Diversity-driven training of machine-learned force fields

Connor Ganley
Kevin T. Crofton Department of Aerospace and Ocean Engineering
Virginia Tech
Blacksburg, VA, USA
cganley2@vt.edu

Maitreyee Sharma Priyadarshini
Kevin T. Crofton Department of Aerospace and Ocean Engineering
Virginia Tech
Blacksburg, VA, USA
msharmap@vt.edu

Abstract

We analyze the importance of training dataset diversity when training machine-learned force fields (MLFFs) with the goal of accelerating their development for AI-driven materials design frameworks. We specifically focus on ceramic systems (3C-SiC) relevant to thermal protection applications in hypersonic flight. We use the MACE model to represent our MLFFs. Each MACE model is trained on datasets sampled from *ab initio* molecular dynamics (AIMD) trajectories simulated at multiple temperatures. By diversity-driven sampling of different training datasets, we investigate the role of training set diversity in constructing an accurate force field with reduced data requirements. The material's structural environment is encoded using the many-body tensor representation (MBTR), and similarity between configurations is quantified via a radial basis function (RBF) kernel and the Vendi score. Our results reveal that greater diversity in the sampled datasets yields more accurate force predictions even with smaller dataset size. These findings underscore the importance of systematically quantifying dataset diversity for efficient MLFF training and highlight a pathway for scalable force field development for automated materials design workflows.

1 Introduction

One major challenge in the design of complex advanced materials is the lack of discovery tools that can efficiently navigate the vast composition space. A traditional Edisonian experimental search for optimal composition relies on expert knowledge. On the other hand, molecular simulation approaches such as density functional theory (DFT) and AIMD can offer insights into microscopic behaviors, but are also prohibitively expensive for an exhaustive search of the large combinatorial search spaces. For many materials applications such as catalysis, solar energy conversion, thermoelectrics, thermal protection systems etc., it is also important to study the dynamical behavior of candidate materials in their operating conditions which requires long-time molecular dynamics (MD) simulations. As such, there remains a need to design methods and workflows that can accelerate materials design by efficient exploration of vast compositional spaces and efficient simulations of the material's dynamical behavior. Here, we explore the application of diversity-driven training dataset curation for efficient training of generalizable MLFFs.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

As a test case, we choose to train a force field for Silicon Carbide (3C-SiC), a well known ultra-high temperature ceramic (UHTC) used in hypersonic vehicles for thermal protection systems. UHTCs are becoming increasingly popular choices for reusable thermal protection system (TPS) materials due to their ability to form protective oxide layers that significantly reduce oxidation rates under extreme thermal loads (Wyatt et al. (2024); Konnik et al. (2025)). However, a major challenge in designing ceramics for TPS applications is studying their degradation in the extreme hypersonic environment. This includes studying the formation and growth of surface oxide scales which form protective layers to prevent further oxidation and degradation of the material. An important computational tool to study these processes on the atomistic scale is molecular dynamics which requires construction of force fields.

Recently, MLFFs have been shown to achieve high accuracy for a number of molecular and solid-state systems (Chmiela et al. (2017); Smith et al. (2019); Zhang et al. (2018b,a); Batatia et al. (2023)). These force fields are revolutionizing the path towards running longer times in MD at an accuracy comparable to first principles calculations such as DFT. However, force field development typically requires a time-consuming fitting process, often involving manual or random selection of thousands of reference configurations from first-principles datasets. In this paper, we aim to develop a methodology to intelligently sample reference configurations for training highly accurate force fields. We achieve this by maximizing a chosen diversity metric in the training dataset.

Several metrics have been proposed to quantify and enforce diversity during data selection for efficient training. For example, kernel-based similarity measures such as SOAP kernel distances (Bartók et al. (2013)) or RBF kernels can capture structural similarity between atomic environments, while descriptors such as MBTR enable comparison of complex configurations in an invariant feature space. Entropy-based approaches, such as the Vendi score (Friedman & Dieng (2023)), provide information-theoretic estimates of the "effective number" of distinct configurations in a dataset. Vendi scores have been applied previously to accelerate molecular simulations (Pasarkar et al. (2023)), polymer design (Jiang et al. (2024)) and MOF design (Liu et al. (2024)). In this work, we develop a diversity-driven data curation methodology to accelerate MLFF development with smaller yet more informative datasets using the Vendi score (Friedman & Dieng (2023)).

2 Methods

2.1 Dataset Generation

The AIMD data was generated with Quantum ESPRESSO (QE) 7.4 using the PBE exchange-correlation functional (Giannozzi et al. (2017)). The kinetic energy and charge density cutoffs were set to 50 and 200 Ry, as recommended by the pseudopotential files used, which were sourced from the Standard Solid State Pseudopotentials (SSSP) library, and confirmed by separate convergence tests. Convergence tests on *k*-point sampling revealed that 8 irreducible *k*-points in the primitive cell and 10 in the supercell were sufficient to describe energy to within 10 meV (Prandini et al. (2023)). Each AIMD trajectory consisted of a $3 \times 3 \times 2$ supercell constructed from a fully relaxed primitive cell downloaded from Materials Project. Here, "full relaxation" refers to the *vc-relax* routine within QE, which allows the lattice parameters and the atomic positions to relax and is distinct from the *relax* routine, which holds lattice parameters constant. Once assembled, the 36-atom supercell was allowed to fully relax once again before dynamics began. We simulated AIMD trajectories at three temperatures - 500, $1\,500$ and $2\,500$ K. For each temperature, 5 picoseconds of AIMD data was generated within the NVT ensemble using 1 femtosecond time steps.

Associated with each AIMD frame is the overall system energy, stress on the cell, and a set of atomic positions and forces. This information is embedded in Atomic Simulation Environment (ASE) Atoms objects, which serve as individual data points for model training. However, because Cartesian coordinate data is not invariant to basic symmetry operations, we employ the MBTR descriptor to transform local atomic environments using kernel density estimation over a geometry function's distribution to represent structure (Laakso et al. (2023); Barnard et al. (2023)). The MBTR descriptors for configurations serve as inputs to our sampling methods discussed in the following section.

2.2 Diversity-based Sampling

From the generated dataset, our goal is to sample diverse configurations that can be used to train the MLFF. We use the Vendi score as the metric to quantify the training dataset diversity. The Vendi score is defined as the exponential of the Shannon entropy of the eigenvalues of a similarity matrix (see Equation 1), and it can be used to quantify the "effective number of data points" in a set (Friedman & Dieng (2023)). Here, the entries to the similarity matrix represent pairwise similarity in the training dataset configurations.

$$VS_k(x_1, \ldots, x_n) = \exp\left(-\sum_{i=1}^{n} \lambda_i \log \lambda_i\right). \tag{1}$$

We used the RBF kernel to generate the pairwise similarity matrix whose eigenvalues are λ_i in Equation 1. The RBF kernel is defined in Equation 2.

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{|\mathbf{x} - \mathbf{x}'|^2}{2\sigma^2}\right). \tag{2}$$

In the RBF kernel, $\mathbf{x}$ and $\mathbf{x}'$ represent two configurations and σ represents the kernel width, or the smoothness of the similarity relation. It is related to another parameter $\gamma = \frac{1}{2\sigma^2}$; the latter is used as the input to the Python package used for computation. σ was set to 0.0001 because it allowed for adequate resolution of the difference between AIMD frames 1 femtosecond apart, where we typically expect atomic environment changes to be small.

A sensitivity analysis for γ was conducted for trajectories generated at $2\,500$ K since they have the most variation in energy values. These results are shown in Appendix A. For lower γ values, the distribution of similarity values is concentrated toward 1.0, meaning that a large portion of the frames in the trajectory are deemed "similar." At large γ, the distribution shifts toward 0.0, indicating the frames are deemed more different from one another. To reflect a balance of similarity and difference, we selected an intermediate value of $\gamma = 1 \times 10^6$, or $\sigma = 0.0001$. This same value was used to calculate the similarity matrices for the other trajectories. This analysis yielded Vendi scores of 2.6, 12.7, and 53.9 for the 500 K, $1\,500$ K, and $2\,500$ K trajectories, respectively. This is made more intuitive by the velocity autocorrelation (VAC) plots shown in Appendix B. The VAC plots show that the variations in the $2\,500$ K trajectory are the largest, and therefore we expect trajectories from $2\,500$ K to have the largest number of "effective" samples when compared to the trajectories from the other two temperatures.

2.3 MLFF Training

The MLFF studied in this work is MACE (Batatia et al. (2023)). This model is one of the models available in the *mlip* Python package published by Instadeep AI (Brunken et al. (2025)). The material of interest is (cubic) 3C-SiC (*mp-8062*), the SiC polymorph with the smallest bandgap, which has been alloyed with other ceramics to achieve ultra-high heat resistance (Jain et al. (2013); Sarikov et al. (2019); Zimmermann et al. (2008)).

The objective of a MLFF is to learn the relationship between an atom's local environment and the forces on that atom, as well as the atom's contribution to the overall system energy. Thus, the training data was comprised of bulk SiC AIMD at three different temperatures: 500, $1\,500$, and $2\,500$ K. A wider range of temperatures exposes the model to a greater variety of atomic configurations which improves its ability to accurately predict forces and energies across different thermodynamic states. Due to atomic vibrations, AIMD trajectories are highly correlated, and there is a characteristic vibrational period that each atom in the system experiences. Plots of the velocity autocorrelation function for each AIMD trajectory are shown in Appendix B. The VAC plots show that there is enough sampling of the trajectories to account for the different portions of each autocorrelation period.

For each 5000-frame trajectory, whose atomic environments vary with time, we calculated a RBF-based similarity matrix with $\sigma = 0.0001$ on all possible pairs, ranked the values, and identified the indices associated with the lowest similarity scores. The unique entries in this ranked assortment corresponded to the order in which frames were added to the training set. The present work refers to this as "RBF-based similarity sampling," and it is contrasted with purely random sampling of frames

from each trajectory. We explored 3 separate training sets, where 300, 400, and 500 frames were sampled from each AIMD trajectory, corresponding to MACE models trained on 900, 1200, and 1500 AIMD frames. For each model, a random set of 1200 frames was used for validation, and the remainder were all added to the test set. The training points sampled using the two methodologies mentioned above are shown on the Energy versus time plots in Appendix C. These training points are used to train the MACE model, where each model is trained with a graph cutoff of 5 Å, 128 channels, and a maximum correlation order of 2. The plots in Appendix C show that the RBF-based sampling is heavily biased towards sampling configurations from the initial part of the trajectory when the system is not yet equilibrated. This has consequences on the accuracy of the MACE model trained using the RBF-based training samples as discussed in the following results section.

3 Results and Discussion

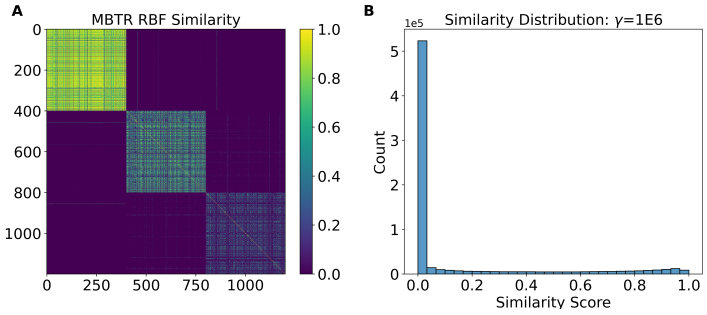

Figure 1: (A) Heatmap of the similarity matrix calculated using the 1200 MBTR representations of atomic configurations used to train the MACE model. The similarity is computed using the RBF kernel with a kernel width of $\sigma = 0.0001$. (B) Histogram of similarity scores shown in (A). There is a large concentration at 0.0 due to the dissimilarity between the three AIMD trajectories.

Figure 1 shows the heatmap and histogram of similarity values for the training set composed of 1200 total frames (400 from each AIMD trajectory). The regularity of the heatmap in (A) is expected given our prior physical knowledge about the system. In the lowest temperature trajectory, the frames are more similar to one another than in the highest temperature. Moreover, the frames in the initial equilibration period of each AIMD trajectory ($\sim$1000 frames) are less similar to one another than the equilibrated frames are to one another. The histogram in (B) is heavily concentrated at 0.0, but this is an effect of combining the three training distributions into one training set. There is inherently minimal similarity between points in different data sets.

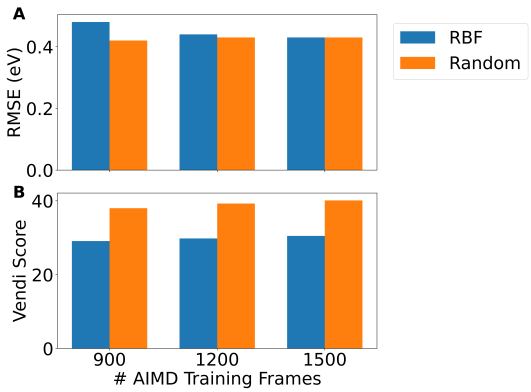

Figure 2: (A) Comparison of Energy RMSE for MACE models trained on 900, 1200, and 1500 AIMD frames using two different training data selection methods: RBF and Random. (B) Vendi scores of the training sets used in MACE model training.

Figure 2A shows the energy RMSE for the MACE models trained and the number of frames used in training according to two different training data sampling methods: random sampling with high dataset Vendi score and RBF-based similarity sampling. The datasets with higher diversity are observed to have outperformed training datasets with lower diversity scores for two of the three cases studied. The models show a trend of "minimizing" RMSE with increasing training set size as expected. However, RMSE values plateau and this may be indicative of the maximum amount of information that can be gleaned from bulk AIMD. We note that for two of the models, the dataset with greater diversity (a higher Vendi score) has a lower RMSE. To better understand this correlation, we evaluated each training set's Vendi score according to Equation 1 and plotted the results in Figure 2B. The lower RMSE values in the models trained with random sampling are likely due to the greater diversity in their training data, as demonstrated in their Vendi scores. Moreover, the RMSE accuracy of the models approaches a constant value as the overlap between the frames sampled by each method increases. This is a consequence of a finite training set.

The energy parity plot for the model with the lowest energy RMSE is shown in Figure 3. Its training losses by epoch are shown in Appendix D. It exhibits an energy recovery of 12 meV/atom and a R^2 value greater than 0.98 on both training and test sets. This energy recovery is adequate for describing simple bulk dynamics.

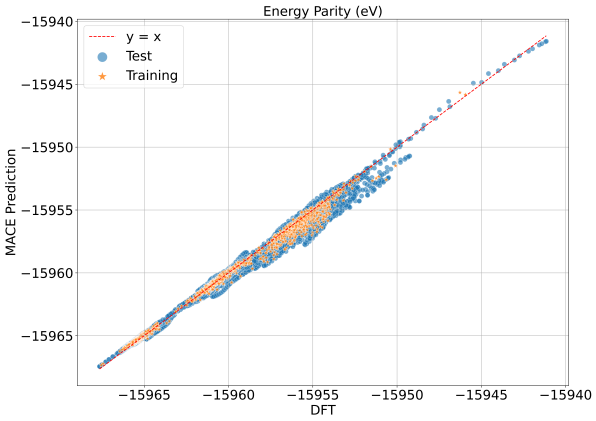

Figure 3: Energy parity plot of the MACE model trained on 1200 random frames (400 per training trajectory) and evaluated on the remainder from each trajectory.

4 Conclusion

This work demonstrates that diversity-aware data selection can accelerate the construction of machine-learned force fields in materials design frameworks. By leveraging the Vendi score to quantify configuration space diversity, we identified cases where sampling strategies that increase dataset diversity improve the force field accuracy. These insights have direct implications for the development of force fields for complex systems such as ultra-high temperature ceramics or multi-principal element alloys, where the combinatorial design space is vast and generating high-fidelity reference data is costly. Although the computational bottleneck for MLFFs is primarily in the acquisition of *ab initio* training data, knowledge of training set diversity can help intelligently acquire further training data. Moreover, training a MLFF on a maximally-informative, diverse dataset enhances the efficiency of transfer learning processes by identifying a maximally-potent training set for model performance in certain conditions. A model would then have more "room" in its training set capacity (before it became too large to be feasible) to transfer learn the behavior of different atomic environments. Future work will extend this framework to integrate active-learning strategies with uncertainty quantification, and explore connections between data diversity, underlying lattice dynamics, and transferability across thermodynamic conditions. Ultimately, we show that diversity-driven training offers a promising route toward building efficient and generalizable MLFFs that can accelerate the discovery and design of advanced materials for extreme environments.

Acknowledgments

The work is supported by start-up funds at Virginia Tech. We also acknowledge the support of VT Advanced Research Computing resources that were used to run the calculations presented in this work.

References

Trent Barnard, Steven Tseng, James P. Darby, Albert P. Bartók, Anders Broo, and Gabriele C. Sosso. Leveraging genetic algorithms to maximise the predictive capabilities of the SOAP descriptor. *Molecular Systems Design & Engineering*, 8(3):300–315, 2023. doi: 10.1039/D2ME00149G. URL https://pubs.rsc.org/en/content/articlelanding/2023/me/d2me00149g. Publisher: Royal Society of Chemistry.

Albert P Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Physical Review B—Condensed Matter and Materials Physics*, 87(18):184115, 2013.

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, January 2023. URL http://arxiv.org/abs/2206.07697. arXiv:2206.07697 [stat].

Martin Brehm and Barbara Kirchner. TRAVIS - A Free Analyzer and Visualizer for Monte Carlo and Molecular Dynamics Trajectories. *Journal of Chemical Information and Modeling*, 51(8): 2007–2023, August 2011. ISSN 1549-9596. doi: 10.1021/ci200217w. URL https://doi.org/10.1021/ci200217w. Publisher: American Chemical Society.

Christoph Brunken, Olivier Peltre, Heloise Chomet, Lucien Walewski, Manus McAuliffe, Valentin Heyraud, Solal Attias, Martin Maarand, Yessine Khanfir, Edan Toledo, Fabio Falcioni, Marie Bluntzer, Silvia Acosta-Gutiérrez, and Jules Tilly. Machine Learning Interatomic Potentials: library for efficient training, model development and simulation of molecular systems, August 2025. URL http://arxiv.org/abs/2505.22397. arXiv:2505.22397 [physics].

Stefan Chmiela, Alexandre Tkatchenko, Huziel E Sauceda, Igor Poltavsky, Kristof T Schütt, and Klaus-Robert Müller. Machine learning of accurate energy-conserving molecular force fields. *Science advances*, 3(5):e1603015, 2017.

Dan Friedman and Adji Bousso Dieng. The Vendi Score: A Diversity Evaluation Metric for Machine Learning, July 2023. URL http://arxiv.org/abs/2210.02410. arXiv:2210.02410 [cs].

P. Giannozzi, O. Andreussi, T. Brumme, O. Bunau, M. Buongiorno Nardelli, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, M. Cococcioni, N. Colonna, I. Carnimeo, A. Dal Corso, S. De Gironcoli, P. Delugas, R. A. Distasio, A. Ferretti, A. Floris, G. Fratesi, G. Fugallo, R. Gebauer, U. Gerstmann, F. Giustino, T. Gorni, J. Jia, M. Kawamura, H. Y. Ko, A. Kokalj, E. Kücükbenli, M. Lazzeri, M. Marsili, N. Marzari, F. Mauri, N. L. Nguyen, H. V. Nguyen, A. Otero-De-La-Roza, L. Paulatto, S. Poncé, D. Rocca, R. Sabatini, B. Santra, M. Schlipf, A. P. Seitsonen, A. Smogunov, I. Timrov, T. Thonhauser, P. Umari, N. Vast, X. Wu, and S. Baroni. Advanced capabilities for materials modelling with Quantum ESPRESSO. *Journal of Physics: Condensed Matter*, 29(46): 465901, October 2017. ISSN 0953-8984. doi: 10.1088/1361-648X/AA8F79. URL https://iopscience.iop.org/article/10.1088/1361-648X/aa8f79. arXiv: 1709.10010 Publisher: IOP Publishing.

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, July 2013. ISSN 2166-532X. doi: 10.1063/1.4812323. URL https://doi.org/10.1063/1.4812323.

Shengli Jiang, Adji Bousso Dieng, and Michael A Webb. Property-guided generation of complex polymer topologies using variational autoencoders. *npj Computational Materials*, 10(1): 139, 2024.

Matthew T. Konnik, Lorenzo Capponi, Trey Oldham, Francesco Panerai, and Kelly A. Stephani. Environmental response characteristics of substoichiometric ZrCx exposed to inductively coupled air plasma. *Journal of the European Ceramic Society*, pp. 117491, May 2025. ISSN 0955-2219. doi: 10.1016/j.jeurceramsoc.2025.117491. URL https://www.sciencedirect.com/science/article/pii/S0955221925003115.

Jarno Laakso, Lauri Himanen, Henrietta Homm, Eiaki V. Morooka, Marc O. J. Jäger, Milica Todorović, and Patrick Rinke. Updates to the DScribe library: New descriptors and derivatives. *The Journal of Chemical Physics*, 158(23):234802, June 2023. ISSN 0021-9606. doi: 10.1063/5.0151031. URL https://doi.org/10.1063/5.0151031.

Tsung-Wei Liu, Quan Nguyen, Adji Bousso Dieng, and Diego A Gómez-Gualdrón. Diversity-driven, efficient exploration of a mof design space to optimize mof properties. *Chemical Science*, 15(45):18903–18919, 2024.

Amey P Pasarkar, Gianluca M Bencomo, Simon Olsson, and Adji Bousso Dieng. Vendi sampling for molecular simulations: Diversity as a force for faster convergence and better exploration. *The Journal of chemical physics*, 159(14), 2023.

Gianluca Prandini, Antimo Marrazzo, Ivano E. Castelli, Nicolas Mounet, Elsa Passaro, Yu, and Nicola Marzari. A Standard Solid State Pseudopotentials (SSSP) library optimized for precision and efficiency. 2023. doi: 10.24435/materialscloud:f3-ym. URL https://archive.materialscloud.org/records/rcyfm-68h65.

Andrey Sarikov, Anna Marzegalli, Luca Barbisan, Emilio Scalise, Francesco Montalenti, and Leo Miglio. Molecular dynamics simulations of extended defects and their evolution in 3C–SiC by different potentials. *Modelling and Simulation in Materials Science and Engineering*, 28(1): 015002, November 2019. ISSN 0965-0393. doi: 10.1088/1361-651X/ab50c7. URL https://dx.doi.org/10.1088/1361-651X/ab50c7. Publisher: IOP Publishing.

Justin S Smith, Benjamin T Nebgen, Roman Zubatyuk, Nicholas Lubbers, Christian Devereux, Kipton Barros, Sergei Tretiak, Olexandr Isayev, and Adrian E Roitberg. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nature communications*, 10(1):2903, 2019.

Brian C. Wyatt, Srinivasa Kartik Nemani, Gregory E. Hilmas, Elizabeth J. Opila, and Babak Anasori. Ultra-high temperature ceramics for extreme environments. *Nature Reviews Materials*, 9(11): 773–789, November 2024. ISSN 2058-8437. doi: 10.1038/s41578-023-00619-0. URL https://www.nature.com/articles/s41578-023-00619-0. Publisher: Nature Publishing Group.

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, 2018a.

Linfeng Zhang, Jiequn Han, Han Wang, Wissam Saidi, Roberto Car, et al. End-to-end symmetry preserving inter-atomic potential energy model for finite and extended systems. *Advances in neural information processing systems*, 31, 2018b.

James W. Zimmermann, Gregory E. Hilmas, William G. Fahrenholtz, Ralph B. Dinwiddie, Wallace D. Porter, and Hsin Wang. Thermophysical Properties of ZrB2 and ZrB2–SiC Ceramics. *Journal of the American Ceramic Society*, 91(5):1405–1411, 2008. ISSN 1551-2916. doi: 10.1111/j.1551-2916.2008.02268.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1551-2916.2008.02268.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1551-2916.2008.02268.x.

A Sensitivity Analysis: γ in RBF kernel

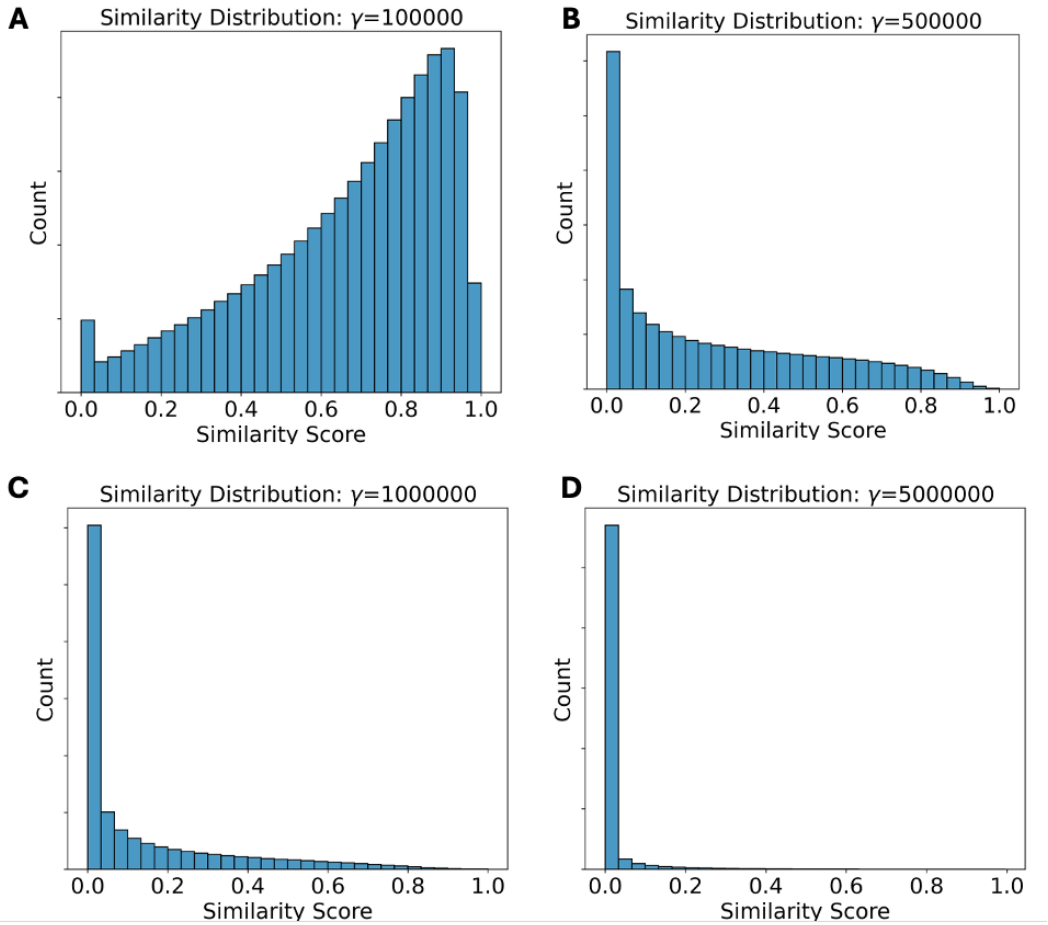

Figure 4: Sensitivity analysis of different values of γ in the RBF kernel on the 2 500 K trajectory.

At low values of γ, the kernel width is larger such that more frames are identified as "similar" to one another. At high values of gamma, the kernel width shrinks such that more frames are recognized as different from one another, shifting the similarity distribution toward 0.

B Velocity Autocorrelation: bulk SiC AIMD trajectories

The velocity autocorrelation functions of the three training trajectories exhibit some periodicity after an equilibration period. This periodicity is longer for the higher-energy trajectory (2 500 K) because of the amount of energy present in the system. This greater variation in atomic configurations and longer periodicity is related to the higher Vendi score for 2 500 K trajectory when compared to the other two trajectories at lower temperatures.

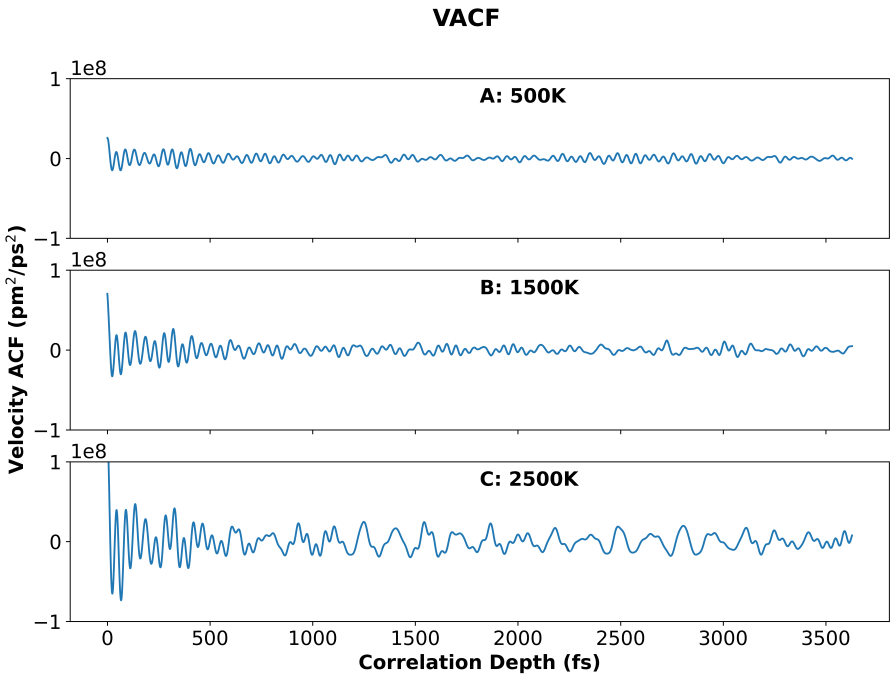

Figure 5: Velocity correlation timeseries for each AIMD trajectory, as analyzed by TRAVIS (Brehm & Kirchner (2011)).

C Energy Distributions of AIMD Trajectories

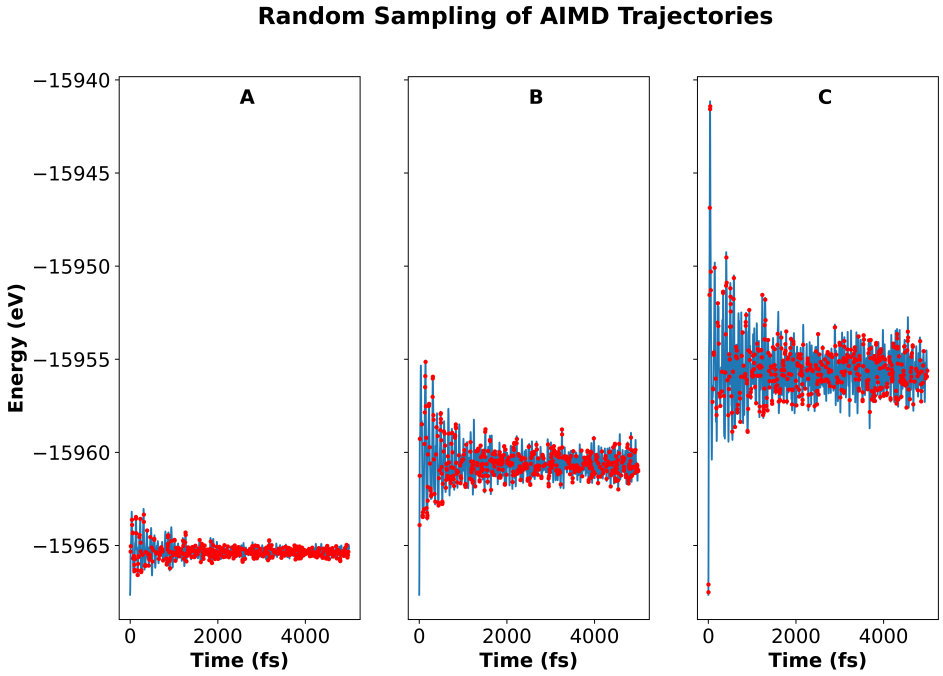

Figure 6: Energy timeseries of each 5-picosecond AIMD trajectory (A: 500K, B: 1500K, C: 2500K). Red dots indicate frames chosen randomly for model training.

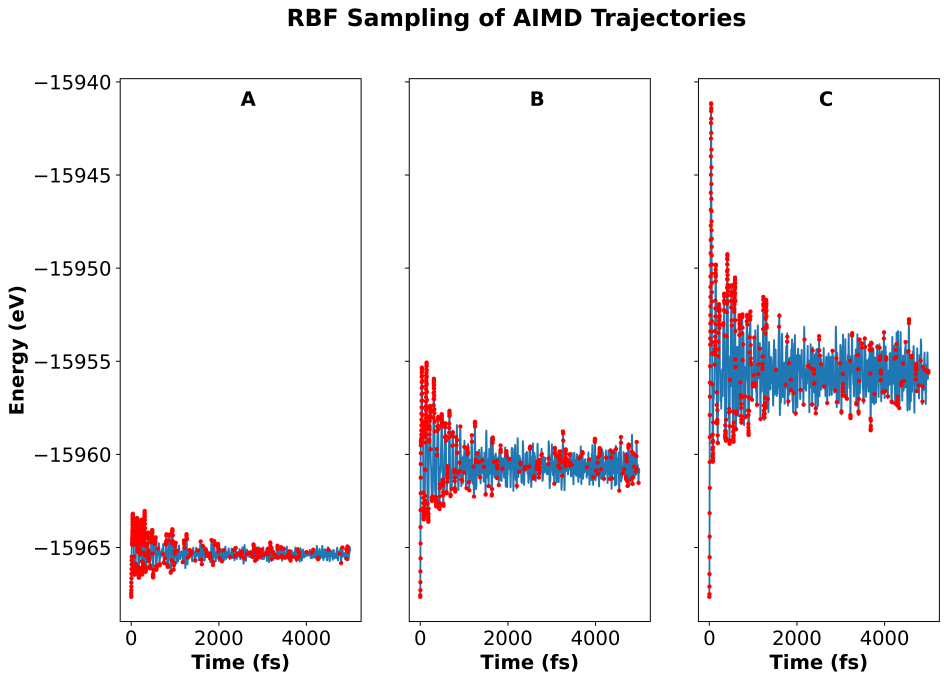

Figure 7: Energy timeseries of each 5-picosecond AIMD trajectory (A: 500K, B: 1500K, C: 2500K). Red dots indicate frames chosen with RBF-based sampling for model training.

The AIMD trajectories experience an equilibration period that lasts roughly 1 picosecond, during which the system energy fluctuates more than it will at equilibrium. The red dots indicate frames that were used in model training based on their method of selection.

D Training Loss by Epoch Curve

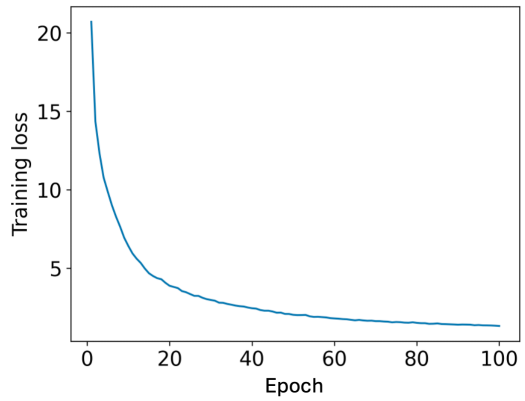

Figure 8: MSE Training loss by epoch for the MACE model trained on 1200 random AIMD frames (400 per trajectory).

The MACE models were trained for 100 epochs and did not exhibit the classical signs of overfitting.

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: The abstract and introduction accurately describe the contributions and scope of the paper. For this work, we analyzed the performance of MACE, a MLFF model, when trained on datasets with different diversity values and found that greater diversity in training data can yield a model with better prediction capability.

   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: This work discusses briefly that the computational bottleneck associated with MLFF training is the acquisition of ab initio training data. Thus, curating a maximally-diverse training data set is not immediately beneficial to the model development workflow, particularly for learning equilibrium bulk dynamics. Although efficient model training has implications for more complex transfer learning frameworks, that was outside the scope of this work.

   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: This work did not include any proofs or theoretical results.

   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: The Methods section described the acquisition of the training data via Quantum ESPRESSO, the MLFF model architecture used (MACE), and its implementation in the *mlip* Python package.

   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: [Yes]

   Justification: The work provides access to the data and code used for it at https://github.com/cganley2/ai4mat2025.

   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: This work varied the size of three subsets of AIMD data when training MACE models with certain hyperparameters. This work also demonstrated a sensitivity analysis of the RBF kernel on the kernel width parameter in Appendix C.

   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: [No]

Justification: This work could have included an analysis of different random seeds used for dataset selection and reported on the statistical significance of the MACE models trained thereon.

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 conducted some computational scaling studies as part of this work. The results are shown in the table below.

N_frames	Epochs	Cores per node	Memory (GB)	Walltime
1000	100	16	21.33	11:12:04
2000	100	16	21.25	21:07:21
3000	100	16	21.12	39:10:02

   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: This work conforms in every respect with the 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: [No]

    Justification: The scope of this work was in the training of machine learned force fields which are of interest for a small subset of researchers and not the general population.

    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: This paper did not develop data or models that have a high risk for misuse.

    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: This work cited Quantum ESPRESSO, ASE, MACE, TRAVIS, and the mlip Python package.

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: [NA]

Justification: This work did not introduce any new assets.

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: This work did 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.
- 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: This work did not require IRB approval.

    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: No LLMs were used in the development of any part of this work.

    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.