When Forces Disagree: A Data-Free Fast Uncertainty Estimate for Direct-Force Pre-trained Neural Network Potentials

Anonymous Author(s)
Affiliation
Address
email

Abstract

Neural Network Interatomic Potentials (NNIPs) are a cornerstone of modern atomistic simulations, but their reliability is limited by the difficulty in quantifying prediction uncertainty. Current uncertainty quantification (UQ) methods present a trade-off: model ensembles offer a robust, data-free metric based on model disagreement but are computationally expensive, while faster single-model methods typically require access to the original training data which can be practically inconvenient and chemically sparse. This paper introduces a novel differentiable UQ metric for direct-force pre-trained models that combines the strengths of both paradigms, offering the data-free reliability of ensembles with the computational speed of a single model. Our metric is derived from the internal disagreement between two force predictions from a single NNIP—the directly predicted (non-conservative) force and the energy-gradient-derived (conservative) force. We show a strong monotonic correlation between this force disagreement and the true force error against Density Functional Theory calculations. This relationship is robust across a diverse set of materials and holds even for out-of-distribution structures generated via adversarial attacks. Because the method is computationally cheap and requires no training data, it offers a powerful, out-of-the-box tool for on-the-fly assessment of model confidence with wide-ranging applications for reliable atomistic modeling.

1 Introduction

Machine-learned interatomic potentials (MLIPs), particularly those based on neural networks (NNIPs), have become essential tools in computational materials science, bridging the accuracy of quantum mechanics with the efficiency of classical force fields [1–4]. Despite their success, NNIPs can fail catastrophically on out-of-distribution (OOD) structures, leading to unstable simulations and incorrect scientific conclusions [5–7]. Robust uncertainty quantification (UQ) is therefore critical for their trustworthy application [8–11].

The dominant UQ methods for MLIPs fall into two main families. The first, deep ensembles, trains multiple models and uses their prediction variance as a robust, data-free uncertainty estimate, but at a high computational cost [5, 12, 13]. The second family comprises single-model methods. Many of these, such as Bayesian Neural Networks (BNNs), require specialized and often complex training procedures to learn an approximate posterior distribution over model weights [14, 15]. Other single-model approaches, including those based on distance metrics or density estimators like Gaussian Mixture Models (GMMs), but these are data-dependent, requiring access to the original training set to assess novelty [2, 16–18]. This is a major bottleneck for the growing ecosystem of large, pre-trained

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

"foundation models" as access to the massive training datasets can be time-consuming and the datasets themselves can have low utility for a specific system of interest [19–22].

Modern NNIPs can predict forces in two ways: (1) **conservative forces** ($\hat{\mathbf{F}}_c$), calculated as the negative gradient of the predicted energy ($\hat{\mathbf{F}}_c = -\nabla\hat{E}$), which is physically rigorous but slower [23, 24]; and (2) **non-conservative forces** ($\hat{\mathbf{F}}_{nc}$), predicted directly as a vector output, which is faster but violates energy conservation [25–27]. We propose that the disagreement between these two predictions, a quantity we term the "Force Delta," can be used as a powerful, data-free UQ metric. Our method captures the data-free benefit of ensembles, which rely on internal model disagreement, while retaining the computational efficiency of single-model UQ, offering an out-of-the-box tool with efficient implementation for on-the-fly model evaluation (see Appendix for Computational Cost).

2 Methods

2.1 The Force Delta Uncertainty Metric

A fundamental property of a physical force field is that it must be conservative, meaning the forces are the negative gradient of a potential energy, $\mathbf{F} = -\nabla_{\mathbf{R}}E$. A direct mathematical consequence is that the curl of a conservative force field is zero ($\nabla \times \mathbf{F} = 0$). Any violation of this condition signals a failure to represent the true underlying physics.

An NNIP is a function $\mathcal{F}_{NN}$ that maps an atomic configuration $\mathbf{R} = \{\mathbf{r}_i, Z_i\}$ to a predicted potential energy $\hat{E}_{\text{NN}}(\mathbf{R})$ and a set of atomic forces. The conservative force on atom i is $\hat{\mathbf{F}}_{c,i}(\mathbf{R}) = -\nabla_{\mathbf{r}_i}\hat{E}_{\text{NN}}(\mathbf{R})$, computed via automatic differentiation. The non-conservative force, $\hat{\mathbf{F}}_{nc,i}$, is predicted directly by a separate output head for direct-force NNIPs.

In a perfectly learned model, these two forces would be identical. Therefore, any disagreement between them is a direct measure of the model's **physical inconsistency** and a local violation of energy conservation. A non-zero difference implies that the directly predicted force field has a non-zero curl, a clear indicator of the model's failure to capture the true potential energy surface. This interpretability is a significant advantage over more abstract metrics like ensemble variance.

We quantify this physical violation by defining our uncertainty metric, the **Force Delta** (U_Δ), as the root-mean-square (RMS) of the vector difference between these two force predictions, averaged over all $3N$ force components:

$$U_\Delta(\mathbf{R}) = \sqrt{\frac{1}{3N}\sum_{i=1}^{N} \|\hat{\mathbf{F}}_{nc,i}(\mathbf{R}) - \hat{\mathbf{F}}_{c,i}(\mathbf{R})\|^2} \tag{1}$$

This metric is applicable to any model architecture that provides both an invariant scalar energy output (for $\hat{\mathbf{F}}_c$) and a separate equivariant vector output (for $\hat{\mathbf{F}}_{nc}$). To validate this metric, we compare it against the **true error**, $\varepsilon_{\text{direct}}$, which is the RMS difference between the model's non-conservative force and the ground-truth DFT force, $\mathbf{F}_{\text{DFT},i}$:

$$\varepsilon_{\text{direct}}(\mathbf{R}) = \sqrt{\frac{1}{3N}\sum_{i=1}^{N} \|\hat{\mathbf{F}}_{nc,i}(\mathbf{R}) - \mathbf{F}_{\text{DFT},i}(\mathbf{R})\|^2} \tag{2}$$

We validate against $\varepsilon_{\text{direct}}$ because $\hat{\mathbf{F}}_{nc}$ is often preferred in production simulations for speed, making its error the most relevant quantity to estimate. Our central claim is that a strong, predictive monotonic relationship exists between U_Δ and $\varepsilon_{\text{direct}}$.

2.2 Adversarial Generation of OOD Structures

To rigorously test our metric on challenging OOD configurations, we employ an adversarial attack strategy [28–31]. Starting from equilibrium structures, we iteratively perturb the atomic positions $\mathbf{r}$ to find configurations that are both physically plausible (low energy) and maximally uncertain. This is achieved by updating the atomic positions along a composite gradient that simultaneously maximizes our uncertainty metric U_Δ while minimizing the predicted potential energy $\hat{E}_{\text{NN}}$ [32]:

$$\mathbf{r}_{\text{new}} = \mathbf{r}_{\text{old}} + \alpha\nabla_{\mathbf{r}}U_\Delta - \beta\nabla_{\mathbf{r}}\hat{E}_{\text{NN}} \tag{3}$$

where α and β is the learning rates for the attack and energy minimization (to ensure the generated OOD configurations still conforms to the Boltzmann distribution), respectively. This differentiable process efficiently drives the system towards high-uncertainty, but low-energy scenarios where the model's internal predictions disagree most strongly [6, 33].

3 Results and Discussion

3.1 Validation on Equilibrium Structures

We first evaluated the Force Delta on stable, in-distribution structures to establish a baseline. For each of the 10 material systems (see Appendix), we used an equilibrium configuration and calculated the average Force Delta (U_Δ) and average true error ($\varepsilon_{\text{direct}}$) across all 15 of the pre-trained models (see Appendix). Figure 1a shows a remarkably strong monotonic association, confirmed by a Spearman's rank correlation coefficient of $r_s = 0.98$. This indicates that for well-behaved structures, the Force Delta is an initial powerful indicator of the underlying model error.

3.2 Comparison with an Ad-Hoc Ensemble Baseline

To benchmark the Force Delta against a standard ensemble-based approach, we performed a head-to-head comparison within each model family since Orb and EquiformerV2 were trained using different ground-truth DFT methods (w/ vs. w/o D3). We compared how well our single-model Force Delta predicts the error of its individual model against how well the ad-hoc ensemble variance predicts the error of the ensemble's average prediction.

It is crucial to note that these ad-hoc collection of models are not "deep ensembles" in the strictest sense, as they were not co-trained with varied initializations on an identical dataset. However, they represent the most direct ensemble-based UQ approach available to a user working with publicly available pre-trained models.

The results, summarized in Table 1, reveal the remarkable effectiveness of the Force Delta. For the EquiformerV2 family, the single-model Force Delta significantly outperforms the 10-model ensemble variance, achieving a much higher average Spearman correlation. This demonstrates that for this diverse set of models, probing the internal physical consistency is a fundamentally more reliable UQ strategy than measuring external disagreement.

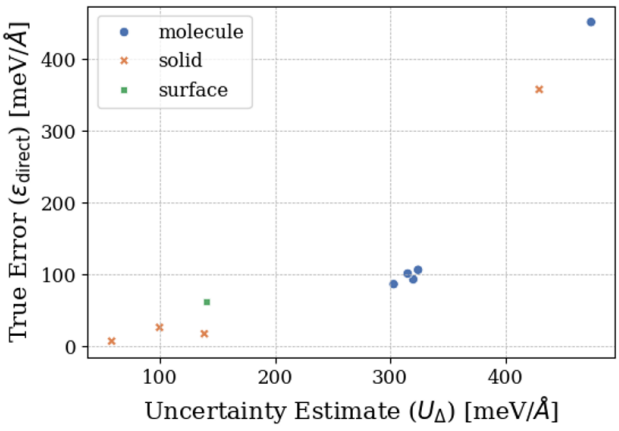
For the Orb family, the ensemble variance shows a slightly stronger correlation on average, suggesting that the optimal UQ strategy can be model-dependent. Nonetheless, the Force Delta still provides a robust and reliable uncertainty estimate. This confirms that our method provides UQ of a quality comparable to an expensive ensemble, while retaining the out-of-the-box efficiency of a single-model approach in any practical application.

Table 1: Comparison of UQ strategies on equilibrium structures. The single-model Force Delta (U_Δ) is benchmarked against the ad-hoc ensemble force variance (U_{var}). The Force Delta demonstrates significantly superior performance for the EquiformerV2 family and competitive performance for the Orb family, while being far more computationally efficient in practice.

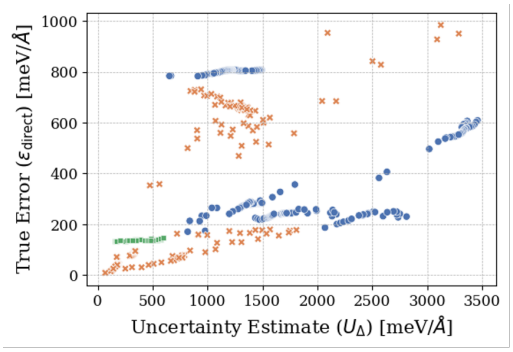
Model Family	Avg. r_s (Force Delta, U_Δ)	r_s (Ensemble Variance, U_{var})
Orb (5 models)	0.70 ± 0.04	**0.73**
EquiformerV2 (10 models)	$\mathbf{0.91 \pm 0.02}$	0.78

3.3 Robustness under Adversarial Attack

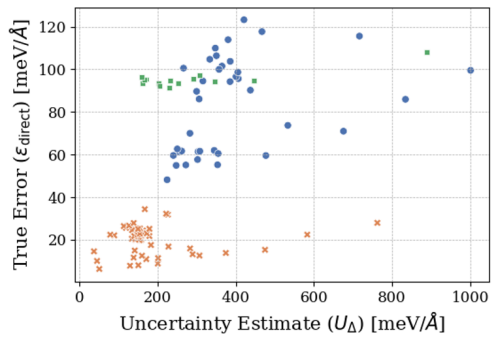
The strong monotonic correlation generally holds even for OOD structures generated via adversarial attacks. Figure 1b shows a parity plot for individual structures (both standard and adversarial) for the Orb potential across all 10 systems. The data points cluster tightly along a monotonic curve, demonstrating a direct correspondence between the internal force disagreement and the actual error. Because the relationship is not strictly linear, we use Spearman's rank correlation, which is a more robust measure of association.



(a) Equilibrium structures



(b) Adversarial (Orb)



(c) Adversarial (eqV2)

Figure 1: Correlation between Force Delta (U_Δ) and true error ($\varepsilon_{\text{direct}}$). (a) Model-averaged values for equilibrium structures ($r_s = 0.98$). Adversarial structures for (b) Orb and (c) EquiformerV2, demonstrating generality across different model architectures and material systems

The correlation for Orb is very strong ($r_s > 0.9$) for most systems (see Appendix). For a few systems (e.g., MoF5, aspirin), the correlation is weaker or even negative. This is because the true error of the initial equilibrium structure was already substantial (see Appendix). Consequently, the adversarial attack, while still finding high-uncertainty configurations, did not produce as dramatic an *increase* in error, which can weaken the calculated correlation coefficient. Crucially, the Force Delta for these points is consistently high, correctly flagging them as unreliable. This shows the metric functions as an effective "failure detector" for applications like active learning or molecular dynamics (MD) monitoring, where identifying failure is often more important than perfect error prediction.

To test generality, we performed the same analysis on EquiformerV2 for six systems. While the Spearman correlation is more modest (see Appendix), a clear positive monotonic trend remains for most systems, as shown in Figure 1c. This demonstrates that the underlying principle—that internal force disagreement tracks with true error—is not unique to one model architecture. The weaker correlation suggests that the *quality* of the UQ metric may be model-dependent (different architectures or training strategies), but the metric itself is still present and useful.

Our results confirm the Force Delta is an excellent metric for **ranking** uncertainty, making it ideal for applications like active learning. However, the metric is not **calibrated**: its magnitude does not directly predict the magnitude of the true error, as shown in Figure 1. Therefore, its primary role is as a robust and efficient criterion for identifying unreliable predictions, not as a precise error estimator.

3.4 The Data-Free Advantage at a Single-Model Efficiency

Our method combines the strengths of the two dominant UQ families, extending the ensemble principle to a single model, where disagreement among diverse models is a robust, data-free estimate for epistemic uncertainty [5, 12]. It measures the disagreement between two physically-motivated predictive pathways within a single model, providing a similar estimate of internal inconsistency but at the computational cost of a single model (see Appendix), avoiding the substantial expense of training and running multiple large models [13, 11].

This data-free nature is not merely a convenience but a critical advantage, essentially eliminating setup costs associated with often proprietary or intractable training data of large-scale foundation models. Moreover, the implementation is computationally efficient, requiring only a single additional backpropagation pass per structure, in contrast to the substantial cost of training an entirely separate statistical model on a large training set. Data-dependent UQ also suffers from a more fundamental data utility problem—a universal potential's training set may be vast but sparse for a specific system [1, 34]. Furthermore, on heterogeneous data, these methods are known to underestimate errors and can fail counterintuitively in OOD settings, where uncertainty may decrease as error grows [6, 35–37]. Our method avoids these pitfalls by directly probing the model's physical inconsistency—the inability of the model to perfectly represent the true physics—which is the dominant source of error in MLIPs and is often ignored by standard Bayesian UQ frameworks [38–40].

3.5 Applications in Atomistic Modeling

This work provides an essential out-of-the-box estimate of model reliability with wide-ranging applications for reliable atomistic modeling. In **high-throughput screening**, it can act as a filter to flag unreliable predictions for more expensive validation, focusing resources where they are most needed [41, 42]. For **molecular dynamics**, it provides an on-the-fly safeguard to detect when a simulation could enter an OOD region, preventing numerical instabilities and enabling more stable long-timescale simulations [43, 44]. In **active learning**, it provides a highly efficient differentiable sampling strategy to guide the selection of new training data, improving the data-efficiency of model training and accelerating the development of a robust potential [45, 46]. Finally, the method enables new paradigms for **benchmarking and model development**. It allows for the data-free selection of the best pre-trained model for a specific task and can be used as a physics-informed regularization term during training to improve generalization. For the growing number of foundation models, this gives users an essential tool to evaluate model reliability on their own systems.

4 Conclusion

We have introduced the Force Delta, a fast, accurate, and data-free uncertainty metric for NNIPs based on internal model disagreement. We demonstrated a strong monotonic correlation between this metric and the true DFT error across diverse materials, models, and for both equilibrium and adversarially generated OOD structures. The method's key advantage is its data-free and single-model nature with efficient implementation, overcoming the severe practical and theoretical limitations of data-dependent UQ and enables reliable uncertainty estimation for any direct-force pre-trained model. By providing a new paradigm for high-throughput screening, molecular dynamics simulations, active learning, model benchmarking, and physics-informed training, this work provides a computationally lightweight and out-of-the-box framework for assessing the predictive reliability of the next generation of direct-force foundation models for materials discovery.

References

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

[2] 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.

[3] Emir Kocer, Tsz Wai Ko, and Jörg Behler. Neural network potentials: A concise overview of methods. *Annual Review of Physical Chemistry*, 73:163–186, 2022.

[4] Oliver Unke, Stefan Chmiela, Michael Gastegger, Kristof T Schütt, H E Sauceda, and Klaus-Robert Müller. Machine learning force fields. *Chemical Reviews*, 121(17):10142–10186, 2021.

[5] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pages 6402–6413, 2017.

[6] 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. *arXiv preprint arXiv:2305.01754*, 2023.

[7] Linfeng Zhang, Jiequn Han, Han Wang, Roberto Car, and Weinan E. Deep potential molecular dynamics: a scalable model with the accuracy of quantum mechanics. *Physical review letters*, 120(14):143001, 2018.

[8] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Ling Liu, Mohammad Ghasemian, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021.

[9] Khachik Sargsyan, Logan Williams, and Habib N. Najm. Uncertainty quantification of machine learning interatomic potential models. USACM Thematic Conference on Uncertainty Quantification for Machine Learning Integrated Physics Modeling (UQ-MLIP), 2022.

[10] Shuo Chen, Suk-Wah Lee, Chi Chen, Weile Ji, Liping Zhang, Mohan Chen, et al. Uncertainty quantification for atomic-scale machine learning. *npj Computational Materials*, 5(1):118, 2019.

[11] Brooke E Husic, Nicholas E Charron, D Lemm, Jiang Wang, Adrià Pérez, Maciej Majewski, et al. Coarse-graining, machine learning, and the curse of dimensionality. *The Journal of chemical physics*, 153(19), 2020.

[12] Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019.

[13] Fredrik K Gustafsson, Martin Danelljan, and Thomas B Schon. Evaluating scalable bayesian deep learning methods for robust computer vision. pages 320–321, 2020.

[14] Radford M Neal. *Bayesian Learning for Neural Networks*, volume 118. Springer Science & Business Media, 1996.

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

[16] Christopher M Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[17] Douglas A Reynolds and Richard C Rose. Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Transactions on speech and audio processing*, 3(1):72–83, 1995.

[18] Cas van der Hirschfeld, Giulio Imbalzano, and Michele Ceriotti. Uncertainty quantification in atomistic machine learning. *The Journal of Chemical Physics*, 153(10), 2020.

[19] Kevin Maik Jablonka, Greeshma M Jothiappan, Shen Wang, Berend Smit, and Berend Yoo. Bias free multiobjective active learning for materials design and discovery. *Nature Communications*, 12(1):2312, 2021.

[20] Amil Merchant, Simon Batzner, Samuel S Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin D Cubuk. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85, 2023.

[21] Chengcheng Lu, Peigang Huang, Weike Liu, Bang Liu, et al. A survey of ai for materials science: Foundation models, llm agents, datasets, and tools. *arXiv preprint arXiv:2506.20743*, 2024.

[22] Nhat-Duc Tran, Chi Chen, S Joshua Swamidass, and Shyue Ping Ong. Towards a universal neural network potential for material discovery. *Chemistry of Materials*, 34(19):8569–8580, 2022.

[23] Mark E Tuckerman. *Statistical Mechanics: Theory and Molecular Simulation*. Oxford University Press, 2023.

[24] Atilim Gunes Baydin, Barak A Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. Automatic differentiation in machine learning: a survey. *Journal of Machine Learning Research*, 18(153):1–43, 2018.

[25] Daniele Bigi, Marco F Langer, and Michele Ceriotti. The dark side of the forces: assessing non-conservative force models for atomistic machine learning. *arXiv preprint arXiv:2401.06208*, 2024.

[26] Yuchao Fu, Tian Lan, et al. Learning smooth and expressive interatomic potentials for physical property prediction. *arXiv preprint arXiv:2502.12147*, 2024.

[27] Andrea Grisafi and Michele Ceriotti. Incorporating long-range physics in atomic-scale machine learning. *The Journal of chemical physics*, 151(20), 2019.

[28] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[29] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[30] Justin Gilmer, Andrew Ilyas, Logan Engstrom, and Aleksander Madry. Adversarial examples are a natural consequence of test error in noise. *arXiv preprint arXiv:1901.10513*, 2019.

[31] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

[32] Daniel Schwalbe-Koda, Aik Rui Tan, and Rafael Gómez-Bombarelli. Differentiable sampling of molecular geometries with uncertainty-based adversarial attacks. *Nature Communications*, 12(1):5035, 2021.

[33] Heiko Stuke, Kristof T Schütt, Michael Gastegger, and Klaus-Robert Müller. Chemical exploration with active learning and adversarial attacks. *Machine Learning: Science and Technology*, 2(2):025034, 2021.

[34] Chi Chen, Kevin Gan, et al. Improved materials 3-body graph network universal potential with direct sampling. *npj Computational Materials*, 10(1):1–10, 2024.

[35] Fei Shuang, Zixiong Wei, Kai Liu, Wei Gao, and Poulumi Dey. Model accuracy and data heterogeneity shape uncertainty quantification in machine learning interatomic potentials. *arXiv preprint arXiv:2508.03405*, 2025.

[36] Yonatan Kurniawan, Mingjian Wen, Ellad B Tadmor, and Mark K Transtrum. Comparative study of ensemble-based uncertainty quantification methods for neural network interatomic potentials. *arXiv preprint arXiv:2508.06456*, 2025.

[37] Andrej Nistranec, A Gilad Kusne, et al. Uncertainty-aware machine learning for autonomous experiments. *Current Opinion in Solid State and Materials Science*, 28(1):100112, 2024.

[38] T. D. Swinburne and D. Perez. Uncertainty quantification for misspecified machine learned interatomic potentials. *arXiv preprint arXiv:2502.07104*, 2025.

[39] Houman Owhadi and Clint Scovel. Kernel-based uncertainty quantification and calibration of black-box models. *arXiv preprint arXiv:1901.07821*, 2019.

[40] Volker L Deringer, Miguel A Caro, and Gábor Csányi. Gaussian process regression for materials and molecules. *Advanced Materials*, 33(46):2002456, 2021.

[41] Jeff Greeley, Thomas F Jaramillo, Jacob Bonde, Ib Chorkendorff, and Jens K Nørskov. Computational high-throughput screening of electrocatalytic materials for hydrogen evolution. *Nature Materials*, 8(10):809–813, 2009.

[42] Y Zhang and C Ling. Machine learning: a powerful tool for facilitating materials discovery and design. *MRS Communications*, 10(1):1–14, 2020.

[43] Shunzhou Wan, Robert C Sinclair, and Peter V Coveney. Uncertainty quantification in classical molecular dynamics. *Philosophical Transactions of the Royal Society A*, 379(2197):20200082, 2021.

[44] Xiaowo Li, Julia A Plews, Guang Lin, and Dallas R Trinkle. Uncertainty quantification of interatomic potentials. *Physical Review B*, 104(1):014105, 2021.

[45] Burr Settles. Active learning literature survey. Technical Report 1648, University of Wisconsin-Madison Department of Computer Sciences, 2009.

[46] Kevin Maik Jablonka, Giriprasad M Jothiappan, Shen Wang, Berend Smit, and Berend Yoo. Bias free multiobjective active learning for materials design and discovery. *Nature Communications*, 12(1):2312, 2021.

[47] 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.

[48] Yi-Lun Liao, Brandon Liu, H. T. Pao, Yuchao Zhao, Bryan Wood, C. Lawrence Zitnick, and Tess Smidt. Equiformerv2: Improved equivariant transformer for scaling to higher-degree representations. In *The Twelfth International Conference on Learning Representations*, 2024.

[49] 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.

[50] Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P Mailoa, et al. E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature communications*, 13(1):2453, 2022.

[51] 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.

[52] 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.

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

[54] 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.

A Appendix

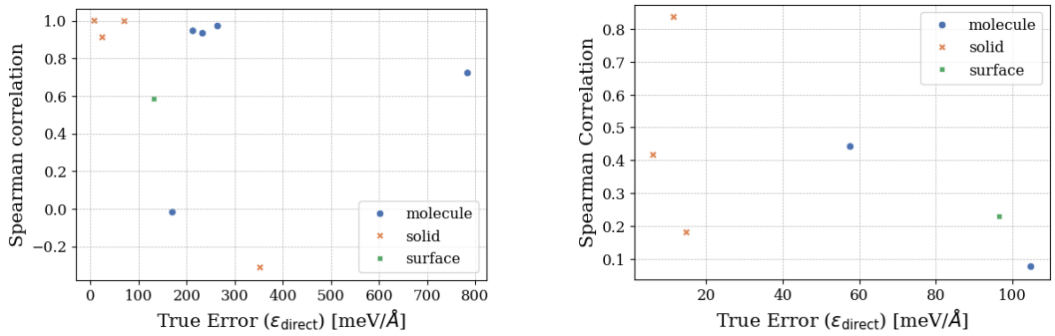
A.1 Correlation between UQ and True Error on OOD adversarial test set

A.2 Analysis of Correlation Strength vs. Initial Model Error

Figure 2 directly visualizes the argument made in the main text: that systems with a high initial error on their equilibrium structure tend to exhibit weaker Spearman correlations during adversarial attacks. This supports our conclusion that a weak correlation coefficient does not necessarily indicate a failure of the UQ metric, but can be an artifact of the model already being highly uncertain.

Table 2: Spearman's rank correlation (r_s) between U_Δ and $\varepsilon_{\text{direct}}$ on adversarial test sets.

Orb v3			**EquiformerV2**		
System	Group	r_s	System	Group	r_s
$Mg_{17}Al_{12}$	Solid	1.00	ice	Solid	0.84
LGPS	Solid	1.00	$Mg_{17}Al_{12}$	Solid	0.42
ice	Solid	0.91	LGPS	Solid	0.18
MoF5	Solid	-0.31	$CaPd\text{-}NH_2$	Surface	0.23
$CaPd\text{-}NH_2$	Surface	0.58	aspirin	Molecule	0.44
paracetamol	Molecule	0.97	paracetamol	Molecule	0.08
stachyose	Molecule	0.93			
Ac-Ala3-NHMe	Molecule	0.95			
DHA	Molecule	0.72			
aspirin	Molecule	-0.02			



(a) Orb v3 (b) EquiformerV2

Figure 2: Analysis of Spearman's rank correlation (r_s) from adversarial attacks versus the initial true error ($\varepsilon_{\text{direct}}$) of the equilibrium structure for each system. For systems with low initial error, the adversarial attack creates a wide range of errors, leading to strong correlations. For systems where the model is already highly inaccurate, the dynamic range is smaller, weakening the calculated correlation.

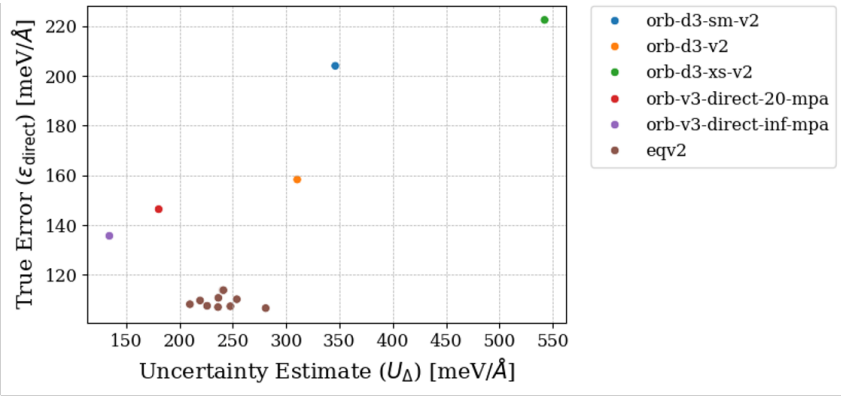


Figure 3: Correlation between Force Delta (U_Δ) and true error ($\varepsilon_{\text{direct}}$). (a) Model-averaged values for equilibrium structures

A.3 Benchmarking Potential of the Force Delta on Equilibrium Structures

To demonstrate the potential of the Force Delta as a tool for benchmarking and model selection, this section provides a plot (Figure 3) showing the Force Delta versus the true error for each of the 15 models in our ensemble, evaluated on the same equilibrium structures. The results show a perfect correlation for the Orb models, indicating that the Force Delta can distinguish between

the performance of different model versions. The EquiformerV2 models, however, show no clear correlation in this test, suggesting that this benchmarking capability may also be model-dependent.

A.4 Computational Details

A.4.1 Model Details

We performed validations using 15 state-of-the-art NNIPs representing diverse architectural classes: five versions of Orb, all of which are attention-augmented Graph Neural Networks, and ten versions of EquiformerV2, an E(3)-Equivariant Transformer [47–50]. This diversity allows us to test the generality of our findings. The specific model versions are listed below.

A.4.2 Orb Models

The five Orb models used were:

- `orb-d3-xs-v2`
- `orb-d3-v2`
- `orb-d3-sm-v2`
- `orb-v3-direct-inf-mpa`
- `orb-v3-direct-20-mpa`

EquiformerV2 Models

The ten EquiformerV2 models used were:

- `eqV2_dens_31M_mp`
- `eqV2_dens_153M_mp`
- `eqV2_dens_86M_mp`
- `eqV2_31M_mp`
- `eqV2_31M_omat`
- `eqV2_153M_omat`
- `eqV2_86M_omat`
- `eqV2_31M_omat_mp_salex`
- `eqV2_153M_omat_mp_salex`
- `eqV2_86M_omat_mp_salex`

Adversarial attacks were performed on all 10 systems for the Orb potential (`orb-v3-direct-20-mpa`) and on six representative systems for `eqV2_dens_31M_mp` model.

A.4.3 Materials Details

Our test set comprised 10 systems spanning solids ($Mg_{17}Al_{12}$, LGPS, ice, and MoF-5), surfaces (CaPd-NH_2), and molecules (Ac-Ala3-NHMe, stachyose, aspirin, paracetamol, and DHA taken md22 dataset [51]).

A.4.4 DFT Calculation Details

All ground-truth Density Functional Theory (DFT) calculations were performed with the Vienna Ab initio Simulation Package (VASP). We used the Perdew-Burke-Ernzerhof (PBE) exchange-correlation functional [52, 53]. All calculation parameters including k-point mesh densities were chosen to be consistent with the Materials Project [54], ensuring convergence. For surface and molecular calculations, structures were placed in a large simulation box with at least 15 Å of vacuum to ensure no spurious interactions between periodic images.

A.4.5 Computational Cost

The Force Delta offers significant speed advantages over ensembles (5-10x faster). Compared to single-model methods, the cost depends on the context. If a simulation typically uses the fast $\hat{\mathbf{F}}_{nc}$, calculating U_Δ requires computing $\hat{\mathbf{F}}_c$ (the backpropagation), which roughly doubles the computational cost per step ($\sim$2x overhead). If the simulation already uses $\hat{\mathbf{F}}_c$, the overhead is negligible if $\hat{\mathbf{F}}_{nc}$ is computed during the initial forward pass. This overhead is significantly less than the cost of data-dependent methods, which require searching large training databases or training separate statistical models.