WHEN FORCES DISAGREE: AN OUT-OF-THE-BOX STABILITY AND UNCERTAINTY ESTIMATE FOR DIRECT-FORCE MACHINE LEARNING INTERATOMIC POTENTIALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine learning interatomic potentials (MLIPs) that predict forces directly, rather than as the gradient of a predicted energy, accelerate atomistic simulations but are unstable in molecular dynamics (MD). They also lack an out-of-the-box single-model uncertainty quantification (UQ) metric, since existing single-model methods require either non-standard MLIP training or post-hoc fitting beyond energy and force matching. Here, we address both problems with the Force Delta (U_Δ), the magnitude of the difference between the direct and conservative forces of a single model, which is available out of the box for any direct-force MLIP. We first find that MD stability depends more on conservativity than on force accuracy. We then show that U_Δ tracks both the exact but expensive Jacobian-based measure of non-conservativity and the resulting instability. Further, U_Δ correlates well with the true force error across equivariant and non-equivariant architectures on in- and out-of-distribution systems, rivaling or outperforming other single-model UQ metrics and ensembles without any additional training or fitting beyond standard MLIP training. Finally, finetuning on only 200 structures that maximize U_Δ improves both the stability and the accuracy of direct forces in pretrained universal MLIPs. U_Δ therefore offers an out-of-the-box way to monitor and improve fast direct-force MLIPs, enabling their safe application to materials discovery.

1 INTRODUCTION

Machine learning interatomic potentials (MLIPs) approach near-density functional theory (DFT) accuracy at a fraction of its cost (Behler, 2016), and universal models trained on large DFT databases now enable atomistic simulation across the periodic table (Merchant et al., 2023). Most of these models obtain forces as the gradient of a predicted energy, which guarantees energy conservation but requires an extra backward pass at inference and expensive double backpropagation during training. Direct-force models bypass both the backward pass at inference and the double backpropagation during training by predicting non-conservative forces directly, accelerating training and inference by 2–3$\times$ while also lowering memory use (Liao et al., 2024; Rhodes et al., 2025; Bigi et al., 2026).

However, since the direct force is not the gradient of any energy, it does not conserve energy in molecular dynamics (MD) simulations, producing energy drift in NVE and artifacts in NVT that disrupt properties (Bigi et al., 2025). Direct forces also tend to underperform conservative forces in predicting properties that depend on higher-order derivatives of the potential energy surface (PES) (Fu et al., 2025). Smoother embeddings have narrowed this gap, but direct forces still slightly underperform and remain unstable in MD simulations (Rhodes et al., 2025). Moreover, it remains unknown whether MD simulation stability is governed more by conservativity or by force accuracy.

Existing solutions for simulation instability are computationally expensive: monitoring non-conservativity requires the Jacobian of the direct force field, which is too costly to evaluate during a simulation, while correcting it through multiple-time-stepping (MTS) requires supervising conservative forces in addition to direct forces on the full training set (Bigi et al., 2025). Training both forces sacrifices the speed advantage of direct-force training and is especially costly for universal MLIPs trained on large datasets. The required data may also not exist for a specific system of interest.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Furthermore, because direct forces lack the physical prior of energy conservation and become less reliable on structures absent from the training set, a fast uncertainty quantification (UQ) metric is needed to monitor their accuracy. Ensembles are a reliable approach but require training and evaluating several models (Lakshminarayanan et al., 2017). Single-model metrics avoid this cost but require either a non-standard MLIP training procedure or post-hoc fitting, each adding architectural and training complexity with more hyperparameter tuning. Thus, whether UQ is available at all depends on decisions made when the model is trained or on access to the training data, and always at an additional cost. To our knowledge, no single-model UQ metric is available out of the box from the standard MLIP training of energy and force matching.

Here, we address both the stability and the uncertainty problems with the Force Delta U_Δ, defined as the magnitude of the difference between the non-conservative forces $\hat{\mathbf{F}}_{\mathrm{nc}}$ and conservative forces $\hat{\mathbf{F}}_{\mathrm{c}}$, which can be computed out of the box for any direct-force model trained with standard energy and force matching. Since $\hat{\mathbf{F}}_{\mathrm{c}}$ is exactly conservative, U_Δ can conceptually be seen as the magnitude of the non-conservative component of $\hat{\mathbf{F}}_{\mathrm{nc}}$. Our contributions are as follows:

- We find that MD simulation stability depends more on conservativity than on static force accuracy, and show that U_Δ tracks both the non-conservativity of $\hat{\mathbf{F}}_{\mathrm{nc}}$ and the resulting instability at one instead of $3N$ backward that the exact Jacobian-based measure requires.
- We introduce U_Δ as **the first out-of-the-box, single-model UQ metric** to our knowledge for direct-force MLIPs, requiring no additional training or fitting beyond standard direct-force MLIP training. It correlates well with the true force error across equivariant and non-equivariant architectures on in- and out-of-distribution systems, rivaling or outperforming other single-model UQ metrics and ensembles, and extends to hybrid models in which both forces are supervised.
- We show that finetuning on only 200 U_Δ-maximized structures improves both the stability and the accuracy of direct forces in pretrained universal MLIPs, without supervising both forces on the full training set.

2 Background and Related Work

2.1 Machine Learning Interatomic Potentials

Machine learning interatomic potentials (MLIPs) approximate the potential energy surface (PES) from first-principles reference data, reaching near-DFT accuracy at a fraction of its cost by scaling linearly with system size (Behler, 2016). Early models combined hand-crafted descriptors with regression models (Behler & Parrinello, 2007; Bartók et al., 2010; Shapeev, 2016), and were followed by end-to-end graph neural networks (Gilmer et al., 2017; Schütt et al., 2018) and E(3)-equivariant message-passing architectures that build physical symmetries into the network and improve data efficiency (Batzner et al., 2022; Batatia et al., 2022). Trained on large DFT databases, universal MLIPs now predict energies and forces across the periodic table without system-specific fitting (Chen & Ong, 2022; Deng et al., 2023; Merchant et al., 2023; Batatia et al., 2025a; Rhodes et al., 2025).

2.2 Direct-Force Architectures

Conventionally, MLIPs obtain atomic forces as the negative spatial gradient of the predicted energy, $\hat{\mathbf{F}}_{\mathrm{c}} = -\nabla_{\mathbf{r}}\hat{E}$, which guarantees energy conservation but requires a backward pass at inference and double backpropagation during training. Direct-force models instead attach two heads to the shared backbone, one predicting the scalar energy $\hat{E}$ and the other directly predicting the non-conservative force vector $\hat{\mathbf{F}}_{\mathrm{nc}}$, accelerating training and inference by 2–3$\times$ while lowering memory use (Gasteiger et al., 2021; Liao et al., 2024; Neumann et al., 2024; Rhodes et al., 2025; Bigi et al., 2025; 2026). Direct-force training minimizes a joint loss on the energy and the direct force against the ground-truth DFT energy E_{DFT} and forces $\mathbf{F}_{\mathrm{DFT}}$,

$$\mathcal{L} = \lambda_E \mathcal{L}_E(\hat{E}, E_{\mathrm{DFT}}) + \lambda_F \mathcal{L}_F(\hat{\mathbf{F}}_{\mathrm{nc}}, \mathbf{F}_{\mathrm{DFT}}), \tag{1}$$

where λ_E and λ_F are the respective loss weights and $\hat{\mathbf{F}}_{\mathrm{c}}$ is constrained only by the supervision of $\hat{E}$.

However, since $\hat{\mathbf{F}}_{\mathrm{nc}}$ is not the exact spatial gradient of any scalar energy, it produces unphysical energy drift in NVE simulations which becomes artifacts in NVT simulations that thermostats cannot absorb

without disrupting structural and dynamical properties (Bigi et al., 2025). Non-conservative forces also tend to underperform conservative forces on predicting PES-curvature-dependent properties such as phonon spectra, thermal conductivity, and elastic moduli (Póta et al., 2024; Loew et al., 2025; Zhao et al., 2025; Fu et al., 2025; Bigi et al., 2026). Smoother edge embeddings substantially narrow this gap in Orb-v3, bringing direct forces to performance competitive with conservative models of comparable size (Rhodes et al., 2025), though they remain slightly behind them in general and still suffer from MD simulation instability. Existing solutions for instability are expensive: Bigi et al. (2025) correct $\hat{\mathbf{F}}_{\mathrm{nc}}$ with $\hat{\mathbf{F}}_{\mathrm{c}}$ during simulation through a multiple-time-stepping (MTS) scheme, which requires supervising both forces on the full training set, while their non-conservativity metric λ requires the Jacobian of the direct force field, assuming that non-conservativity is the primary driver of instability. Moreover, although Fu et al. (2023) showed more broadly that static force accuracy is an insufficient predictor of simulation stability, it remains unknown whether the MD simulation stability is governed more by non-conservativity or by static force accuracy.

2.3 Uncertainty Quantification for MLIPs

Since MLIPs become less reliable on structures absent from the training set, and direct forces additionally lack the physical prior of energy conservation, a fast uncertainty quantification (UQ) metric is needed to monitor their accuracy. UQ flags unreliable predictions during simulation (Schran et al., 2020; Vandermause et al., 2020; Imbalzano et al., 2021) and drives active learning loops that select structures for DFT labeling and retraining (Podryabinkin & Shapeev, 2017; Smith et al., 2018; Zhang et al., 2019; Schwalbe-Koda et al., 2021; Kulichenko et al., 2023).

Deep ensembles remain the gold standard for estimating epistemic uncertainty (Lakshminarayanan et al., 2017; Musil et al., 2019; Tan et al., 2023), but training and evaluating several models is computationally expensive, especially at foundation scale (Wen & Tadmor, 2020). Single-model alternatives avoid this cost, but they require a step beyond standard MLIP training that adds architectural and training complexity with more computational cost and hyperparameter tuning. Many methods modify the training procedure itself, including Monte Carlo dropout (Gal & Ghahramani, 2016; Wen & Tadmor, 2020), Bayesian potentials (Vandermause et al., 2020; Coscia et al., 2025), mean-variance estimation (Kahle & Zipoli, 2022), evidential regression (Amini et al., 2020; Xu et al., 2026), shallow ensembles (Kellner & Ceriotti, 2024), and learned functional perturbations (Zaghen et al., 2026), which require a specialized loss or additional architecture and can be difficult to optimize (Seitzer et al., 2022). A few of these can still be applied to a released model, but not without backbone modification and retraining (Kellner & Ceriotti, 2024; Coscia et al., 2025; Zaghen et al., 2026).

Other single-model methods instead fit a separate post-hoc model after training, including distance-based methods (Janet et al., 2019; Hirschfeld et al., 2020; Zhu et al., 2023; Willimetz & Grajciar, 2025), prediction rigidity (Bigi et al., 2024), and residual or reliability modeling (Huang et al., 2025; Mehdi et al., 2026), which require access to the training set and descriptor calculation on all training structures, scaling poorly with dataset size and become impractical for foundation-scale MLIPs trained on millions of structures. Conformal prediction similarly requires a held-out calibration set (Hu et al., 2022; Angelopoulos & Bates, 2021). Some methods, such as loss trajectory analysis (Vita et al., 2025), require both modified training and post-hoc fitting, and therefore cannot be applied to pretrained models.

Whether any of these UQ methods is available therefore depends on a decision made at training time or on a fitting step afterwards. Users of a released model are limited to whichever uncertainty estimate its authors chose to provide, and model developers must perform non-standard training or additional post-hoc fitting. Tan et al. (2023) further found that no single-model method consistently outperforms ensembles. To our knowledge, no single-model UQ metric is currently available out of the box from the energy and force matching that any direct-force model already requires.

3 Method

A direct-force model produces two forces: the direct force $\hat{\mathbf{F}}_{\mathrm{nc}}$, the vector output of the force head, and the conservative force $\hat{\mathbf{F}}_{\mathrm{c}} = -\nabla_{\mathbf{r}}\hat{E}$, the negative spatial gradient of the predicted energy. Only $\hat{\mathbf{F}}_{\mathrm{nc}}$ is supervised under the standard direct-force loss in Equation 1. We therefore define the Force Delta U_{Δ} as the mean absolute difference between these two forces over the $3N$ force components of

an N-atom structure,

$$U_\Delta(\mathbf{r}) = \frac{1}{3N}\sum_{i=1}^{N}\sum_{\alpha\in\{x,y,z\}}\left|\hat{\mathbf{F}}_{\mathrm{nc},i\alpha}(\mathbf{r}) - \hat{\mathbf{F}}_{\mathrm{c},i\alpha}(\mathbf{r})\right|. \tag{2}$$

To assess how well U_Δ reflects the true errors, we evaluate U_Δ against the errors of the two forces with respect to the ground-truth DFT forces $\mathbf{F}_{\mathrm{DFT}}$, defined using the same convention,

$$\varepsilon_{\mathrm{nc}}(\mathbf{r}) = \frac{1}{3N}\sum_{i,\alpha}\left|\hat{\mathbf{F}}_{\mathrm{nc},i\alpha}(\mathbf{r}) - \mathbf{F}_{\mathrm{DFT},i\alpha}(\mathbf{r})\right|, \qquad \varepsilon_{\mathrm{c}}(\mathbf{r}) = \frac{1}{3N}\sum_{i,\alpha}\left|\hat{\mathbf{F}}_{\mathrm{c},i\alpha}(\mathbf{r}) - \mathbf{F}_{\mathrm{DFT},i\alpha}(\mathbf{r})\right|. \tag{3}$$

Because $\hat{\mathbf{F}}_{\mathrm{c}}$ is exactly conservative, the residual $\hat{\mathbf{F}}_{\mathrm{nc}} - \hat{\mathbf{F}}_{\mathrm{c}}$ can conceptually be seen as the non-conservative component of $\hat{\mathbf{F}}_{\mathrm{nc}}$, and U_Δ measures the magnitude of this component. Moreover, evaluating $\hat{\mathbf{F}}_{\mathrm{c}}$ requires only a single backward pass through the energy head at inference, but not the double backpropagation needed to supervise it during training. Direct-force training therefore remains around three times cheaper than conservative training (Bigi et al., 2025), and U_Δ is available for any model trained with the standard energy and direct force matching loss.

4 Results

We evaluate U_Δ in three areas. First, section 4.1 establishes that MD simulation stability depends more on conservativity than on force accuracy, and section 4.2 then shows that U_Δ tracks the exact but expensive Jacobian-based measure of non-conservativity and the resulting instability well. Second, sections 4.3 to 4.5 evaluate U_Δ as a UQ metric, on pretrained universal models and on models trained from scratch to enable comparison with other UQ methods that are unavailable for pretrained models. Third, section 4.6 uses U_Δ to generate data for finetuning, improving both the stability and the accuracy of direct forces in pretrained universal models.

4.1 Simulation Instability from Non-Conservativity

To determine whether MD simulation stability is governed more by conservativity or by static accuracy against the ground truth, we compare NVE simulations driven by both the conservative $\hat{\mathbf{F}}_{\mathrm{c}}$ and non-conservative $\hat{\mathbf{F}}_{\mathrm{nc}}$ forces of pretrained direct-force models, tracking the total energy drift $\Delta E_{\mathrm{drift}}$ along each trajectory. We evaluate static accuracy using the mean absolute error of both forces against DFT references, $\varepsilon_{\mathrm{nc}}$ for $\hat{\mathbf{F}}_{\mathrm{nc}}$ and ε_{c} for $\hat{\mathbf{F}}_{\mathrm{c}}$. Full simulation details are in Appendix A.3.

As shown in Figure 1, the trajectory driven by $\hat{\mathbf{F}}_{\mathrm{nc}}$ drifts by 0.8 eV/atom over 50 ps, while the one driven by $\hat{\mathbf{F}}_{\mathrm{c}}$ shows no measurable drift, even though $\hat{\mathbf{F}}_{\mathrm{nc}}$ is over three times more accurate against DFT+D3 than $\hat{\mathbf{F}}_{\mathrm{c}}$, which is never directly supervised on DFT+D3 forces, matching the level of theory Orb was pretrained on (Neumann et al., 2024). However close $\hat{\mathbf{F}}_{\mathrm{nc}}$ is to the DFT+D3 forces, which are conservative by construction, it is not the exact gradient of any scalar energy, and thus the MD integrator cannot produce a constant-energy trajectory (Hairer et al., 2006).

We also find larger temperature fluctuations in NVT simulations driven by $\hat{\mathbf{F}}_{\mathrm{nc}}$ than by $\hat{\mathbf{F}}_{\mathrm{c}}$ under the same thermostat (Appendix B.1). This agrees with Bigi et al. (2025), who find that the thermostat needs to perform continuous work to counteract the artificial energy the non-conservative forces inject, and that substantial thermostat interference needed to absorb this injected energy can significantly disrupt structural and dynamical properties.

These results confirm that a conservative force gives stable dynamics even when substantially less accurate than a non-conservative one, and MD simulation stability therefore depends more on conservativity than on force accuracy. This justifies monitoring the non-conservativity of $\hat{\mathbf{F}}_{\mathrm{nc}}$ during a simulation. However, the exact measure of non-conservativity, the antisymmetric part of the Jacobian of $\hat{\mathbf{F}}_{\mathrm{nc}}$ (Bigi et al., 2025), requires $3N$ backward passes per evaluation and is impractical to monitor during a simulation. We therefore propose U_Δ as a cheaper alternative and evaluate it against both the exact measure and the resulting simulation instability in Section 4.2.

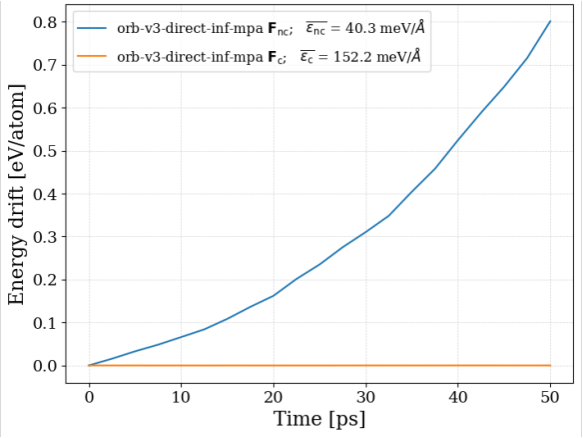


Figure 1: MD simulation stability depends more on conservativity than on force accuracy. Total energy drift over 50 ps of NVE for liquid water driven by both forces of the pretrained `orb-v3-direct-inf-mpa`. $\hat{\mathbf{F}}_{\mathrm{nc}}$ is over three times more accurate against DFT+D3 than $\hat{\mathbf{F}}_{\mathrm{c}}$, but still produces energy drift by 0.8 eV/atom, while the less accurate $\hat{\mathbf{F}}_{\mathrm{c}}$ shows no measurable drift. The same results also hold for other systems (Appendix B.1).

4.2 Tracking Non-Conservativity and Simulation Instability via U_Δ

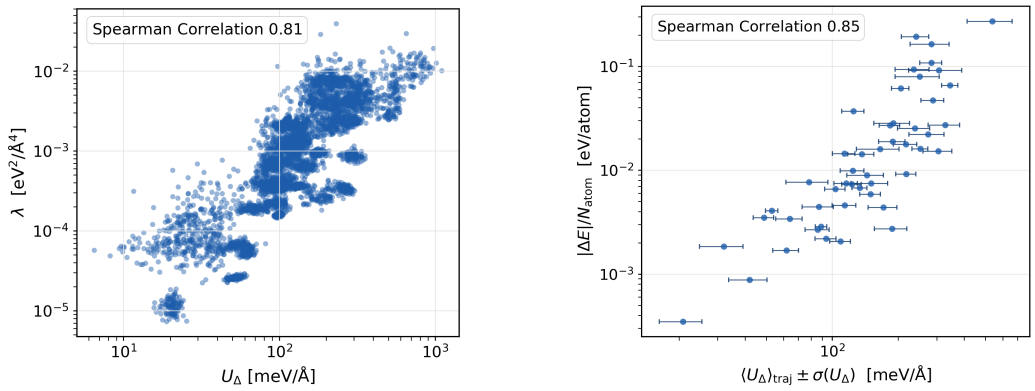
We validate U_Δ against λ, an exact measure of the non-conservativity of the direct force field, and against the resulting simulation instability. The decomposition of $\hat{\mathbf{F}}_{\mathrm{nc}}$ in Section 3 conceptually mirrors that of the force field Jacobian $\mathbf{J}_{i\alpha,j\beta} = \partial \hat{\mathbf{F}}^{i\alpha}_{\mathrm{nc}} / \partial r_{j\beta}$ into a symmetric conservative part $(\mathbf{J} + \mathbf{J}^\top)/2$ and an antisymmetric part $\mathbf{A} = (\mathbf{J} - \mathbf{J}^\top)/2$, which measures non-conservativity exactly (Bigi et al., 2025). We compute $\mathbf{A}$ by automatic differentiation and define λ as the mean of its squared entries, matching the per-component mean convention of U_Δ.

We perform NVE simulations using $\hat{\mathbf{F}}_{\mathrm{nc}}$ of pretrained direct-force models on 50 MPtrj systems (Deng et al., 2023), and track U_Δ and λ along the trajectories. The simulation instability, $\Delta E_{\mathrm{drift}}$, is calculated as the total absolute energy change normalized by the number of atoms. More details are in Appendix A.3. As shown in Figure 2a, U_Δ correlates well with λ, demonstrating that it successfully captures the non-conservative component of $\hat{\mathbf{F}}_{\mathrm{nc}}$ without expensive Jacobian evaluation. Moreover, Figure 2b shows a strong correlation between U_Δ and $\Delta E_{\mathrm{drift}}$, confirming that the simulation instability primarily arises from the non-conservativity of the direct force. U_Δ captures and monitors both non-conservativity and the resulting instability at a fraction of the cost of the Jacobian-based λ.

4.3 U_Δ as an Out-of-the-Box UQ Metric for Pretrained Universal Models

We next evaluate U_Δ as a UQ metric for the error of $\hat{\mathbf{F}}_{\mathrm{nc}}$ of pretrained universal MLIPs. For users of pretrained models, the only uncertainty estimate available with standard MLIP training of energy and force matching is the variance U_{var} of a heterogeneous ensemble of publicly released models (Liu et al., 2026), which requires evaluating every member at inference. Single-model alternatives instead depend on choices made at training time and do not transfer between models. Orb-v3 (Rhodes et al., 2025) provides a confidence head that predicts a binned per-atom uncertainty score, following the lDDT score (Mariani et al., 2013) used as the pLDDT confidence estimate in AlphaFold (Jumper et al., 2021), but was not trained with dropout, which would otherwise enable MC-dropout uncertainty (Gal & Ghahramani, 2016; Wen & Tadmor, 2020). EquiformerV2 (Liao et al., 2024) was trained with dropout but provides no confidence head, and MC dropout requires multiple stochastic forward passes at inference, avoiding the training cost of an ensemble but not its inference cost. U_Δ does not require training-time choice and is available out of the box for both model families.

We evaluate twelve pretrained universal MLIPs from the Orb and EquiformerV2 families on diverse in- and out-of-distribution systems spanning bulk solids, surfaces, and molecules, using the Spearman rank correlation ρ between each metric and the true direct force error $\varepsilon_{\mathrm{nc}}$. More details are in



(a) (b)

Figure 2: U_Δ tracks non-conservativity and simulation instability at a fraction of the cost of the exact Jacobian-based measure λ, shown for `orb-v3-direct-inf-mpa`. (a) U_Δ tracks λ across 5000 frames of 50 different systems while requiring a single backward pass per evaluation, against the $3N$ passes required for λ. (b) U_Δ averaged over each trajectory tracks the total energy drift ΔE_{drift} , with error bars giving the standard deviation of U_Δ along each trajectory. The same results also hold for other models (Appendix B.2).

Table 1: U_Δ delivers ensemble-level error correlation as the only single-model metric available out of the box for both model families, since the confidence head and MC dropout each require a training-time choice the other model family did not make. Spearman ρ between each uncertainty metric and the true direct force error ε_{nc}, averaged over the models in each family. U_Δ rivals the five-model Orb heterogeneous ensemble and outperforms the seven-model EquiformerV2 heterogeneous ensemble.

Method	Orb (5 models)	EquiformerV2 (7 models)
U_Δ	0.70 ± 0.04	$\mathbf{0.91 \pm 0.02}$
Heterogeneous Ensemble U_{var}	$\mathbf{0.73}$	0.79
Confidence Head	0.61 ± 0.05	—
MC Dropout	—	0.74 ± 0.08

Appendix A.5. As shown in Table 1, U_Δ is competitive with U_{var} on the Orb family and outperforms U_{var} on EquiformerV2, while requiring only a single model. It also outperforms the model-specific UQ methods wherever those are available.

4.4 U_Δ AS A UQ METRIC ON ADVERSARIAL OUT-OF-DISTRIBUTION (OOD) STRUCTURES

We evaluate U_Δ on OOD structures generated by the adversarial attack of Schwalbe-Koda et al. (2021), which perturbs the atomic positions $\mathbf{r}$ to maximize U_Δ while keeping the predicted energy low. More details of the attack are in Appendix A.4. Figure 3 shows that U_Δ retains a strong positive correlation with ε_{nc} across different systems, so it continues to track the error of the direct force for structures far out of distribution. U_Δ also correlates almost perfectly with the conservative force error ε_{c} in this regime (Appendix B.3), which follows from the asymmetric supervision in Equation 1, where $\hat{\mathbf{F}}_{\text{nc}}$ is supervised directly on DFT forces while $\hat{\mathbf{F}}_{\text{c}}$ is constrained only via the scalar energy.

4.5 BENCHMARKING U_Δ AGAINST TRAINED UQ BASELINES

We train MLIP models from scratch to benchmark U_Δ against deep ensembles (Lakshminarayanan et al., 2017), mean-variance estimation (MVE) (Kahle & Zipoli, 2022), and the distance-based Gaussian mixture model (GMM) (Zhu et al., 2023) on the ammonia dataset of Tan et al. (2023), which is designed to evaluate how well an uncertainty metric generalizes to the extrapolative OOD

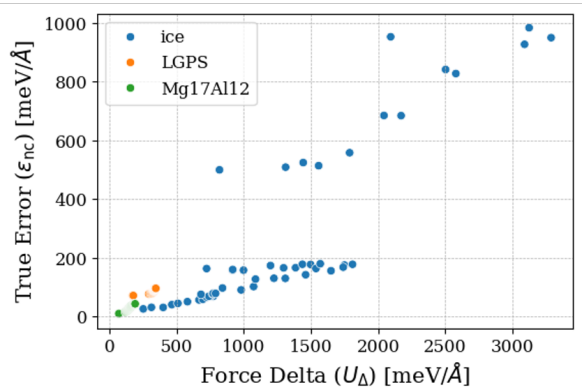


Figure 3: U_Δ correlates well with the direct force error far out of distribution. Correlation between U_Δ and the true direct force error ε_{nc} for adversarial OOD structures of three systems, shown for `orb-v3-direct-20-mpa`. The same results hold for other models and systems (Appendix B.3).

regime. We train two architectures using standard direct-force MLIP training of energy and $\hat{\mathbf{F}}_{\text{nc}}$ matching in Equation 1: the non-equivariant Orb-v3 (Rhodes et al., 2025) and the equivariant MACE (Batatia et al., 2022) with an added direct force head. MVE instead has to be trained with its own heteroscedastic loss. We evaluate the Spearman correlation ρ with ε_{nc} and the area under the receiver operating characteristic curve (AUROC) for OOD detection, averaged over five random initialization seeds. More details on model training, datasets, and metrics are in Appendix A.6.

Table 2 shows that U_Δ is competitive with all three UQs while coming out of the box with standard direct-force MLIP training, without the non-standard training procedure required by MVE or the post-hoc fitting required by GMM. It gives the highest ρ on Orb and rivals the deep ensemble on MACE. GMM's strong performance on MACE could be attributed to the more expressive equivariant embedding, which separates in- and out-of-distribution structures better than Orb's. However, GMM requires fitting the MLIP embeddings of the training set with the expectation-maximization algorithm using full-rank covariance matrices (Zhu et al., 2023; Tan et al., 2023). This scales poorly with data set size and embedding dimension, becoming impractical for universal MLIPs trained on millions of structures, a cost that U_Δ avoids entirely with no post-hoc fitting.

MVE performs worst on both architectures and varies far more across seeds than any other method. Its heteroscedastic loss is difficult to optimize (Seitzer et al., 2022), which the original benchmark of Tan et al. (2023) noted but could not expose from a single training run. Multi-seed training shows that this optimization difficulty cannot be avoided by seed choice, making MVE the least reliable of all UQ methods we evaluate. This also illustrates a cost of UQ methods that require their own loss or architecture: the added training complexity introduces failure modes that a single run does not reveal. U_Δ, by contrast, requires no training beyond standard direct-force MLIP training.

We also find that U_Δ correlates well with the error of both forces in hybrid models where $\hat{\mathbf{F}}_{\text{c}}$ is also supervised on the DFT forces, since it then acts as a two-member ensemble over the two forces (Appendix B.4). Such models are useful for alternating between the fast $\hat{\mathbf{F}}_{\text{nc}}$ and the physically correct $\hat{\mathbf{F}}_{\text{c}}$ where an exact potential energy surface is required, as in the MTS scheme of Bigi et al. (2025) or in predicting properties that depend on higher-order derivatives of the PES (Bigi et al., 2026). Nevertheless, supervising $\hat{\mathbf{F}}_{\text{c}}$ reintroduces the double backpropagation that direct-force training avoids and is costly at foundation scale.

4.6 Data-Efficient Finetuning with U_Δ

Since U_Δ tracks both non-conservativity and force error, maximizing it should generate structures that are useful for improving both simulation stability and accuracy. This offers a cheaper alternative to Bigi et al. (2025), who stabilize direct forces by supervising $\hat{\mathbf{F}}_{\text{c}}$ on the full training set, which is costly for universal MLIPs and requires data that may not exist for a specific system of interest. We generate 100 adversarial structures per generation and label them with DFT+D3. We evaluate four

Table 2: U_Δ is competitive out of the box with other UQ metrics, requiring no additional training or fitting beyond the standard MLIP loss. Spearman ρ against $\varepsilon_{\mathrm{nc}}$ and AUROC for OOD detection on the ammonia benchmark, averaged over five seeds with standard deviations across seeds. Deep ensembles use five independently trained models, and the other baselines require either a non-standard training procedure or post-hoc fitting on the training set. MVE is the least reliable method on both architectures, reflecting the difficulty of optimizing its heteroscedastic loss.

	Orb		MACE	
Method	ρ	AUROC	ρ	AUROC
U_Δ	$\mathbf{0.832 \pm 0.026}$	0.969 ± 0.007	0.861 ± 0.055	0.966 ± 0.003
GMM	0.819 ± 0.070	0.962 ± 0.023	$\mathbf{0.898 \pm 0.030}$	$\mathbf{0.982 \pm 0.003}$
MVE	0.692 ± 0.120	0.897 ± 0.076	0.675 ± 0.241	0.791 ± 0.129
Deep ensemble	0.825	**0.982**	0.873	0.976

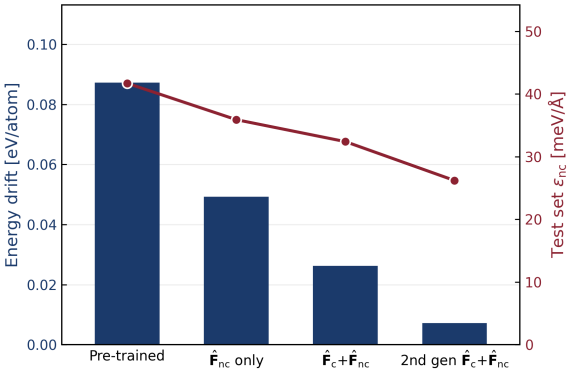


Figure 4: Finetuning on only 200 U_Δ-maximized structures improves both the stability and the accuracy of the direct force in a pretrained model, without supervising both forces on the full training set as in Bigi et al. (2025). NVE energy drift and direct force error on a held-out liquid water test set after each finetuning stage, for `orb-v3-direct-inf-mpa` on ice. The same results also hold for other models and systems (Appendix B.5).

sequential settings: the pretrained model, finetuning $\hat{\mathbf{F}}_{\mathrm{nc}}$ only, finetuning $\hat{\mathbf{F}}_{\mathrm{nc}}$ and $\hat{\mathbf{F}}_{\mathrm{c}}$ together, and a second generation of attack and finetuning on the combined 200 structures. Note that $\hat{\mathbf{F}}_{\mathrm{nc}}$, which drives the simulations, is always supervised to prevent its degradation (Bigi et al., 2025; 2026). More details are in Appendix B.5.

Figure 4 shows that both stability and accuracy of $\hat{\mathbf{F}}_{\mathrm{nc}}$ improve at every stage. Finetuning $\hat{\mathbf{F}}_{\mathrm{nc}}$ alone reduces the drift, and finetuning both forces reduces it further even though the simulation is still driven by $\hat{\mathbf{F}}_{\mathrm{nc}}$ alone, which is consistent with the hybrid model results in Appendix B.4, where supervising $\hat{\mathbf{F}}_{\mathrm{c}}$ also improves $\hat{\mathbf{F}}_{\mathrm{nc}}$. A second generation nearly removes the remaining drift. The direct force error on a held-out liquid water system also falls, indicating that finetuning on U_Δ-generated data improves accuracy on systems not used for finetuning.

5 Discussion

U_Δ requires no additional training or fitting beyond the standard energy and force matching already needed to train any direct-force MLIP, which makes it, to our knowledge, **the first out-of-the-box, single-model UQ metric** for these models. The same quantity serves two roles: it tracks the non-conservativity of direct forces and the resulting simulation instability, and it correlates well with the force error across multiple equivariant and non-equivariant architectures on in- and out-of-distribution systems. U_Δ also works under both standard direct-force training and hybrid training where $\hat{\mathbf{F}}_{\mathrm{c}}$ is also supervised, correlating well with the error of both forces in the latter case (Appendix B.4).

For model users, this means that U_Δ is available out of the box on any direct-force model, whether or not its authors chose to train a UQ metric. As Section 4.3 shows, other single-model metrics are available only on the models whose training happened to enable them. For model developers, U_Δ removes the need to add a UQ head, adopt a non-standard training procedure, or run post-hoc fitting or training at additional cost. It can monitor stability and accuracy of an MD trajectory on the fly, and select structures for active learning. It is also useful for hybrid models with both forces being supervised, such as within the multiple-time-stepping scheme of Bigi et al. (2025), which relies on both forces being accurate: since U_Δ correlates well with the error of both forces in these models, a large value signals that both are likely to be unreliable.

A possible explanation for these results is that both force heads make predictions from a shared representation. When a structure moves out of distribution, the representation becomes less reliable and both heads deviate from the ground truth, but in different ways, since one is a direct output and the other a gradient of the predicted energy. Under standard direct-force training this deviation is also asymmetric, with ε_c growing faster than ε_nc and dominating the magnitude of U_Δ, because the loss in Equation 1 constrains the two forces differently. U_Δ nonetheless correlates well with ε_nc, since $\hat{\mathbf{F}}_\mathrm{nc}$ also deviates alongside $\hat{\mathbf{F}}_\mathrm{c}$. The same explanation also covers hybrid models, where both forces are supervised and U_Δ correlates well with the error of both, since the two heads are computed differently from the shared representation and diverge in different ways as the representation becomes less reliable (Appendix B.4). The same relationship holds in reverse during finetuning: supervision on structures where the heads disagree should improve the shared representation, which in turn improves both heads. This agrees with Bigi et al. (2025; 2026) and Fu et al. (2025), who find that direct-force training improves the conservative force, whereas our finetuning and hybrid model results show the reverse direction, where supervising $\hat{\mathbf{F}}_\mathrm{c}$ improves $\hat{\mathbf{F}}_\mathrm{nc}$. More broadly, the multi-head replay finetuning of Batatia et al. (2025b) shows that heads trained on different levels of electronic structure theory transfer knowledge through a shared backbone, indicating that such a backbone learns representations common to all of its heads.

Lastly, we discuss the limitations of U_Δ and opportunities for future work. **A direct force head is required.** U_Δ does not apply to purely conservative potentials. A shallow direct force head could be added to a frozen conservative backbone and trained, giving a model close to the hybrid setting, but this requires post-hoc training and loses the out-of-the-box advantage. $\hat{\mathbf{F}}_\mathrm{c}$ **must be evaluated at inference.** Computing U_Δ requires one backward pass through the energy head, which slows inference relative to a direct forward pass alone, although this can be reduced by evaluating U_Δ every M steps of a simulation. The training advantage of direct-force models is unaffected since no double backpropagation is needed. U_Δ **correlates well with error but is not calibrated.** Its magnitude is not the magnitude of either force error, and quantitative bounds would require post-hoc calibration or conformal methods (Kuleshov et al., 2018; Angelopoulos & Bates, 2021). The true error ε_nc cannot serve this purpose, since computing it requires DFT labels for the configuration being simulated, whereas U_Δ is obtained from the model alone. U_Δ **can also be small when both forces agree but are both wrong.** Because U_Δ measures disagreement, it stays small when both forces deviate from the ground truth in the same direction, as we observe for one model and system in Appendix B.3.

AI Use Statement

For the paper, an LLM was employed solely as a grammar-checking and writing refinement tool. Its use was limited to improving the clarity and coherence of the written language.

References

Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. In *Advances in Neural Information Processing Systems*, volume 33, pp. 14927–14937, 2020.

Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Albert P. Bartók, Mike C. Payne, Risi Kondor, and Gábor Csányi. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Physical Review Letters*, 104(13): 136403, 2010. doi: 10.1103/PhysRevLett.104.136403.

Ilyes Batatia, Dávid Péter Kovács, Gregor N. C. Simm, Christoph Ortner, and Gábor Csányi. MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. In *Advances in Neural Information Processing Systems*, volume 35, pp. 11423–11436. Curran Associates, Inc., 2022.

Ilyes Batatia, Philipp Benner, Yuan Chiang, Alin M Elena, Dávid P Kovács, Janosh Riebesell, Xavier R Advincula, Mark Asta, Matthew Avaylon, William J Baldwin, et al. A foundation model for atomistic materials chemistry. *The Journal of chemical physics*, 163(18), 2025a.

Ilyes Batatia, Chen Lin, Joseph Hart, Elliott Kasoar, Alin M. Elena, Sam Walton Norwood, Thomas Wolf, and Gábor Csányi. Cross learning between electronic structure theories for unifying molecular, surface, and inorganic crystal foundation force fields. *arXiv preprint arXiv:2510.25380*, 2025b.

Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P. Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E. Smidt, and Boris Kozinsky. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications*, 13(1):2453, 2022. doi: 10.1038/s41467-022-29939-5.

Jörg Behler. Perspective: Machine learning potentials for atomistic simulations. *The Journal of chemical physics*, 145(17), 2016.

Jörg Behler and Michele Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical Review Letters*, 98(14):146401, 2007.

Filippo Bigi, Sanggyu Chong, Michele Ceriotti, and Federico Grasselli. A prediction rigidity formalism for low-cost uncertainties in trained neural networks. *Machine Learning: Science and Technology*, 5(4):045018, 2024. doi: 10.1088/2632-2153/ad805f.

Filippo Bigi, Marcel F. Langer, and Michele Ceriotti. The dark side of the forces: assessing non-conservative force models for atomistic machine learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 4384–4414, 2025.

Filippo Bigi, Paolo Pegolo, Arslan Mazitov, Jonathan Schmidt, and Michele Ceriotti. Pushing the limits of unconstrained machine-learned interatomic potentials. *Machine Learning: Science and Technology*, 7(3):035051, 2026. doi: 10.1088/2632-2153/ae6417.

Erik Bitzek, Pekka Koskinen, Franz Gähler, Michael Moseler, and Peter Gumbsch. Structural relaxation made simple. *Physical Review Letters*, 97(17):170201, 2006. doi: 10.1103/PhysRevLett.97.170201.

Chi Chen and Shyue Ping Ong. A universal graph deep learning interatomic potential for the periodic table. *Nature Computational Science*, 2(11):718–728, 2022. doi: 10.1038/s43588-022-00349-3.

Stefan Chmiela, Valentin Vassilev-Galindo, Oliver T Unke, Adil Kabylda, Huziel E Sauceda, Alexandre Tkatchenko, and Klaus-Robert Müller. Accurate global machine learning force fields for molecules with hundreds of atoms. *Science Advances*, 9(2):eadf0873, 2023.

Dario Coscia, Pim de Haan, and Max Welling. BLIPs: Bayesian learned interatomic potentials. *arXiv preprint arXiv:2508.14022*, 2025.

Bowen Deng, Peichen Zhong, KyuJung Jun, Janosh Riebesell, Kevin Han, Christopher J. Bartel, and Gerbrand Ceder. CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nature Machine Intelligence*, 5(9):1031–1041, 2023. doi: 10.1038/s42256-023-00716-3.

Xiang Fu, Zhenghao Wu, Wujie Wang, Tian Xie, Sinan Keten, Rafael Gomez-Bombarelli, and Tommi Jaakkola. Forces are not enough: Benchmark and critical evaluation for machine learning force fields with molecular simulations. *Transactions on Machine Learning Research*, 2023.

Xiang Fu, Brandon M. Wood, Luis Barroso-Luque, Daniel S. Levine, Meng Gao, Misko Dzamba, and C. Lawrence Zitnick. Learning smooth and expressive interatomic potentials for physical property prediction. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 17875–17893. PMLR, 2025.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pp. 1050–1059. PMLR, 2016.

Johannes Gasteiger, Florian Becker, and Kristof T Schütt. Gemnet: Universal directional graph neural networks for molecules. *Advances in Neural Information Processing Systems*, 34:6790–6800, 2021.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1263–1272. PMLR, 2017.

Stefan Grimme, Jens Antony, Stephan Ehrlich, and Helge Krieg. A consistent and accurate ab initio parametrization of density functional dispersion correction (dft-d) for the 94 elements h-pu. *The Journal of chemical physics*, 132(15):154104, 2010.

Ernst Hairer, Christian Lubich, and Gerhard Wanner. *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*, volume 31 of *Springer Series in Computational Mathematics*. Springer, 2 edition, 2006.

Lior Hirschfeld, Kyle Swanson, Kevin Yang, Regina Barzilay, and Connor W. Coley. Uncertainty quantification using neural networks for molecular property prediction. *Journal of Chemical Information and Modeling*, 60(8):3770–3780, 2020. doi: 10.1021/acs.jcim.0c00502.

Yuge Hu, Joseph Musielewicz, Zachary W Ulissi, and Andrew J Medford. Robust and scalable uncertainty estimation with conformal prediction for machine-learned interatomic potentials. *Machine Learning: Science and Technology*, 3(4):045028, 2022.

Shih-Peng Huang, Nontawat Charoenphakdee, Yuta Tsuboi, Yong-Bin Zhuang, and Wenwen Li. P-drum: Post-hoc descriptor-based residual uncertainty modeling for machine learning potentials. *arXiv preprint arXiv:2509.02927*, 2025.

Giulio Imbalzano, Yong-Bin Zhuang, Venkat Kapil, Kevin Rossi, Edgar A. Engel, Federico Grasselli, and Michele Ceriotti. Uncertainty estimation for molecular dynamics and sampling. *The Journal of Chemical Physics*, 154(7):074102, 2021. doi: 10.1063/5.0036522.

Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A. Persson. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, 2013. doi: 10.1063/1.4812323.

Jon Paul Janet, Chenru Duan, Tzuhsiung Yang, Aditya Nandy, and Heather J. Kulik. A quantitative uncertainty metric controls error in neural network-driven chemical discovery. *Chemical Science*, 10(34):7913–7922, 2019. doi: 10.1039/C9SC02298H.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

Leonid Kahle and Federico Zipoli. Quality of uncertainty estimates from neural network potential ensembles. *Physical Review E*, 105(1):015311, 2022.

Matthias Kellner and Michele Ceriotti. Uncertainty quantification by direct propagation of shallow ensembles. *Machine Learning: Science and Technology*, 5(3):035006, 2024. doi: 10.1088/2632-2153/ad594a.

Georg Kresse and Jürgen Furthmüller. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Physical Review B*, 54(16):11169, 1996.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 2796–2804, 2018.

Maksim Kulichenko, Kipton Barros, Nicholas Lubbers, Ying Wai Li, Richard Messerly, Sergei Tretiak, Justin S Smith, and Benjamin Nebgen. Uncertainty-driven dynamics for active learning of interatomic potentials. *Nature computational science*, 3(3):230–239, 2023.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Ask Hjorth Larsen, Jens Jørgen Mortensen, Jakob Blomqvist, Ivano E Castelli, Rune Christensen, Marcin Dułak, Jesper Friis, Michael N Groves, Bjørk Hammer, Cory Hargus, et al. The atomic simulation environment—a python library for working with atoms. *Journal of Physics: Condensed Matter*, 29(27):273002, 2017.

Yi-Lun Liao, Brandon M. Wood, Abhishek Das, and Tess Smidt. EquiformerV2: Improved equivariant transformer for scaling to higher-degree representations. In *The Twelfth International Conference on Learning Representations*, 2024.

Kai Liu, Zixiong Wei, Wei Gao, Poulumi Dey, Marcel H. F. Sluiter, and Fei Shuang. Heterogeneous ensemble enables a universal uncertainty metric for atomistic foundation models. *npj Computational Materials*, 12(1):34, 2026. doi: 10.1038/s41524-025-01905-x.

Antoine Loew, Dewen Sun, Hai-Chen Wang, Silvana Botti, and Miguel AL Marques. Universal machine learning interatomic potentials are ready for phonons. *npj Computational Materials*, 11 (1):178, 2025.

Valerio Mariani, Marco Biasini, Alessandro Barbato, and Torsten Schwede. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, 29(21):2722–2728, 2013.

Shams Mehdi, Ilkwon Cho, and Olexandr Isayev. Knowing when to trust machine-learned interatomic potentials. *arXiv preprint arXiv:2605.00640*, 2026.

Amil Merchant, Simon Batzner, Samuel S. Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85, 2023. doi: 10.1038/s41586-023-06735-9.

Jason M Munro, Katherine Latimer, Matthew K Horton, Shyam Dwaraknath, and Kristin A Persson. An improved symmetry-based approach to reciprocal space path selection in band structure calculations. *npj Computational Materials*, 6(1):112, 2020.

Felix Musil, Michael J. Willatt, Mikhail A. Langovoy, and Michele Ceriotti. Fast and accurate uncertainty estimation in chemical machine learning. *Journal of Chemical Theory and Computation*, 15(2):906–915, 2019. doi: 10.1021/acs.jctc.8b00959.

Mark Neumann, James Gin, Benjamin Rhodes, Steven Bennett, Zhiyi Li, Hitarth Choubisa, Arthur Hussey, and Jonathan Godwin. Orb: A fast, scalable neural network potential. *arXiv preprint arXiv:2410.22570*, 2024.

John P Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple. *Physical Review Letters*, 77(18):3865, 1996.

Evgeny V Podryabinkin and Alexander V Shapeev. Active learning of linearly parameterized interatomic potentials. *Computational Materials Science*, 140:171–180, 2017.

Balázs Póta, Paramvir Ahlawat, Gábor Csányi, and Michele Simoncelli. Thermal conductivity predictions with foundation atomistic models. *arXiv preprint arXiv:2408.00755*, 2024.

Benjamin Rhodes, Sander Vandenhaute, Vaidotas Šimkus, James Gin, Jonathan Godwin, Tim Duignan, and Mark Neumann. Orb-v3: atomistic simulation at scale. *arXiv preprint arXiv:2504.06231*, 2025.

Christoph Schran, Krystof Brezina, and Ondrej Marsalek. Committee neural network potentials control generalization errors and enable active learning. *The Journal of Chemical Physics*, 153 (10):104105, 2020. doi: 10.1063/5.0016004.

Kristof T. Schütt, Huziel E. Sauceda, Pieter-Jan Kindermans, Alexandre Tkatchenko, and Klaus-Robert Müller. SchNet – a deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24):241722, 2018. doi: 10.1063/1.5019779.

Kristof T. Schütt, Oliver T. Unke, and Michael Gastegger. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9377–9388. PMLR, 2021.

Daniel Schwalbe-Koda, Aik Rui Tan, and Rafael Gómez-Bombarelli. Differentiable sampling of molecular geometries with uncertainty-based adversarial attacks. *Nature Communications*, 12(1): 5104, 2021. doi: 10.1038/s41467-021-25342-8.

Maximilian Seitzer, Arash Tavakoli, Dimitrije Antic, and Georg Martius. On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks. *arXiv preprint arXiv:2203.09168*, 2022.

Alexander V Shapeev. Moment tensor potentials: A class of systematically improvable interatomic potentials. *Multiscale Modeling & Simulation*, 14(3):1153–1173, 2016.

Justin S. Smith, Ben Nebgen, Nicholas Lubbers, Olexandr Isayev, and Adrian E. Roitberg. Less is more: Sampling chemical space with active learning. *The Journal of Chemical Physics*, 148(24): 241733, 2018. doi: 10.1063/1.5023802.

William C Swope, Hans C Andersen, Peter H Berens, and Kent R Wilson. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *The Journal of chemical physics*, 76(1):637–649, 1982.

Aik Rui Tan, Shingo Urata, Samuel Goldman, Johannes CB Dietschreit, and Rafael Gómez-Bombarelli. Single-model uncertainty quantification in neural network potentials does not consistently outperform model ensembles. *npj Computational Materials*, 9(1):225, 2023.

Richard Tran, Janice Lan, Muhammed Shuaibi, Brandon M Wood, Siddharth Goyal, Abhishek Das, Javier Heras-Domingo, Adeesh Kolluru, Ammar Rizvi, Nima Shoghi, et al. The open catalyst 2022 (oc22) dataset and challenges for oxide electrocatalysts. *ACS Catalysis*, 13(5):3066–3084, 2023.

Jonathan Vandermause, Steven B Torrisi, Simon Batzner, Yu Xie, Lixin Sun, Alexie M Kolpak, and Boris Kozinsky. On-the-fly active learning of interpretable bayesian force fields for atomistic rare events. *npj Computational Materials*, 6(1):20, 2020.

Joshua A. Vita, Amit Samanta, Fei Zhou, and Vincenzo Lordi. LTAU-FF: Loss trajectory analysis for uncertainty in atomistic force fields. *Machine Learning: Science and Technology*, 6(1):015048, 2025. doi: 10.1088/2632-2153/adb4b9.

Mingjian Wen and Ellad B Tadmor. Uncertainty quantification in molecular simulations with dropout neural network potentials. *npj Computational Materials*, 6(1):124, 2020.

Daniel Willimetz and Lukáš Grajciar. A simple and scalable kernel density approach for reliable uncertainty quantification in atomistic machine learning. *The Journal of Physical Chemistry Letters*, 16(42):11081–11086, 2025.

Han Xu, Taoyong Cui, Chenyu Tang, Ji Ma, Dongzhan Zhou, Yuqiang Li, Xiang Gao, Xingao Gong, Wanli Ouyang, Shufei Zhang, and Mao Su. Evidential deep learning for interatomic potentials. *Nature Communications*, 17(1):937, 2026. doi: 10.1038/s41467-025-67663-y.

Olga Zaghen, Maksim Zhdanov, Dario Coscia, David R. Wessels, and Erik J. Bekkers. Uncertainty-aware machine learning interatomic potentials via learned functional perturbations. *arXiv preprint arXiv:2605.19939*, 2026.

Linfeng Zhang, De-Ye Lin, Han Wang, Roberto Car, and Weinan E. Active learning of uniformly accurate interatomic potentials for materials simulation. *Physical Review Materials*, 3(2):023804, 2019. doi: 10.1103/PhysRevMaterials.3.023804.

Qiyuan Zhao, Yunhong Han, Duo Zhang, Jiaxu Wang, Peichen Zhong, Taoyong Cui, Bangchen Yin, Yirui Cao, Haojun Jia, and Chenru Duan. Harnessing machine learning to enhance transition state search with interatomic potentials and generative models. *Advanced Science*, 12(34):e06240, 2025.

Albert Zhu, Simon Batzner, Albert Musaelian, and Boris Kozinsky. Fast uncertainty estimates in deep learning interatomic potentials. *The Journal of Chemical Physics*, 158(16):164111, 2023. doi: 10.1063/5.0136574.

Appendix

A Experimental Details **15**

A.1 Models Details 15

A.2 Systems and DFT Calculations Details 16

A.3 Simulation Details 16

A.4 Adversarial Attack 16

A.5 UQ Baselines and Metrics 17

A.6 Ammonia Dataset 17

B Additional Results **18**

B.1 Simulation Instability 18

B.2 Non-Conservativity and Instability Tracking 18

B.3 U_Δ on Adversarial Out-of-Distribution Structures 18

B.4 Hybrid Model Results 19

B.5 Finetuning Experiments 21

A Experimental Details

A.1 Models Details

We evaluate twelve pretrained universal MLIPs, all of which include MPtrj (Deng et al., 2023) in their pretraining data for section 4.3. The five Orb models (Neumann et al., 2024; Rhodes et al., 2025) are `orb-d3-xs-v2`, `orb-d3-v2`, `orb-d3-sm-v2`, `orb-v3-direct-inf-mpa`, and `orb-v3-direct-20-mpa`. The seven EquiformerV2 models (Liao et al., 2024) are `eqV2_dens_31M_mp`, `eqV2_dens_86M_mp`, `eqV2_dens_153M_mp`, `eqV2_31M_mp`, `eqV2_31M_omat_mp_salex`, `eqV2_86M_omat_mp_salex`, and `eqV2_153M_omat_mp_salex`. Only Orb-v3 provides a confidence head, and only EquiformerV2 was trained with dropout that enables MC dropout UQ, so neither UQ is available for both model families.

For Section 4.5 we train two architectures from scratch. The first is Orb-v3-direct-20 (Rhodes et al., 2025), a non-equivariant model with a 256-dimensional embedding. The second is MACE (Batatia et al., 2022), an equivariant model with a 128-dimensional embedding. We use the medium foundation configuration: `ScaleShiftMACE` with two interaction layers, a 6.0 Å cutoff, spherical harmonics up to $\ell_{\max} = 3$, correlation order 3, 128 channels, and hidden irreps $128 \times 0e + 128 \times 1o$. The parameter count is far below that of the foundation checkpoints because the ammonia dataset contains only two elements, and the element-dependent blocks scale with the number of species.

We add a direct force head of the same structure as Orb's to MACE backbone, a two-layer perceptron with a SiLU activation mapping to three outputs, resized to MACE's node feature width (128). It reads the per-atom node features concatenated across both interaction layers, and is therefore non-equivariant by construction, as is the case Orb's own force head. MACE also outputs conservative forces $\hat{\mathbf{F}}_{\mathrm{c}}$ as the gradient of its energy head.

Both architectures use the same training procedure in the standard $\hat{\mathbf{F}}_{\mathrm{nc}}$-only direct-force training and $\hat{\mathbf{F}}_{\mathrm{nc}} + \hat{\mathbf{F}}_{\mathrm{c}}$ hybrid settings: 500 epochs of Adam with loss weights $\lambda_E = 0.1$ and $\lambda_F = 1.0$, and a ReduceLROnPlateau schedule with factor 0.5. Orb uses a learning rate of 10^{-4}, patience 30, and batch size 8. MACE uses a learning rate of 3×10^{-4}. These differ slightly from the PaiNN (Schütt et al., 2021) training hyperparameters of Tan et al. (2023), which does not transfer to either

architecture. All runs converge within 500 epochs, although MVE converges to a poor optimum on some seeds.

A.2 Systems and DFT Calculations Details

Section 4.2 uses fifty structures randomly sampled from MPtrj (Deng et al., 2023), listed in Table 3.

Table 3: The 50 Materials Project structure IDs sampled from MPtrj and used in Section 4.2.

mp-1025458	mp-1025768	mp-1039650	mp-1044784	mp-1046472
mp-1073728	mp-1096858	mp-1097794	mp-1100678	mp-1101094
mp-1136	mp-1183626	mp-1188989	mp-1192108	mp-1196079
mp-1204176	mp-1208115	mp-1208209	mp-1209582	mp-1210593
mp-1217136	mp-1233565	mp-1246564	mp-1266342	mp-1275604
mp-1402005	mp-1403324	mp-1800	mp-18684	mp-18983
mp-20831	mp-21549	mp-27996	mp-30709	mp-540651
mp-540767	mp-542699	mp-555387	mp-560064	mp-569060
mp-601228	mp-754672	mp-757220	mp-764375	mp-769501
mp-771059	mp-772995	mp-7976	mp-979948	mp-983556

The UQ benchmark for pretrained models covers the following systems spanning three structural classes. The bulk solids are $Mg_{17}Al_{12}$, LGPS, ice, and MOF-5, taken from the Materials Project (Jain et al., 2013). The surface slab with adsorbate is CaPd-NH_2, taken from OC22 (Tran et al., 2023). The molecules are Ac-Ala3-NHMe, stachyose, aspirin, paracetamol, and DHA, taken from MD22 (Chmiela et al., 2023). The simulations in Sections 4.1 and 4.6 additionally use a liquid water box and a solid ice crystal.

All ground-truth forces are computed with VASP (Kresse & Furthmüller, 1996) using the PBE functional (Perdew et al., 1996), matching the level of theory of the pretraining datasets. For the Orb family we add the D3 dispersion correction (Grimme et al., 2010), since these models were pretrained on DFT+D3 labels (Neumann et al., 2024). The remaining calculation parameters follow Materials Project protocols (Jain et al., 2013; Munro et al., 2020).

A.3 Simulation Details

All MD simulations use the velocity Verlet integrator (Swope et al., 1982) and are performed in ASE (Larsen et al., 2017).

For Section 4.1, each system is first relaxed by geometry optimization with FIRE to a maximum force of 0.02 eV/Å, then equilibrated in the NPT ensemble at 300 K, or 200 K for ice which would otherwise melt, for 10 ps. The NVE simulation then runs for 50 ps with a 0.5 fs timestep, and we take the total absolute energy change per atom over the trajectory as $\Delta E_{\mathrm{drift}}$. To evaluate static accuracy along the same trajectory, we extract a frame every 100 steps and compute $\varepsilon_{\mathrm{nc}}$ and ε_{c} against the DFT references described in Appendix A.2.

For Section 4.2, each MPtrj structure is relaxed with FIRE (Bitzek et al., 2006) to a maximum force of 0.02 eV/Å, equilibrated for 1 ps in NPT at 300 K and 0 bar, then run for 5 ps of NVE with a 1 fs timestep. Trajectories are driven by $\hat{\mathbf{F}}_{\mathrm{nc}}$, so that they sample the instability the metric is meant to detect. We compute U_{Δ} every 10 steps and λ every 50, since the latter is far more expensive, requiring $3N$ backward passes per evaluation.

The NVT simulations in Appendix B.1 use a Nosé–Hoover thermostat with $\tau = 200$ fs at 300 K, with the same timestep and system as the corresponding NVE run.

A.4 Adversarial Attack

We generate OOD structures with the differentiable adversarial attack of Schwalbe-Koda et al. (2021), updating the atomic positions along a composite gradient that maximizes U_{Δ} while minimizing the

predicted energy, to obtain low-energy and high-uncertainty configurations,

$$\mathbf{r} \leftarrow \mathbf{r} + \alpha \nabla_{\mathbf{r}} U_{\Delta} - \beta \nabla_{\mathbf{r}} \hat{E}, \tag{4}$$

where α and β are the respective learning rates. Both range from 10^{-3} to 10^{-5} and are tuned per system, since the two terms have to be balanced against each other and their relative scale varies. Each attack runs for 10^3 to 10^4 steps, and structures are taken every 10 steps.

We stop collecting structures before $\hat{E}$ and U_{Δ} spike, which they usually do together and which indicates that the structures have become unphysical. Structures beyond that point are unreliable because no real system visits them, and are not worth labeling with DFT.

Selecting structures where a model is least reliable, that is, where it is highly uncertain, is a standard strategy for building training data efficiently, and has been shown to outperform random or MD-sampled configurations (Podryabinkin & Shapeev, 2017; Smith et al., 2018; Schwalbe-Koda et al., 2021; Kulichenko et al., 2023). MD-sampled frames are of limited value here, since a pretrained model already predicts them accurately and they carry little information about where it fails. Schwalbe-Koda et al. (2021) attacked ensemble disagreement, whereas we attack U_{Δ}, which Sections 4.2 to 4.5 show to track both non-conservativity and force error while requiring only a single model.

A.5 UQ Baselines and Metrics

We compare U_{Δ} against the uncertainty estimates available for pretrained models and, in Section 4.5, against methods that require their own training or fitting.

For the pretrained universal models, the heterogeneous ensemble uncertainty U_{var} is the variance of the force predictions across the members of each model family trained on the same dataset (Liu et al., 2026). We ensemble within model families rather than pooling all twelve models, since the variance of a mixed ensemble would be dominated by the systematic offset between the DFT and DFT+D3 references rather than by epistemic uncertainty. Orb-v3's confidence head predicts a binned per-atom uncertainty score, and we take the softmax-weighted bin value averaged over atoms. For MC dropout on EquiformerV2, we enable attention-weight dropout at inference with the rate of 0.1 used during training, and take the variance of the force predictions across five stochastic passes, averaged over atoms. We use five passes to match the five-member ensemble, following Huang et al. (2025) and Xu et al. (2026).

For the models trained from scratch, the deep ensemble is five independently seeded models, and its uncertainty is the variance of their force predictions. The GMM follows Tan et al. (2023): we extract the per-atom latent features from the layer preceding the energy prediction, fit a Gaussian mixture with full-rank covariance by expectation-maximization with k-means initialization, select the number of components from a trade-off between the Bayesian information criterion and the silhouette score, and score a structure by the negative log-likelihood of its features. MVE replaces the force head with one predicting both a mean and a variance per force component, trained with the Gaussian negative log-likelihood as in Tan et al. (2023), and its uncertainty is the predicted variance.

All uncertainties are aggregated to one value per structure by averaging over atoms. We evaluate them against the true force error with the Spearman rank correlation ρ, and with the AUROC for classifying structures as OOD using a cutoff at the 20th percentile of the true error, as in Tan et al. (2023).

A.6 Ammonia Dataset

The ammonia dataset of Tan et al. (2023) contains 78 near-equilibrium geometries for training and 129 adversarially sampled geometries for testing, the latter spanning a far wider energy range. We use their splits without modification: 62 training and 16 validation structures, with the full test set held out and never trained on. Energies and forces are converted from kcal/mol to eV, which scales both by the same constant and therefore preserves the ratio of the energy and force loss weights.

We follow their protocol throughout, including the data splits, the loss weights, the learning-rate schedule, and the OOD cutoff at the 20th percentile of the true force error. The only deviation is in the training settings, given in Appendix A.1, since their PaiNN (Schütt et al., 2021) architecture is different from both of our architectures.

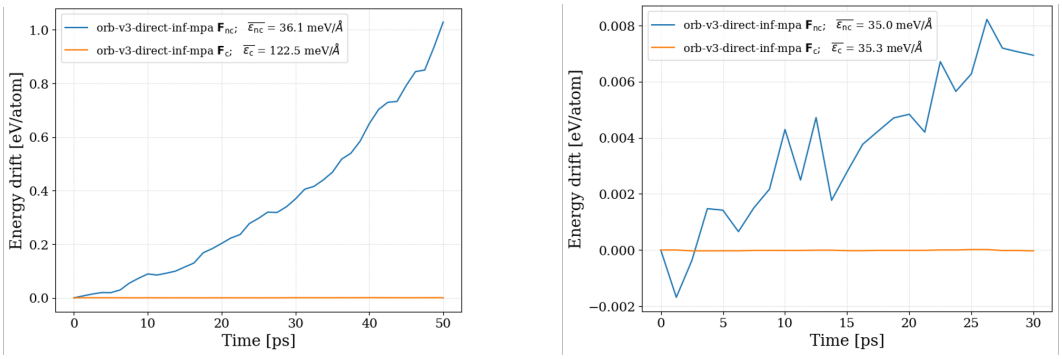
Because the original benchmark reports a single training run per method, we train five seeds of every method and report the mean and standard deviation across them. The deep ensemble is a single five-model predictor by construction, so we report one value for it rather than a spread.

B Additional Results

B.1 Simulation Instability

Figure 5 shows NVE simulation results for solid ice crystal and for $Mg_{17}Al_{12}$, with the same trend: the more accurate $\hat{\mathbf{F}}_{\mathrm{nc}}$ drifts, while the less accurate $\hat{\mathbf{F}}_{\mathrm{c}}$ of the same model does not.

We also run the liquid water box in NVT with $\hat{\mathbf{F}}_{\mathrm{nc}}$ and, separately, with $\hat{\mathbf{F}}_{\mathrm{c}}$ from the same model under identical thermostat settings. We find that non-conservativity also produces artifacts in thermostatted dynamics, where the accurate $\hat{\mathbf{F}}_{\mathrm{nc}}$ gives a temperature standard deviation of 19.9 K, against 10.6 K for the less accurate $\hat{\mathbf{F}}_{\mathrm{c}}$. This agrees with Bigi et al. (2025), who report that the thermostat performs continuous work to counteract the energy the non-conservative forces inject, and that this interference can disrupt structural and dynamical properties.



(a) Solid ice (b) $Mg_{17}Al_{12}$

Figure 5: Total energy drift over 50 ps of NVE driven by $\hat{\mathbf{F}}_{\mathrm{nc}}$ and $\hat{\mathbf{F}}_{\mathrm{c}}$ of the pretrained `orb-v3-direct-inf-mpa`, with each force's MAE against DFT+D3. The result of Section 4.1 holds for both systems.

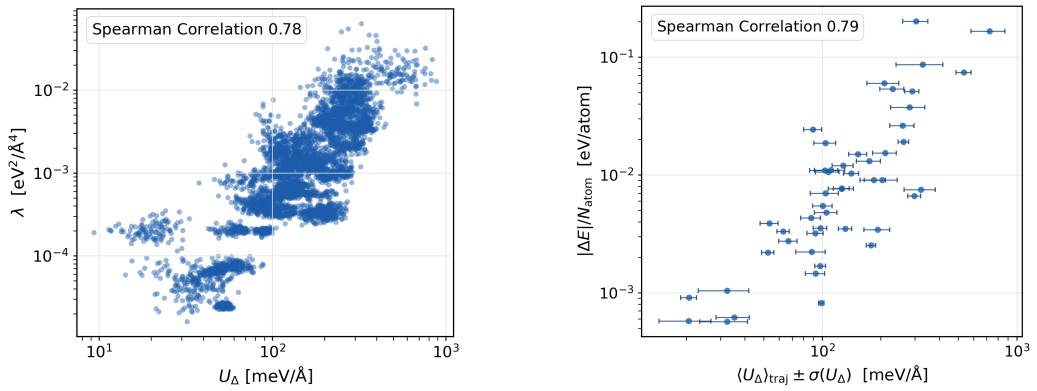
B.2 Non-Conservativity and Instability Tracking

We compute the Jacobian $\mathbf{J}$ of the direct force field by automatic differentiation, taking one backward pass per output component and $3N$ per evaluation, and define λ as the mean of the squared entries of its antisymmetric part $\mathbf{A} = (\mathbf{J} - \mathbf{J}^{\top})/2$. We use the exact Jacobian rather than a stochastic estimator, since the reference used to validate U_{Δ} should not itself be approximate.

Figure 6 shows the same correlations as Section 4.2 for `orb-v3-direct-20-mpa`, which give $\rho = 0.78$ against λ and $\rho = 0.79$ against $\Delta E_{\mathrm{drift}}$. Both are slightly weaker than for `orb-v3-direct-inf-mpa`, whose infinite neighborlist removes the effect of the neighborlist limit on the simulation instability. The two models agree closely, showing that U_{Δ} captures the non-conservativity and the resulting instability of multiple models.

B.3 U_{Δ} on Adversarial Out-of-Distribution Structures

Table 4 gives the per-system correlations between U_{Δ} and the two force errors, for OOD structures generated by adversarial attack and by NVE during energy drift. Across the 22 model–system pairs under attack, $\rho(U_{\Delta}, \varepsilon_{\mathrm{c}}) \geq 0.98$ in 19 systems and ≥ 0.93 in 21 systems. The one exception is MOF-5 under `orb-d3-xs-v2`, where U_{Δ} tracks neither error, indicating the U_{Δ} limitation discussed in Section 5: both forces deviate from the ground truth in the same direction, and their disagreement stays small while both errors are large.

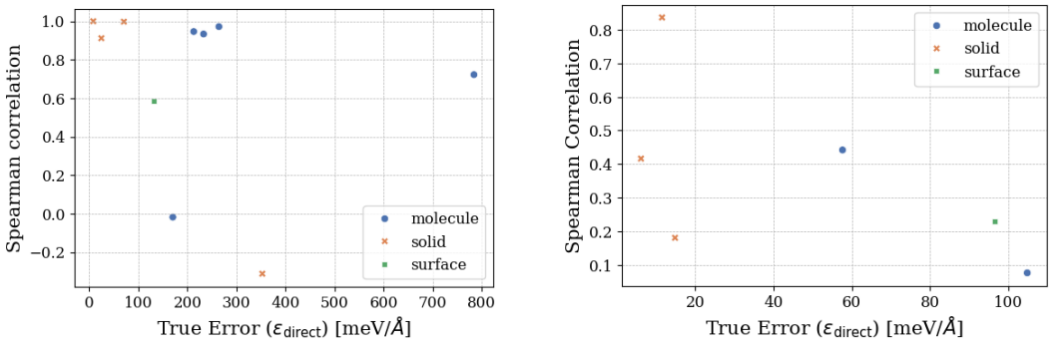


(a) (b)

Figure 6: U_Δ against (a) the exact Jacobian-based measure of non-conservativity λ and (b) the total NVE energy drift ΔE_{drift}, for `orb-v3-direct-20-mpa`. Error bars in (b) give the standard deviation of U_Δ along each trajectory.

Moreover, U_Δ tracks ε_{c} more closely than ε_{nc} because the two forces fail asymmetrically out of distribution. Under the loss in Equation 1, $\hat{\mathbf{F}}_{\text{nc}}$ is supervised directly on DFT forces, while $\hat{\mathbf{F}}_{\text{c}}$ is constrained only through the supervision of $\hat{E}$, which fixes the values of the predicted energy surface but not its curvature. Differentiation then acts as a high-pass filter, amplifying small high-frequency ripples in $\hat{E}$ into large errors in its gradient. As a model extrapolates, ε_{c} therefore grows faster than ε_{nc} and comes to dominate U_Δ.

The correlation with ε_{nc} is weaker for a few model–system pairs, and Figure 7 shows the reason being the initial error of the equilibrium structure, which predicts the resulting correlation with a negative trend. Where the model was already inaccurate before the attack began, the attack cannot increase the error enough to produce a wide range to rank over. Since U_Δ is uniformly high for these structures, they are still flagged as unreliable. Additionally, correlations are weaker for EquiformerV2 than for Orb, consistent with its force-to-energy loss weight ratio of 25 against Orb's 1, which makes $\hat{\mathbf{F}}_{\text{nc}}$ more robust and ε_{nc} correspondingly smaller and harder to rank.



(a) Orb (b) EquiformerV2

Figure 7: Initial ε_{nc} of the equilibrium structure against the resulting $\rho(U_\Delta, \varepsilon_{\text{nc}})$ under adversarial attack, for each model family.

B.4 Hybrid Model Results

Both forces of a direct-force model can be trained, so that a single model provides the fast $\hat{\mathbf{F}}_{\text{nc}}$ and the physically correct $\hat{\mathbf{F}}_{\text{c}}$. The two are then used interchangeably, with $\hat{\mathbf{F}}_{\text{nc}}$ for high-throughput

Table 4: Spearman ρ between U_Δ and the two force errors, per model and system. Upper blocks: adversarially attacked structures. Lower block: structures from NVE during energy drift.

System	Group	$\rho(U_\Delta, \varepsilon_{\mathrm{nc}})$	$\rho(U_\Delta, \varepsilon_{\mathrm{c}})$
`orb-v3-direct-inf-mpa`, adversarial attack			
$Mg_{17}Al_{12}$	Solid	1.00	1.00
LGPS	Solid	1.00	1.00
ice	Solid	0.91	0.99
MOF-5	Solid	−0.31	0.98
CaPd-NH_2	Surface	0.58	1.00
paracetamol	Molecule	0.97	1.00
stachyose	Molecule	0.93	1.00
Ac-Ala3-NHMe	Molecule	0.95	1.00
DHA	Molecule	0.72	0.99
aspirin	Molecule	−0.02	0.98
`orb-v3-direct-20-mpa`, adversarial attack			
LGPS	Solid	0.78	0.99
ice	Solid	0.96	1.00
`orb-d3-xs-v2`, adversarial attack			
$Mg_{17}Al_{12}$	Solid	0.73	0.94
LGPS	Solid	0.70	0.93
ice	Solid	0.70	0.99
MOF-5	Solid	−0.24	0.05
`eqV2-dens-31M-mp`, adversarial attack			
ice	Solid	0.84	0.99
$Mg_{17}Al_{12}$	Solid	0.42	0.98
LGPS	Solid	0.18	0.99
CaPd-NH_2	Surface	0.23	1.00
aspirin	Molecule	0.44	0.99
paracetamol	Molecule	0.08	0.98
`orb-v3-direct-inf-mpa`, NVE frames during energy drift			
$Mg_{17}Al_{12}$	Solid	0.85	0.99
LGPS	Solid	0.92	0.93
ice	Solid	0.71	0.98
water	Liquid (periodic)	0.88	0.94

screening or long trajectories and $\hat{\mathbf{F}}_{\mathrm{c}}$ where an exact potential energy surface is required, and the multiple-time-stepping scheme of Bigi et al. (2025) is one example that combines them to stabilize a direct-force-driven simulation. However, training both forces is more expensive than direct-force training alone, since supervising $\hat{\mathbf{F}}_{\mathrm{c}}$ reintroduces double backpropagation.

Tables 5 and 6 show U_Δ performance benchmark as a UQ for both forces against other metrics on the ammonia benchmark, following the same five-seed procedure as Section 4.5 and restricted to the $\hat{\mathbf{F}}_{\mathrm{nc}} + \hat{\mathbf{F}}_{\mathrm{c}}$ setting. U_Δ correlates well with the error of both forces, and therefore also serves as an out-of-the-box UQ metric covering both errors of hybrid models.

Moreover, supervising $\hat{\mathbf{F}}_{\mathrm{c}}$ reduces the in-domain magnitude of U_Δ, from 624.7 ± 13.5 to 50.8 ± 14.3 meV/Å on Orb and from 589.3 ± 8.6 to 40.3 ± 14.3 meV/Å on MACE. This increases the separation between in-distribution and out-of-distribution values, and accounts for the higher AUROC of U_Δ in the $\hat{\mathbf{F}}_{\mathrm{nc}} + \hat{\mathbf{F}}_{\mathrm{c}}$ setting than in the $\hat{\mathbf{F}}_{\mathrm{nc}}$-only setting in Table 2. Additionally, supervising $\hat{\mathbf{F}}_{\mathrm{c}}$ improves $\hat{\mathbf{F}}_{\mathrm{nc}}$ itself in domain, from 60.2 ± 9.7 to 51.8 ± 3.7 meV/Å on Orb and from 43.3 ± 2.5 to 30.2 ± 3.8 meV/Å on MACE, which agrees with Bigi et al. (2025; 2026) and Fu et al. (2025).

B.4.1 A Lower Bound on the Larger Force Error

Because both forces are compared against the same reference, their difference is the difference of their error fields componentwise. Taking the mean absolute value over the $3N$ components and

Table 5: U_Δ still serves as an out-of-the-box UQ metric with standard energy and force matching, correlating well with the error of both forces in hybrid Orb models. Spearman ρ and AUROC against the error of $\hat{\mathbf{F}}_{\mathrm{nc}}$ and of $\hat{\mathbf{F}}_{\mathrm{c}}$ on the ammonia benchmark, averaged over five seeds with standard deviations across seeds.

	$\hat{\mathbf{F}}_{\mathrm{nc}}$		$\hat{\mathbf{F}}_{\mathrm{c}}$	
Method	ρ	AUROC	ρ	AUROC
U_Δ	0.834 ± 0.067	0.982 ± 0.003	0.841 ± 0.062	0.985 ± 0.003
GMM	0.821 ± 0.065	0.968 ± 0.015	0.827 ± 0.061	0.971 ± 0.014
Deep ensemble	$\mathbf{0.858}$	$\mathbf{0.985}$	$\mathbf{0.864}$	$\mathbf{0.987}$

Table 6: U_Δ still serves as an out-of-the-box UQ metric with standard energy and force matching, correlating well with the error of both forces in hybrid MACE models. Spearman ρ and AUROC against the error of $\hat{\mathbf{F}}_{\mathrm{nc}}$ and of $\hat{\mathbf{F}}_{\mathrm{c}}$ on the ammonia benchmark, averaged over five seeds with standard deviations across seeds.

	$\hat{\mathbf{F}}_{\mathrm{nc}}$		$\hat{\mathbf{F}}_{\mathrm{c}}$	
Method	ρ	AUROC	ρ	AUROC
U_Δ	0.865 ± 0.077	0.973 ± 0.005	0.871 ± 0.073	0.976 ± 0.005
GMM	$\mathbf{0.895 \pm 0.032}$	0.981 ± 0.004	0.899 ± 0.030	0.983 ± 0.004
Deep ensemble	0.882	$\mathbf{0.986}$	$\mathbf{0.904}$	$\mathbf{0.989}$

applying the triangle inequality gives

$$\left|\varepsilon_{\mathrm{nc}} - \varepsilon_{\mathrm{c}}\right| \le U_\Delta \le \varepsilon_{\mathrm{nc}} + \varepsilon_{\mathrm{c}}, \qquad \text{and therefore} \qquad \max(\varepsilon_{\mathrm{nc}}, \varepsilon_{\mathrm{c}}) \ge \tfrac{1}{2} U_\Delta. \tag{5}$$

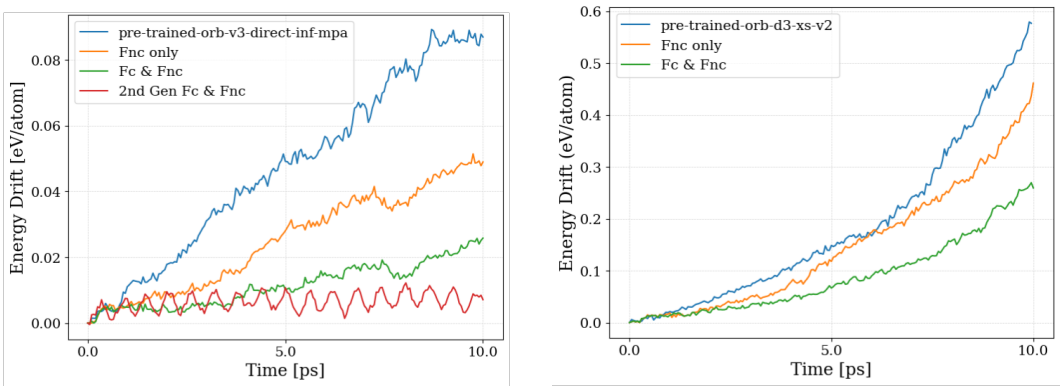
A large U_Δ therefore guarantees that at least one force is wrong by at least $U_\Delta/2$, without calibration or reference data.

The bound constrains $\max(\varepsilon_{\mathrm{nc}}, \varepsilon_{\mathrm{c}})$ rather than $\varepsilon_{\mathrm{nc}}$ alone. For a standard direct-force model, where only $\hat{\mathbf{F}}_{\mathrm{nc}}$ drives the simulation, a large U_Δ can be satisfied entirely by a large ε_{c} while $\varepsilon_{\mathrm{nc}}$ stays small, which Section 4.4 shows to be the typical case out of distribution. The bound is therefore useful for hybrid models, where both forces are used in production, while U_Δ's value for standard direct-force models rests on the correlations in Appendix B.3 instead.

B.5 Finetuning Experiments

We finetune the whole model with a learning rate of 10^{-5} for 200 epochs under a cosine annealing schedule, on 100 adversarial structures per generation labeled with DFT at the PBE+D3 level. $\hat{\mathbf{F}}_{\mathrm{nc}}$ is always supervised to prevent its degradation, since the production simulation is driven by $\hat{\mathbf{F}}_{\mathrm{nc}}$ (Bigi et al., 2025; 2026). The second generation continues from the first-generation model and trains on the combined 200 structures, following standard active learning practice.

Figure 8 shows the full NVE trajectories behind Figure 4, for `orb-v3-direct-inf-mpa` on ice and for `orb-d3-xs-v2` on LGPS. Both short-term drift and long-term instability fall at each stage, and the same trend holds across the two models and systems.



(a) Ice, `orb-v3-direct-inf-mpa` (b) LGPS, `orb-d3-xs-v2`

Figure 8: Energy evolution during NVE for the pretrained and finetuned models, showing reduced drift at each finetuning stage for both model–system pairs.