P-DRUM: Post-hoc Descriptor-based Residual Uncertainty Modeling for Machine Learning Potentials

Shih-Peng Huang*
Massachusetts Institute of Technology
sphuang@mit.edu

Nontawat Charoenphakdee
Preferred Networks
nontawat@preferred.jp

Yuta Tsuboi
Preferred Networks
tsuboi@preferred.jp

Yong-Bin Zhuang
Preferred Networks
ybzhuang@preferred.jp

Wenwen Li
Preferred Networks
wenwenli@preferred.jp

Abstract

Ensemble method is considered the gold standard for uncertainty quantification (UQ) in machine learning interatomic potentials (MLIPs). However, their high computational cost can limit its practicality. Alternative techniques, such as Monte Carlo dropout and deep kernel learning, have been proposed to improve computational efficiency; however, some of these methods cannot be applied to already trained models and may affect the prediction accuracy. In this paper, we propose a simple and efficient post-hoc framework for UQ that leverages the descriptor of a trained graph neural network potential to estimate residual errors. We refer to this method as post-hoc descriptor-based residual uncertainty modeling (P-DRUM). P-DRUM models the discrepancy between MLIP predictions and ground truth values, allowing these residuals to act as proxies for prediction uncertainty. We explore multiple variants of P-DRUM and benchmark them against established UQ methods, evaluating both their effectiveness and limitations.

1 Introduction

Machine learning interatomic potentials (MLIPs) are transforming materials science by enabling atomic-scale simulations with the accuracy of quantum mechanical methods but at orders-of-magnitude higher computational efficiency [1–10]. Despite their promise, the predictive reliability of MLIPs remains a critical concern, especially for atomic configurations outside the training data distribution [11–14]. Robust uncertainty quantification (UQ) is essential for assessing model reliability, guiding decision-making, and ensuring trustworthy simulation outcomes.

Several methods for UQ have been explored for MLIPs [15–17]. Ensemble approaches, which aggregate predictions from multiple independently trained models, are widely regarded as the gold standard due to their effectiveness and their simplicity to implement [12, 18]. However, ensembles are computationally expensive, particularly in settings with large-scale data or complex models like graph neural networks (GNN) [8, 19–21]. Several alternative techniques have been proposed to address these limitations. Monte Carlo (MC) dropout [22] suggests to enable dropout during inference to introduce stochasticity into predictions at test time, and deep kernel learning combines neural networks with Gaussian processes [11, 23, 24]. While these approaches improve efficiency, many of them require modification of the training pipeline or are incompatible with post-hoc applications, limiting their flexibility in scenarios where models are pre-trained and fixed.

*This work was conducted while the author was an intern at Preferred Networks.
Contact: nontawat@preferred.jp, tsuboi@preferred.jp

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Machine Learning and the Physical Sciences (ML4PS).

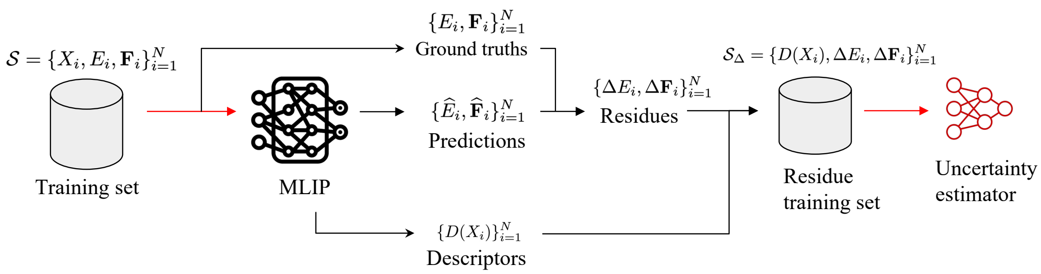

Figure 1: Overview of the P-DRUM. Red arrow indicates "model training using supervised learning".

Model descriptors have been shown to be effective in various downstream applications such as chemical property prediction [25, 26]. Also, they have been combined with prediction errors for UQ, under the assumption that structures with higher prediction errors tend to exhibit greater uncertainty on average. Vita et al. [17] proposed loss trajectory analysis for uncertainty (LTAU), which leverages per-atom force error predictions to train an uncertainty model. While effective, this method requires logging the loss trajectory for every atom and is limited to capturing uncertainty in per-atom force predictions. In Orb-v3 [27], prediction errors are discretized in a manner inspired by pLDDT in Alphafold [28], enabling joint optimization of the UQ objective alongside the prediction objective during model training. In contrast to post-hoc methods, UQ objective of pLDDT is optimized jointly with the model training process. For methods that are post-hoc that only utilizes model descriptor features, Janet et al. [15] proposed to calculate the average feature distance between a test point and its k-nearest neighbors in the training data using revised autocorrelation descriptors. Zhu et al. [16] demonstrated the utility of fitting a Gaussian mixture model (GMM) for UQ in NequIP [29]. To the best of our knowledge, there has been limited study of post-hoc methods that explicitly estimate prediction errors using only the trained model, without requiring detailed training logs or specialized optimization procedures during training.

In this paper, we explore post-hoc descriptor-based residual uncertainty modeling (P-DRUM) that utilize model descriptors of a trained graph neural network potential to estimate prediction error. These residuals act as proxies for prediction error, enabling efficient and scalable UQ without modifying the original model architecture or training process. We investigate multiple variants of P-DRUM for energy and force errors: error-norm learning and deviation learning. Error-norm learning predicts a scalar representing the norm of the error, while deviation learning models the intrinsic error directly, maintaining the same format as the original prediction.

2 Notation

Let $X \in \mathcal{X}$ be a molecular or material structure in the input space $\mathcal{X}$. To learn an MLIP, it is common that we use a training dataset with N molecular or meterial structures: $\mathcal{S} = \{X_i, E_i, \mathbf{F}_i\}_{i=1}^N$. Each structure X_i consists of n_i atoms and is represented as $X_i = \{\mathbf{R}_i, \mathbf{z}_i\}$. Here, $\mathbf{R}_i \in \mathbb{R}^{n_i \times 3}$ represents the atomic position information in three-dimensional space, and $\mathbf{z}_i \in \mathbb{Z}^{n_i}$ encodes the atomic numbers corresponding to each atom within the structure. For example, $z_{i1} = 1$ indicates that the first atom in structure i is hydrogen. X_i's energy label $E_i \in \mathbb{R}$ and force label $\mathbf{F}_i \in \mathbb{R}^{n_i \times 3}$ are also provided for each structure. With this training data, the goal of MLIP training is to accurately learn a real-valued function $f^{\text{energy}} : \mathcal{X} \to \mathbb{R}$, which predicts energy $\widehat{E} \in \mathbb{R}$ given a structure $X = \{\mathbf{R}, \mathbf{z}\}$ of interest. Not only energy, we also expect that the force prediction is correctly predicted, where one can obtain the force information of atom i given a predicted energy E by calculating a negative gradient of E (obtained via f^{energy}) with respect to atomic position. We note that some MLIPs directly predict forces (e.g., ForceNet [30]), but these are beyond the scope of this paper.

Table 1: Dataset statistics used in this paper. HME21 consists of 37 different elements: H, Li, C, N, O, F, Na, Mg, Al, Si, P, S, Cl, K, Ca, Sc, Ti, V, Cr, Mn, Fe, Co, Ni, Cu, Zn, Mo, Ru, Rh, Pd, Ag, In, Sn, Ba, Ir, Pt, Au, and Pb.

	rMD17 [31]			Ni_3Al	HME21 [9]
	uracil	salicylic	malondialdehyde		
Elements	C,H,O,N	C,H,O	C,H,O	Ni, Al	37
Structure size (atom numbers)	12	16	9	32	8–32
Number of training data	800	800	800	480	19956
Number of validation data	200	200	200	120	2498
Number of test data	1000	1000	1000	600	2495

3 Post-hoc descriptor-based residual uncertainty modeling (P-DRUM)

Given structure $X = \{\mathbf{R}, \mathbf{z}\}$, in general, we can extract its descriptor in graph neural networks. In this paper, we focus on a message-passing atomic cluster expansion (MACE) [8] model, where its descriptor (aka. features) can be extracted. For X, we denote a function $D : \mathcal{X} \to \mathbb{R}^{d_{\text{desc}} \times n}$, which maps a structure to a descriptor for each atom, where d_{desc} indicates the dimension of the descriptor. we denote $D_{ij} \in \mathbb{R}^{d_{\text{desc}}}$ a descriptor of atom j in structure i. Given a structure along with its energy and force X, E, $\mathbf{F}$, we can calculate energy residual $\Delta E = E - \widehat{E}$ and force residual $\Delta \mathbf{F} = \mathbf{F} - \widehat{\mathbf{F}}$, where $\widehat{E}$ and $\widehat{\mathbf{F}}$ indicate energy and force predictions of an MLIP. Given N structures and a trained model, we can prepare the training data for P-DRUM: $\mathcal{S}_\Delta = \{D(X_i), \Delta E_i, \Delta \mathbf{F}_i\}_{i=1}^N$. Using the residual training data $\mathcal{S}_\Delta$, we train a supervised model based on a multilayer perceptron (MLP) architecture, which offers significantly faster inference than MLIP models such as MACE. Figure 1 shows the overview of P-DRUM.

3.1 Energy residual learning

A naive approach for residual energy learning might involve designing a single model that processes an entire molecular structure with n atoms and outputs the residual. However, this straightforward approach fails to naturally preserve permutational invariance and lacks the flexibility to handle systems of varying atomic sizes. To address these challenges, we train a multilayered perceptron designed to take as input a descriptor for a single atom and output its corresponding scalar value $r^{\text{s}} : \mathbb{R}^{d_{\text{desc}}} \to \mathbb{R}$. Since the energy depends on the whole structure, but each structure can have different number of atoms, we model the energy residual as the sum of the atom-wise score function. This approach allows the model to operate atom-wise, ensuring invariance properties and scalability to arbitrarily-sized structures. In error norm learning, we can calculate a structure-wise squared loss by $\mathcal{L}_{\text{E-norm}}(X_i) = ((\sum_j^{n_i} r^{\text{s}}_{\text{E-norm}}(D_{ij})) - |\Delta E_i|)^2$. For the energy deviation learning, we learn ΔE directly, i.e., $\mathcal{L}_{\text{E-diff}}(X_i) = ((\sum_j^{n_i} r^{\text{s}}_{\text{E-diff}}(D_{ij})) - \Delta E_i)^2$.

3.2 Force residual learning

Unlike energy, we designed a model that directly predicts force residuals using descriptors of individual atoms. In error norm learning, we learn a function to estimate the Euclidean norm of the force error by minimizing the atomwise loss $\mathcal{L}_{\text{F-norm}}(X_{ij}) = (r^{\text{s}}_{\text{F-norm}}(D_{ij}) - ||\Delta \mathbf{F}_{ij}||)^2$, where $r^{\text{s}}_{\text{F-norm}}$ is a real-valued function similarly to energy residual learning and $|| \cdot ||$ denotes the euclidean norm. In deviation learning, for each atom in a structure, we minimize the average of the coordinate-wise force-error loss: $\mathcal{L}_{\text{F-diff}}(X_{ij}) = \frac{1}{3} \sum (r^{\text{v}}_{\text{F-diff}}(D_{ij}) - \Delta \mathbf{F}_{ij})^2$, where $r^{\text{v}}_{\text{F-diff}} : \mathbb{R}^{d_{\text{desc}}} \to \mathbb{R}^3$ is a vector-valued function. Uncertainty can be calculated by the Euclidean norm of the 3-dimensional output of $r^{\text{v}}_{\text{F-diff}}$. Similarly to energy residual learning, we use a multilayered perceptron to model the force residual function.

4 Experimental results

We compare P-DRUM against ensemble, MC-dropout [22], k-nearest neighbors of descriptors proposed by Janet et al. [15], and GMM of descriptors proposed by Zhu et al. [16]. We used HME21

Table 2: Five-trial average and standard deviation of Spearman correlation between prediction error and uncertainty of in-domain test data. The highest correlation values are highlighted in bold.

Error type	Method	Uracil	Salicylic	Malondi-aldehyde	Ni_3Al	HME21
Energy	**Ensemble**	0.04 (0.04)	0.08 (0.04)	-0.01 (0.07)	0.39 (0.05)	0.27 (0.02)
	MC-dropout [22]	-0.02 (0.02)	-0.02 (0.02)	-0.03 (0.02)	-0.05 (0.05)	0.20 (0.03)
	GMM [16]	0.07 (0.05)	0.07 (0.09)	0.13 (0.08)	0.64 (0.05)	0.06 (0.03)
	kNN [15]	0.06 (0.04)	0.06 (0.07)	0.09 (0.06)	0.64 (0.05)	-0.05 (0.03)
	P-DRUM-norm	0.12 (0.13)	-0.01 (0.04)	-0.09 (0.03)	0.62 (0.07)	**0.30 (0.03)**
	P-DRUM-diff	**0.18 (0.14)**	**0.16 (0.16)**	**0.21 (0.14)**	**0.87 (0.03)**	0.26 (0.02)
Force	**Ensemble**	**0.68 (0.01)**	0.65 (0.01)	0.69 (0.01)	0.97 (0.00)	0.78 (0.00)
	MC-dropout [22]	0.24 (0.03)	0.27 (0.02)	0.27 (0.06)	0.87 (0.01)	0.68 (0.01)
	GMM [16]	0.58 (0.01)	0.67 (0.03)	0.68 (0.02)	0.96 (0.01)	0.64 (0.04)
	kNN [15]	0.52 (0.02)	0.61 (0.04)	0.65 (0.02)	0.96 (0.01)	0.54 (0.01)
	P-DRUM-norm	0.67 (0.02)	**0.71 (0.04)**	**0.69 (0.02)**	**0.98 (0.00)**	**0.92 (0.00)**
	P-DRUM-diff	0.53 (0.02)	0.58 (0.04)	0.57 (0.01)	0.96 (0.01)	0.85 (0.01)

dataset which contains 37 elements [9], and three datasets from rMD17 datasets [31]: malondialdehyde, salicylic acid, uracil. Table 1 shows statistics of all datasets used in this paper. Additionally, we assess out-of-distribution (OOD) robustness with a benchmark on the nickel aluminide (Ni_3Al) dataset, generated using Matlantis [32]. We obtained the MACE model from the official MACE repository [2] and used its implementation of the descriptor. Since MACE does not support MC-dropout by default, we implemented dropout in the fully connected layers after the activation function of both the interaction block and the readout block. Additional training details for our experiments are provided in Appendix B. In terms of computational efficiency, ensemble and MC-dropout require five forward passes of MACE, whereas kNN, GMM, and P-DRUM need only a single pass, with only negligible additional overhead compared to MACE forward pass cost.

4.1 Uncertainty-error correlation evaluation

In this section, we evaluate how uncertainty scores relate to prediction errors using Spearman correlation. This evaluation is based on the idea that structures with higher uncertainty should tend to exhibit larger errors. We analyze both energy and force errors in our experiments. Note that only the kNN and GMM uncertainties do not differentiate between energy and force uncertainties. As a result, we used the same uncertainty estimates for evaluating both energy and force performance for them.

Table 2 reports the 5-trial average Spearman correlation for each method across datasets. The results indicate that P-DRUM-diff performs relatively well in predicting energy uncertainties, whereas P-DRUM-norm is more effective for force uncertainties. Notably, the performance gap between GMM and kNN is larger in HME21. For readers who are interested, we provide an additional analysis using principal component analysis of HME21 in Appendix D.

4.2 OOD detection of Ni_3Al

In this section, we evaluate the performance in the task of out-of-distribution (OOD) detection. We used the Ni_3Al dataset, which was generated using Preferred Potential (PFP) [9] within the Matlantis platform [32], where the initial structure is collected from Materials Project (mp-2593) [3] [33]. To assess OOD detection, we prepared several distinct OOD datasets for Ni_3Al: (1) High-temperature OOD: structures generated via molecular dynamics simulations at temperatures higher than those used in the training dataset. The training data included temperatures of 500K, 1000K, and 1500K, while the OOD data was derived from simulations at 2000K and 3000K. (2) Hexagonal: Ni_3Al with different phase from original dataset (mp-1183232).(3) Cubic: Ni_3Al with different phase from the original dataset (mp-672232).(4) Swap: randomly swap positions of Ni and Al in the structures for 2, 4, and 8 pairs. We used PFP predictions as ground truths and force uncertainty for evaluation.

[2] `https://github.com/ACEsuit/mace`

[3] https://next-gen.materialsproject.org/materials/mp-2593

Table 3: Five-trial average of Spearman correlation and AUC performance in Ni_3Al OOD detection suite for each method. Standard deviation is omitted due to space constraint. **Force uncertainty** is used for ensemble, dropout, and P-DRUM.

Method	**High temp.**		**Hexagonal**		**Cubic**		**Swap**		**All**	
	Corr.	AUC	Corr.	AUC	Corr.	AUC	Corr.	AUC	Corr.	AUC
Ensemble	0.98	**1.00**	0.88	0.94	**0.95**	**1.00**	0.76	**1.00**	**0.90**	0.99
MC-dropout [22]	0.92	**1.00**	0.54	0.63	0.81	0.84	0.62	0.82	0.72	0.82
GMM [16]	0.98	**1.00**	0.74	**1.00**	0.73	**1.00**	0.77	**1.00**	0.81	**1.00**
kNN [15]	0.98	**1.00**	0.80	0.99	0.75	**1.00**	0.76	**1.00**	0.82	**1.00**
P-DRUM-norm	**0.99**	**1.00**	0.82	0.82	0.78	0.82	0.70	0.99	0.82	0.91
P-DRUM-diff	0.97	**1.00**	**0.89**	0.97	0.80	**1.00**	**0.81**	**1.00**	0.87	0.99

Table 3 summarizes the performance of various OOD detection methods. Ensemble, kNN, GMM, and P-DRUM-diff consistently achieve strong OOD detection performance across the benchmark. In comparison, P-DRUM-norm performs less effectively, while MC-dropout yields the lowest performance in terms of both Spearman correlation and AUC. Although P-DRUM-norm achieves the best performance in Table 2, it proves less effective in the OOD setting in our experiments.

5 Discussions

When P-DRUM outperform kNN, GMM? While descriptor-based methods perform competitively, P-DRUM clearly outperforms kNN and GMM in error–uncertainty correlation on the HME21 dataset. We hypothesize that when a dataset contains many elements, descriptor information alone may be difficult to capture the error correlation, and incorporating the prediction error signal can enhance the uncertainty estimation by aligning it more closely with the true prediction error.

Which approach is better: error norm learning or deviation learning? In Table 2, deviation learning outperforms error-norm learning for energy Spearman correlation, whereas the opposite trend is observed for force correlation. This difference likely arises from the nature of the targets: energy deviation is a scalar, where retaining the error sign aids learning, while force deviation is a three-dimensional vector, making direct estimation more challenging. Using the force error norm reduces this complexity, resulting in an improvement in the Spearman correlation. However, as shown in Table 3, using force uncertainty, error-norm learning underperforms deviation learning in OOD detection, indicating the need for further investigation into the advantages and limitations of these approaches. We hypothesize that compressing errors (3-dimensional for forces) into a norm (1-dimensional) might be detrimental for OOD detection.

6 Conclusion and future work

We investigated the effectiveness of post-hoc descriptor-based residual uncertainty modeling (P-DRUM) for machine learning interatomic potentials (MLIPs). P-DRUM achieved a higher Spearman correlation with prediction errors compared to other methods. However, P-DRUM-norm showed inferior out-of-distribution (OOD) detection performance compared to Gaussian mixture model (GMM) and k-nearest neighbor (kNN) descriptor-based approaches, revealing its potential limitations. Future work will explore extending P-DRUM to active learning pipelines and broader MLIP applications to further evaluate its versatility, as well as developing improved training strategies to enhance its performance. In addition, assessing the reliability of reusing the training set to construct the residual training set is an important direction. Using separate dataset splits for training the MLIP and the uncertainty estimation model can reduce bias from the original data but also limits the available training samples. Exploring this trade-off is useful to developing a data-efficient P-DRUM strategy.

Acknowledgments and Disclosure of Funding

We would like to thank Professor Ju Li for helpful comments and suggestions.

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

[4] Oliver T Unke, Stefan Chmiela, Huziel E Sauceda, Michael Gastegger, Igor Poltavsky, Kristof T Schutt, Alexandre Tkatchenko, and Klaus-Robert Muller. Machine learning force fields. *Chemical Reviews*, 121(16):10142–10186, 2021.

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

[6] Volker L Deringer, Miguel A Caro, and Gábor Csányi. Machine learning interatomic potentials as emerging tools for materials science. *Advanced Materials*, 31(46):1902765, 2019.

[7] Lowik Chanussot, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, Morgane Riviere, Kevin Tran, Javier Heras-Domingo, Caleb Ho, Weihua Hu, et al. Open catalyst 2020 (oc20) dataset and community challenges. *Acs Catalysis*, 11(10):6059–6072, 2021.

[8] Ilyes Batatia, David P Kovacs, Gregor Simm, Christoph Ortner, and Gábor Csányi. Mace: Higher order equivariant message passing neural networks for fast and accurate force fields. *NeurIPS*, 35:11423–11436, 2022.

[9] So Takamoto, Chikashi Shinagawa, Daisuke Motoki, Kosuke Nakago, Wenwen Li, Iori Kurata, Taku Watanabe, Yoshihiro Yayama, Hiroki Iriguchi, Yusuke Asano, et al. Towards universal neural network potential for material discovery applicable to arbitrary combination of 45 elements. *Nature Communications*, 13(1):2991, 2022.

[10] 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. *arXiv preprint arXiv:2401.00096*, 2023.

[11] Tom Wollschläger, Nicholas Gao, Bertrand Charpentier, Mohamed Amine Ketata, and Stephan Günnemann. Uncertainty estimation for molecules: Desiderata and methods. In *ICML*, pages 37133–37156. PMLR, 2023.

[12] 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.

[13] Matthias Kellner and Michele Ceriotti. Uncertainty quantification by direct propagation of shallow ensembles. *Machine Learning: Science and Technology*, 5(3):035006, 2024.

[14] Jin Dai, Santosh Adhikari, and Mingjian Wen. Uncertainty quantification and propagation in atomistic machine learning. *Reviews in Chemical Engineering*, 41(4):333–357, 2025.

[15] 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.

[16] Albert Zhu, Simon Batzner, Albert Musaelian, and Boris Kozinsky. Fast uncertainty estimates in deep learning interatomic potentials. *The Journal of Chemical Physics*, 158(16), 2023.

[17] 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.

[18] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *NeurIPS*, 30, 2017.

[19] Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *NeurIPS*, 30, 2017.

[20] Johannes Gasteiger, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs. In *ICLR*, 2020. URL `https://openreview.net/forum?id=B1eWbxStPH`.

[21] Kristof Schütt, Oliver Unke, and Michael Gastegger. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *ICML*, pages 9377–9388. PMLR, 2021.

[22] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, pages 1050–1059. PMLR, 2016.

[23] Kishansingh Rajput, Malachi Schram, and Karthik Somayaji. Uncertainty aware deep learning for particle accelerators. *arXiv preprint arXiv:2309.14502*, 2023.

[24] Jeremiah Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax Weiss, and Balaji Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *NeurIPS*, 33:7498–7512, 2020.

[25] Tomoya Shiota, Kenji Ishihara, and Wataru Mizukami. Universal neural network potentials as descriptors: towards scalable chemical property prediction using quantum and classical computers. *Digital Discovery*, 3(9):1714–1728, 2024.

[26] Zetian Mao, Wenwen Li, and Jethro Tan. Dielectric tensor prediction for inorganic materials using latent information from Preferred Potential. *npj Computational Materials*, 10(1):265, 2024.

[27] 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.

[28] 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.

[29] 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.

[30] Weihua Hu, Muhammed Shuaibi, Abhishek Das, Siddharth Goyal, Anuroop Sriram, Jure Leskovec, Devi Parikh, and C Lawrence Zitnick. Forcenet: A graph neural network for large-scale quantum calculations. *arXiv preprint arXiv:2103.01436*, 2021.

[31] Anders S Christensen and O Anatole Von Lilienfeld. On the role of gradients for machine learning of molecular energies and forces. *Machine Learning: Science and Technology*, 1(4): 045018, 2020.

[32] Matlantis, software as a service style material discovery tool. `https://matlantis.com/`.

[33] A Jain, SP Ong, G Hautier, W Chen, WD Richards, S Dacek, S Cholia, D Gunter, D Skinner, G Ceder, et al. The Materials Project: a materials genome approach to accelerating materials innovation, apl mater. 1 (2013) 011002. *Available from DOI*, 10(1.4812323).

Table 4: Five-trial average and standard deviation of AUC using uncertainty to classify low error and high error class on in-domain test data. The highest correlation values are highlighted in bold.

Error type	Method	Uracil	Salicylic	Malondialdehyde	Ni_3Al	HME21
Energy	**Ensemble**	0.48 (0.02)	0.53 (0.02)	0.50 (0.03)	0.58 (0.03)	0.58 (0.01)
	MC-dropout [22]	-0.48 (0.01)	0.49 (0.02)	0.50 (0.01)	0.49 (0.04)	0.55 (0.02)
	GMM [16]	0.47 (0.06)	0.46 (0.07)	0.55 (0.04)	0.71 (0.04)	0.54 (0.01)
	kNN [15]	0.48 (0.05)	0.47 (0.05)	0.54 (0.03)	0.70 (0.04)	0.51 (0.01)
	P-DRUM-norm	0.51 (0.03)	0.48 (0.01)	0.46 (0.02)	0.72 (0.03)	**0.63 (0.02)**
	P-DRUM-diff	**0.56 (0.07)**	**0.57 (0.06)**	**0.59 (0.06)**	**0.90 (0.04)**	0.61 (0.01)
Force	**Ensemble**	0.83 (0.01)	0.83 (0.01)	0.83 (0.01)	**0.98 (0.00)**	0.93 (0.00)
	MC-dropout [22]	0.62 (0.03)	0.63 (0.02)	0.63 (0.04)	0.95 (0.01)	0.83 (0.01)
	GMM [16]	0.79 (0.00)	0.84 (0.02)	0.82 (0.02)	0.98 (0.01)	0.85 (0.03)
	kNN [15]	0.77 (0.01)	0.81 (0.02)	0.81 (0.02)	0.98 (0.01)	0.84 (0.00)
	P-DRUM-norm	**0.85 (0.01)**	**0.86 (0.03)**	**0.83 (0.01)**	**0.98 (0.00)**	**0.98 (0.00)**
	P-DRUM-diff	0.76 (0.01)	0.78 (0.02)	0.77 (0.02)	0.96 (0.06)	0.92 (0.01)

A Broader Impact

This work focuses on developing uncertainty estimation techniques called post-hoc descriptor-based residual uncertainty modeling (P-DRUM) for machine learning interatomic potentials. P-DRUM is designed to improve the reliability and robustness of simulations in chemistry, materials science, and related fields. Importantly, our research does not involve human subjects, personal data, or any form of unethical experimentation. The techniques we propose are purely computational and are evaluated on standard benchmark datasets and simulated molecular systems.

That said, as with many advances in machine learning and computational modeling, there is the possibility that the methods we develop could be misused. More accurate and reliable atomistic simulations may be applied in contexts that could lead to harmful outcomes, for instance in the design of materials for military applications or environmentally damaging technologies. We strongly discourage the use of our methods in ways that could contribute to unethical purposes.

B Training details and hyperparameters

All models were trained using the default MACE architecture with 32 channels, a radial cutoff of 5 Å, a batch size of 50, and two interaction layers (RealAgnosticInteractionBlock and RealAgnosticInteractionResidualBlock), yielding 64 descriptor dimensions. Training ran for 100 epochs: for the first 75 epochs, energy and force loss weights were 1 and 100, respectively; for the final 25 epochs, the energy weight was increased to 1000, with the force weight unchanged.

During the P-DRUM experiments, validation set was used for learning rate scheduling and early stopping. Training began with a learning rate of 10^{-3} and patience of 10 epochs, halving the learning rate whenever the validation error failed to improve for 10 consecutive epochs. Training was terminated after a maximum of 1000 epochs or once the learning rate decreased to 10^{-7}.

Except for Ni_3Al and HME21 dataset, in which we used a batch size of 2048 for P-DRUM-diff forces prediction, the batch size was set to 64 atoms in all the other P-DRUM forces training. For P-DRUM energy training, the batch size was set to 64 chemical structures. The larger dataset sizes in these cases lead to a greater number of total atoms per batch, and increasing the batch size helped stabilize training. For P-DRUM-norm energy and forces prediction, we employed the ReLU activation function with a single hidden layer, along with a softplus activation right before output. P-DRUM-diff uses similar but softplus-removed MLP architecture, where one hidden layer was used for energy and two hidden layers were used for forces. For MC-dropout, we set the dropout ratio to 10%.

For computing resources, we used an NVIDIA V100 GPU (32 GB) of memory for training different trials. The execution time depends on the dataset. Although we did not precisely measure the training time, each trial of each method can be completed within 7 hours on a single GPU, including training a MACE model and uncertainty estimation method.

Table 5: Five-trial average and standard error of Spearman correlation performance in Ni_3Al OOD detection suite for each method. **Force uncertainty** is used for ensemble, dropout, and P-DRUM.

Method	High temp.	Hexagonal	Cubic	Swap	Average
Ensemble	0.98 (0.00)	0.88 (0.01)	**0.95 (0.01)**	0.76 (0.01)	**0.90**
MC-dropout [22]	0.92 (0.01)	0.54 (0.10)	0.81 (0.03)	0.62 (0.06)	0.72
GMM [16]	0.98 (0.00)	0.74 (0.05)	0.73(0.02)	0.77 (0.04)	0.81
kNN [15]	0.98 (0.01)	0.80 (0.03)	0.75 (0.03)	0.76 (0.07)	0.82
P-DRUM-norm	**0.99 (0.00)**	0.82 (0.10)	0.78 (0.26)	0.70 (0.06)	0.82
P-DRUM-diff	0.97 (0.00)	**0.89 (0.04)**	0.80 (0.08)	**0.81 (0.03)**	0.87

Table 6: Five-trial average and standard error of AUC performance in Ni_3Al OOD detection suite for each method. **Force uncertainty** is used for ensemble, dropout, and P-DRUM.

Method	High temp.	Hexagonal	Cubic	Swap	Average
Ensemble	**1.00 (0.00)**	0.94 (0.02)	**1.00 (0.00)**	**1.00 (0.00)**	0.99
MC-dropout [22]	**1.00 (0.00)**	0.63 (0.04)	0.84 (0.04)	0.82 (0.08)	0.82
GMM [16]	**1.00 (0.00)**	**1.00 (0.00)**	**1.00 (0.00)**	**1.00 (0.00)**	**1.00**
kNN [15]	**1.00 (0.00)**	0.99 (0.01)	**1.00 (0.00)**	**1.00 (0.00)**	**1.00**
P-DRUM-norm	**1.00 (0.00)**	0.82 (0.09)	0.82 (0.25)	0.99 (0.02)	0.91
P-DRUM-diff	**1.00 (0.00)**	0.97 (0.04)	**1.00 (0.00)**	**1.00 (0.00)**	0.99

C Additional experimental results

C.1 AUC evaluation of in-domain dataset

Here, we show the results of AUC where we split test data into two classes: low error and high error classes. We put lowest 20% error as low error class and high error otherwise. Table 4 shows the comparison of the AUC performance across all different methods.

C.2 OOD detection results with standard deviation

Table 5 shows the spearman correlation comparisons and Table 6 shows the AUC comparisons.

D When does P-DRUM outperform kNN and GMM?: an analysis based on principle component analysis (PCA)

In the in-domain setup, we observed that the P-DRUM-norm method performed the best across all methods; however, the baseline descriptor methods (kNN and GMM) also performed comparably on almost all datasets except HME21. Taking the Ni_3Al dataset as an example (Figure 2), the descriptors in PC space for each atom in the train and test set were plotted, and each cluster in the plot represents the atoms of Ni or Al. The lower force error atoms in the figure represent atoms with lower force prediction error or low uncertainty. In the training set distribution, denser regions on the PCA plot correspond to lower force errors, i.e., smaller error magnitudes. This indicates that descriptor-based baseline methods such as GMM and kNN can achieve good performance without requiring prior knowledge of the force errors.

However, comparing to rMD17 elements and Ni_3Al that contain at most 4 types of elements in each dataset, the HME21 dataset contains 37 different elements and a more diverse interaction between different elements. The landscape of prediction error is more complicated and difficult to learn without having the error information of the training set. Figure 3 shows the oxygen atoms, which is the most common atom in HME21 descriptors in PC space. We observed that in the train set, the circled area is the densest while having slightly higher force prediction error than the area on the right of the circle. kNN and GMM was unable to capture this and predicted the circled area as lowest uncertainty, while P-DRUM-norm method learned this from the forces prediction error during the training step.

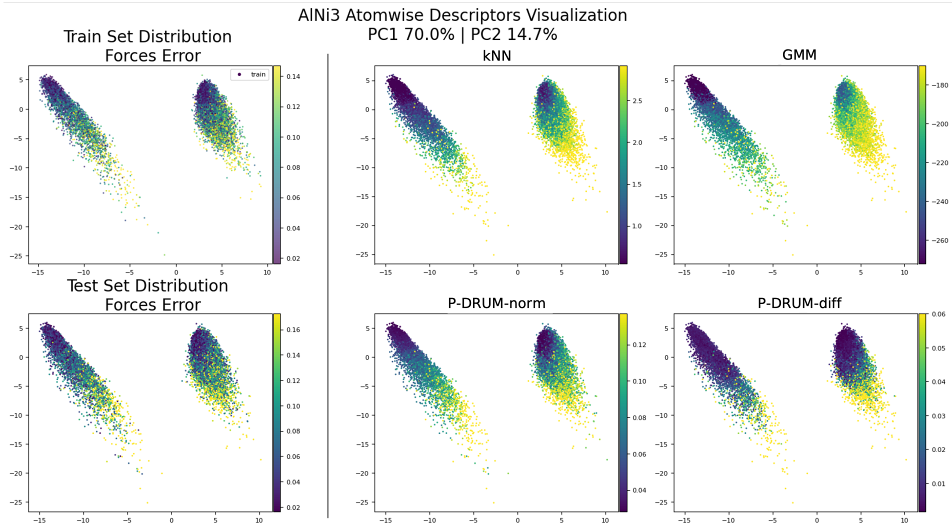

Figure 2: PCA visualization of the Ni_3Al dataset. The left subplots show the prediction error of train and test set in PC space, while the uncertainty metrics of the test set on the right subplots.

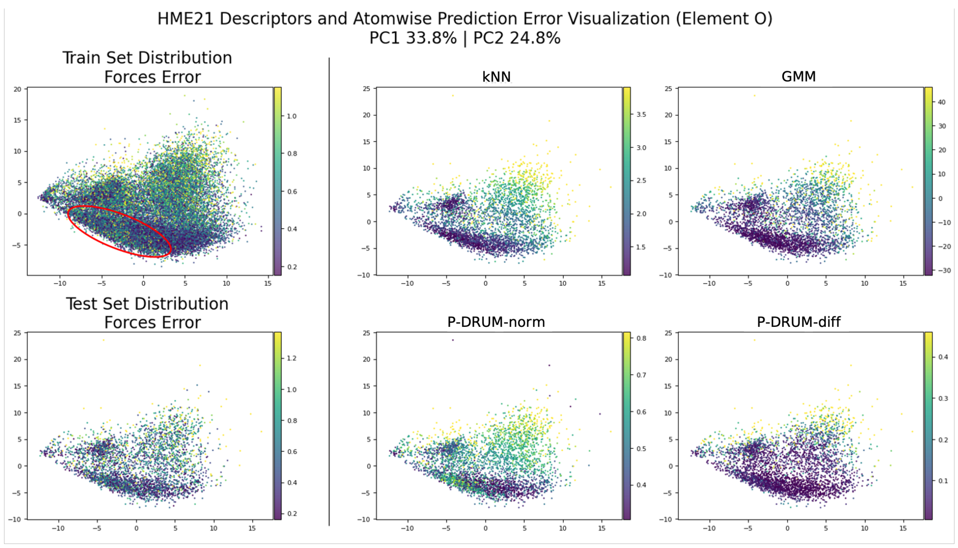

Figure 3: PCA visualization of the oxygen atoms in HME21 dataset. The left subplots show the prediction error of train and test set in PC space, while the uncertainty metrics of the test set on the right subplots.

Similar trends are observed in Figure 4, which only shows the calcium atoms in HME21. The top-left region of the PCA plot is densely populated, yet exhibits relatively high force-prediction errors. Both of our P-DRUM methods successfully capture this behavior, as the prediction error was explicitly incorporated into the training process. In contrast, the kNN and GMM approaches rely solely on descriptor information and therefore incorrectly assign low uncertainty to the same tail region in the top left. These results suggest that P-DRUM methods have strong potential for uncertainty prediction in more diverse and complex datasets, paving the way toward the development of universal potentials uncertainty estimation.

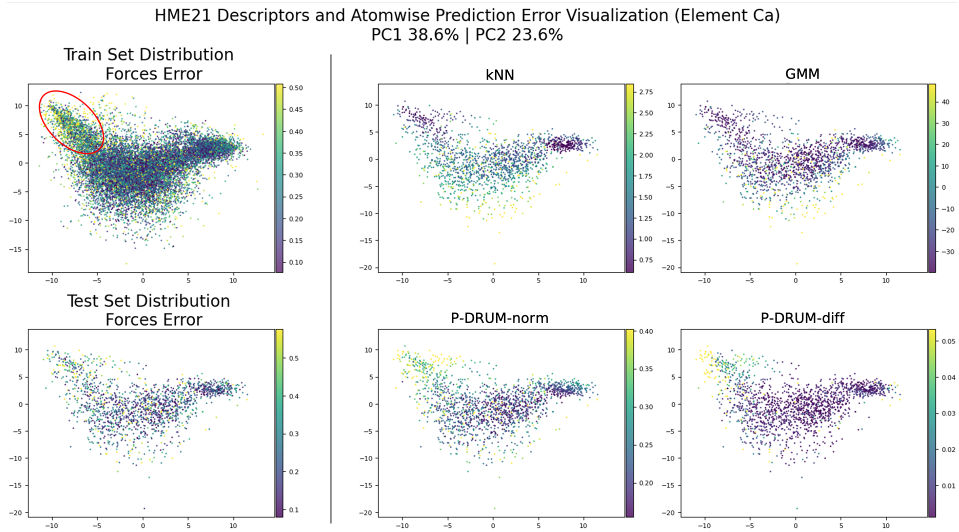

Figure 4: PCA visualization of the calcium atoms in HME21 dataset. The left subplots show the prediction error of train and test set in PC space, while the uncertainty metrics of the test set on the right subplots.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We propose P-DRUM and evaluate its performance to show the capability of our method.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discussed limitations of our method in the conclusion and future work in the Section 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.

- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: [NA]

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We share the training information and disclose training hyperparameters in Appendix B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often

one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We do not plan to release the code for the workshop submission version.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We listed hyperparameters in Appendix B and describe training dataset in the main body. However, we also used Ni_3Al dataset which is not open source but we believe we provided enough information for readers to understand the results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We conducted experiments for 5 trials to produce mean and standard deviation of the results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We described this information in Appendix B.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: We have read the code of ethics and believe our research conforms to this NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We wrote a broader impact section in Appendix A.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the source of datasets and code of the MACE repository in the paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provided information of how Ni_3Al dataset was generated in the experiment section.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.