DATA CURATION FOR MACHINE LEARNING INTERATOMIC POTENTIALS BY DETERMINANTAL POINT PROCESSES

Joanna Zou
Computational Science & Engineering
Massachusetts Institute of Technology
Cambridge, MA, 02139, USA
{jjzou}@mit.edu

Youssef Marzouk
Computational Science & Engineering
Massachusetts Institute of Technology
{ymarz}@mit.edu

ABSTRACT

The development of machine learning interatomic potentials faces a critical computational bottleneck with the generation and labeling of useful training datasets. We present a novel application of determinantal point processes (DPPs) to the task of selecting informative subsets of atomic configurations to label with reference energies and forces from costly quantum mechanical methods. Through experiments with hafnium oxide data, we show that DPPs are competitive with existing approaches to constructing compact but diverse training sets by utilizing kernels of molecular descriptors, leading to improved accuracy and robustness in machine learning representations of molecular systems. Our work identifies promising directions to employ DPPs for unsupervised training data curation with heterogeneous or multimodal data, or in online active learning schemes for iterative data augmentation during molecular dynamics simulation.

1 INTRODUCTION

A primary challenge in the development of machine learning interatomic potentials (MLIPs) for atomistic simulation is the high degree of sensitivity of model performance to the choice of training set. In practice, parameters of MLIPs are learned from a training set of atomic configurations labeled with reference ground state energy and force values obtained from higher-order quantum mechanical (QM) calculations such as density functional theory (DFT). Due to the significant cost of QM calculations, training datasets must be limited in size to keep the data generation task computationally tractable and reduce overfitting to redundant data, but also retain representation of conformational diversity in order to produce robust MLIPs capable of capturing chemical processes of interest.

In this work, we propose using *determinantal point processes* (DPPs) for automated training set construction for machine learning-driven atomistic simulation. A DPP is an efficient probabilistic model over subsets of discrete sets which assigns greater likelihood to subsets with diverse elements, as determined by the determinant of a kernel matrix measuring the similarity between elements. Our work is one of the first to compare state-of-the-art data subselection algorithms on the task of training MLIPs, assessed in terms of diversity of atomic environments sampled, accuracy of the MLIP as it varies with training set size, and transferrability of the MLIP to predicting energies of atomic environments unseen during training.

2 BACKGROUND AND RELATED WORK

Data subselection is one form of *active learning*, in which an algorithm queries informative samples from a large pool of unlabeled data to label using a cost-intensive process for regression tasks. For MLIP training, the initial pool of unlabeled data is generally sourced from 1) hand-crafted datasets using expert judgment for each system of study, such as those of the QM9 database (Ramakrishnan et al., 2014); or 2) from time steps of molecular dynamics (MD) simulation which are efficiently evaluated using an empirical potential or initial iterate of the MLIP to approximately sample from

the Gibbs distribution of the system. However, since the step size must be sufficiently small for stable numerical simulation (on the order of 10^{-12} or 10^{-15} seconds), configurations generated by MD simulation are highly correlated and lead to high redundancy in the dataset.

Data subselection techniques seek to assemble compact training sets which improve on hand-crafted datasets as well as the naive approach of uniform sampling from the MD trajectory. In Huan et al. (2017), the descriptor space is partitioned into clusters using Euclidean distances via k-means which are each subsampled randomly. However, the k-means algorithm is sensitive to the geometry of the descriptor space and performs poorly in the high dimensional regime, which is often the setting for molecular descriptors. In Podryabinkin & Shapeev (2017); Lysogorskiy et al. (2023), the MaxVol algorithm identifies an optimal subset of configurations whose descriptors span the largest volume, based on the D-optimality criterion in linear experimental design. This approach is limited to choosing a fixed number of data, namely the same number as the dimension of the descriptors, to form a square D-optimal matrix. In the entropy-maximization method in Karabin & Perez (2020); de Oca Zapiain et al. (2022), local maxima of a chosen entropy function are assumed to be sufficiently non-redundant and included in the training dataset. This method heavily relies on the assumption that the full support of an effective potential function can be adequately sampled such that the local maxima correspond to well-separated modes of the entropy function. Each method utilizes a different similarity metric to compare atomic configurations, summarized in Table 1 in Appendix A.1, with the aim to reduce redundancy in the training set.

In the absence of labels, samples must be selected in an unsupervised manner utilizing only information on the atomic environment, which is summarized using *descriptors* which characterize multi-body interactions while preserving invariance properties of the representation. Examples of such descriptors include symmetric polynomials (ACE potential, Drautz (2019)), bispectrum components (SNAP potential, Thompson et al. (2015)), SOAP descriptors (GAP potential, Bartók et al. (2010)), atom-centered symmetry functions (Behler, 2011), SMILES strings used in graph neural network (NN) potentials (Weininger, 1988), or other latent representations learned simultaneously by NN potentials such as NequIP (Batzner et al., 2022) and Allegro (Bartók et al., 2022). In principle, the choice of descriptors for active learning may be independent from the choice of MLIP architecture which is trained, though the expressiveness of the descriptor – its ability to distinguish between atomic environments – will affect the efficiency of the data subselection algorithm.

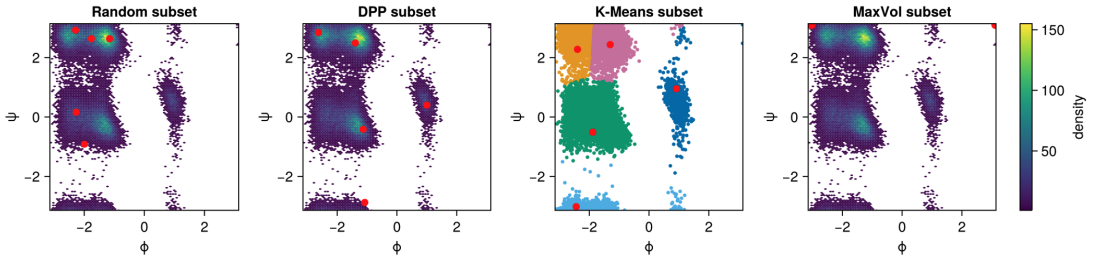

Figure 1: Sample subsets drawn with each method from a MD trajectory of alanine dipeptide.

To illustrate the data subselection methods, Figure 1 shows subsets of $N = 5$ configurations of alanine dipeptide, taken from a 50-ns Langevin dynamics simulation at 300 K using the AMBER ff-99SB-ILDN force field, sourced from Wehmeyer et al. (2023). The subsets are drawn using two collective variables (backbone dihedral angles ϕ, ψ) as the descriptors of the atomic configurations. Unlike the uniform random subset, the subsets drawn with the DPP and k-means more consistently distribute the five selected points to distinct regions of descriptor space, utilizing similarity-based metrics and distance-based clustering respectively. The MaxVol algorithm is limited to selecting 2 points (equal to the collective variable dimension), and these configurations are located at the bounds of the descriptor space in order to maximize the volume enclosed by their span.

3 METHODOLOGY

DPPs have seen a recent surge of interest in applications to machine learning (Kulesza & Taskar, 2012), experimental design (Dereziński et al., 2020), and dimensionality reduction (Belhadji et al., 2020). This section gives an intuitive introduction to DPPs, with details in Appendix A.4. If the data

pool is modeled by a point process, where each configuration constitutes a point in descriptor space, then a DPP is a type of repulsive point process which favors the sampling of distinctly unique and non-clustered points. Unlike other repulsive point processes, DPPs have several features, including closed-form probabilities and efficient sampling algorithms, which make them attractive in ML applications (Lavancier et al., 2015). Along this vein, simple random sampling can be interpreted as a homogeneous Poisson process where points are purely independent and uniformly distributed.

As a point process model, a DPP places a random measure over all subsets of a discrete dataset $\mathcal{Y}$ of M elements. The degree of repulsion between elements is controlled by a kernel matrix $K \in \mathbb{R}^{M\times M}$, where $K_{ij} = \kappa(Y_i, Y_j)$ for some positive semidefinite (PSD) kernel function $\kappa : \mathcal{Y}\times\mathcal{Y} \to \mathbb{R}$. The DPP places greater probability on subsets which have a higher degree of repulsion corresponding to a larger determinant of a kernel matrix, which can be interpreted as a measure of volume spanned by the matrix columns. For a subset $\{Y_i\}_{i\in\boldsymbol{\alpha}} \subseteq \mathcal{Y}$ indexed by $\boldsymbol{\alpha} = \{\alpha_1, ..., \alpha_m\}$ with $m \leq M$ and the kernel matrix $[K_{ij}]_{i,j\in\boldsymbol{\alpha}}$ restricted to rows and columns indexed by $\boldsymbol{\alpha}$, the probability of the subset being drawn by the DPP is given by:

$$\mathcal{P}(\{Y_i\}_{i\in\boldsymbol{\alpha}}) = \det\left([K_{ij}]_{i,j\in\boldsymbol{\alpha}}\right) \tag{1}$$

A DPP is defined by a choice of PSD kernel function. In this work, we employ a linear kernel equivalent to the cosine similarity of the descriptor vectors. Given a configuration of J_i atoms with Cartesian positions $\mathbf{x}_i \in \mathbb{R}^{3J_i}$, a second configuration of J_j atoms with positions $\mathbf{x}_j \in \mathbb{R}^{3J_j}$, and a global q-dimensional descriptor $\phi : \mathbb{R}^{3J} \to \mathbb{R}^q$ as a function of the positions of an arbitrary number of atoms J, the normalized linear kernel is given by:

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \frac{\phi(\mathbf{x}_i)\cdot\phi(\mathbf{x}_j)}{\|\phi(\mathbf{x}_i)\|\,\|\phi(\mathbf{x}_j)\|} \tag{2}$$

A DPP models probabilities over all 2^M subsets of $\mathcal{Y}$ without constraints on the size of the subset; however, subsets of a pre-specified size are often desired in applications. Therefore, we employ fixed-size DPPs (Kulesza & Taskar, 2011; Barthelmé et al., 2019) which provide a probability measure over subsets of $\mathcal{Y}$ of the same cardinality. Efficient sampling algorithms have been developed for both regular and fixed-size DPPs which perform sampling by decomposing the DPP into a mixture of elementary DPPs; refer to Kulesza & Taskar (2011); Barthelmé et al. (2019) and Appendix A.4. In general, the probabilistic nature of DPPs can offer substantially improved computational efficiency over brute force or optimization-based methods to data subselection.

4 EXPERIMENT

Reference dataset. We utilize hafnium oxides as systems of study for benchmarking the performance of the data subselection algorithms. Training and validation data belong to a 45,201-element set of HfO_2 and HfO configurations generated from six simulation sets to represent a range of states across the compression curve of each system, starting from crystal structures obtained from the Materials Project database (Jain et al., 2013). A 67% partition of this set, chosen proportionally from each simulation setting, is taken to be candidate data from which training data are subselected, while the remaining 33% constitutes the validation set. The test data consist of 67,219 configurations of either hafnium (Hf) or oxygen (O) atoms generated across 12 simulation sets by a similar procedure. Using this test set, we evaluate the ability of the MLIP trained on multi-species configurations of hafnium oxides to extrapolate to single-species configurations of Hf_2, Hf, O_2, and O, which are not explicitly seen during training. Reference QM values of energies and forces are computed with DFT using Quantum ESPRESSO (Giannozzi et al., 2017). Further details are provided in Appendix A.2.

Experimental setup. Data subselection is performed with four methods, by taking 1) a uniform sample of the candidate data, referred to as a simple random sample (SRS); 2) a random sample from clusters identified with k-means using $k = 5$ and covariance-weighted Euclidean distances between descriptors, where clusters are uniformly sampled proportionally to the cluster size; 3) the fixed set of basis vectors selected by the MaxVol algorithm; and 4) a random sample from a fixed-sized DPP computed from a linear kernel matrix of the descriptors. The molecular descriptors used to represent configurations are global ACE descriptors of Hf-O systems with body order $N = 5$ and polynomial degree $p = 8$ (Drautz, 2019), leading to descriptors of dimension $d = 1160$. Once training data are selected, reference QM values of energy are queried and an ACE potential is trained on the set by taking the least squares estimator of the model parameters.

Accuracy vs. training set size. We compare the accuracy of a MLIP trained on data subselected with each method using two common accuracy metrics: root mean square error (RMSE) in energy predictions capturing bias and variance information, and the coefficient of determination (R^2) measuring goodness of fit. For algorithms which sample variable set sizes (SRS, k-means, and DPP), accuracy statistics over 100 trials on the validation data are reported in Figure 2 for set sizes $N \in \{100, 200, 400, 800, 1600, 3200, 6400\}$, whereas the accuracy is reported for a single set size $d = 1160$ for MaxVol, as by construction it solves for a deterministic subset with cardinality equal to the dimension of the descriptor vector. The DPP-trained MLIP is comparable to that trained with k-means in both the data-poor regime (<200 samples), where small training sets limit model accuracy, as well as in the data-rich regime (>6400 samples), where large training sets have reduced distinction from one another. DPP outperforms the other variable-size data subselection methods, achieving low RMSE and high R^2 for each training set size N with less dispersion as represented by the interquartile range. MaxVol achieves an R^2 value closest to 1. This study demonstrates that DPP-based subselection is effective at constructing training datasets which summarize the Hf-O data with $\mathcal{O}(10^2 - 10^3)$ elements.

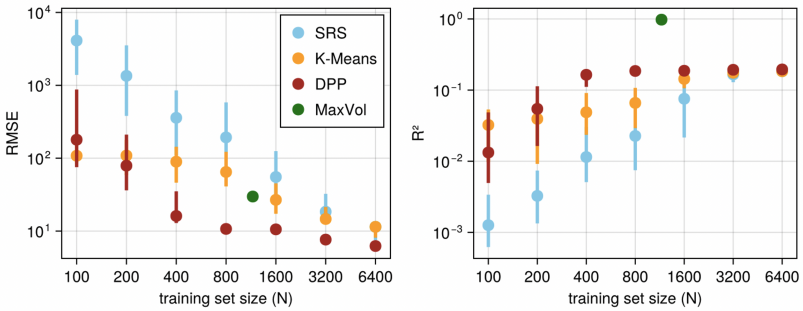

Figure 2: RMSE and R^2 values of energy predictions on the validation set by the MLIP trained on set size N. The center point denotes the median value and rangebars denote the 25th to 75th percentile over 100 trials.

Although SRS, k-means, and DPP are flexible to select variable set sizes, we fix the set size sampled by each of these methods to $N = d = 1160$ for the remainder of the studies, in order to maintain one-to-one comparison with the MaxVol algorithm.

Diversity of training set. The subsets chosen by each method is assessed for diversity, which is measured in terms of reference energies and force amplitudes (averaged over all atoms) associated with configurations in the subset, following de Oca Zapiain et al. (2022). In Figure 3, the subsets drawn with DPP and MaxVol cover a greater range of energies and force amplitudes, whereas the subset drawn with k-means has marginal distinction from that drawn with SRS, indicating that distance-based clustering in descriptor space does not necessarily correspond to better coverage of output quantities. This study provides evidence of the expressiveness of product-based similarity metrics used by DPPs and MaxVol to choose subsets with a greater range in the output space.

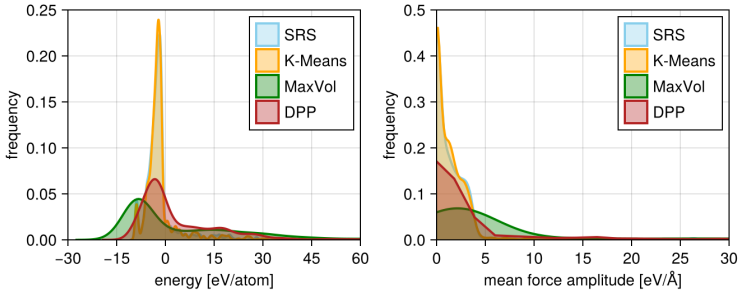

Figure 3: Distribution of energies and force amplitudes of 1160 selected configurations.

Prediction error on unseen data. We evaluate the generalization capabilities of the MLIP, trained on atomic interactions with HfO_2 and HfO data, to predict energies of single-species Hf and O

systems. Figure 4 shows the relative error in energy predictions over test configurations categorized into 5 molecule types, as summarized in Appendix A.2. In general, the prediction error by the DPP-trained MLIP is close to that of the MaxVol-trained MLIP, which achieves the lowest distribution of error for the bulk 128-atom Hf system.

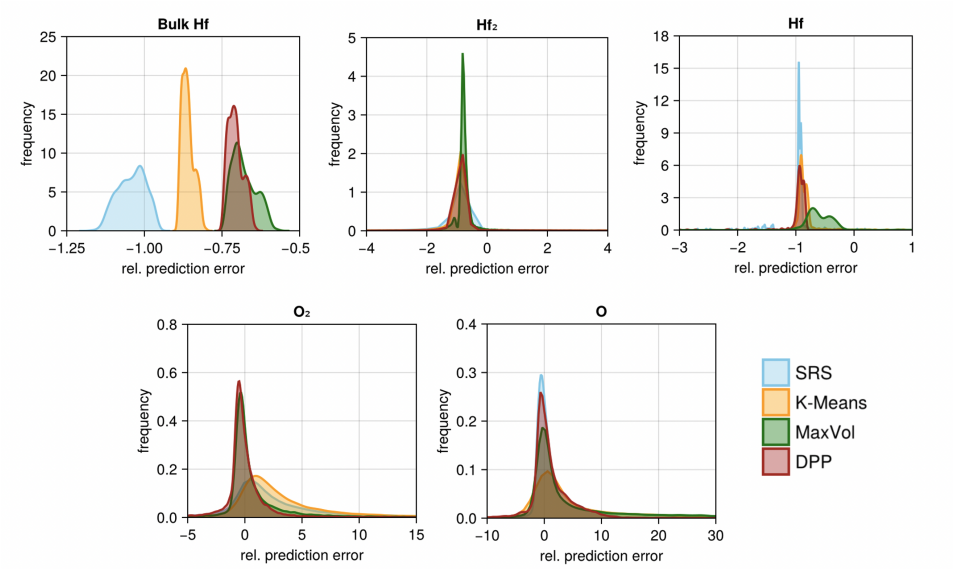

Figure 4: Relative error in energy predictions on the test set by an MLIP trained on a 1160-element set. Errors in Hf_2, Hf, O_2, and O have been truncated to visualize the main mass of the distribution.

5 Discussion

We present a new approach to training data curation for MLIPs leveraging DPPs and provide one of the first studies to benchmark the performance of different data subselection algorithms, advancing the state of the art in computational materials characterization. Our work demonstrates the competitiveness of the DPP-based approach with respect to existing approaches to variable-size data subselection in terms of accuracy of the trained MLIP, diversity of sampled configurations, and generalization to predictions on unseen data. If the kernel matrix is chosen to be the Fisher information matrix, then DPPs can be viewed as a principled probabilistic counterpart to the MaxVol algorithm, as both approaches rely on the D-optimality principle of maximizing the matrix determinant (Dereziński et al., 2020). This technique may complement several existing efforts to accelerate active learning in an online setting, where data are simultaneously sampled via molecular dynamics simulation of a fast potential and assessed for labeling with calls to a costly QM method. The next step of this work is to employ conditional DPP sampling for data augmentation tasks amenable to online active learning, to draw subsets of novel configurations conditioned on an existing set of training configurations, and compare this kernel-based strategy to uncertainty-based strategies proposed in Vandermause et al. (2020); van der Oord et al. (2023); Kulichenko et al. (2023). Another future direction is to apply this method to heterogeneous or multimodal datasets, such as simulation data of multiple element species or combinations of simulation and experimental data, as an automated technique for extracting informative data subsets which are most efficient for model development.

Acknowledgments

The authors would like to sincerely thank Dallas Foster for discussions on this work and Dionysios Sema for supplying the reference dataset of DFT calculations. JZ and YM gratefully acknowledge support from the United States Department of Energy, National Nuclear Security Administration under Award Number DE-NA0003965.

REFERENCES

Simon Barthelmé, Pierre-Olivier Amblard, and Nicolas Tremblay. Asymptotic equivalence of fixed-size and varying-size determinantal point processes. *Bernoulli*, 25 (4B):3555–3589, 2019.

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

Albert P. Bartók, Matthias Rupp, Heng Huo, Miguel A. Caro, Anton Götz, and Michele Ceriotti. Allegro: A fast and accurate machine learning model for interatomic potentials. *Nature Communications*, 13:3130, 2022. doi: 10.1038/s41467-022-30943-0.

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*, 13:2453, 2022. doi: 10.1038/s41467-022-29939-5.

Jörg Behler. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *The Journal of Chemical Physics*, 134(7):074106, 02 2011. ISSN 0021-9606. doi: 10.1063/1.3553717.

Ayoub Belhadji, Rémi Bardenet, and Pierre Chainais. A determinantal point process for column subset selection. *Journal of Machine Learning Research*, 21:1–62, 2020.

David Montes de Oca Zapiain, Mitchell A. Wood, Nicholas Lubbers, Carlos Z. Pereyra, Aidan P. Thompson, and Danny Perez. Training data selection for accuracy and transferability of interatomic potentials. *npj Computational Materials*, 8, 12 2022. ISSN 20573960. doi: 10.1038/s41524-022-00872-x.

Michał Dereziński, Feynman Liang, and Michael W. Mahoney. Bayesian experimental design using regularized determinantal point processes. *AISTATS 2020*, 2020.

Ralf Drautz. Atomic cluster expansion for accurate and transferable interatomic potentials. *Physical Review B*, 99(1):014104, 2019.

Alan Edelman. *Random Matrix Theory*. (Work in progress), 2024.

Fredrik Eriksson, Elias Fransson, and Paul Erhart. The hiphive package for the extraction of high-order force constants by machine learning. *Advanced Theory and Simulations*, 2(5):1800184, 2019.

Paolo Giannozzi, Oliviero Andreussi, Thomas Brumme, Olivier Bunau, Marco Buongiorno Nardelli, Matteo Calandra, Roberto Car, Carlo Cavazzoni, Davide Ceresoli, Matteo Cococcioni, et al. Advanced capabilities for materials modelling with quantum espresso. *Journal of Physics: Condensed Matter*, 29(46):465901, 2017.

Tran Doan Huan, Rohit Batra, James Chapman, Sridevi Krishnan, Lihua Chen, and Rampi Ramprasad. A universal strategy for the creation of machine learning-based atomistic force fields. *npj Computational Materials*, 3, 12 2017. ISSN 20573960. doi: 10.1038/s41524-017-0042-y.

Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Weike Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, 2013.

Mariia Karabin and Danny Perez. An entropy-maximization approach to automated training set generation for interatomic potentials. *The Journal of Chemical Physics*, 153(9):094110, 2020.

Alex Kulesza and Ben Taskar. k-dpps: Fixed-sized determinantal point processes. *ICML 2011*, pp. 1193–1200, 2011.

Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 5(2–3):123–286, 2012.

Maksim Kulichenko, Kipton Barros, Nicholas Lubbers, Ying Wai Li, Richard Messerly, Sergei Tretiak, Justin S. Smith, and Benjamin Nebgen. Uncertainty-driven dynamics for active learning of interatomic potentials. *Nature Computational Science*, 3 2023. ISSN 26628457. doi: 10.1038/s43588-023-00406-5.

Frédéric Lavancier, Jesper Møller, and Ege Rubak. Determinantal point process models and statistical inference. *Journal of the Royal Statistical Society Series B*, 77(4):853–877, 2015.

Yury Lysogorskiy, Anton Bochkarev, Matous Mrovec, and Ralf Drautz. Active learning strategies for atomic cluster expansion models. *Physical Review Materials*, 7:043801, Apr 2023.

Nicola Marzari, David Vanderbilt, Alessandro De Vita, and Mike C. Payne. Thermal contraction and disordering of the al(110) surface. *Physical Review Letters*, 82(16):3296–3299, April 1999.

Evgeny V. Podryabinkin and Alexander V. Shapeev. Active learning of linearly parametrized interatomic potentials. *Computational Materials Science*, 140:171–180, 12 2017. ISSN 09270256. doi: 10.1016/j.commatsci.2017.08.031.

Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1:140022, 2014.

Ganesh Sivaraman, Anand Narayanan Krishnamoorthy, Matthias Baur, Christian Holm, Marius Stan, Gábor Csányi, Chris Benmore, and Álvaro Vázquez-Mayagoitia. Machine-learned interatomic potentials by active learning: amorphous and liquid hafnium dioxide. *npj Computational Materials*, 6(1):104, 2020.

Aidan P Thompson, Laura P Swiler, Christian R Trott, Stephen M Foiles, and Garritt J Tucker. Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials. *Journal of Computational Physics*, 285:316–330, 2015.

Cas van der Oord, Matthias Sachs, Dávid Péter Kovács, Christoph Ortner, and Gábor Csányi. Hyperactive learning for data-driven interatomic potentials. *npj Computational Materials*, 9, 12 2023. ISSN 20573960. doi: 10.1038/s41524-023-01104-6.

Jonathan Vandermause, Steven B Torrisi, Simon Batzner, Yu Xie, Lixin Sun, Alexie M Kolpak, and Boris Kozinsky. On-the-fly active learning of interpretable bayesian force fields for atomistic rare events. *npj Computational Materials*, 6(1):1–11, 2020.

Christoph Wehmeyer, Martin K. Scherer, Yaoyi Chen, Moritz Hoffmann, and Tim Hempel. markovmodel/mdshare, 11 2023. URL `https://github.com/markovmodel/mdshare`.

David Weininger. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988. doi: 10.1021/ci00057a005.

A APPENDIX

A.1 COMPARISON OF DATA SUBSAMPLING ALGORITHMS

Table 1 provides a summary of the literature review from Section 2. A primary distinction among the algorithms for data subselection is the similarity metric used to compare two configurations. While the entropy maximization method is amenable to offline active learning schemes for data subselection, the entropy function depends on local (per-atom) descriptors, and is therefore excluded from comparison studies of methods depending on global (molecular) descriptors.

Table 1: Review of data subselection algorithms for ML-IPs in literature.

Algorithm	Primary Reference	Similarity Metric	Subset Size	Type
SRS	–	None	Variable	Probabilistic
k-means clustering	Huan et al. (2017)	Euclidean distance in descriptor space	Variable	Probabilistic
MaxVol	Podryabinkin & Shapeev (2017)	Fisher information matrix	Fixed	Deterministic
Entropy max.	Karabin & Perez (2020)	Entropy function	Variable	Deterministic
fixed-size DPP	This work	PSD kernel	Variable	Probabilistic

A.2 REFERENCE DATASET

The reference data used to validate energy predictions in the experiments of Section 4 consist of quantum-level energies and forces of atomistic configurations of hafnium (Hf), oxygen (O), hafnium dioxide (HfO_2), and hafnium oxide (HfO) systems. The configurations are generated using a systematic approach to explore the descriptor space of interest: for each system, the experimentally observed crystal structures are obtained from Materials Project (Jain et al., 2013). Using the compression curves of each material, Latin Hypercube Sampling (LHS) is performed, treating the lattice parameters (length and angles) and density as random variables, to draw random samples ranging from highly compressed states to melting point states. The atomic positions are then minimized for each configuration to obtain the equation of state (EOS). Additional simulations are carried out for bulk Hf systems to sample configurations along the phase diagram from the convex hull to high energy states In particular, a modified Monte Carlo rattling procedure using the hiPhive package (Eriksson et al., 2019) is employed on supercells replicated from the unit cell of each structure to generate LHS samples emulating the phase diagram. For each configuration, quantum-level energies are computed with Quantum ESPRESSO (Giannozzi et al., 2017) utilizing the PBE functional with the Optimized Norm Conserving Vanderbilt (ONCV) pseudopotential. All calculations used the Marzari-Vanderbilt method (Marzari et al., 1999) for electron smearing and the electronic temperature was set to 0.06 Ry. The kinetic energy cutoff was 90 Ry and the k-point spacing of 0.025 $Å^{-1}$ was adopted to ensure all configurations have consistent spacing.

Table 2 summarizes the configurations of HfO_2, HfO used for training and validation and Table 3 summarizes the configurations of Hf and O systems used for the test set, categorized into simulation sets. The simulation sets are labeled with their chemical species; Materials Project ID number (MP ID), if one exists; the number of translation and/or rotational degrees of freedom (1D, 3D, 6D); phase (primitive, gas); whether the Monte Carlo rattling procedure is performed (MC); and whether the atomic positions have been minimized (EOS). Set 1 of Table 2 (labeled with "figshare") corresponds to hafnium dioxide configurations sourced from Sivaraman et al. (2020) where electronic structure calculations are recomputed with the ONCV pseudopotential. The number of Hf or O atoms per configuration are reported in the last two columns.

Table 2: Training/validation set of hafnium dioxide (HfO_2) and hafnium monoxide (HfO) configurations.

index	MP ID	set	no. configurations	no. Hf	no. O
1	–	HfO_2 figshare	2052	36	72
2	352	HfO_2 EOS 1D	300	4	8
3	550893	HfO_2 EOS 1D	131	1	2
4	550893	HfO_2 EOS 6D	27620	1	2
5	–	HfO gas	129	1	1
6	–	HfO EOS 1D	14969	1	1

Note that in Figure 4, the bulk Hf test data corresponds to sets 1-3, the Hf_2 test data to sets 4-7, the Hf test data to sets 8-9, the O_2 test data to sets 10-11, and the O test data to set 12, as listed in Table 3.

Table 3: Test set of hafnium (Hf) and oxygen (O) configurations.

index	MP ID	set	no. configurations	no. Hf	no. O
1	–	Hf_2 gas	63	2	0
2	103	Hf_2 EOS 1D	202	2	0
3	103	Hf_2 EOS 3D	9377	2	0
4	103	Hf_2 EOS 6D	17205	2	0
5	100	Hf EOS 1D	201	1	0
6	100	Hf 1D primitive	201	1	0
7	100	bulk Hf MC	306	128	0
8	103	bulk Hf MC	50	128	0
9	–	bulk Hf MC	498	128	0
10	607540	O_2 EOS 6D	19223	0	2
11	–	O_2 gas	204	0	2
12	–	O EOS 6D	19689	0	1

A.3 Additional experiments on hafnium data

In another study, we consider single-species hafnium (Hf) data to evaluate the performance of DPP, k-means ($k = 5$), and SRS to draw variable-sized training datasets. We take the candidate training data and validation data from a 27,249-element set of 1-atom and 2-atom Hf configurations, generated from simulation sets 1-6 detailed in Table 3. The training sets, subselected from a 70% split of the data, are used to learn a 5-body 6-degree ACE potential of hafnium, which have a descriptor dimension of $d = 35$. Figure 5 shows the statistics of RMSE and R^2 values of the energy predictions on the validation set by MLIPs trained on each subset. Figure 6 shows the distribution of per-atom energies and force magnitudes associated with a 400-sample set drawn by each of the three methods. While the results using the Hf data largely mirror the trends observed with the Hf-O data in Section 4, the improvements of the DPP-based approach in terms of both prediction accuracy and diversity of configurations are more pronounced in this study. Further experiments can be conducted to assess the dependency of the data subselection algorithms on the choice of descriptor, in particular the dimension and expressivity of the descriptor for a given system of study.

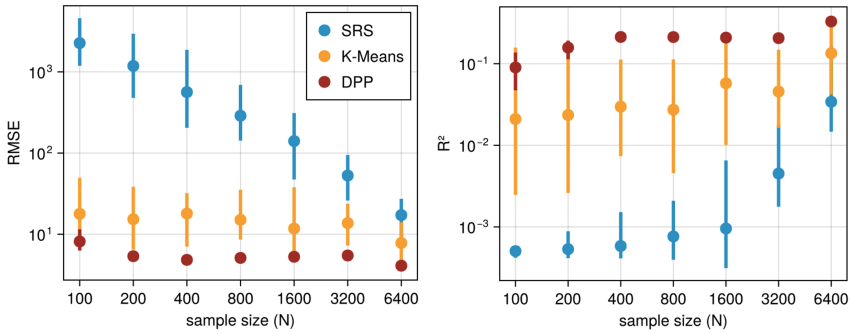

Figure 5: RMSE and R^2 values of energy predictions on the validation set by the hafnium MLIP for training set size N. The center point denotes the median value and rangebars denote the interquartile range (25th to 75th percentile) over 100 trials.

In addition, we assess diversity of subsets selected by DPP and k-means in terms of the rate at which the algorithm samples each of the 6 simulation datasets composing the candidate training data (the first six rows of Table 3). Over 100 trials, subsets of 800 elements are drawn from the data pool of $M_{\text{tot}} = 19,090$ elements and the number of instances of each simulation dataset is counted and normalized by the batch size and total number of elements per dataset, such that the normalized rate can be compared to the uniform sampling rate of $1/M_{\text{tot}}$ from SRS. Figure 7 shows that the DPP leads to differential rates of sampling between datasets, with Sets 1 and 6 favored relative to the SRS baseline. This indicates that these two simulation sets, while relatively small in number compared to other sets, are important to include in the DPP subset to represent diversity in the data pool. Therefore, DPPs can also be utilized as a diagnostic for categories of data which

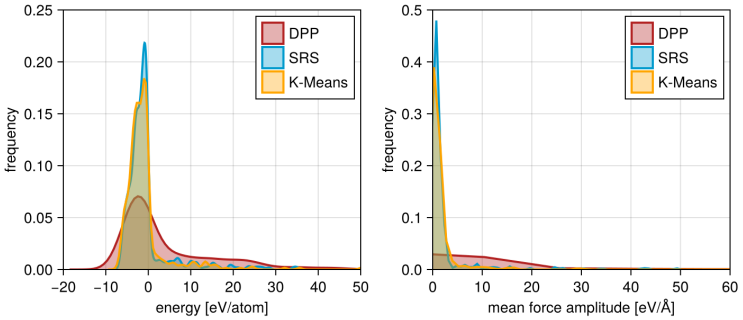

Figure 6: Distribution of energies and force amplitudes of 400 configurations selected by each algorithm.

have proportionally greater influence in training, which can inform the collection of additional data. In contrast, the k-means clustering method does not lead to substantially different sampling rates compared to SRS.

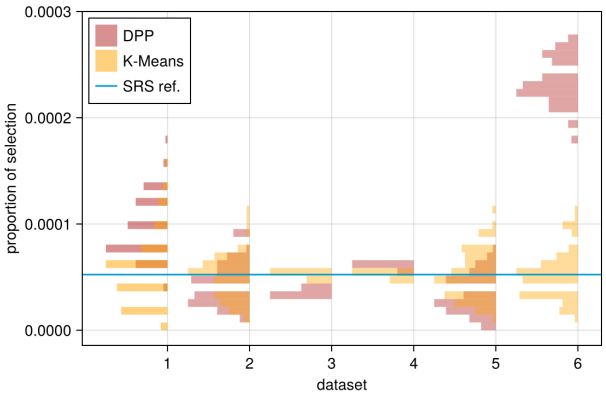

Figure 7: Representation from simulation datasets 1-6 (Table 3) in the 800-element subsets selected by DPPs and k-means, with the uniform rate from SRS as reference.

A.4 Determinantal point processes

Consider a set of M discrete elements represented by their indices $\mathcal{Y} = \{1, ..., M\}$. A determinantal point process (DPP) is a probability measure placed over all 2^M subsets of $\mathcal{Y}$, where probabilities are determined by the kernel matrix $K \in \mathbb{R}^{M \times M}$ associated with the process. In practice, the kernel matrix is constructed from evaluations of a positive semidefinite (PSD) kernel function $\kappa : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ between each pair of elements in the set, where $K_{ij} = \kappa(Y_i, Y_j)$ for $Y_i, Y_j \in \mathcal{Y}$. If the kernel matrix satisfies conditions for the existence of the L formulation (namely, that $P(\emptyset) \neq 0$ and K has no eigenvalue equal to 1 (Kulesza & Taskar, 2012)), then the L ensemble corresponding to K is given by:

$$L = K(I - K)^{-1} \tag{3}$$

The DPP is then defined equivalently by the following, for random subsets $Y \subseteq \mathcal{Y}$ and a fixed subset $A \subseteq \mathcal{Y}$ (Edelman, 2024):

PDF definition. The probability density function of the DPP is given by:

$$\mathcal{P}(Y = A) = \frac{\det(L_A)}{\sum_{A' \subseteq \mathcal{Y}} \det(L_{A'})} \tag{4}$$

where $L_A = [L_{ij}]_{i,j \in A}$ denotes the matrix restricted to entries indexed by the elements of A. The PDF definition is also referred to as the "L formulation" of the DPP (Edelman, 2024).

CCDF definition. The complementary cumulative density function of the DPP is given by:

$$\mathcal{P}(Y \supseteq A) = \det(K_A) \tag{5}$$

In other words, the probability that A is a subset of the randomly drawn set Y is given by the determinant of the kernel matrix restricted to entries indexed by A. A special case of the CCDF is the marginal probability of each element of the set, which is given by the diagonal of the K matrix:

$$\mathcal{P}(i \in Y) = K_{ii} \tag{6}$$

The marginal probability is also referred to in literature as the "inclusion probability" (Kulesza & Taskar, 2012). The CCDF definition is also referred to as the "K formulation" of the DPP (Edelman, 2024).

CDF definition. The cumulative density function of the DPP is given by:

$$\mathcal{P}(Y \subseteq A) = \det(I - K)_{\bar{A}} \tag{7}$$

where $\bar{A}$ denotes the complement, $\bar{A} = \mathcal{Y} \setminus A$.

Mixture of elementary DPPs. An important property of a DPP is that it can be represented as the mixture of elementary DPPs. Also referred to as *projection DPPs*, an elementary DPP has a kernel matrix which is a projection matrix of rank $r \leq M$, e. g. $K^{\mathrm{T}}K = K$ and $K = VV^{\mathrm{T}}$ for a set of r orthonormal vectors V (Edelman, 2024). Elementary DPPs then have the property that:

$$\mathcal{P}^{V_r}(A) = \begin{cases} \det(K_A) & \text{if } |A| = r \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

Therefore, only $\binom{M}{r}$ subsets of size exactly r have non-zero probability (Edelman, 2024). The PDF of the DPP can be represented as the mixture of elementary DPPs, using the eigendecomposition of the L matrix $L = \sum_{i=1}^{M} \lambda_i \mathbf{v}_i \mathbf{v}_i^{\mathrm{T}}$:

$$\mathcal{P}(Y = A) = \frac{1}{\det(I + L)} \sum_{J \subseteq 1:M} \mathcal{P}^{V_J} \prod_{i \in J} \lambda_i \tag{9}$$

In particular, it can be shown that the normalizing constant of the density can be derived as in Kulesza & Taskar (2011):

$$\sum_{A' \subseteq \mathcal{Y}} \det(L_{A'}) = \det(I + L) = \prod_{i=1}^{M} (\lambda_i + 1) \tag{10}$$

The mixture representation of DPPs lends it to a computationally tractable sampling algorithm. In particular, the DPP can be sampled by drawing samples from each of the elementary DPPs with probability $\frac{\prod_{i \in J} \lambda_i}{\prod_{i=1}^{M} (\lambda_i + 1)}$.

Fixed-size determinantal point processes. A k-DPP is a DPP which produces samples of fixed size $k \leq M$ (not to be misconstrued with the number of clusters in k-means). Unlike elementary DPPs, which are restricted to represent specific probability measures associated with a projection kernel matrix, k-DPPs can represent a more flexible range of probability measures over the subsets. As one example, a k-DPP can be defined to assign a uniform distribution over subsets, whereas a singular elementary DPP cannot (Kulesza & Taskar, 2012). Therefore, elementary DPPs can be considered a subclass of k-DPPs.

A k-DPP can be understood as a special form of conditional DPP, with the following PDF:

$$\mathcal{P}(Y = A| \, |Y| = k) = \frac{\det(L_A)}{\sum_{|A'|=k} \det(L_{A'})} \tag{11}$$

where the normalizing constant is a sum over all subsets $A' \in \mathcal{Y}$ with restricted cardinality $|A'| = k$. It can be shown that the k-DPP can also be expressed as a mixture of elementary DPPs:

$$\mathcal{P}(Y = A| \, |Y| = k) = \frac{1}{e_k^M} \sum_{|J|=k} \mathcal{P}^{V_J} \prod_{i \in J} \lambda_i \tag{12}$$

The normalizing constant of this distribution differs from that of the standard DPP. One can show it is derived as:

$$\begin{aligned} \sum_{|A'|=k} \det(L_{A'}) &= \det(I + L) \sum_{|A'|=k} \mathcal{P}(Y = A') \\ &= \sum_{|J|=k} \prod_{i \in J} \lambda_i \end{aligned} \tag{13}$$

The derivation uses the property that sets drawn from the elementary DPP have cardinality $|J| = k$ with probability 1, such that the expression reduces to the sum of products of eigenvalues indexed by elements in each J subset. One can recognize this term to be the kth elementary symmetric polynomial (Kulesza & Taskar, 2012):

$$e_k^M = e_k(\lambda_1, ..., \lambda_M) = \sum_{\substack{J \subseteq 1:M \\ |J|=k}} \prod_{i \in J} \lambda_i \tag{14}$$

The marginal probability of elements, now considering fixed sizes to the subsets drawn, is proportional to the eigenvalues of L scaled by a ratio of the elementary symmetric polynomials:

$$\mathcal{P}(i \in Y| \, |Y| = k) = \lambda_M \frac{e_{k-1}^{M-1}}{e_k^M} \tag{15}$$

Sampling algorithm for k-DPPs. Algorithm 1 for sampling from k-DPPs is reproduced from Kulesza & Taskar (2011; 2012). The algorithm is composed of two loops: Loop 1 samples eigenvectors of L to form a subspace from which to draw the samples. While the eigenvectors are sampled with probability $\frac{\lambda_n}{\lambda_n+1}$ for standard DPPs, the probability becomes $\lambda_n \frac{e_{l-1}^{n-1}}{e_l^n}$ for k-DPPs. Moreover, sampling is performed until strictly k eigenvectors are obtained. Loop 2 iteratively samples elements from the eigenvectors, orthonormalizing the basis after each sample is drawn. While the cardinality of the basis V varies for standard DPPs, it is kept fixed at k for k-DPPs.

Computationally, the sampling of regular DPPs and k-DPPs differ primarily in the computation of the probabilities in Loop 1. For k-DPPs, the normalization constant of the probability density involves calculating the ratio of elementary symmetric polynomials, the cost of which scales exponentially. For instance, calculation of e_k^N takes on the order of $\mathcal{O}(k\binom{N}{k})$ operations due to the combinatorial problem. To address this cost, Kulesza & Taskar (2011; 2012) recommend implementing a recursive algorithm to construct all symmetric elementary polynomials at once, using the recurrence relation:

$$e_k^N = e_k^{N-1} + \lambda_N e_{k-1}^{N-1} \tag{16}$$

The recursive algorithm has polynomial cost at $\mathcal{O}(Nk)$, significantly reducing the computational overhead of the sampling algorithm.

Algorithm 1 Sampling from a k-DPP

Require: $0 < k \leq N$, eigendecomposition $\{(\mathbf{v}_n, \lambda_n)\}_{n=1}^{N}$ of L from Equation 3

Loop 1: sample eigenvectors to form subspace
$J \leftarrow \emptyset$
$l \leftarrow k$
for n = N, ..., 2, 1 **do**
  if $l = 0$ **then**
    break
  end if
  if $u \sim U[0,1] < \lambda_n \frac{e_{l-1}^{n-1}}{e_l^n}$ **then**
    $J \leftarrow J \cup \{n\}$
    $l \leftarrow l - 1$
  end if
end for

Loop 2: draw samples from orthonormalized subspace
$V \leftarrow \{\mathbf{v}_n\}_{n \in J}$
$Y \leftarrow \emptyset$
while $|V| > 0$ **do**
  Select i from $\mathcal{Y}$ with $Pr(i) - \frac{1}{|V|} \sum_{\mathbf{v} \in V} (\mathbf{v}^{\mathrm{T}} \mathbf{e}_i)^2$
  $Y \leftarrow Y \cup i$
  $V \leftarrow V_\perp$ (orthonormal basis for subspace of V orthogonal to $\mathbf{e}_i$)
end while

Output: Y

A.5 IMPLEMENTATION

Experiments were conducted in Julia using the packages PotentialLearning.jl and InteratomicPotentials.jl, developed Dallas Foster, Emmanuel Lujan, Spencer Wyant, and JZ; Determinantal.jl, developed by Simon Barthelmé; Maxvol.jl, developed by Aleksandr Mikhalev; ACE.jl, developed by Christoph Ortner et al.; and Clustering.jl for the k-means algorithm.