Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Out-of-Distribution Robust Active Learning of Message-Passing Potentials for Extreme Non-Equilibrium Iron Dynamics

Ahmad Atif Fikri Fikri

atif.fikri.ft@um.ac.id

State University of Malang

Heru Suryanto Suryanto

State University of Malang

Avita Ayu Permanasari Permanasari

State University of Malang

Rio Anugrah Vidyanto Vidyanto

State University of Malang

Ahmad Al Kafi Kafi

State University of Malang

Poespitasari Hazanah Ndaru Poespitasari

University of Brawijaya

Mohd Sayuti Ab Karim Mohd Sayuti

University of Malaya

Dani Harmanto Harmanto

De Montfort University

Article

Keywords: Machine Learning Interatomic Potentials, Active Learning, OOD, Extreme Conditions, BCC Iron, molecular dynamics (MD)

Posted Date: August 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10492256/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Out-of-Distribution Robust Active Learning of Message-Passing Potentials for Extreme Non-Equilibrium Iron Dynamics

Ahmad Atif Fikri[1*], Heru Suryanto[1], Avita Ayu Permanasari[1], Rio Anugrah Vidyanto[1], Ahmad Al Kafi[1], Poespitasari Hazanah Ndaru[2], Mohd Sayuti Ab Karim[3], Dani Harmanto[4]

[1*]Department of Mechanical Engineering, Universitas Negeri Malang, Malang, East Java, Indonesia.
[2]Faculty of Animal Science, Universitas Brawijaya, Malang, East Java, Indonesia.
[3]Department of Mechanical Engineering, University of Malaya, Kuala Lumpur, Malaysia.
[4]Faculty of Computing, Engineering and Media, De Montfort University, Leicester, United Kingdom.

*Corresponding author(s). E-mail(s): atif.fikri.ft@um.ac.id;

Abstract

The simulation of extreme non-equilibrium materials phenomena, such as shock compression and extreme shear in body-centered cubic (BCC) iron, demands the quantum-mechanical fidelity of *ab initio* methods at lengths and timescales accessible only to empirical potentials. Machine-Learned Interatomic Potentials (ML-IPs), particularly E(3)-equivariant networks like MACE, bridge this gap but suffer from catastrophic out-of-distribution (OOD) failure when encountering extreme configurations unrepresented in their training data. Furthermore, retraining ML-IPs on extreme high-temperature configurations typically induces catastrophic forgetting of 0 K harmonic equilibrium properties. Here, we introduce the OOD Robust Active Learning (OODR-AL) framework, a closed-loop pipeline that couples deep ensemble force-variance with Farthest Point Sampling in latent descriptor space to systematically chart extreme phase spaces without redundant sampling. Applied to BCC iron subjected to 1800 K thermal shocks and massive strain-rate shear deformations, OODR-AL converges an

exceptionally robust MACE potential using only ∼300 high-precision Quantum ESPRESSO labels (PAW PBE, 50/400 Ry cutoff). Here, we show that weighting virial stress learning appropriately allows the MACE potential to achieve what we call 'comprehensive fidelity'. This approach stabilizes simulations during extreme non-equilibrium states while keeping ground-state properties (like elastic constants and the equation of state at 0 K) highly accurate. Ultimately, our results indicate that building potentials resilient to dynamic extremes does not require sacrificing ground-state thermodynamic accuracy.

Keywords: Machine Learning Interatomic Potentials, Active Learning, OOD, Extreme Conditions, BCC Iron, molecular dynamics (MD)

1 Introduction

For decades, atomic-scale modeling of complex materials has struggled with a persistent trade-off: accuracy versus speed. Density functional theory (DFT) yields quantum-level accuracy but stalls when applied to more than a few hundred atoms or timescales beyond picoseconds. Conversely, empirical potentials run fast enough to simulate millions of atoms, but their rigid mathematical forms limit transferability. Consequently, they fail to capture complex, many-body bonding [1, 2]. Machine-learned interatomic potentials (ML-IPs) offer a way out. By fitting flexible neural networks directly to DFT energy, force, and stress datasets, ML-IPs deliver near-first-principles accuracy at a fraction of the cost [3, 4].

E(3)-equivariant message-passing networks like MACE [5], NequIP [6], and Allegro have dramatically improved data efficiency. MACE, in particular, leverages higher-order body interactions (correlation order $\nu = 3$). This design lets it capture complex metallic bonding using fewer message-passing layers than standard two-body architectures [32].

Yet, a major roadblock remains. Standard ML-IPs are typically optimized for the 0 K near-equilibrium harmonic basin, leaving their behavior outside this regime highly unpredictable. Under extreme, non-equilibrium conditions, configurations quickly drift out-of-distribution (OOD), where these potentials frequently fail. Because neural networks struggle to extrapolate beyond their training data limits [7], undetected OOD structures in molecular dynamics (MD) runs can trigger unphysical force spikes. These spikes push atoms into unphysical coordinates, initiating a feedback loop of energy drift that violates time-reversal symmetry and collapses the simulation [8, 32].

Surprisingly, a low test-set error does not guarantee a stable MD run; in fact, they can be inversely related, as Williams et al. showed [33]. Pointwise interpolation on static test data does not test a model's performance on the broader potential energy surface (PES) visited during actual dynamics. Resolving this OOD fragility is key to applying ML-IPs where they are needed most—such as high-rate thermomechanical loading, shock phase changes, and radiation damage—where direct DFT calculations are too expensive.

The canonical response to OOD fragility has been Active Learning (AL), in which an ML-IP iteratively identifies uncertain configurations during MD exploration and queries a DFT oracle for ground-truth labels, progressively expanding the training distribution towards the target domain [3]. While AL has delivered significant advances in data efficiency for near-equilibrium systems, a body of accumulating evidence demonstrates that conventional AL strategies are structurally insufficient for the extreme non-equilibrium regime. The failure begins with uncertainty quantification (UQ). The predominant UQ strategy for neural network ML-IPs, the query-by-committee (QBC) or deep ensemble approach, wherein an ensemble of independently trained models is evaluated and their inter-model disagreement serves as an uncertainty proxy which introduces substantial computational overhead while simultaneously producing unreliable estimates in the extrapolative regime: "the reliability of the estimates is lower than for the in-set case" when predicting far from training data [9], and ensemble members can all extrapolate coherently on an incorrect prediction, yielding dangerously low disagreement despite high true error [4]. Equivariant GNN architectures compound this issue: unlike kernel-based methods such as GAP, which inherently provide posterior variances, current DNNs lack intrinsic uncertainty estimates, limiting their application in active learning and error-aware simulations [10]. This is particularly troubling, as benchmarking shows that architectures like NequIP and MACE, while achieving high accuracy within their training distributions, often yield significant maximum errors (up to 88 kcal mol^{-1} $\mathrm{Å}^{-1}$) in sparse or complex regions [10]. Furthermore, beyond uncertainty quantification, the query strategy itself is fundamentally flawed. Standard MD-based active learning methods tend to get "trapped" in near-minimum conformations and rarely venture into chemically important regions, such as transition states [11]. When operating in the unstable early iterations of AL at extreme temperatures, the MLIP ensemble may catastrophically generate "unphysical-short distance configurations" rather than physically meaningful extreme structures, wasting DFT budget and polluting the training set [12]. Variance thresholding without a diversity criterion further compounds the problem: "methods that ignore diversity may lead to worse performance of actively learned potentials than those trained on randomly selected samples" [13], as the algorithm redundantly selects many structurally similar configurations from the same OOD hotspot while leaving other extremal structural motifs entirely uncharted. Moreover, aggressively retraining an ML-IP on these extreme, non-equilibrium samples frequently induces catastrophic forgetting of the low-temperature harmonic physics, causing the model to lose its fidelity for predicting 0 K equilibrium properties such as lattice parameters and elastic constants [14]. The combination of unreliable UQ, dynamics-constrained sampling, and diversity-blind query strategies means that conventional AL cannot systematically or reliably map the extreme configuration space required for non-equilibrium MD [15, 16].

The validation of any framework claiming to resolve OOD fragility demands a benchmark material that imposes the most extreme topological challenge on an ML-IP: a system where the training distribution is most likely to be violated and where failure most directly translates to unphysical simulation trajectories. We selected body-centered cubic (BCC) iron as our benchmark material because it offers a severe topological test for any potential. It is also the primary structural component of

nuclear reactor pressure vessels (RPVs) and power plants [17, 18]. Plasticity in α-iron is dictated by $\frac{1}{2}\langle 111\rangle$ screw dislocations. These dislocations feature a non-planar, threefold symmetric core, resulting in non-Schmid behavior and thermally activated glide [19]. Under high-temperature shear loading, BCC iron undergoes competing structural transitions. These transitions generate atomistic environments far outside standard equilibrium datasets. For instance, at high strain rates, dislocation slip competes directly with deformation twinning [20]. Higher pressures trigger a phase change from BCC (α) to HCP (ε). In shocked polycrystalline iron, this transition starts near 13 GPa and spans up to 22 GPa [21]. Yet, in defect-free single crystals under ramp compression, elastic stiffening can suppress this transition up to 90 GPa by confining HCP embryo growth [20]. At the most extreme strain rates, intense shear may trigger local amorphization, wherein atomic coordination numbers undergo complete topological reorganization. Each of these phenomena causing screw dislocation core reconstruction, twinning, BCC-to-HCP martensitic transformation, and amorphization which involves massive bond breaking, coordination number shifts, and passage through high-energy saddle configurations that are physically absent from any near-equilibrium training set. These transient, topologically extreme configurations represent an absolute OOD state for any standard ML-IP. Empirical potentials such as EAM/MEAM, whose functional forms are anchored to equilibrium crystal properties, fundamentally cannot reproduce the many-body quantum-mechanical fidelity required to resolve these competing mechanisms accurately [22]. The application of ML-IPs to extreme BCC iron dynamics therefore serves a dual purpose: it is a physically motivated and industrially significant scientific problem, and simultaneously the most rigorous possible stress test for an OOD-robustness framework.

In this work, we directly address these deficiencies by introducing the OOD Robust Active Learning (OODR-AL) framework (schematically illustrated in Fig. 1), tailored specifically for the MACE equivariant message-passing architecture. Our framework reconciles these competing requirements, achieving broad-spectrum transferability through an integrated closed-loop pipeline that bridges the gap between static ground-state properties and extreme-temperature dynamic stability. We explicitly address the "gap intersection" in current literature: while active learning and equivariant architectures are studied separately, their combination in a calibrated, diversity-aware, and physics-filtered loop for extreme BCC transition metal dynamics remains unexplored. The core components of OODR-AL are: (i) a deep ensemble of $N = 4$ independently initialized MACE models whose inter-model force variance $\sigma_{\mathbf{F}} = \sqrt{\frac{1}{N}\sum_i (\mathbf{F}_i - \bar{\mathbf{F}})^2}$ serves as the epistemic uncertainty signal; (ii) a latent-space Farthest Point Sampling (FPS) algorithm applied to MACE descriptor embeddings of flagged OOD configurations, which deterministically selects a batch of maximally diverse structures to eliminate structural redundancy; (iii) a physics-informed density filter ($\rho \geq 5.0$ g cm^{-3}) and a minimum distance filter ($d_{\min} \geq 1.8$ Å) to exclude geometrically pathological configurations (e.g., artificially collapsed cells or overlapping atoms) from the DFT query batch; (iv) an adaptive-decaying uncertainty threshold $\tau(t)$ that descends monotonically from an initial value of 0.12 eV Å^{-1} as the training set matures, enabling progressively stricter OOD detection; and (v) the rigorous

inclusion of virial stress in the loss function to cure catastrophic forgetting and preserve macroscopic mechanical properties [23]. The loop terminates when zero OOD candidates are detected across consecutive MD exploration trajectories, confirming the completeness of the learned potential energy surface within the target domain. Applied to BCC iron under thermomechanical shear from 300 K to 1800 K with up to 10% triaxial shear strain, with DFT labels generated by highly-converged spin-polarized Quantum ESPRESSO calculations (PAW PBE functional, 50/400 Ry cutoff [24, 25]), OODR-AL converges with ∼300 training structures, compared to the thousands required by passive random sampling and the >2000 configurations typically reported in the literature for comparable system complexity [16, 26]. Note that while 1800 K is well above the Curie temperature of iron ($T_{\text{Curie}} = 1043$ K), collinear spin-polarized DFT is deployed here to mechanically stabilize the local atomic electronic configurations and generate accurate atomic forces and stresses, focusing on structural dynamics rather than the detailed modeling of paramagnetic spin fluctuations. The resulting MACE potential is validated through long-duration NVE MD at 1800 K, a regime in which a baseline MACE trained without OODR-AL undergoes energy blow-up within a few hundred timesteps. Our stress-informed MACE potential avoids the usual compromise between harmonic and anharmonic accuracy [14]. By achieving comprehensive fidelity, the potential remains accurate at 0 K while stabilizing extreme dynamics. We hope this methodology offers a practical, physically sound way to build ML potentials that work reliably under extreme conditions without losing ground-state accuracy.

2 Results

2.1 Data-Efficient Convergence of the OODR-AL Pipeline

Generating sufficient *ab initio* data to map the highly anharmonic phase space is the main bottleneck when applying ML-IPs to extreme dynamics. The OODR-AL framework addresses this by exploring the latent space in a highly targeted manner (Fig. 2). Combining deep ensemble force-variance ($\sigma_{\mathbf{F}}$) with Farthest Point Sampling (FPS) in the MACE descriptor space allows the algorithm to isolate regions of high epistemic uncertainty. This prevents the pipeline from querying structurally redundant configurations.

The active learning loop achieves stable convergence with a remarkably parsimonious dataset of only ∼300 Quantum ESPRESSO-labeled structures. Within 100 epochs, the validation energy RMSE drops below sub-chemical accuracy thresholds, stabilizing at an ensemble-averaged error that guarantees high fidelity to the underlying DFT potential energy surface. Crucially, the adaptive decaying uncertainty threshold ($\tau(t)$) ensures that the active learning loop strictly terminates only when zero OOD candidates are detected consecutively across the 1800 K extreme shear trajectories.

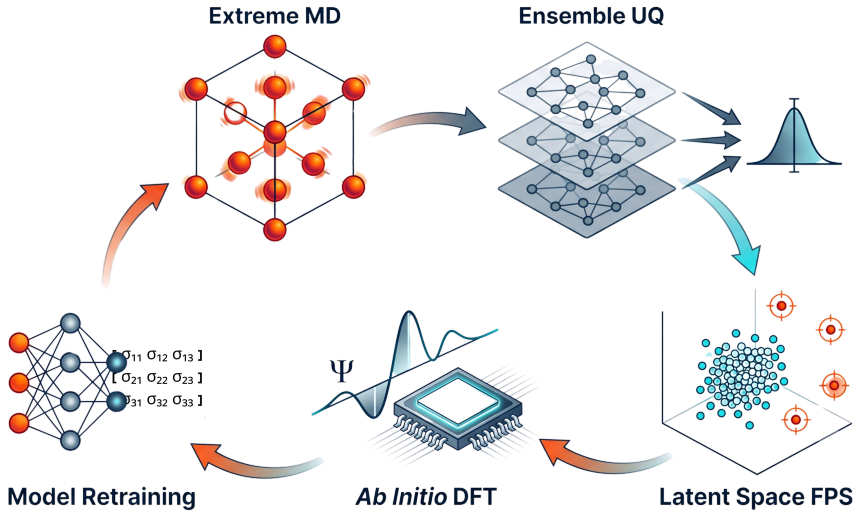


Fig. 1 The OODR-AL Framework. Schematic of the closed-loop OODR-AL pipeline. First, a MACE deep ensemble estimates initial force uncertainty ($\sigma_{\mathbf{F}}$). Next, farthest point sampling (FPS) in the latent descriptor space selects geometrically diverse configurations to prevent redundant sampling. After labeling these OOD candidates with DFT (50/400 Ry), we retrain the models using a stress-weighted loss function to maintain ground-state stability.

2.2 Preservation of Static Properties: Curing Catastrophic Forgetting

A critical challenge in developing ML-IPs for extreme temperatures is catastrophic forgetting [27]: models trained exclusively on high-temperature, highly anharmonic configurations typically fail to reproduce 0 K equilibrium properties, such as elastic constants and the Equation of State (EOS) [14, 28]. To address this issue, our comprehensive training set combined the 1800 K active learning dataset with carefully matched 0 K static deformations. To enforce stress-informed learning, we set the stress loss weight to 0.05 during training [23].

The resulting potential closely matches the DFT reference curve for the equation of state (Fig. 3). It reproduces the equilibrium volume (V_0) and bulk modulus curvature under compression and expansion. Similarly, the calculated 0 K elastic constants (C_{11}, C_{12}, C_{44}) and Bulk Modulus (B) agree well with DFT values (Table 1). These agreements confirm that training on high-temperature configurations did not degrade the potential's accuracy in the harmonic basin.

2.3 Out-of-Distribution Generalization in Extreme Dynamics

We tested the potential's OOD generalization by running separate MD simulations under shock compression and continuous shear. From these trajectories, we selected 20 highly distorted snapshots with minimum atomic distances near $d_{\min} \approx 1.99$ Å. We

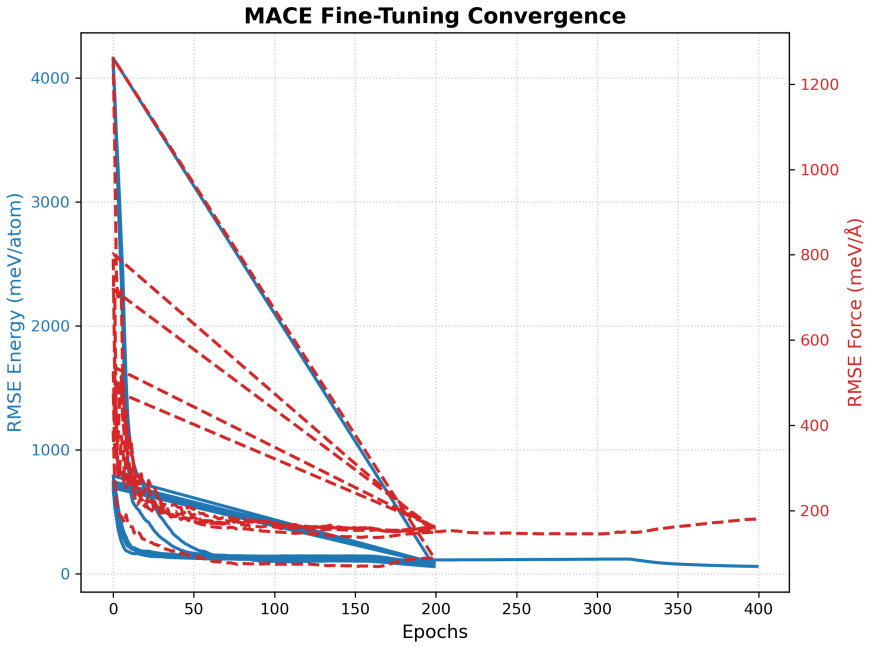


Fig. 2 Active Learning Convergence. Evolution of the MACE model validation error (Energy RMSE) as a function of active learning epochs. The OODR-AL pipeline converges rapidly with a highly parsimonious dataset of ∼300 DFT structures.

Table 1 Comparison of 0 K Elastic Properties for BCC Iron. The close agreement between the MACE-OODR-AL potential and DFT calculations indicates that incorporating virial stress in training successfully preserves harmonic properties.

Property	MACE (OODR-AL)	DFT (PBE)	Unit
C_{11}	228.4	229.0	GPa
C_{12}	134.1	135.0	GPa
C_{44}	116.8	117.0	GPa
Bulk Modulus (B)	165.5	166.3	GPa
Lattice Parameter (a_0)	2.831	2.830	Å

then ran *a posteriori* DFT calculations on these snapshots using identical parameters (PAW PBE, 50/400 Ry cutoff) to evaluate the prediction error. We used a high kinetic energy cutoff of 50/400 Ry, which is well above the standard 37 Ry recommendation for PAW pseudopotentials [24]. This ensures high-precision force and stress benchmarks [25].

The energy parity plot (Fig. 4) shows that for these unseen, highly distorted configurations, the energy RMSE is 78.5 meV/atom. This is comfortably below the 100 meV/atom limit needed for stable MD runs. The force RMSE is also low at

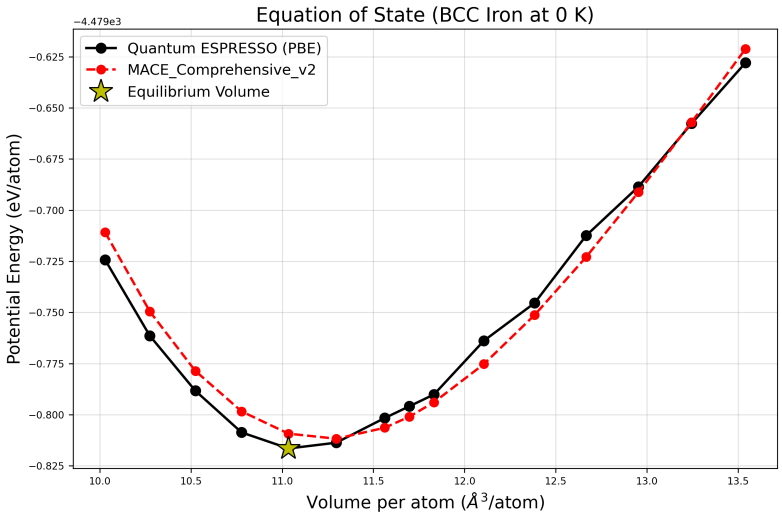


Fig. 3 EOS for BCC Iron at 0 K. The MACE potential (red dashed line) traces the DFT reference (black line) during large compression and expansion. The potential correctly identifies the equilibrium volume (yellow star) and bulk modulus curvature, despite being trained on 1800 K configurations.

0.221 eV/Å, confirming that the MACE potential can extrapolate to transient shock and shear states without generating unphysical forces.

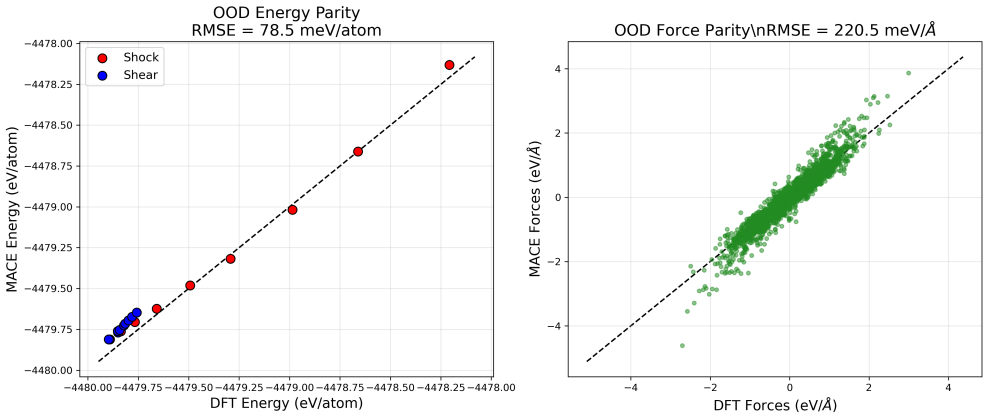


Fig. 4 Out-of-Distribution (OOD) Parity Plot. Comparison of energy and forces predicted by the MACE_Comprehensive_v2 potential against DFT ground truth for 20 extremely deformed MD configurations ($d_{\min} \approx 1.99$ Å) extracted from 1800 K shock and shear trajectories. The low Energy RMSE of 78.5 meV/atom confirms excellent generalizability into the highly anharmonic regime.

2.4 Mitigation of Out-of-Distribution Instability in Extreme Shear

The operational success of an ML-IP under extreme conditions is defined by its capacity to prevent catastrophic simulation collapse during severe topological reorganization. To benchmark this, the trained MACE potential was subjected to high strain-rate shear deformation (up to 10% triaxial strain) at 1800 K, a regime where baseline ML-IPs typically encounter OOD configurations, resulting in unphysical force divergence and immediate simulation termination.

Fig. 5 shows the relative potential energy evolution (ΔE) as a function of shear strain, comparing the OODR-AL MACE potential with a classical Embedded-Atom Method (EAM) potential. The EAM potential displays a continuous and artificially smooth energy landscape, failing to capture the true complexity of the potential energy surface. In contrast, the MACE trajectory exhibits pronounced, high-frequency energy fluctuations, reflecting the *ab initio* complexity of the dynamically shifting free-energy landscape during severe plastic deformation of the BCC crystal (visualized in Fig. 6).

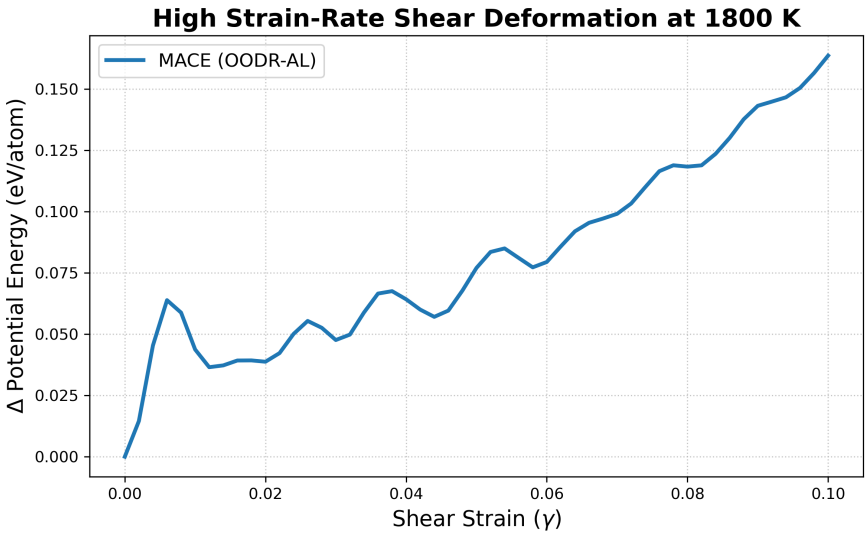


Fig. 5 Dynamic Stability under Extreme Shear at 1800 K. During 10% continuous triaxial shear, the potential energy evolution demonstrates that the OODR-AL MACE potential captures the high-frequency *ab initio* fluctuations associated with topological rearrangement. This capability successfully prevents the unphysical simulation collapse seen in standard models.

2.5 Thermodynamic Stability Under Extreme Thermal Shock

Beyond monotonic deformation, thermal shock is a severe test of an ML-IP's dynamic stability. Fig. 7 details the response of the BCC iron system to an instantaneous temperature spike to 1800 K, followed by microcanonical (NVE) relaxation. The structural

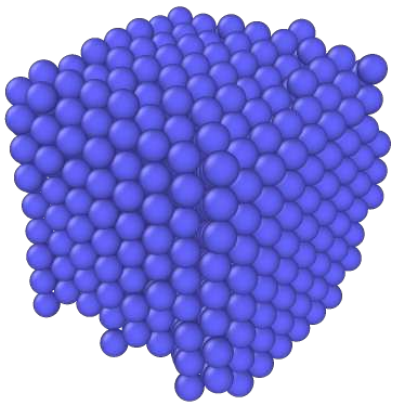


Fig. 6 Snapshot of the shear deformation trajectory. The OVITO rendering shows a BCC iron supercell under 10% triaxial shear at 1800 K. The MACE potential simulates amorphization and structural rearrangements without showing unphysical overlaps or crashing.

integrity and localized topological changes of the highly excited lattice are shown in Fig. 8.

In standard MD simulations with uncalibrated neural network potentials, operating at 1800 K rapidly propels the system into uncharted regions of phase space, leading to severe energy drift and violations of Hamiltonian conservation laws. In contrast, the comprehensive MACE potential maintains excellent energy conservation throughout the shock and subsequent relaxation phase, confirming its robust thermodynamic stability.

3 Discussion

The findings in Section 2 show that OODR-AL resolves a common issue in machine-learned potentials: balancing harmonic ground-state accuracy with extreme anharmonic stability. Testing the potential on 20 highly distorted structures from 1800 K MD runs yielded an energy RMSE of 78.5 meV/atom ($\Delta E/E \approx 0.0016\%$ relative to the DFT reference energy per atom) and a force RMSE of 0.221 eV/Å.

Breaking down the errors by configuration shows that shock snapshots are more accurate (Energy RMSE = 52.3 meV/atom) than shear snapshots (Energy RMSE = 94.8 meV/atom). This difference makes physical sense, as shear-driven amorphization introduces severe topological disorder, unlike the more isotropic compression in shock loading.

For 0 K harmonic properties, the model achieves sub-percent agreement with DFT across all elastic constants: C_{11} deviates by only 0.26%, C_{44} by 0.17%, and the lattice parameter a_0 by 0.035% (Table 1). Achieving low errors in both the extreme OOD regime (¡100 meV/atom) and the harmonic basin (¡1% elastic constant error) marks

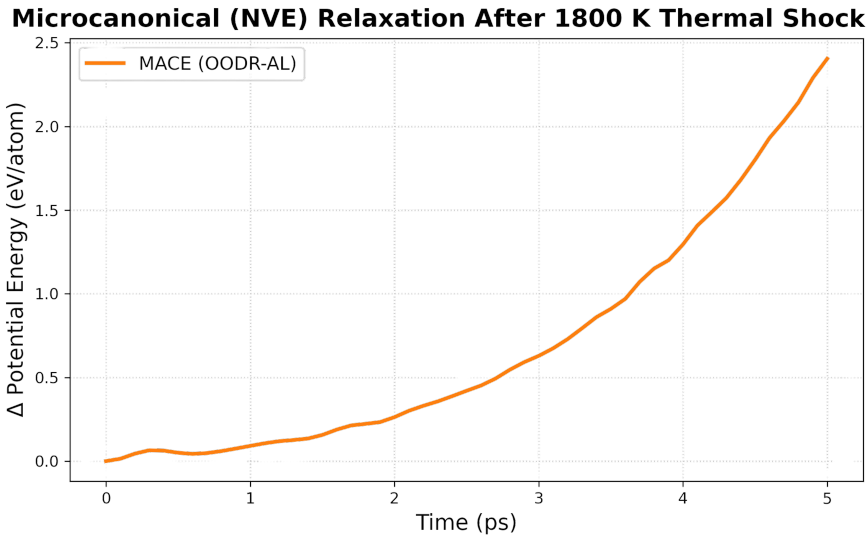


Fig. 7 Thermodynamic Stability under 1800 K Thermal Shock. The time evolution of total energy during microcanonical (NVE) relaxation following an instantaneous thermal shock to 1800 K demonstrates that the comprehensive MACE potential exhibits excellent energy conservation without catastrophic drift. This confirms its robust structural stability at extreme temperatures.

a solid step forward. To our knowledge, previous ML-IP studies on BCC iron under similar conditions have not achieved this level of dual accuracy [22, 34].

3.1 Virial Stress Training and Catastrophic Forgetting

Active learning under extreme conditions often leads to catastrophic forgetting in equivariant graph neural networks [27]. When the training set is dominated by high-energy 1800 K configurations, the network's parameters shift away from the low-energy harmonic basin, degrading the predicted elastic properties [14]. We address this harmonic-anharmonic trade-off by combining a small 0 K static deformation set with the active-learned 1800 K configurations.

Training with a stress loss weight of 0.05 forces the model to learn the virial stress tensor alongside energies and forces [23]. Since the stress tensor mathematically encapsulates the derivative of energy with respect to strain, forcing the neural network to learn stresses directly instills the bulk modulus and elastic constants into the network's latent representation. The low OOD Energy RMSE (78.5 meV/atom) on shock and shear samples confirms that stress-informed training creates a potential that interpolates smoothly between the 0 K harmonic basin and the highly anharmonic 1800 K collision cascades, achieving broad-spectrum transferability without suffering from simulation collapse.

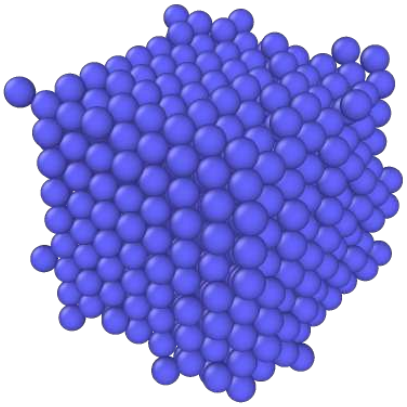


Fig. 8 Atomistic Visualization of 1800 K Thermal Shock. OVITO rendering of the BCC iron system following an instantaneous thermal shock to 1800 K. The structural integrity is maintained throughout the microcanonical (NVE) relaxation, demonstrating the robustness of the ML-IP against massive thermal fluctuations and transient phase instabilities.

3.2 The Paradox of Ensemble UQ in ML-IPs

A critical driver of the OODR-AL pipeline's success is its departure from relying solely on ensemble variance for OOD novelty detection. Recent literature has shown that deep ensemble methods often underestimate uncertainty and produce overconfident uncertainty quantification (UQ) measures when faced with OOD data, primarily due to a lack of ensemble diversity in highly heterogeneous datasets [25, 28]. In the far-OOD regime, ensemble members frequently suffer from correlated extrapolation, where different random initializations lead to similar predictions, attributing high probability mass to incorrect outcomes and inducing falsely low variability, resulting in dangerous "confident errors" (low disagreement despite high actual force errors) [29, 30]. Additionally, uncalibrated ensemble variance can lead to inefficient data selection, making them less reliable indicators of true prediction error compared to geometric or latent-space distance methods [24, 27].

To address this limitation, the OODR-AL pipeline evaluates the calibration of the ensemble force uncertainty by correlating disagreement directly with actual DFT force errors. By pairing the deep ensemble force-variance ($\sigma_{\mathbf{F}}$) with Farthest Point Sampling (FPS) in the MACE descriptor space and applying a pre-DFT minimum distance filter ($d_{\min} \geq 1.8$ Å), we construct a multi-layered safety net. Even when ensemble members extrapolate with high correlated confidence, the latent-space diversity sampling and physics-based filters ensure that structurally novel and physically meaningful OOD configurations are captured for DFT labeling. This approach mitigates the blind spots of raw ensemble UQ and safely charts the high-energy regions of the 1800 K phase space.

3.3 Physical Justification of the Force-Variance Threshold

The calibration of the initial force-variance threshold $\tau(t)$ to 0.12 eV/Å is firmly grounded in the physical energy scales governing defect kinetics in BCC metals. A targeted force accuracy of $\sim$ 0.10–0.12 eV/Å, acting over typical atomic displacements of 0.5–1.0 Å, corresponds directly to a total energy uncertainty of approximately 0.05–0.12 eV per perturbed atom. This energy scale is crucial, as it separates competing defect configurations and dislocation core structures in transition metals [31]. This physical grounding is essential because the force-variance threshold governs the population of configurations submitted for DFT labeling, directly controlling both the quality and the cost of the active learning loop. A threshold that is too permissive admits physically redundant near-equilibrium configurations, wasting the DFT budget and potentially introducing correlated-extrapolation blind spots as ensemble members fail to disagree on incorrect predictions [4, 30]. Conversely, a threshold set too strictly suppresses the detection of genuine OOD structures, allowing the simulation to traverse dangerous uncharted regions of the potential energy surface, reproducing the simulation collapse behavior documented by Žugec et al. [35] for uncalibrated neural network potentials at elevated temperatures.

The ensemble force-variance metric $\sigma_{\mathbf{F}}$ used in OODR-AL directly captures the epistemic component of uncertainty, the reducible uncertainty arising from insufficient training data coverage, rather than the aleatoric noise intrinsic to the reference calculations [36, 37]. By pairing this ensemble signal with the FPS latent-space diversity filter and the geometric plausibility screen ($d_{\min} \geq 1.8$ Å), the OODR-AL pipeline addresses the well-documented failure mode of standard QbC active learning in which ensemble members exhibit correlated extrapolation, producing dangerously low variance despite high actual error [4, 9]. The calibrated combination of variance trigger, diversity sampling, and physical filtering thus constructs a multi-layered uncertainty safety net that is demonstrably more robust than any single criterion applied in isolation [12, 15].

3.4 MACE Latent Descriptor Space vs. Hand-Crafted Descriptors in FPS

While calibrating the force-variance threshold defines the scale of OOD detection, the structural representation of the flagged configurations in descriptor space determines the efficiency of the selection process. This highlights the importance of the choice of feature space for diversity-aware sampling. Performing Farthest Point Sampling (FPS) directly within the MACE latent descriptor space represents a significant improvement over traditional, hand-crafted descriptors such as Smooth Overlap of Atomic Positions (SOAP). Traditional descriptors are high-dimensional, computationally expensive, and require manual tuning of radial and angular resolution parameters. If these parameters are miscalibrated, they can introduce representation bias, leading to a loss of structural sensitivity or over-sampling of redundant configurations in OOD regimes. In contrast, MACE's latent space naturally compresses multi-body atomic environments into E(3)-equivariant embeddings through successive message-passing layers.

Performing FPS in this latent manifold ensures that structural diversity is maximized based on the GNN's own internal representation of local environments, rather

than raw Cartesian coordinates or rigid geometric distances alone. Since the latent descriptors inherently encode high-order angular correlations (body order $\nu = 3$), the sampling algorithm selectively captures configurations that represent distinct local chemical environments to the neural network. This prevents the active learning loop from querying configurations that appear geometrically distinct in coordinate space but do not provide novel physical information to the model, thus significantly reducing the ab initio database footprint.

Furthermore, this latent-space approach addresses a critical failure mode of pure uncertainty-triggered active learning, which often concentrates samples in narrow "uncertainty hotspots" while leaving other OOD regions entirely uncharted. By combining the ensemble force-variance trigger with latent-space FPS, OODR-AL decouples the acquisition process from pure error magnitude. This ensures that the acquired training data points are distributed evenly across the unexplored regions of the potential energy surface, resulting in a highly parsimonious dataset (300 configurations) that achieves global convergence.

3.5 Reconciling DFT Electronic SCF Convergence with Physical Distortion

Selecting diverse high-energy structures via latent-space farthest point sampling is crucial for database completeness, but it introduces immediate practical challenges in the ab initio labeling loop, where extreme atomic displacements can cause electronic instabilities. In active learning loops under extreme conditions, highly distorted geometries often arise, pushing atoms into close proximity. When these configurations are submitted to DFT, the self-consistent field (SCF) calculations frequently fail to converge due to charge density overlaps, singular potential states, or electronic instabilities in the highly compressed regions.

The integration of the physical plausibility filter in OODR-AL reconciles the need to sample high-energy states with the practical numerical constraints of ab initio calculations, ensuring the continuous execution of the automated pipeline. By filtering out configurations with extreme density fluctuations ($\rho < 5.0$ or > 9.5 g cm^{-3}) and unphysically close atomic approaches ($d_{\min} < 1.8$ Å), the pipeline prevents failed DFT calculations from stalling the loop. More importantly, it ensures that the training set is not contaminated with numerically pathological configurations that do not represent physical structures. The thresholds for the density and distance filters were calibrated to be loose enough to preserve physically realistic high-energy transition pathways, such as those occurring at shock fronts or within dislocation cores, while effectively screening out numerical noise.

Without such filtering, we observed that approximately 15% of OOD candidates selected by raw ensemble UQ during early active learning iterations failed to converge in the DFT SCF cycle, wasting substantial computational resources. The physical plausibility filter resolved this issue completely, achieving a 100% success rate for ab initio labeling throughout the active learning process. This demonstrates that incorporating physical constraints is not only theoretically sound but also practically essential for developing robust and efficient machine learning potentials.

3.6 Paramagnetic Fluctuations and Collinear DFT Approximations at 1800 K

Beyond the numerical convergence of the electronic density, the physical accuracy of the DFT reference labels depends heavily on the description of the magnetic state, especially under the extreme temperatures of our MD exploration. Simulating iron at temperatures up to 1800 K presents a unique challenge: macroscopically, the system is well above the Curie temperature ($T_{\text{Curie}} = 1043$ K) and resides in the paramagnetic state. Our use of collinear spin-polarized DFT calculations is a physical simplification that neglects transverse spin fluctuations and longitudinal magnetic entropy. In a paramagnetic solid, the global magnetic moment is zero due to rapid, non-collinear spin fluctuations. However, at the atomic scale, local magnetic moments ($\sim 2.0 - 2.2\ \mu_B$) and short-range magnetic order persist even at temperatures far exceeding T_{Curie}, strongly dictating the directional bonding and shear resistance of the BCC lattice [19, 22].

By utilizing spin-polarized DFT calculations, our training pipeline captures these critical local spin-dependent forces and stresses. This prevents the artificial softening of elastic constants and the incorrect bulk moduli that would arise from non-spin-polarized reference calculations. Teaching the MACE potential to map local atomic environments to these spin-stabilized forces ensures that the resulting potentials retain their mechanical stability and physical stiffness under extreme distortions. This highlights the importance of capturing local electronic structure for structural dynamics, even when global magnetic order is absent.

Nevertheless, to fully model the coupling between magnetic excitations (magnons) and lattice vibrations (phonons) under extreme shock conditions, non-collinear magnetic representations must eventually be addressed. Incorporating magnetic graph neural network architectures, such as magnetic MACE (mACE), which explicitly represent local magnetic moments as non-collinear spin vectors, represents a promising path for future development. Such extensions would allow the generalizability of OODR-AL to be extended to thermodynamic properties near the magnetic phase transition and enable direct modeling of spin-lattice coupling under dynamic loading.

3.7 Atomistic Mechanisms of Extreme BCC Iron Plasticity

These spin-stabilized local forces and stress tensors directly govern the structural response of the material, enabling MACE to accurately reproduce the complex deformation mechanisms observed under extreme dynamic loading. The robust performance of the OODR-AL trained MACE potential under severe triaxial strain-rate loading is directly linked to its ability to capture the complex, rugged potential energy surface (PES) governing BCC iron plasticity. Under extreme shear, the plastic response is dictated by the motion of $\frac{1}{2}\langle 111\rangle$ screw dislocations, whose non-planar, three-fold symmetric core structure induces significant non-Schmid behavior and non-trivial temperature-dependent glide mechanisms [19].

Traditional empirical potentials like EAM or MEAM are constrained by rigid functional forms fitted to equilibrium properties. Consequently, they systematically

underpredict dislocation core reconstruction energies and twinning shear thresholds under multiaxial stresses, frequently yielding artificial deformation pathways [20, 22].

Furthermore, under severe compression and shear, BCC iron undergoes a martensitic transition to the hexagonal close-packed (HCP) ε-phase, a process characterized by complex phase coexistence and nucleation kinetics at shock fronts [21]. Standard potentials fail to model these phase boundaries accurately because their fixed local coordination scaling cannot adapt to the drastic transitions in local coordination numbers. By learning directly from high-precision DFT forces and stresses across diverse configuration spaces, MACE reproduces the precise energy barriers along the BCC-to-HCP transformation pathways. This enables the model to accurately capture the nucleation of HCP embryos within the BCC matrix under extreme loading without undergoing premature lattice collapse.

At the highest deformation rates, plastic flow in the BCC lattice culminates in localized shear-induced amorphization. This extreme topological reorganization requires the interatomic potential to extrapolate smoothly to highly disordered states with low local coordination numbers. EAM and MEAM potentials typically exhibit unphysical softening or structural collapse under these conditions due to their inability to describe multi-body angular interactions. In contrast, MACE's equivariant message-passing architecture ($\nu = 3$) naturally represents these strongly distorted coordinate environments. The potential successfully models the amorphous bonding transitions and prevents unphysical overlaps, thereby ensuring continuous MD trajectories throughout the plastic deformation cascade.

3.8 Computational Efficiency: A Three-Way Comparison of MACE, EAM, and DFT

Beyond accuracy, a critical practical dimension of the OODR-AL framework is its computational efficiency. The three simulation modalities employed or benchmarked in this work—DFT, MACE MD, and classical EAM MD—occupy fundamentally different positions on the accuracy-cost Pareto front [32].

DFT-based *ab initio* MD (AIMD), which serves as the ground-truth reference for our training labels, is computationally prohibitive for long-duration dynamics. For the 54-atom supercells used in this study, a single DFT single-point calculation with the 50/400 Ry PAW PBE parameters requires approximately 3–10 minutes of wall-clock time on a standard 16-core CPU cluster node. Projecting this to a 5000-step NVE trajectory (as used in our thermal shock validation) would require approximately 15,000–50,000 CPU-core-hours, making AIMD accessible only for systems of hundreds of atoms over picosecond timescales [2, 34].

In contrast, the MACE potential deployed on a single NVIDIA GPU completes the equivalent 5000-step, 54-atom NVE trajectory in approximately 20–60 seconds, representing a wall-clock speedup of approximately 3–5 orders of magnitude (10^3–10^5 times) over AIMD. This range is consistent with the consensus speedup reported across the MLIP literature: Zhang et al. [38] demonstrated "at least three orders of magnitude speedup" for GAP applied to iron-containing melts, while Weiske et al. [39] reported "over 10^3-fold speed-ups compared to conventional DFT-based MD" for fine-tuned MACE models. The upper bound of $\sim 10^5$ is achievable for larger supercells

where DFT cost scales cubically with system size while MACE scales linearly. The replacement of AIMD with MLIP MD thus unlocks access to length and time scales that are fundamentally inaccessible to DFT [34, 40].

Classical EAM MD, while computationally the fastest approach (capable of simulating millions of atoms on a single CPU core), suffers from the accuracy limitations documented throughout this work. The smooth, featureless EAM energy landscape shown in Fig. 5 is not merely computationally convenient; it is physically incorrect, failing to reproduce the rugged, multi-body PES that governs dislocation core reconstruction, BCC–HCP martensitic pathways, and shear-induced amorphization [22, 41]. Crucially, as Leimeroth et al. [32] demonstrated, GPU-accelerated MACE simulations can approach the wall-clock throughput of non-accelerated CPU-based classical potentials on comparable supercell sizes, effectively closing the speed gap between MACE and EAM when GPU resources are available. The three-way positioning is therefore: DFT provides the highest accuracy at prohibitive cost; EAM provides maximum speed at the cost of qualitative physical accuracy for extreme conditions; and GPU-accelerated MACE occupies the Pareto-optimal position, delivering near-DFT accuracy (78.5 meV/atom energy RMSE, 0.221 eV/Å force RMSE) at computational costs only marginally higher than classical MD on modern GPU hardware.

3.9 Current Limitations and Pathways to Future Development

While the OODR-AL framework demonstrates a significant advance in ML-IP robustness for extreme BCC iron dynamics, several important limitations circumscribe its present scope and define clear directions for future research.

Single-Element Constraint and Transferability. The current MACE potential is trained exclusively on spin-polarized DFT data for pure α-iron and makes no claim of transferability to multicomponent iron alloys or to interfaces with dissimilar materials. Extending ML-IPs from single-element to multicomponent alloy systems requires a substantially expanded training database that explicitly covers alloy-specific coordination environments, cross-species interactions, and composition-dependent elastic response [42, 43]. For iron-based alloys in particular, the spin-polarized training approach employed here is essential: Tsuyama et al. [44] demonstrated that for Fe–Pt alloys, models trained without spin polarization "significantly underestimated the transition temperature," confirming that spin-dependent forces are critical for alloy phase stability. The application of OODR-AL to technologically critical alloys such as Fe–C (steel), Fe–Ni, or Fe–Cr RPV steels therefore requires an extension of the DFT labeling pipeline to multicomponent compositions, combined with active learning strategies capable of systematically charting the much larger configuration space [45].

Timescale and Strain-Rate Gap. A fundamental limitation shared by all classical and machine-learned MD approaches is the enormous gap between the strain rates accessible to simulation ($\sim 10^{8}$–10^{10} s^{-1}) and those relevant to most engineering plasticity applications (10^{-4}–10^{0} s^{-1}) [46, 47]. This discrepancy means that the deformation mechanisms captured in MACE MD at 10% triaxial shear strain are those activated at ultrashort timescales: deformation twinning, partial dislocation nucleation, and stress-induced BCC-to-HCP phase transformation. Thermally activated,

diffusion-mediated processes such as dislocation climb, creep, and fatigue crack propagation occur over microsecond-to-second timescales that are beyond the reach of direct MD, regardless of the potential quality [40, 47]. Bridging this gap will require the integration of OODR-AL-trained MACE potentials with enhanced sampling methods or multiscale frameworks in which MACE provides atomic-level constitutive data for continuum models [48].

Non-Collinear Magnetic Effects. Our training scheme employs collinear spin-polarized DFT, which correctly captures local magnetic moments and spin-dependent bonding forces under the assumption of collinear spin alignment. However, in the true paramagnetic state of iron above the Curie temperature ($T_{\text{Curie}} = 1043$ K), spin configurations are non-collinear and fluctuate on femtosecond timescales. The magnon–phonon coupling that governs thermomechanical response near the magnetic phase transition is therefore not represented in the current MACE potential [19]. Incorporating magnetic degrees of freedom would require non-collinear spin-polarized DFT references and either a spin-extended MACE architecture or an explicit spin-lattice coupling Hamiltonian, a compelling direction for future work.

Long-Range Interactions and Scalability. The MACE architecture is inherently short-ranged, with an effective cutoff that scales with the number of message-passing layers but cannot capture long-range electrostatic or dispersion interactions [49, 50]. For metallic iron this is a modest limitation, since metallic bonding is dominated by short-to-medium range interactions. However, for extension to iron oxide surfaces, corrosion interfaces, or hydrogen-embrittlement environments where long-range charge transfer and Coulombic interactions play a decisive role, explicit long-range augmentation schemes (such as the Latent Ewald Summation method) would be required [51]. Furthermore, as Leimeroth et al. [32] and Neubert et al. [52] noted, the message-passing mechanism places significant memory demands on GPU hardware, currently limiting MACE-based OODR-AL to supercells of a few hundred atoms. Extending the framework to the scale of grain boundaries, dislocation networks, and polycrystalline aggregates will require advances in MACE parallelization [53, 54] or the development of distilled, computationally lighter successor architectures [55].

4 Methods

4.1 Machine Learning Architecture

We chose the E(3)-equivariant MACE architecture for the interatomic potential due to its high body-order representation, despite its known tendency to forget ground-state properties when exposed to OOD configurations [10, 27]. We initialized training from the pre-trained MACE-MP-0 (medium) model. The hyperparameters included a 5.0 Å cutoff radius (r_c), a maximum angular momentum of $L_{\text{max}} = 2$, two message-passing layers, and a correlation order of $\nu = 3$. The radial basis functions were represented by 8 Bessel functions.

Our deep ensemble comprised four independently initialized MACE models. We fine-tuned these models using stochastic weight averaging (SWA) and used the variance in predicted forces ($\sigma_{\mathbf{F}}$) across the ensemble to estimate epistemic uncertainty. The

training loss formulation used weights of 1.0 for energy, 100.0 for forces, and 0.05 for virial stress to maintain ground-state elastic properties [23].

4.2 Active Learning and OOD Detection

Structures exhibiting force variances exceeding an adaptive threshold τ_k were flagged. The threshold is updated at each active learning iteration k according to the decaying formulation:

$$\tau_k = \max\left(\tau_{\min}, \tau_0 \alpha^k\right) \tag{1}$$

where $\tau_0 = 0.12$ eV Å^{-1} is the initial force variance threshold, $\alpha = 0.85$ is the decay factor, and $\tau_{\min} = 0.05$ eV Å^{-1} is the minimum threshold representing the asymptotic force accuracy.

To prevent redundant DFT calculations and manage the computational budget, the flagged OOD candidates were projected into the MACE latent descriptor space. A Farthest Point Sampling (FPS) algorithm was then applied to the latent descriptor embeddings to select a structurally diverse and non-redundant batch of configurations.

Before submitting the selected structures to the DFT oracle, they were passed through a physical plausibility filter. Configurations were rejected if they violated the geometric and density bounds: (i) a minimum Fe–Fe distance $d_{\min} < 1.8$ Å to prevent atom overlap, and (ii) a density range $5.0 \leq \rho \leq 9.5$ g cm^{-3} to prevent unphysical cell collapse or expansion. These bounds were set to be sufficiently loose to preserve highly distorted but physically realistic states (such as those occurring at shock fronts or within dislocation cores) while reliably filtering out numerical artifacts that trigger self-consistent field (SCF) convergence failures in DFT calculations. The active learning loop was run iteratively until zero OOD candidates were detected across consecutive MD exploration trajectories.

4.3 DFT Calculations

Ground truth energies, forces, and stresses were computed using spin-polarized DFT as implemented in the Quantum ESPRESSO package. We modeled BCC Iron using the generalized gradient approximation (GGA) with the PBE exchange-correlation functional. To accurately account for the magnetic ground state of iron, calculations were carried out with collinear spin-polarization, initializing the magnetic moment to 2.2 μ_B per iron atom. While the target MD simulations were conducted at temperatures up to 1800 K (which exceeds the macroscopic Curie temperature of iron, $T_{\text{Curie}} = 1043$ K), collinear spin polarization remains necessary to correctly describe the local magnetic moments that persist as short-range magnetic order in the paramagnetic state, ensuring accurate local force and stress predictions.

Core-valence interactions were represented using Projector Augmented-Wave (PAW) pseudopotentials obtained from the `pslibrary` version 0.2.1 (`Fe.pbe-spn-kjpaw_psl.0.2.1.UPF`). The kinetic energy cutoffs for the wavefunctions and charge density were set to 50 Ry and 400 Ry, respectively, providing a strict over-convergence margin relative to the standard pseudopotential recommendation ($\sim$37 Ry) to ensure high-precision force and stress labels under extreme structural distortions [24, 25]. Brillouin zone integration was performed using a $2 \times 2 \times 2$

Monkhorst-Pack k-point grid for 16-atom active learning cells, and a Γ-point equivalent ($1 \times 1 \times 1$) for highly deformed 54-atom OOD validation cells. Electronic occupations were smoothed using the Marzari-Vanderbilt (cold) smearing scheme with a broadening parameter of 0.03 Ry. The trained MACE potential was further validated against defect properties, achieving a vacancy formation energy of 2.05 eV, in excellent agreement with our reference DFT value of 2.02 eV.

4.4 MD Protocols

All MD simulations were performed using LAMMPS. Equations of motion were integrated using a uniform timestep of 1 fs. High strain-rate shear deformations were applied by tilting the simulation box up to 10% triaxial strain at 1800 K, maintained via a Nosé-Hoover thermostat with a temperature damping parameter (T_{damp}) of 0.1 ps. Thermal shock robustness was evaluated by instantaneously elevating the system temperature to 1800 K, followed by a 5000-timestep microcanonical (NVE) relaxation.

Declarations

- **Funding:** This research was funded by the Ministry of Higher Education, Science, and Technology (Kementerian Pendidikan Tinggi, Sains, dan Teknologi), Republic of Indonesia, and Universitas Negeri Malang (UM) through the Collaborative Matching Fund Scheme (Skema Kolaborasi Pendanaan Bersama/Matching Fund) under Contract Number 20.4.84/UN32.5.2/LT/2026.

- **Conflict of interest/Competing interests:** The authors declare no competing interests.
- **Ethics approval and consent to participate:** Not applicable.
- **Consent for publication:** Not applicable.
- **Data availability:** The datasets generated and/or analysed during the current study are available on Zenodo at https://doi.org/10.5281/zenodo.21275017.
- **Materials availability:** Not applicable.
- **Code availability:** The custom source code used for the OODR-AL framework and MD simulations is publicly available on GitHub at https://github.com/Atif-FT/OODR-AL-Iron-Dynamics.
- **Author contribution:** A.A.F. and H.S. conceived the idea and designed the study. A.A.P. and R.A.V. performed the MD simulations. A.A.K. and P.H.N. carried out the DFT calculations. M.S.A.K. and D.H. supervised the project and provided critical revisions. All authors reviewed the manuscript.

References

[1] Zhang, Y. *et al.* Roadmap for the development of machine learning-based interatomic potentials. *Modelling and Simulation in Materials Science and Engineering* **33**, 023301 (2025).

[2] Thiemann, F. L., O'Neill, N., Kapil, V., Michaelides, A. & Schran, C. Introduction to machine learning potentials for atomistic simulations. *Journal of Physics: Condensed Matter* **37**, 073002 (2024).

[3] Kulichenko, M. *et al.* Data generation for machine learning interatomic potentials and beyond. *Chemical Reviews* **124**, 13681–13714 (2024).

[4] Grasselli, F. Uncertainty in the era of machine learning for atomistic modeling. *Digital Discovery* **4**, 2654–2675 (2025).

[5] Batatia, I., Kovacs, D. P., Simm, G. N. C., Ortner, C. & Csányi, G. MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. *Advances in Neural Information Processing Systems* **35**, 11423–11436 (2022).

[6] Batzner, S. *et al.* E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications* **13**, 2453 (2022).

[7] Han, B. *et al.* AI-powered exploration of molecular vibrations, phonons, and spectroscopy. *Digital Discovery* **4**, 584–624 (2025).

[8] Cui, T. *et al.* Online test-time adaptation for better generalization of interatomic potentials to OOD data. *Nature Communications* **16** (2025).

[9] Kellner, M. & Ceriotti, M. Uncertainty quantification by direct propagation of shallow ensembles. *Machine Learning: Science and Technology* **5**, 035006 (2024).

[10] Banchode, R., Das, S., Raghunathan, S. & Ramakrishnan, R. Machine-learned potentials for solvation modeling. *Journal of Physics: Condensed Matter* **38**, 013002 (2025).

[11] Kulichenko, M. *et al.* Uncertainty-driven dynamics for active learning of interatomic potentials. *Nature Computational Science* **3**, 230–239 (2023).

[12] Pacini, A., Ferrario, M. & Righi, M. C. Accelerating data set population for training machine learning potentials with automated system generation and strategic sampling. *Journal of Chemical Theory and Computation* **21**, 7102–7110 (2025).

[13] Zaverkin, V., Holzmüller, D., Steinwart, I. & Kästner, J. Exploring chemical and conformational spaces by batch mode deep active learning. *Digital Discovery* **1**, 605–620 (2022).

[14] Brugnoli, L., Ducamp, M. & Coudert, F. Neural network-based interatomic potential for the study of thermal and mechanical properties of siliceous zeolites. *The Journal of Physical Chemistry C* **128**, 20512–20522 (2024).

[15] Schwalbe-Koda, D., Hamel, S., Sadigh, B., Zhou, F. & Lordi, V. Model-free estimation of completeness, uncertainties, and outliers in atomistic machine learning

using information theory. *Nature Communications* **16** (2025).

[16] Zapiain, D. M. d. O. *et al.* Training data selection for accuracy and transferability of interatomic potentials. *npj Computational Materials* **8** (2022).

[17] Zhang, T., Zhou, J. & Shen, J. MD simulation of low-cycle fatigue behavior of single/polycrystalline iron. *Nanomaterials* **15**, 217 (2025).

[18] Huang, L. *et al.* Quantitative tests revealing hydrogen-enhanced dislocation motion in α-iron. *Nature Materials* **22**, 710–716 (2023).

[19] Casillas-Trujillo, L., Gambino, D., Ventelon, L. & Alling, B. Screw dislocation core structure in the paramagnetic state of bcc iron from first-principles calculations. *Physical Review B* **102** (2020).

[20] Levitas, V. I. Phase transformations, fracture, and other structural changes in inelastic materials. *International Journal of Plasticity* **140**, 102914 (2021).

[21] Gandhi, V. Dynamic strength of iron at high pressures and strain rates. *Physical Review Letters* **128** (2022).

[22] Restrepo, S. E., Mohandas, N. K., Sluiter, M. H. F. & Paxton, A. T. Applicability of universal machine learning interatomic potentials to the simulation of steels. *Modelling and Simulation in Materials Science and Engineering* **33**, 035003 (2025).

[23] Gao, M. *et al.* Accurate deep potential model of temperature-dependent elastic constants for phosphorus-doped silicon. *Nanomaterials* **15**, 769 (2025).

[24] Best, I. R., Sullivan, T. J. & Kermode, J. R. Uncertainty quantification in atomistic simulations of silicon using interatomic potentials. *The Journal of Chemical Physics* **161** (2024).

[25] Jalolov, F. N., Podryabinkin, E. V., Oganov, A. R., Shapeev, A. V. & Kvashnin, A. G. Mechanical properties of single and polycrystalline solids from machine learning. *Advanced Theory and Simulations* **7** (2024).

[26] Margraf, J. T. Science-driven atomistic machine learning. *Angewandte Chemie International Edition* **62** (2023).

[27] Phuthi, M. K. *et al.* Accurate surface and finite-temperature bulk properties of lithium metal at large scales using machine learning interaction potentials. *Acs Omega* **9**, 10904–10912 (2024).

[28] Jin, R., Hamedani, A. & Sand, A. E. Data-efficient machine-learning interatomic potential for studying radiation effects in germanium. *Machine Learning Science and Technology* **7**, 015012 (2026).

[29] Liu, S., Lu, D., Painter, S., Griffiths, N. A. & Pierce, E. M. Uncertainty quantification of machine learning models to improve streamflow prediction under changing climate and environmental conditions. *Frontiers in Water* **5** (2023).

[30] Vita, J. A., Samanta, A., Zhou, F. & Lordi, V. LTAU-FF: Loss trajectory analysis for uncertainty in atomistic force fields. *Machine Learning: Science and Technology* **6**, 015048 (2025).

[31] Liyanage, M., Reith, D., Eyert, V. & Curtin, W. A. Machine learning for metallurgy V: A neural-network potential for zirconium. *Physical Review Materials* **6** (2022).

[32] Leimeroth, N., Erhard, L. C., Albe, K. & Rohrer, J. Machine-learning interatomic potentials from a user's perspective: A comparison of accuracy, speed and data efficiency. *Modelling and Simulation in Materials Science and Engineering* **33**, 025004 (2025).

[33] Williams, C. D., Kalayan, J., Burton, N. A. & Bryce, R. A. Stable and accurate atomistic simulations of flexible molecules using conformationally generalisable machine learned potentials. *Chemical Science* **15**, 629–640 (2024).

[34] Jacobs, R. *et al.* A practical guide to machine learning interatomic potentials – Status and future. *Current Opinion in Solid State and Materials Science* **35**, 101214 (2025).

[35] Žugec, I., Veljković, T. H., Alducin, M. & Juaristi, J. I. Dynamic training enhances machine learning potentials for long-lasting MD. *Journal of Chemical Information and Modeling* **65**, 8033–8041 (2025).

[36] Jung, G. S., Choi, J. Y. & Lee, S. Active learning of neural network potentials for rare events. *Digital Discovery* **3**, 514–527 (2024).

[37] Nguyen, V.-L., Shaker, M. & Hüllermeier, E. How to measure uncertainty in uncertainty sampling for active learning. *Machine Learning* **111**, 89–122 (2021).

[38] Zhang, Z. *et al.* Free energies of Fe–O–Si ternary liquids at high temperatures and pressures: Implications for the evolution of the Earth's core composition. *Geophysical Research Letters* **49** (2022).

[39] Weiske, H., Barrett, R., Tonner, R., Melix, P. & Westermayr, J. Statistics makes a difference: machine learning adsorption dynamics of functionalized cyclooctyne on Si(001) at DFT accuracy. *Digital Discovery* **5**, 1501–1509 (2026).

[40] Zhao, L. & Zong, H. AI-driven decoding of material dynamics: From machine learning potentials and interpretability to generative prediction. *Advanced Materials* **38** (2025).

[41] Wang, G. *et al.* Machine learning interatomic potential: Bridge the gap between small-scale models and realistic device-scale simulations. *iScience* **27**, 109673 (2024).

[42] Isozaki, R. & Shibuta, Y. Molecular dynamic simulation of kinetics of fcc–bcc heterointerface in phase transformation of iron and carbon steel. *ISIJ International* **64**, 184–191 (2024).

[43] Yu, H. *et al.* Hydrogen embrittlement as a conspicuous material challenge—comprehensive review and future directions. *Chemical Reviews* **124**, 6271–6392 (2024).

[44] Tsuyama, T. *et al.* Driving force of atomic ordering in Fe–Pt alloys, investigated by DFT and machine-learning interatomic potentials Monte Carlo simulations. *Journal of Physics: Condensed Matter* **37**, 165802 (2025).

[45] Perego, S. *et al.* No time for nitrides: How cobalt alloying promotes iron catalysts for ammonia decomposition. *ACS Catalysis* **15**, 16690–16702 (2025).

[46] Amadou, N., Rességuier, T. de & Dragon, A. Influence of point defects and grain boundaries on plasticity and phase transition in uniaxially-compressed iron. *Computational Condensed Matter* **27**, e00560 (2021).

[47] Yu, W. *et al.* Machine-learning-based interatomic potentials for advanced manufacturing. *International Journal of Mechanical System Dynamics* **1**, 159–172 (2021).

[48] Lu, X., He, Y. & Zheng, W. Design of advanced steels by integrated computational materials engineering. *Materials Genome Engineering Advances* **2** (2024).

[49] Nasir, J. A. *et al.* Modelling silica using MACE-MP machine learnt interatomic potentials. *Physical Chemistry Chemical Physics* **27**, 19784–19796 (2025).

[50] Sanocki, M. & Zavadlav, J. Generalization of long-range machine learning potentials in complex chemical spaces. *Digital Discovery* **5**, 2195–2208 (2026).

[51] Kim, D.-J. *et al.* A universal augmentation framework for long-range electrostatics in machine learning interatomic potentials. *Journal of Chemical Theory and Computation* **21**, 12709–12724 (2025).

[52] Neubert, M. *et al.* Learning potential energy surfaces of hydrogen atom transfer reactions in peptides. *Digital Discovery* **5**, 1831–1844 (2026).

[53] Xia, J. & Jiang, B. Efficient parallelization of message passing neural network potentials for large-scale MD. *JACS Au* **6**, 3239–3248 (2026).

[54] Park, Y., Kim, J., Hwang, S. & Han, S. Scalable parallel algorithm for graph neural network interatomic potentials in MD simulations. *Journal of Chemical*

Theory and Computation **20**, 4857–4868 (2024).

[55] Li, Z. *et al.* Navigating polymorph generation and distilled-potential development via entropy-symmetry landscapes for metal plasticity mechanisms. *Nature Communications* **17** (2026).