

ORIGINAL RESEARCH

Open access

Self-Consistency Is Not Enough: A Framework for Error-Correcting ML Potentials with On-the-Fly Residual Learning

Siti Rahman^{1*}, Ahmad Zaki¹, Nurul Huda², Amir Faisal¹, Lim Wei²

Abstract

Machine learning interatomic potentials now enable molecular dynamics simulations with near-density-functional-theory accuracy at scales inaccessible to conventional quantum methods. Yet most remain fundamentally static: once trained, they are deployed without adaptation, even as simulations enter configurations outside the original training distribution. Although self-consistency ensures that predicted forces remain exact derivatives of the learned energy surface, it guarantees only internal coherence, not fidelity to the true reference landscape. As a result, small local errors can accumulate into substantial long-timescale drift. This paper proposes a conceptual framework for error-correcting machine learning potentials based on on-the-fly residual learning. The architecture combines a self-consistent base predictor with an error detector, a lightweight residual corrector, an online updater, and a memory manager. Embedded directly within the molecular dynamics loop, these components enable the system to identify unreliable predictions, apply immediate corrections, selectively request sparse density-functional-theory labels, and retain corrective knowledge during continuous adaptation. By shifting from static deployment to simulation-aware error correction, the framework addresses the central limitations of extrapolation failure and accumulated drift. It therefore outlines a path toward adaptive machine learning potentials capable of sustaining reliable long-timescale materials simulations with controlled computational overhead.

Keywords Active learning, Machine learning potentials, Error correction, On-the-fly learning, Residual learning, Self-consistency

*Correspondence:

Siti Rahman

siti.rahman@outlook.com

¹ Department of Computational Materials Engineering, Faculty of Engineering, University of Malaya, Kuala Lumpur, Malaysia

² Department of Materials Data Analytics, Faculty of Science and Technology, Universiti Teknologi Malaysia, Johor Bahru, Malaysia

Introduction

Machine learning potentials have revolutionized molecular dynamics. Zhang *et al.* [1] introduced Deep Potential Molecular Dynamics, a scalable model that achieves the accuracy of quantum mechanics while enabling simulations of systems containing thousands of atoms [2]. Batzner *et al.* [3] later demonstrated that E(3)-equivariant graph neural networks can reach comparable accuracy with far fewer training data. Bartók *et al.* [4] showed how Gaussian process regression can unify the modeling of materials and molecules within a single framework. These and related advances [5–10] have made it possible to study phase transitions, defect dynamics, and reaction pathways that were previously intractable.

Yet all such models share a fundamental limitation: they are static. Once trained on an initial density-functional-theory dataset, they are deployed forever without further adaptation [11–13]. During long molecular dynamics simulations, errors

that are negligible on short timescales accumulate and cause the trajectory to drift away from the true quantum-mechanical behavior [14-16]. The community has long relied on self-consistency—ensuring that forces are exact derivatives of the predicted energy—as the primary safeguard against unphysical behavior [1, 3]. Self-consistency guarantees energy conservation in the NVE ensemble and satisfies basic physical plausibility constraints. It is necessary. It is not sufficient [17].

Table 1 distinguishes the properties that self-consistency secures from the additional capabilities required for a machine learning potential to remain externally correct, error-detectable, and corrigible during long-timescale molecular dynamics.

Table 1. Analytical Separation of Internal Consistency, External Correctness, Detectability, and Corrigibility in Machine Learning Potentials

Analytical dimension	What it means in this manuscript	Satisfied by a conventional self-consistent ML potential?	Why it matters for long-timescale MD	Failure if absent	Framework component that addresses it
Internal consistency	Forces are exact analytical derivatives of the predicted energy surface	Yes	Preserves conservative dynamics and prevents unphysical numerical energy drift	Integration remains stable but may still follow the wrong energy landscape	Base predictor
External correctness	Predicted energy/force landscape remains close to the reference quantum-mechanical surface	No	Determines whether conserved dynamics are physically meaningful rather than merely mathematically coherent	Trajectory can conserve a systematically shifted or distorted surface	Residual corrector plus sparse reference updating
Error detectability	System can recognize when current predictions are likely unreliable	No	Prevents silent extrapolation into unsafe regions of configuration space	Model remains confidently wrong with no trigger for intervention	Error detector
Immediate corrigibility	System can repair an unreliable prediction during the same simulation stage	No	Limits error accumulation before it propagates over many time steps	Small force errors accumulate into positional and thermodynamic drift	Residual corrector
Adaptive reference acquisition	System can selectively request new high-fidelity labels when uncertainty remains unresolved	No	Keeps DFT cost sparse while expanding trust only where needed	Either excessive DFT expense or uncorrected extrapolation persists	Online updater
Knowledge retention under	New corrections do not erase earlier useful corrections	No	Supports stable performance across recurring motifs and	Catastrophic forgetting degrades	Memory manager

streaming updates			revisited regions of configuration space	previously corrected behavior	
Long-timescale reliability	Accuracy remains acceptable across nanosecond-to-microsecond trajectories, not just at isolated configurations	No	Aligns evaluation with actual materials-simulation use cases	Static benchmark success masks deployment failure	Full five-component architecture

A self-consistent model can still be precisely wrong. Its energy surface may possess the correct local curvature everywhere (hence forces match the derivative) while remaining globally offset from the true surface by a constant or slowly varying shift [17–19]. The simulation appears energetically stable yet evolves on an incorrect landscape. Worse, when the system explores atomic configurations far from the original training distribution, the model must extrapolate [4, 5, 13]. No amount of internal consistency can detect or repair the resulting inaccuracies [15, 20].

This paper therefore proposes a conceptual framework for error-correcting machine learning potentials that employ on-the-fly residual learning. The framework treats error correction as an intrinsic, continuously operating part of the simulation loop rather than a post hoc retraining step [11, 13, 17]. It builds directly on established foundations: the scalable base predictors of Zhang et al. [1], the uncertainty-aware active-learning protocols of Jinnouchi et al. [12, 14] and Vandermause et al. [15], the residual-correction ideas explored in recent work on error-aware force fields [17], and the on-the-fly generation strategies reviewed by Deringer et al. [5] and Unke et al. [6].

The central insight is that residual learning—training a lightweight model to predict the difference between the base predictor and the true reference—can be performed incrementally during the simulation itself [17, 21, 22]. By coupling this residual corrector with a fast error detector and a memory-efficient online updater [11, 12, 23], the potential becomes capable of detecting its own mistakes, repairing them in real time, and continuously improving its accuracy as the simulation proceeds. The framework is deliberately model-agnostic: any self-consistent base predictor (DeepMD [1], NequIP [3], or Gaussian-process-based potentials [4, 5]) can serve as the foundation. The additional components add only modest overhead while delivering the missing capability of lifelong adaptation.

The remainder of this article develops the framework in detail. Section 2 explains why self-consistency, although essential, cannot solve the drift and extrapolation problems [17, 18]. Section 3 defines the five core components and their information flows. Section 4 presents the operational protocol that governs on-the-fly residual learning [11, 14]. Section 5 articulates seven operational principles required for practical deployment in production molecular dynamics workflows. Together these sections provide a complete conceptual blueprint for constructing the next generation of truly adaptive machine learning potentials.

Why Self-Consistency is Not Enough

Self-consistency in machine learning potentials means that the force on each atom is obtained as the negative gradient of the predicted total energy with respect to atomic positions [1, 3]. This property delivers three critical benefits. First, it ensures that the forces are conservative, eliminating spurious energy creation or destruction during integration of the equations of motion. Second, it guarantees perfect energy conservation in microcanonical (NVE) simulations, allowing long trajectories without artificial thermostating artifacts. Third, it satisfies fundamental physical requirements such as translational and rotational invariance when the underlying architecture is properly designed [1, 3, 24].

These advantages explain why nearly every modern potential—from Deep Potential Molecular Dynamics [1] to equivariant graph networks [3]—enforces self-consistency by construction. Yet self-consistency addresses only the

internal mathematical structure of the model. It provides no information about how closely the predicted energy surface matches the true quantum-mechanical surface [6, 17, 19]. A model can be perfectly self-consistent while still being systematically inaccurate [13, 17].

Consider a hypothetical but realistic scenario. The base predictor learns an energy surface whose local curvature (second derivatives) matches the reference Hessian everywhere within the training domain. Forces therefore equal the analytic gradient. During a long simulation, however, the surface sits 0.1 eV/atom above the true surface. Every force evaluation is correct relative to the model's own energy, yet the entire trajectory evolves on a parallel but offset landscape. The simulation conserves energy perfectly; it simply conserves the wrong energy. Self-consistency checks pass, yet the physics is incorrect [17, 18].

Even when the average offset is small, local force errors of only 0.01 eV/Å—well below typical training tolerances—accumulate over nanosecond timescales [14, 16]. The resulting positional drift can exceed several angstroms, rendering free-energy calculations or diffusion coefficients meaningless. Self-consistency offers no protection against this gradual divergence because it contains no external reference [17].

The extrapolation problem compounds the issue. Molecular dynamics inevitably drives the system into configurations distant from the original training set. Bartók *et al.* [4] demonstrated that even sophisticated kernels struggle when the system leaves the convex hull of the training data. Jinnouchi *et al.* [12, 14] and Vandermause *et al.* [15] showed that on-the-fly active learning can mitigate this by adding new points, yet their methods still treat the potential as a static entity between queries. No mechanism exists inside the model to recognize that a prediction is unreliable and to correct it immediately [7, 10, 20, 25].

In short, self-consistency ensures that the model is internally coherent. It does not ensure that the model is right [17]. Error detection, error correction, and continuous adaptation must be added as first-class citizens of the potential architecture [11, 17]. Only then can machine learning potentials remain reliable across the full duration of realistic materials simulations.

Components of an Error-Correcting ML Potential

An error-correcting machine learning potential is constituted by interdependent modules that co-evolve throughout molecular dynamics, forming a closed-loop system in which prediction, evaluation, and adaptation remain tightly coupled. At its core lies a base predictor, instantiated through established self-consistent architectures such as Deep Potential Molecular Dynamics [1], E(3)-equivariant graph neural networks [3], or Gaussian-process regression models [2, 4, 5, 24]. This component sustains the primary estimation of energies and forces at each time step, preserving its role as a computationally efficient and internally coherent surrogate anchored in the original training distribution. Its stability, however, is complemented by a continuously operating error detector that interrogates each prediction with minimal overhead, yielding either a binary signal or a graded confidence measure. Implementations range from ensemble variance estimates [15, 26] to distance-based metrics relative to the training manifold [13, 25], as well as physically informed diagnostics sensitive to deviations in bond-order or coordination patterns. Under these conditions, the detector functions as a gatekeeper, identifying configurations where the base model's extrapolation risk becomes non-negligible [20].

This selective intervention activates a residual corrector that introduces an immediate, additive refinement to both energy and force predictions [17]. Designed to remain computationally subordinate to the base predictor, the corrector typically adopts lightweight forms such as sparse Gaussian processes or compact neural networks, and is trained to approximate the residual between reference values and base outputs using data accrued during the simulation [17, 22, 27]. Because it operates within the same descriptor space, the resulting corrections preserve differentiability and maintain thermodynamic consistency. Beyond this immediate adjustment, an online updater governs the adaptive learning process by monitoring the corrector's uncertainty; once predefined thresholds are exceeded, it invokes density-functional-theory

evaluations, computes updated residuals, and incrementally refines model parameters through controlled learning schedules [11, 12, 14, 16]. This adaptive mechanism introduces a dynamic interplay between accuracy and computational cost. Sustaining this process over extended trajectories requires a memory manager that mitigates catastrophic forgetting by retaining a strategically curated subset of past configurations and associated residuals. Through mechanisms such as reservoir sampling or importance weighting, it preserves representational diversity within fixed resource constraints, ensuring that corrections remain valid across disparate regions of configuration space [5, 18, 23].

Figure 1 illustrates the hierarchical architecture through which self-consistent base predictions are augmented by error detection, residual correction, and online adaptation to prevent drift and maintain long-timescale fidelity.

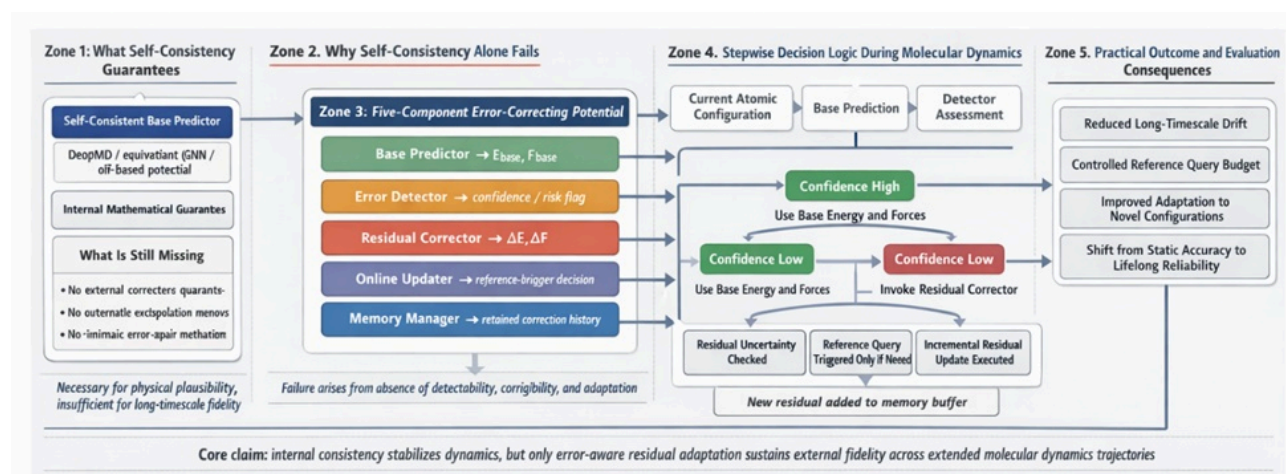


Figure 1. Hierarchical Architecture of Error Detection, Residual Correction, and Online Adaptation in Self-Consistent Machine Learning Potentials

On-The-Fly Residual Learning Protocol

The operational protocol embeds the five components into the molecular dynamics time loop. It begins with a well-trained, self-consistent base predictor initialized from an initial density-functional-theory dataset [1-3, 7]. No residual corrector exists at time zero.

At each integration step the base predictor evaluates energy and forces. The error detector immediately assesses reliability [15, 20, 26]. If the confidence score exceeds a predefined threshold the simulation advances using the base values alone. If confidence falls below the threshold the residual corrector is queried [17, 27]. Using the current atomic environment and any relevant history stored in the memory manager, the corrector predicts the residual correction. The corrected energy and forces are applied for integration.

If, after correction, the residual corrector's internal uncertainty remains above a second threshold, the online updater flags the configuration for a density-functional-theory query [11, 12, 16]. The expensive reference calculation is performed (typically in parallel on a separate compute resource), the true residual is computed, and the new data point is added to the memory manager's replay buffer. The online updater then performs one or more incremental gradient steps on the residual corrector, using a small batch drawn from the buffer to maintain stability [14].

Several design choices govern the protocol's efficiency and robustness. The error detector can be uncertainty-based, distance-based, or consistency-based; each offers different sensitivity versus false-positive trade-offs [15, 25]. The correction trigger can be a fixed threshold or an adaptive schedule that tightens as the simulation progresses. The

corrector architecture—Gaussian process with sparse inducing points, small neural network, or even linear model—balances accuracy against update speed [5, 17, 22, 27]. Update frequency can be every flagged step or accumulated into mini-batches to amortize overhead. Memory management employs reservoir sampling or importance weighting to keep the buffer size constant while preserving coverage of the explored configuration space [5, 18, 23].

The active-learning trigger for density-functional-theory queries is deliberately conservative. Queries occur only when (i) epistemic uncertainty in the base predictor is high [15, 20], (ii) the configuration lies far from previously seen data [4, 25], (iii) the residual corrector itself expresses high uncertainty [17], or (iv) a physical inconsistency is detected. This multi-criteria gating keeps the number of expensive calculations to a minimum—typically a few per nanosecond of simulated time—while still supplying the residual corrector with the information it needs to improve [11, 12, 14, 16].

The protocol is fully online: no offline retraining phase is required after the simulation begins. The residual corrector continuously refines its understanding of the base predictor's systematic errors, turning the entire molecular dynamics run into an active, self-improving learning process [11, 17].

Operational Principles for Deployment

Robust deployment of the error-correcting framework in production molecular dynamics hinges on a set of design commitments that preserve both numerical stability and computational efficiency under sustained operation. Central to this architecture is the requirement that correction mechanisms augment rather than compensate for deficiencies in the underlying model, which necessitates a base predictor that is already self-consistent and exhibits reliable accuracy across the anticipated configuration space [1, 3, 24]. This foundational constraint places immediate pressure on the surrounding modules to operate with minimal overhead, particularly the error detector, whose continuous invocation at each time step demands that its computational footprint remain tightly bounded relative to the base evaluation. In practice, this favors parsimonious strategies such as distance-based measures or low-cost ensemble statistics that can signal extrapolative risk without perturbing overall throughput [15, 20, 26]. A similar economy governs the residual corrector, whose contribution must remain subordinate in cost while still capturing systematic deviations; lightweight formulations, including sparse Gaussian processes or shallow neural architectures, achieve this balance by delivering targeted corrections without introducing prohibitive latency [5, 17, 22, 27].

Sustained accuracy under streaming data conditions further depends on controlled adaptation, where online updates are carefully regulated to avoid destabilizing previously acquired knowledge. Mechanisms such as replay buffers or elastic weight consolidation become essential in maintaining continuity across evolving regions of configuration space, particularly as new corrections are assimilated [5, 18, 23]. This adaptive loop is constrained by the need to preserve computational tractability, which in turn imposes a strict budget on density-functional-theory evaluations, ensuring that high-fidelity queries remain sufficiently sparse to prevent them from dominating wall-clock time [11, 12, 14, 16]. Within this bounded regime, periodic validation against held-out reference configurations provides a necessary diagnostic, enabling the system to recalibrate thresholds or request additional data when correction fidelity begins to degrade [17]. Even under adverse conditions, operational resilience is maintained through fallback behavior that privileges continuity over optimality: when memory limits are reached or updates become unstable, the framework reverts to the base predictor while logging the deviation, thereby preserving simulation integrity without interruption [18].

Table 2 consolidates the central engineering trade-offs of the framework by showing how detector design, residual-model choice, update frequency, memory size, and reference-query policy jointly determine the balance between reliability, stability, and computational overhead.

Table 2. Component-Level Design Trade-Offs for Deploying On-the-Fly Residual Learning in Production Molecular Dynamics

Design choice	Primary benefit	Principal trade-off	Risk if underdesigned	Risk if overdesigned	Recommended evaluation criterion in this framework
Error detector based on ensemble variance	Strong epistemic sensitivity and interpretable risk signaling	Requires multiple model evaluations	Missed unreliable states if ensemble is too small or poorly calibrated	Excessive overhead and false positives if ensemble is too large	Detection precision/recall against reference-triggered failure cases
Error detector based on distance-to-training-set metrics	Very low per-step computational cost	Proxy may not track true physical error well	Silent failure in novel but geometrically nearby states	Over-triggering in benign but distributionally unusual states	Correlation between distance score and reference residual magnitude
Physical-consistency checks in detector	Captures chemistry-aware anomalies missed by pure statistics	Requires hand-crafted or domain-specific thresholds	Important anomalies remain invisible	Overconstrained detector blocks legitimate rare events	Rate of physically meaningful interventions per 10,000 steps
Sparse Gaussian-process residual corrector	Fast uncertainty-aware local correction	Limited scalability in very high-volume streaming settings	Corrections remain too local and undergeneralized	Inducing-point growth increases latency	Correction gain per query and latency per flagged step
Small neural residual corrector	Flexible nonlinear residual modeling	Harder uncertainty calibration than GP-based correctors	Underfits systematic residual structure	Unstable online learning and oscillatory corrections	Post-correction force/energy error plus update stability
Per-step online updates	Fastest adaptation to newly encountered states	Greater risk of instability and noisy parameter drift	Important new errors persist too long before learning occurs	Correction model becomes reactive and unstable	Error reduction half-life after first flagged novel event
Mini-batch periodic updates	Better optimization stability and amortized cost	Slower response to emerging failure modes	Delayed adaptation allows drift accumulation	Oversized batches reduce responsiveness	Accuracy-overhead balance across fixed wall-clock budget
Small replay buffer	Low memory footprint	Limited diversity retention	Earlier corrections are forgotten quickly	—	Retention of correction quality on revisited motifs

Large replay buffer	Better coverage of explored space	Higher memory and sampling cost	—	Slower updates and redundant data accumulation	Marginal gain in retained accuracy per added memory unit
Aggressive DFT query threshold	Strong protection against extrapolation failure	High reference cost	—	Query budget dominates simulation runtime	Total DFT calls per nanosecond at target drift tolerance
Conservative DFT query threshold	Preserves efficiency	Greater chance of uncorrected error propagation	Drift persists before intervention occurs	—	Drift accumulation versus query count
Graceful fallback to base predictor	Preserves simulation continuity during instability	Sacrifices adaptive benefit temporarily	No safe recovery path if updater destabilizes	Too frequent fallback nullifies the framework's value	Fraction of simulation completed under safe degraded mode

Collectively these principles transform the conceptual framework into a practical engineering solution. They guarantee that error correction adds value without introducing unacceptable overhead or instability, making the approach viable for the long-timescale simulations that materials scientists require [6, 8, 10].

Relation to Existing Methods

The proposed framework builds upon and extends several established strands of research in machine learning potentials while introducing a distinct emphasis on continuous, on-the-fly residual correction.

Active learning has been a cornerstone for improving interatomic potentials by strategically selecting new density-functional-theory calculations to augment training data. Jinnouchi *et al.* [12] demonstrated on-the-fly active learning of interatomic potentials for melting-point calculations, showing how uncertainty-driven queries can expand the applicability domain during simulation. Their later work [14] further refined the active-learning loop to handle thousands of atoms efficiently. Vandermause *et al.* [15] introduced interpretable Bayesian force fields that incorporate active learning for rare events, while Sivaraman *et al.* [25] applied similar strategies to amorphous and liquid systems. These approaches, however, focus on expanding the training set for a single static potential [6, 7, 10]. The present framework extends active learning into a continuous correction layer: rather than retraining the entire base model, it maintains the base predictor unchanged and learns only the residuals on the fly [17, 27], thereby achieving adaptation without the cost of full model retraining.

Ensemble-based uncertainty estimation provides another natural connection. Vandermause *et al.* [15] and related uncertainty-guided methods use variance across multiple models to flag unreliable predictions [26]. The error detector in the current framework can directly incorporate such ensemble signals or simpler distance-to-training metrics, turning uncertainty from a diagnostic tool into a trigger for residual correction [20].

On-the-fly learning literature offers the closest precedent. The placeholder protocol for on-the-fly generation of machine learning potentials during molecular dynamics [11] and the residual-based error correction explored in recent work [17] already hint at incremental updates. Deringer *et al.* [5] reviewed Gaussian process regression techniques that support online adaptation, and Unke *et al.* [6] surveyed machine learning force fields with adaptive elements. The present

framework generalizes these as special cases: any on-the-fly method that updates a single model can be recast as a base predictor plus a lightweight residual corrector managed by the online updater [11, 13, 16, 17].

Finally, the residual-learning aspect resonates with meta-learning concepts, in which a model learns how to learn quickly from limited examples. By training the residual corrector to rapidly capture systematic errors of the base predictor, the framework implicitly performs a form of meta-adaptation during the simulation [17, 22, 23]. This distinguishes it from purely static or periodically retrained potentials [1, 3] and positions error-correcting potentials as a bridge between active learning, on-the-fly generation, and lifelong adaptation [12, 15].

Challenges and Open Questions

Despite its conceptual elegance, the error-correcting framework faces several practical and theoretical challenges that must be addressed before widespread adoption.

The first challenge is computational overhead. Error detection, residual correction, and incremental updates all add latency to each molecular dynamics step. Even when the residual corrector is deliberately lightweight, the cumulative cost over microsecond-scale simulations can become significant [8, 11, 17]. Efficient implementations that keep overhead below 50 % of the base predictor remain an open engineering task.

Stability of online learning constitutes the second challenge. Incremental gradient updates on streaming data can destabilize the residual corrector, producing oscillating corrections that introduce artificial forces. Guarantees of convergence and bounded error growth during continuous training are still lacking and will require new theoretical tools tailored to molecular dynamics trajectories [18, 21, 28].

Catastrophic forgetting is the third critical issue. As the simulation explores new regions, the memory manager's replay buffer may discard earlier corrections that remain relevant for recurring motifs. Strategies such as importance sampling or elastic weight consolidation, already explored in related adaptive potentials [5, 23], must be refined specifically for residual learning.

Error detector calibration presents a fourth difficulty. Poorly calibrated detectors generate either excessive false positives—triggering unnecessary density-functional-theory queries—or false negatives that allow uncorrected errors to propagate. Achieving well-calibrated uncertainty estimates that generalize across different materials systems is an open research question [15, 20, 26].

The fifth challenge concerns correction generalization. Residuals learned on one trajectory may not transfer to another simulation even within the same material, because error patterns can be trajectory-specific [13, 17]. Techniques to encourage broader generalization of the residual corrector without sacrificing on-the-fly speed need to be developed.

Finally, the cost of density-functional-theory queries, even when limited to a few per nanosecond, remains a bottleneck for truly long simulations [12, 14, 16]. Developing cheaper surrogate oracles or hierarchical correction strategies that reduce reliance on full reference calculations is essential for scalability [11, 27].

Addressing these challenges will require concerted efforts in algorithm design, memory management, and uncertainty quantification [5–7, 10]. Until they are solved, the framework's promise of fully adaptive potentials will remain partially unrealized.

Implications for ML Potential Development

The error-correcting framework carries concrete implications for three stakeholder groups: model developers, practitioners, and benchmark designers.

For model developers, the framework encourages designing base predictors with built-in support for error detection. Architectures that natively output uncertainty estimates or expose intermediate descriptors will integrate more seamlessly with the error detector and residual corrector [13, 15, 20, 26]. Developers should also provide standardized interfaces for lightweight residual modules so that existing potentials such as Deep Potential Molecular Dynamics [1] or E(3)-equivariant networks [3] can be upgraded without architectural overhaul [9, 24]. Most importantly, future models should be benchmarked not only on static force and energy errors but on long-timescale stability metrics that quantify drift over nanoseconds [6, 10, 14].

Practitioners running production molecular dynamics simulations gain a clear decision guideline. For short runs below 100 ps, a well-trained base model may still suffice [1, 3]. For simulations exceeding 1 ns—common in studies of diffusion, phase transformations, or defect migration—the error-correcting layer becomes essential [9, 11, 12, 16]. Practitioners should allocate a modest density-functional-theory budget (1–10 queries per nanosecond) and monitor validation metrics periodically to ensure the residual corrector improves rather than degrades performance [17, 29].

Benchmark designers must evolve current evaluation protocols. Existing static test sets capture accuracy at fixed configurations but miss the drift and extrapolation problems that dominate real workflows [4, 5]. New long-MD benchmarks spanning nanosecond timescales, with explicit reporting of trajectory drift and cumulative error, will be required. Such benchmarks will drive community progress toward potentials that remain reliable throughout extended simulations rather than merely accurate at the outset [6, 7, 10, 14].

Collectively, these implications shift the focus of machine learning potential research from one-time training to lifelong adaptation, aligning development incentives with the actual demands of materials engineering applications [11, 17].

Toward Fully Adaptive ML Potentials

The ultimate vision enabled by this framework is a class of fully adaptive machine learning potentials that continuously improve during deployment and never stop learning. Instead of being frozen after initial training, potentials become living systems that detect, correct, and learn from their own errors in real time, adapting to any new physics encountered during simulation [11, 13, 17, 23].

A short-term roadmap (1–2 years) focuses on hybrid implementations: error detection and residual correction run online while periodic offline retraining of the residual corrector occurs between simulation segments [11, 16]. This staged approach allows immediate testing of the conceptual components without requiring fully incremental optimizers.

In the medium term (2–5 years), fully online updates with controlled forgetting become feasible. Advances in memory management and stable incremental learning will allow the residual corrector to evolve smoothly throughout microsecond-scale simulations while preserving knowledge of earlier error patterns [5, 18, 23].

The long-term goal (5–10 years) is autonomous, self-improving potentials that require no human intervention. Such systems will automatically decide when to request reference data, update their correction modules, and even adjust their own error-detection thresholds as simulation conditions change [12, 15, 29]. Success will be measured by a simple yet demanding criterion: a potential that sustains 1 microsecond of molecular dynamics with error drift below 0.1 eV/atom while issuing fewer than 100 density-functional-theory queries in total [8, 14].

Achieving this vision will transform computational materials science. Long-timescale phenomena that today require heroic human-guided active learning will become routine [12, 16, 25]. Error-correcting potentials will enable predictive simulations of complex processes—irradiation damage, glass formation, catalytic cycles—under conditions where current static models inevitably fail [4, 5, 10, 19, 29]. The framework therefore marks not merely an incremental improvement but

a paradigm shift toward potentials that learn continuously, just as the physical systems they model evolve continuously [17].

Conclusion

Self-consistency is not enough. Machine learning potentials have delivered remarkable accuracy and scalability, yet their static nature causes errors to accumulate and trajectories to drift during long molecular dynamics simulations. The conceptual framework presented here addresses this limitation by embedding error detection, residual correction, and on-the-fly learning directly into the simulation loop.

The architecture rests on five components—base predictor, error detector, residual corrector, online updater, and memory manager—that operate as a closed-loop system. The operational protocol triggers corrections only when needed, queries reference data sparingly, and maintains stability through principled memory management. Seven deployment principles ensure that the added capabilities remain practical for production workflows.

Relations to active learning, ensemble uncertainty, and existing on-the-fly methods show that the framework generalizes and unifies prior advances while adding the missing residual-learning layer. Challenges of overhead, stability, forgetting, calibration, generalization, and reference cost are acknowledged and framed as concrete targets for future research. Implications for developers, practitioners, and benchmark designers point toward a new generation of potentials evaluated on long-timescale fidelity rather than static accuracy alone.

The roadmap toward fully adaptive potentials is clear: from hybrid implementations to fully autonomous systems capable of microsecond simulations with minimal intervention. The community is invited to implement, refine, and benchmark error-correcting machine learning potentials so that the next decade of materials modeling can rest on foundations that are not only accurate but truly self-improving.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 17 Dec 2023

Revised: 24 Mar 2024

Accepted: 21 May 2024

Published online: 18 July 2024

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

- 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>.
- Chan H, Narayanan B, Cherukara MJ, Sen FG, Sasikumar K, Gray SK, et al. Machine learning classical interatomic potentials for molecular dynamics from first-principles training data. *J Phys Chem C.* 2019;123(12):6941-57.
<https://doi.org/10.1021/acs.jpcc.8b09917>.
- Batzner S, Musaelian A, Sun L, Geiger M, Mailoa JP, Kornbluth M, et al. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat Commun.* 2022;13(1):2453.
<https://doi.org/10.1038/s41467-022-29939-5>.
- Bartók AP, De S, Poelking C, Bernstein N, Kermode JR, Csányi G, et al. Machine learning unifies the modeling of materials and molecules. *Sci Adv.* 2017;3(12):e1701816.
<https://doi.org/10.1126/sciadv.1701816>.
- Deringer VL, Bartók AP, Bernstein N, Wilkins DM, Ceriotti M, Csányi G. Gaussian process regression for materials and molecules. *Chem Rev.* 2021;121(16):10073-141.
<https://doi.org/10.1021/acs.chemrev.1c00022>.
- 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>.
- Kulichenko M, Nebgen B, Lubbers N, Smith JS, Barros K, Allen AE, et al. Data generation for machine learning interatomic potentials and beyond. *Chem Rev.* 2024;124(24):13681-714.
<https://doi.org/10.1021/acs.chemrev.4c00572>.
- 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>.
- Yu W, Ji C, Wan X, Zhang Z, Robertson J, Liu S, et al. Machine-learning-based interatomic potentials for advanced manufacturing. *Int J Mech Syst Dyn.* 2021;1(2):159-72.
<https://doi.org/10.1002/msd2.12021>.
- Eyert V, Wormald J, Curtin WA, Wimmer E. Machine-learned interatomic potentials: Recent developments and prospective applications. *J Mater Res.* 2023;38(24):5079-94.
<https://doi.org/10.1557/s43578-023-01239-8>.
- Hu D, Xie Y, Li X, Li L, Lan Z. Inclusion of machine learning kernel ridge regression potential energy surfaces in on-the-fly nonadiabatic molecular dynamics simulation. *J Phys Chem Lett.* 2018;9(11):2725-32.
<https://doi.org/10.1021/acs.jpclett.8b00684>.

Jinnouchi R, Karsai F, Kresse G. On-the-fly machine learning force field generation: Application to melting points. *Phys Rev B*. 2019;100(1):014105.
<https://doi.org/10.1103/PhysRevB.100.014105>.

Nguyen NC, Sema D. Environment-adaptive machine learning potentials. *Phys Rev B*. 2024;110(6):064101.
<https://doi.org/10.1103/PhysRevB.110.064101>.

Jinnouchi R, Miwa K, Karsai F, Kresse G, Asahi R. On-the-fly active learning of interatomic potentials for large-scale atomistic simulations. *J Phys Chem Lett*. 2020;11(17):6946-55.
<https://doi.org/10.1021/acs.jpclett.0c01061>.

Vandermause J, Torrisi SB, Batzner S, Xie Y, Sun L, Kolpak AM, et al. On-the-fly active learning of interpretable Bayesian force fields for atomistic rare events. *NPJ Comput Mater*. 2020;6(1):20.
<https://doi.org/10.1038/s41524-020-0283-z>.

Cui C, Zhang Y, Ouyang T, Chen M, Tang C, Chen Q, et al. On-the-fly machine learning potential accelerated accurate prediction of lattice thermal conductivity of metastable silicon crystals. *Phys Rev Mater*. 2023;7(3):033803.
<https://doi.org/10.1103/PhysRevMaterials.7.033803>.

González-Muñiz A, Díaz I, Cuadrado AA, García-Pérez D, Pérez D. Two-step residual-error based approach for anomaly detection in engineering systems using variational autoencoders. *Comput Electr Eng*. 2022;101:108065.
<https://doi.org/10.1016/j.compeleceng.2022.108065>.

Gao A, Remsing RC. Self-consistent determination of long-range electrostatics in neural network potentials. *Nat Commun*. 2022;13(1):1572.
<https://doi.org/10.1038/s41467-022-29243-2>.

Anstine DM, Isayev O. Machine learning interatomic potentials and long-range physics. *J Phys Chem A*. 2023;127(11):2417-31.
<https://doi.org/10.1021/acs.jpca.2c06778>.

Mueller T, Hernandez A, Wang C. Machine learning for interatomic potential models. *J Chem Phys*. 2020;152(5):050902.
<https://doi.org/10.1063/1.5126336>.

Zhou ZH. Machine Learning. Liu S, translator. Singapore: Springer; 2021.
<https://doi.org/10.1007/978-981-15-1967-3>.

Fang W, Yu Z, Chen Y, Huang T, Masquelier T, Tian Y. Deep residual learning in spiking neural networks. *Adv Neural Inf Process Syst*. 2021;34:21056-69.
<https://doi.org/10.5555/3540261.3541872>.

Hie BL, Yang KK. Adaptive machine learning for protein engineering. *Curr Opin Struct Biol*. 2022;72:145-52.
<https://doi.org/10.1016/j.sbi.2021.11.002>.

Takamoto S, Izumi S, Li J. TeaNet: Universal neural network interatomic potential inspired by iterative electronic relaxations. *Comput Mater Sci*. 2022;207:111280.
<https://doi.org/10.1016/j.commatsci.2022.111280>.

Sivaraman G, Krishnamoorthy AN, Baur M, Holm C, Stan M, Csányi G, et al. Machine-learned interatomic potentials by active learning: Amorphous and liquid hafnium dioxide. *NPJ Comput Mater*. 2020;6(1):104.
<https://doi.org/10.1038/s41524-020-00367-7>.

Heid E, Schörghuber J, Wanzenböck R, Madsen GKH. Spatially resolved uncertainties for machine learning potentials. *J Chem Inf Model*. 2024;64(16):6377-87.
<https://doi.org/10.1021/acs.jcim.4c00904>.

Cao L, O'Leary-Roseberry T, Jha PK, Oden JT, Ghattas O. Residual-based error correction for neural operator accelerated infinite-dimensional Bayesian inverse problems. *J Comput Phys.* 2023;486:112104.
<https://doi.org/10.1016/j.jcp.2023.112104>.

Shi Z, Yang W, Deng X, Cai C, Yan Y, Liang H, et al. Machine-learning-assisted high-throughput computational screening of high performance metal–organic frameworks. *Mol Syst Des Eng.* 2020;5(4):725-42.
<https://doi.org/10.1039/D0ME00005A>.

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>.