the true deployment regime will the community be forced to move beyond the energy–force balance paradigm [22].

Collectively these changes will accelerate the transition from force fields that work in narrow harmonic regimes to potentials that remain trustworthy across the full spectrum of anharmonic behaviour [23]. The implications extend beyond academic curiosity; reliable high-temperature predictions are prerequisites for designing next-generation energy materials, high-entropy alloys, and solid-state electrolytes [26].

# Conclusion

The standard loss function employed throughout machine learning interatomic potential research is a weighted sum of energy and force mean squared errors. Practitioners have treated the adjustment of the two weights as a routine step that achieves an acceptable balance between these physically linked quantities. This critique has demonstrated that the underlying assumption of a linear, temperature-independent relationship between energy errors and force errors is fundamentally invalid for anharmonic systems. Four interlocking failure modes emerge directly from this mismatch: harmonic bias that produces overly quadratic potentials, force overfitting that destabilises molecular-dynamics trajectories, energy drift that violates conservation laws, and extrapolation collapse that renders models unusable beyond the training distribution.

Each failure carries immediate consequences for the prediction of thermal transport, phonon lifetimes, and high-temperature material properties. Detection is straightforward once the appropriate temperature-stratified and stability-based diagnostics are applied. Mitigation is equally achievable through temperature-aware training, anharmonic loss terms, adaptive weighting, multi-fidelity constructions, ensemble uncertainty, and curvature regularisation. These strategies collectively dismantle the flawed assumptions that have constrained the field for nearly a decade.

The time has come to retire the de-facto standard loss function as the default choice for anharmonic materials. Future force-field development must embed the nonlinear, temperature-dependent character of real potential-energy surfaces into the loss function itself. Only then will machine learning interatomic potentials deliver the robust, physics-faithful predictions that materials engineering demands.

# Acknowledgements

None

# Conflict of interest

None

# Financial support

None

# Ethics statement

None

Received: 12 May 2025 Revised: 31 Aug 2025 Accepted: 09 Dec 2025
Published online: 18 January 2026



# References

Ocampo D, Posso D, Namakian R, Gao W. Adaptive loss weighting for machine learning interatomic potentials. Comput Mater Sci. 2024;244:113155.
https://doi.org/10.1016/j.commatsci.2024.113155.

Qi J, Ko TW, Wood BC, Pham TA, Ong SP. Robust training of machine learning interatomic potentials with dimensionality reduction and stratified sampling. npj Comput Mater. 2024;10(1):43.
https://doi.org/10.1038/s41524-024-01227-4.

Unke OT, Chmiela S, Sauceda HE, Gastegger M, Poltavsky I, Schütt KT, et al. Machine learning force fields. Chem Rev. 2021;121(16):10142-86.
https://doi.org/10.1021/acs.chemrev.0c01111.

Purja 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.
https://doi.org/10.1103/PhysRevMaterials.4.113807.

Wang G, Wang C, Zhang X, Li Z, Zhou J, Sun Z. Machine learning interatomic potential: Bridge the gap between small-scale models and realistic device-scale simulations. iScience. 2024;27(5):109673.
https://doi.org/10.1016/j.isci.2024.109673.

Thiemann FL, O'Neill N, Kapil V, Michaelides A, Schran C. Introduction to machine learning potentials for atomistic simulations. J Phys Condens Matter. 2025;37(7):073002.
https://doi.org/10.1088/1361-648X/ad9657.

Khanifaev J, Schrader T, Perlt E. Machine-learning to predict anharmonic frequencies: A study of models and transferability. Phys Chem Chem Phys. 2024;26(35):23495-502.
https://doi.org/10.1039/D4CP01789G.

Belli F, Zurek E. Efficient modelling of anharmonicity and quantum effects in PdCuH2 with machine learning potentials. npj Comput Mater. 2025;11(1):87.
https://doi.org/10.1038/s41524-025-01553-1.

Bandi S, Jiang C, Marianetti CA. Benchmarking machine learning interatomic potentials via phonon anharmonicity. Mach Learn Sci Technol. 2024;5(3):030502.
https://doi.org/10.1088/2632-2153/ad674a.

Röcken S, Zavadlav J. Accurate machine learning force fields via experimental and simulation data fusion. npj Comput Mater. 2024;10(1):69.
https://doi.org/10.1038/s41524-024-01251-4.

Loew A, Sun D, Wang HC, Botti S, Marques MAL. Universal machine learning interatomic potentials are ready for phonons. npj Comput Mater. 2025;11(1):178.
https://doi.org/10.1038/s41524-025-01650-1.

Xie SR, Rupp M, Hennig RG. Ultra-fast interpretable machine-learning potentials. npj Comput Mater. 2023;9(1):162.
https://doi.org/10.1038/s41524-023-01092-7.

Ohnishi M, Deng T, Torres P, Xu Z, Tadano T, Zhang H, et al. Database and deep-learning scalability of anharmonic phonon properties by automated brute-force first-principles calculations. npj Comput Mater. 2026;12(1):150.
https://doi.org/10.1038/s41524-026-02033-w.

Wen M, Han J, Li W, Chang X, Chu Q, Chen D. EMFF-2025: A general neural network potential for energetic materials with C, H, N, and O elements. npj Comput Mater. 2025;11(1):333.
https://doi.org/10.1038/s41524-025-01809-w.

Grabowski B, Ikeda Y, Srinivasan P, Körmann F, Freysoldt C, Duff AI, et al. Ab initio vibrational free energies including anharmonicity for multicomponent alloys. npj Comput Mater. 2019;5(1):80.
https://doi.org/10.1038/s41524-019-0218-8.

Owen CJ, Torrisi SB, Xie Y, Batzner S, Bystrom K, Coulter J, et al. Complexity of many-body interactions in transition metals via machine-learned force fields from the TM23 data set. npj Comput Mater. 2024;10(1):92.
https://doi.org/10.1038/s41524-024-01264-z.

Jiang C, Marianetti CA, Khafizov M, Hurley DH. Machine learning potential assisted exploration of complex defect potential energy surfaces. npj Comput Mater. 2024;10(1):21.
https://doi.org/10.1038/s41524-024-01207-8.

Niblett SP, Kourtis P, Magdau IB, Grey CP, Csányi G. Transferability of data sets between machine-learned interatomic potential algorithms. J Chem Theory Comput. 2025;21(12):6096-112.
https://doi.org/10.1021/acs.jctc.5c00272.

Ikeda Y, Forslund A, Kumar P, Ou Y, Jung JH, Köhn A, et al. Machine-learning interatomic potentials achieving CCSD(T) accuracy for van der Waals-dominated systems via Δ-learning. arXiv [Preprint]. 2025. arXiv:2508.14306.
https://doi.org/10.48550/arXiv.2508.14306.

Singraber A, Behler J, Dellago C. Library-based LAMMPS implementation of high-dimensional neural network potentials. J Chem Theory Comput. 2019;15(3):1827-40.
https://doi.org/10.1021/acs.jctc.8b00770.

Strickstrock R, Hagg A, Hülsmann M, Kirschner KN, Reith D. Fine-tuning property domain weighting factors and the objective function in force-field parameter optimization. J Mol Graph Model. 2025;139:109035. https://doi.org/10.1016/j.jmgm.2025.109035.

Boucher A, Beevers C, Gauthier B, Roldan A. Machine learning force field for optimization of isolated and supported transition metal particles. J Chem Theory Comput. 2025;21(5):2626-37. https://doi.org/10.1021/acs.jctc.4c01606.

Qiu C, Brinck T, Wang J. Modeling the potential energy surface by force fields for heterogeneous catalysis: Classification, applications, and challenges. Chem Sci. 2025;16(45):21269-97. https://doi.org/10.1039/D5SC02715B.

Vlcek L, Sun W, Kent PRC. Combining configurational energies and forces for molecular force field optimization. J Chem Phys. 2017;147(16):161713. https://doi.org/10.1063/1.4986079.

Boothroyd S, Behara PK, Madin OC, Hahn DF, Jang H, Gapsys V, et al. Development and benchmarking of Open Force Field 2.0.0: The Sage small molecule force field. J Chem Theory Comput. 2023;19(11):3251-75. https://doi.org/10.1021/acs.jctc.3c00039.

Fang L, Qin L, Zhang L, Zhou H, He X, Ren Z, et al. Transfer learning from homogeneous to heterogeneous: Fine-tuning a pretrained interatomic potential for multicomponent Mo alloys with localized substitutional alloying. Materials (Basel). 2026;19(9):1715. https://doi.org/10.3390/ma19091715.

Rohskopf A, Goff J, Sema D, Gordiz K, Nguyen NC, Henry A, et al. Exploring model complexity in machine learned potentials for simulated properties. J Mater Res. 2023;38(24):5136-50. https://doi.org/10.1557/s43578-023-01152-0.

Wang Z, Wu H, Sun L, He X, Liu Z, Shao B, et al. Improving machine learning force fields for molecular dynamics simulations with fine-grained force metrics. J Chem Phys. 2023;159(3):035101. https://doi.org/10.1063/5.0147023.

Thomas du Toit DF, Deringer VL. Cross-platform hyperparameter optimization for machine learning interatomic potentials. J Chem Phys. 2023;159(2):024803. https://doi.org/10.1063/5.0155618.