An entropy-maximization approach to automated training set generation for interatomic potentials

Perez, Danny
Karabin, Mariia

To be published in: The Journal of Chemical Physics

DOI to publisher's version: 10.1063/5.0013059

Permalink to record: http://permalink.lanl.gov/object/view?what=info:lanl-repo/lareport/LA-UR-20-21236

Disclaimer:
Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by Triad National Security, LLC for the National Nuclear Security Administration of U.S. Department of Energy under contract 89233218CNA000001. By approving this article, the publisher recognizes that the U.S. Government retains nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy. Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

An Entropy-Maximization Approach to Automated Training Set Generation for Interatomic Potentials

Mariia Karabin[1, 2] and Danny Perez[2, a)]

[1)] *Department of Chemistry, Clemson University, Clemson, SC 29634 USA*

[2)] *Theoretical Division T-1, Los Alamos National Laboratory, Los Alamos, NM 87545 USA*

(Dated: 5 August 2020)

Machine learning (ML)-based interatomic potentials are currently garnering a lot of attention as they strive to achieve the accuracy of electronic structure methods at the computational cost of empirical potentials. Given their generic functional forms, the transferability of these potentials is highly dependent on the quality of the training set, the generation of which can be highly labor-intensive. Good training sets should at once contain a very diverse set of configurations while avoiding redundancies that incur cost without providing benefits. We formalize these requirements in a local entropy maximization framework and propose an automated sampling scheme to sample from this objective function. We show that this approach generates much more diverse training sets than unbiased sampling and is competitive with hand-crafted training sets.

[a)] danny_perez@lanl.gov

I. INTRODUCTION

The practical usefulness of atomistic simulations ultimately relies on the availability of interatomic potentials that are able to provide reliable energies and forces at a sufficiently affordable computational cost. Since electronic structure calculations using techniques such as density functional theory (DFT) are often prohibitively expensive, simplified empirical forms have been the norm, especially for molecular dynamics (MD) applications where long simulation times and large systems are often required. Early empirical potentials were traditionally highly computationally efficient[1] but often lacked in accuracy and transferability. Over the last few years, the need to bridge the gap between empirical methods and direct electronic structure calculations has driven the explosive development of machine learning (ML) based approaches that aim to combine the accuracy of the electronic structure methods and the efficiency of the early simplified potentials[2–6]. The two main components of ML-based potentials are the representation of atomic structures with a set of generic descriptors that characterize local atomic environments and a general functional form for the energy of the system as a functions of the descriptors, which can be trained to reproduce the result of large amounts of high-quality electronic structure calculations (energies, forces, stresses, etc.).

While ML-based potentials have proved able to capture subtle features of the training data, their ability to extrapolate to situations that markedly differ from those encountered during training remains limited[7]. Therefore, the accuracy of ML-based potentials is highly dependent on the choice of the training set, which should i) cover as much of the relevant configuration space as possible, and ii) remain sufficiently compact so that the cost of computing the reference values with quantum calculations and training the model remains affordable. Traditionally, training set generation has been a highly labor-intensive activity that relies on physical intuition in order to carefully select the configurations that should be included. The advent of highly-flexible functional forms with many free parameters, which hence require a large amount of data to parameterize, gradually makes such manual curation impractical.

Different approaches have been proposed to address the first objective using sampling strategies[8–13], including evolutionary structural searches[14], normal mode sampling[15], and exploration of the potential energy surface using on-the-fly approximations of the target potential[8]. Training set configurations are also selected from DFT-MD simulations[16].

The second objective is often achieved by sub-sampling from larger data sets. Possible approaches include random selection[17], binning-based sub-sampling to achieve uniform representation of relevant quantities like atomic forces[18], clustering in descriptors space to identify distinct groups[18]. Finally, a number of recent approaches incrementally include data to the training set based on whether the prediction of the properties of new configurations require extrapolation[10,19–24]. These approaches differ by the algorithm type and query strategy[10,21]. These workflows are often driven by MD simulations that use preliminary versions of the potentials, which are continuously refined as the learning proceeds.

It is important to note that both these aspects are critical in practice: sub-sampling approaches can only result in a diverse training set if the larger candidate set it selects from is itself diverse. In that respect, direct-MD-based simulations can be used to generate new configurations, but MD is a notoriously inefficient sampler which can often have a very long correlation time. The hypothesis behind the current work is that the situation can be improved through specially-designed samplers whose objective function is a measure of diversity. We also note that *online* active-learning approaches (see e.g. Ref. 25 for an early incarnation), can be used to refine the potential only as necessary along a given MD trajectory. These approaches can certainly be powerful, but can lead to very high and unpredictable computational cost, and can result in community-wide inefficiencies when many independent groups carry out similar simulations. In the following, we focus on conventional (i.e., *offline*) approaches where training and production simulations occur in distinct phases, but note that our approach could also be used to locally promote diversity in online workflows.

In this manuscript, we unify the diversity and non-redundancy objectives in a simple local approach where the diversity of atomic environment within individual configurations (as measured by an entropy metric) is maximized subject to the constraint that it does not contain unphysical configurations (i.e., overlapping atoms). This objective is embodied in a generic effective potential energy function whose low-lying local minima are good candidates for inclusion in a training set. Such minima are sampled using a simple annealing scheme that can be easily automated. Importantly, this effective energy is *not* meant as an approximation to the energy of the target system; instead it is an abstract construct that enables the creation of material-agnostic training sets. In this sense, our approach aims at creating a "universal" set of configurations that captures a very wide range of local environments and does not focus solely on low-lying energy structures. The large volume of configuration space covered entails a trade-off between the size of the training set and the target accuracy, but the high transferability it affords is important to capture high-energy, far from equilibrium effects that can occur in extreme conditions, such as under irradiation or under shock conditions. A global approach where diversity maximization is carried out globally over the whole training set is currently in development and will be reported in an upcoming manuscript.

We emphasize that the objective of the current manuscript is not to create a training set for a specific material, nor to train an actual potential. Instead, our focus is on the development of a method to generate diverse sets of configurations that could then be characterized for a specific material with quantum methods such as DFT. While actual potentials are not trained, the performance of the generated training set can be assessed on the basis of surrogate metrics that derive from distances in descriptor space between training and testing configurations. As shown below, the training sets generated by maximizing descriptor entropy are more diverse than those generated by conventional unbiased methods, and are even competitive with hand-crafted datasets used in the recent literature. Training of actual potentials based on this method will be reported in an upcoming publication.

II. METHODS

1. *Entropy maximization approach*

Implementing the entropy-maximization idea in practice requires first defining a set of atomic descriptors $\{\mathbf{q}\}$ that characterize the local environment of each atom, and then defining a measure of the diversity of the distribution of these descriptors within a configuration containing multiple atoms. A wide array of atomic descriptors have been proposed in the literature[26], as these form the inputs of many machine learning approaches that learn atomic energies. The method we propose is agnostic to the specific choice of descriptors so as long as they are differentiable functions of atomic positions. In the following, the set of m descriptors $q_{i,k}$ of the local atomic environment of atom i is arranged into a vector $\mathbf{q}_i$ of length m.

As a measure of diversity, we use an approximation to the entropy of the m-dimensional distribution of atomic descriptors $S(\{\mathbf{q}\})$ contained in a given configuration of atoms, which is a natural choice in this case: it is maximized for a uniform distribution of descriptors and minimized for configurations where all environments are identical, i.e., maximizing descriptor entropy promotes diversity and penalizes redundancy. The effective energy we propose is therefore of the form:

$$V = E_{\text{repulsive}} - KS(\{\mathbf{q}\}) \tag{1}$$

where $E_{\text{repulsive}}$ is a short range repulsive term that penalizes very short distances between atoms (so as to enforce an excluded volume around each atom) and K is a so-called entropy scaling coefficient that controls the relative importance of the entropy and of the repulsive contribution. Local minima of this function can therefore be expected to contain a high diversity of different environments without any two atoms being unphysically close. We postulate that low-lying minima of this effective potentials are therefore good targets for inclusion in a training set. We again stress that this effective energy is not meant to approximate the actual energy of a given configuration of atoms. It is simply a formal tool to promote descriptor diversity within atomic configurations.

A number of approaches have been proposed to numerically estimate the entropy of a distribution of descriptors. In the following, we adopt a simple nonparametric form where the local density is approximated using the first neighbor distance (in descriptor space)[27]. In this case, the estimator is of the form:

$$S(\{\mathbf{q}\}) = \frac{1}{n}\sum_{j=1}^{n} \ln(n \min_{l} \Delta q_{\text{j},l}) \tag{2}$$

where $\Delta q_{\text{j},l}$ is the cartesian distance between between the descriptors of atoms j and l (in descriptor space) and n is the number of atoms in the cell. This specific choice of entropy approximation is computationally convenient, but is not expected to be critical and other estimators could be used instead.

We recommend rescaling the different descriptors to a common scale in order to avoid the distance being dominated by only one or a few of them. In the following, typical values of descriptors have here been estimated from a set of preliminary simulations; based on these results, each descriptor has been renormalized to a value on the order of unity.

2. *Computational details*

The training set is incrementally constructed by adding independent local minima of the effective energy Eq. 1. As the effective potential (much like actual potentials) is rough, a simple annealing procedure was introduced, as illustrated in Fig.1. Note that the aim is not to locate the

FIG. 1: Schematic representation of the cyclic annealing procedure. The entropy term is initially turned off (K is set to zero) and the temperature T is set to a very high values. This thoroughly randomizes the configuration with respect to the previous one. Then, the temperature is linearly decreased and the strength of the entropy term is concurrently increased, so as to converge to a high-entropy configuration (and hence a low effective potential configuration), which is harvested at the end of each cycle. The cycle then repeats to generate additional configurations. The repulsive potential strength E_0 remains fixed at all times.

FIG. 2: Radial distribution function for an annealed configuration obtained with $K = 0$ eV (purple), $K = 1000$ eV (green), $K = 2000$ eV (blue), and $K = 3000$ eV (yellow).

global minimum of the effective energy (which would beat the purpose as repeating this procedure would not generate a diverse set) but simply to avoid trapping in low entropy configurations. The annealing procedure proceeds through a simultaneous ramping down of the temperature and ramping up of K. The goal is to initially favor a thorough shuffling of the atomic positions and avoid correlations between successive configurations by using a high temperature ($10,000$K) and no entropy bias. Entropy maximization is then gradually favored by linearly decreasing the temperature down to 0 and ramping up K to 1000 eV. The resulting configuration is then harvested and added to the training set. The cycle then simply repeats as many times as needed.

The maximal value of entropy scaling factor K was empirically tuned so as not to overwhelm $E_{\text{repulsive}}$ while still providing a strong driving force for the maximization of the entropy. As shown in Fig. 2, increasing K too much yields configurations where some pairs of atoms become separated by very short distances. Large values of K also yield stiff effective potentials that are prone to instabilities during annealing. We therefore settled on a maximal value of $K = 1000$ eV. Note that the specific choice of K and E_0 depends on the number of atoms in the simulation cell, as the entropy so-defined is intensive, but the repulsive contribution is extensive. Their value should therefore be readjusted as needed.

The spatial scale of the problem was chosen to be representative of tungsten atoms, but the training set is fully generic, and can therefore be rescaled as needed to describe other elements. In the following, we used a fully-periodic cell containing $n = 39$ atoms with a volume of 9.54x9.54x14.31 Å. This corresponds to a density of 0.03 atoms/Å^3, as compared to a bulk BCC density of 0.062 atoms/Å^3 for tungsten. The number of atoms was chosen so that cubic-scaling DFT calculations would be affordable whereas the volume was chosen so that both high and low density regions could coexist within the same simulation cell, thereby creating configurations that contain to bulk, surfaces, and voids. While even larger volumes allowed for higher descriptor entropy because of additional opportunities to create complex atomic arrangements, local minima of the effective energy at low density tend to contain high proportion of 1D filament-like structures and of gas-like configurations. If such configurations are deemed relevant, a training set can constructed by combining a range of different cell sizes. This possibility will be explored a future study. The simulation cell was chosen to be elongated in one direction, so as to facilitate the formation of free surfaces within the cell (which typically, but not always, form parallel to the long axis). The specific of the cell size, shape, and number of atoms is not critical. The were loosely set in order to discourage the formation of very ordered configurations, which could be promoted e.g., if the cell size was a multiple of the hard core radius, or if the density was so high that only closed-packed configurations could form.

In the following, atomic environments were described in terms of the so-called bispectrum components originally developed in the context of the Gaussian Approximation Potentials (GAP) potentials[28], and then adopted by the SNAP approach[29]. These descriptors are invariants of an expansion of the the density of neighboring atoms around a central atom in terms of hyperspherical harmonics. They are attractive because they are rotationally and permutationally invariant, which facilitates the development of energy expressions that inherit from these same properties. Progressively higher-order components then capture increasingly fine details of the distributions of neighboring atoms. Details of the computation of the bispectrum components can be found in the original publications[28]. The results presented below used the first 6 bispectrum components to characterize each atomic environment.

In the following, $E_{\text{repulsive}}$ follows the form proposed by Clarke and Smith[30]:

$$E_{\text{repulsive}} = \sum_i \sum_j \frac{E_0}{n-m}\left[m\left(\frac{r_0}{r_{ij}}\right)^n - n\left(\frac{r_0}{r_{ij}}\right)^m\right] \quad (3)$$

with $E_0 = 1$ eV, $n = 8$, $r_0 = 2.7$ Å, and $m = 4$. The potential was truncated at $r = 2.71$Å, and shifted to zero at the cutoff so as to capture only the repulsive part of the potential. The results are not expected to be sensitive to the specific form of the repulsive potential, as its only purpose is to enforce excluded volumes around each atom.

FIG. 3: Examplar configurations generated with the entropy maximization approach. Most configurations appear thoroughly random with no clear order.

III. RESULTS

1. *Characterization of the descriptor diversity*

10,000 configurations of $n = 39$ atoms were generated using the procedure described above. A few representative configurations are shown in Fig. 3. The ensemble of these configurations is referred to as the "biased" dataset. As a point of comparison, we compare the results with a so-called "unbiased" reference dataset, where configurations were sampled from an MD simulation at a temperature of $T = 10,000$K with $K = 0$, i.e., without attempting to maximize the entropy but while enforcing excluded volume constraints. This would intuitively roughly correspond to sampling from a high-temperature soft-sphere gas. As expected, the average descriptor entropy of configurations in the biased set ($S \sim 4.4$) is larger than that of the unbiased set ($S \sim 3.2$), which reflects the explicit promotion of diversity enforced by the entropic term.

The consequences of this increase in entropy can be appreciated by contrasting the distribution of individual descriptors over the biased and unbiased sets, as shown in Fig. 4 for the first bispectrum component, which is a measure of the local density around each atom. The distribution over the biased set is clearly much broader than its unbiased counterpart, which was the intended behavior. This also shows the importance of using cells with an overall low density, which allows for the formation of both high and low density regions within the same cell. This shows that, even if the entropy maximization was applied locally to each configuration, the procedure yields broad distributions over the whole training set. Perhaps surprisingly, computing high ($>$ 6th) order descriptors shows that their distribution is also broadened in the biased set. A multiple correlation analysis indicates that this results from linear dependence between descriptors; on average, we observe a correlation coefficient of about 0.7 for high-order descriptors against the first 6. Therefore, in this specific case, the entropy-generated training set exhibits broader descriptor distributions, even for descriptors that were not explicitly biased, which limits the need to extend the dimension of the biased space.

FIG. 4: Distribution of the first descriptor. Left: biased dataset; Right: unbiased dataset. The biased dataset shows a much broader distribution of values, indicative of higher diversity.

FIG. 5: Distribution of the eight descriptor. Left: biased dataset; Right: unbiased dataset. The biased dataset shows a much broader distribution of values, indicative of higher diversity.

2. *Error estimations on trained potentials: biased vs unbiased datasets*

The purpose of this work is not to train an actual potential for W, but to demonstrate that high-diversity training sets can be generated. We therefore did not generate quantum reference data for our training set. The potential impact of the increased diversity on accuracy and transferability can nonetheless be estimated in an ML scenario where energies and forces are computed through a gaussian process regression (GPR) constructed using the training set[31,32] as was done in the GAP approach[7,33]. For the purpose of this simplified analysis, we consider the GPR to act as an interpolator that exactly reproduces reference data at training points and use the distance to the nearest training point (in descriptor space) as a surrogate for the error in predictions at arbitrary test points. In reality, the rate of variation of the energy with respect to position in descriptor space would add an additional contribution to the prediction error of the GPR, as would forms of regularization. The shift and scale transformation that renders the distribution of each descriptors *in the unbiased dataset* mean free and unit variance was applied to both training and testing sets in order to uniformize the scales of each descriptors. In the following, the training sets contains 6000 randomly selected atomic environments, and testing sets 3000. Results were averaged over 1000 random decompositions between testing and training sets. The unit-less distances were measured in the 6-dimensional space spanned by the renormalized descriptors.

This nearest-neighbor distance in descriptor space is used to first compare the quality of the unbiased and biased datasets. Table I shows the mean nearest-neighbor descriptor distances obtained using different combinations of training and testing sets. The absolute value of the reported errors, as it measures a distance in descriptor space does not have a simple physical interpretation, beyond the expectation that they are proportional to the error in a trained ML potential. However, the values can be directly compared between the different choices of training and testing sets. Training on the unbiased set performs well (i.e., the mean error is low) when testing points are also sampled from the unbiased set. The distances however become very large when testing points are sampled from the biased set. This is a reflection of the fact that the distribution of descriptors in the unbiased set has a relatively narrow support: the training distribution is therefore dense over the support, yielding small distances when the test points fall within the support, but potentially very large distances when the test points fall outside of the support (e.g., when test points are sampled from the biased distribution). In this latter case, the GPR is extrapolating, sometimes leading to very large distances, as shown by the mean being significantly larger than the median distance, and by the maximum distance being very large. This illustrates an important trade-off: a potential trained on a narrow set of configurations can be expected to do well when used on configurations close to this narrow set; it will however do poorly when departing from it. In contrast, the distances obtained from training on the biased set show little dependence on the nature of the testing set, as the GPR is not forced to extrapolate outside of the support of the training set. Another apparent trade-off is that the mean distance when training on the biased set and testing on the unbiased set is higher than that observed when training and testing on the unbiased set. This follows from the inverse relationship between the size of the support and the density of points in descriptor space when the number of training points if fixed. If one is aiming at a high-accuracy potential that is only valid in a narrow region of descriptor space, lower errors can be achieved for a given amount of training data. However, if one is instead aiming at transferability, the biased training set is clearly superior to the unbiased one, as it limits opportunities for very large errors that can occur when the ML method is forced to extrapolate.

3. *Error estimations on trained potentials: biased vs hand-crafted datasets*

A more stringent test of our approach is to compare the biased training set with a dataset that was "hand-crafted" by domain experts. To this end, we select two training that were used in the development of several recent potentials for tungsten. The first, hand-crafted #1,[7,33] contains elastically deformed crystalline configurations, configurations harvested from DFT-based MD at high temperature, liquid configurations, various surfaces, and a range of defects (interstitials, dislocations, vacancies, and stacking faults), for a total of about 300,000 local W environment. The second training set (hand-crafted #2) was designed with a particular focus on radiation damage and defects[34]. This second training set includes a subset of hand-crafted set #1, adding isolated atoms and dimers, disordered surfaces, and configuration containing atoms approaching at short range (which is common in radiation damage applications), for a total of about 40,000 local W environments.

Tables II and III reports the results of the distances to the nearest training point for training and testing sets drawn from the hand-crafted sets #1, and #2, respectively, and the biased set. For the purpose of that analysis, the training sets contain 2000 randomly selected atomic environments and 1000 for the testing sets. Results were averaged over 1000 random decompositions between testing and training sets. The results are largely similar to that observed when comparing to the unbiased set. Training and testing from the hand-crafted set yields low errors as the support of both training and testing distributions is very narrow (c.f. purple histogram in Fig. 8), but these increase dramatically upon switching the test set to the unbiased set, again because extrapolation is then required (c.f. the very long tail in the green distribution in Fig. 8). In contrast, training from the biased set yields results that are similar for both testing sets (c.f. blue and red histograms in Fig. 8). These results suggest that the entropy-biased training set should be competitive with or even superior to the hand-crafted set as it contains a more diverse distribution of atomic environments.

FIG. 6: Distribution of the first descriptor. Left: biased dataset; Right: hand-crafted #1 dataset.

FIG. 7: Distribution of the eight descriptor. Left: biased dataset; Right: hand-crafted #1 dataset.

Close analysis reveals that this characterization comes with caveats. Indeed, as shown in Fig. 6, the distribution of descriptors is in general significantly wider in the biased set than in the hand-crafted set. However, the distribution of some descriptors in the hand-crafted set is strongly peaked, as it contains a high proportion of crystalline local environments. In some cases, the peak falls into a region where the density descriptors in the biased dataset is low, c.f. Fig. 7, which can limit the accuracy of predictions carried out using the biased set alone for training. For example, it can be seen in Tables II and III and errors are larger when testing from the hand-crafted sets and training from the biased set than when testing and training on the biased set. This is an indication that some regions that are well represented in the hand-crafted sets are relatively sparser in the biased dataset. This effect becomes stronger when the space of descriptors in which the GPR interpolation is carried out increases to tens or hundreds of dimensions. Note however that even in this case, the errors remain below that of training with the hand-crafted set and testing with the biased set. This limitation could potentially be addressed by increasing the size of the biased space or by also explicitly favoring high-symmetry local order, a strategy that we are currently exploring.

This observation illustrates the tradeoffs discussed above: if one seeks an highly accurate potentials that is valid in a small region of the possible configuration space of the problem (e.g. in BCC crystalline configurations), a narrow but tailored training set, such as the hand-crafted ones, is likely to perform better; on the other hand, if transferability is paramount, an automated approach that explicitly favors diversity as the one proposed here is highly beneficial. Indeed, generating a very large and diverse training set by by hand is extremely challenging, which is why most training sets are restricted to crystalline phases, liquids, and simple defects. Only including these cherry-picked configurations is however unlikely to cover the range of possible uses for interatomic potentials, especially in extreme conditions of temperature, pressure, stress, irradiation, etc, hence the importance of developing automated approaches such as the entropy-maximization method proposed here.

While the ultimate goal is to completely automate the creation of training sets, these two approaches can be bridged to achieve both high accuracy in known low-energy states and transferability to higher-energy configurations. For example, if very high accuracy is paramount in the BCC phase, e.g., for W, but high transferability is required elsewhere (e.g., to properly describe disordered configurations that can be created following radiation damage events), an unbiased dataset created by the entropy-maximization method can directly be combined with a hand-crafted set when training a potential.

FIG. 8: Distribution of the distances to the closest training point for different combinations of training and testing sets. Training from the biased set and testing from the hand-crafted #1 set (red); training and testing from the biased sets (blue); training and testing from the hand-crafted #1 set (purple); training from the hand-crafted #1 set and testing from the biased set (green).

IV. DISCUSSION

While the entropy-maximization method itself is very general, the importance of choosing an appropriate descriptor space should be emphasized. Indeed, the method only generates diversity in

TABLE I: Error estimations: biased vs unbiased datasets. The error metric correspond to unit-less (e.g., renormalized) nearest-neighbor distance in descriptor space between points in the testing set and training sets.

Training set	Testing set	Mean	Median	Max
Unbiased	Unbiased	0.2294	0.1827	5.9339
Unbiased	Biased	4.6949	3.5311	32.0983
Biased	Unbiased	0.9053	0.9068	2.3402
Biased	Biased	0.8541	0.7846	5.7246

TABLE II: Error estimations on trained potentials: biased vs hand-crafted #1 datasets. The error metric correspond to unit-less (e.g., renormalized) nearest-neighbor distance in descriptor space between points in the testing set and training sets.

Training set	Testing set	Mean	Median	Max
Hand-crafted	Hand-crafted	0.2263	0.1562	7.6514
Hand-crafted	Biased	3.6680	2.3153	36.4918
Biased	Hand-crafted	0.9923	1.0014	7.9049
Biased	Biased	0.8541	0.7846	5.7246

TABLE III: Error estimations on trained potentials: biased vs vs hand-crafted #2 datasets. The error metric correspond to unit-less (e.g., renormalized) nearest-neighbor distance in descriptor space between points in the testing set and training sets.

Training set	Testing set	Mean	Median	Max
Hand-crafted	Hand-crafted	0.1220	0.0510	5.1667
Hand-crafted	Biased	9.6168	8.2602	34.7536
Biased	Hand-crafted	2.9559	2.5148	20.4147
Biased	Biased	1.0927	1.0024	7.0593

the space spanned by the descriptors. If these descriptors are insensitive to some important physical or chemical features, improvements in diversity could be limited along these dimensions. The proposed implementation based on the bispectrum components can be expected to perform well for monoatomic systems. Extensions to multi-component materials are in principle straightforward in conjunction with descriptors that are sensitive to local chemical order, such as the explicit multi-element bispectrum components[35]. This generalization is currently underway. Further extension to molecular systems should also be possible in principle, but chemical constraints might have to be introduced to limit sampling to the chemically relevant subspace, thereby avoiding the allocation of resources to chemically inaccessible or irrelevant configurations.

V. CONCLUSIONS

We introduced a sampling-based approach for the automated training set generation of interatomic potentials. Configurations are generated by sampling low-lying minima of an effective potential energy function that explicitly favors the diversity of the local atomic environment through an entropy maximization process. The generated training set is shown to be more diverse than that generated by a random sampling procedure and even compared to hand-crafted sets used in state-of-the-art machine-learned potentials, which promises improved transferability. Extensions to global entropy-maximization over the whole training set (in contrast to the local configuration-by-configuration optimization presented here) is in development and will be reported in an upcoming publication.

VI. ACKNOWLEDGEMENTS

We thank Mitchell Wood and Nicholas Lubbers for stimulating discussions. This research was supported by the Exascale Computing Project (17-SC-20-SC), a collaborative effort of the U.S. Department of Energy Office of Science and the National Nuclear Security Administration. Los Alamos National Laboratory is operated by Triad National Security LLC, for the National Nuclear Security administration of the U.S. DOE under Contract No. 89233218CNA0000001.

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] S. J. Plimpton and A. P. Thompson, MRS Bulletin **37**, 513–521 (2012).

[2] A. P. Bartók and G. Csányi, International Journal of Quantum Chemistry **115**, 1051 (2015).

[3] P. E. Dolgirev, I. A. Kruglov, and A. R. Oganov, AIP Advances **6**, 085318 (2016).

[4] D. Dragoni, T. D. Daff, G. Csányi, and N. Marzari, Physical Review Materials **2**, 013808 (2018).

[5] J. S. Smith, O. Isayev, and A. E. Roitberg, Chemical science **8**, 3192 (2017).

[6] Y. Zuo, C. Chen, X. Li, Z. Deng, Y. Chen, J. Behler, G. Csányi, A. V. Shapeev, A. P. Thompson, M. A. Wood, *et al.*, The Journal of Physical Chemistry A **124**, 731 (2020).

[7] M. A. Wood, M. A. Cusentino, B. D. Wirth, and A. P. Thompson, Phys. Rev. B **99**, 184305 (2019).

[8] V. L. Deringer, C. J. Pickard, and G. Csanyi, Phys. Rev. Lett. **120** (2018).

[9] V. L. Deringer, D. M. Proserpio, G. Csanyi, and C. J. Pickard, Faraday Discuss. **211** (2018).

[10] E. V. Podryabinkin, E. V. Tikhonov, A. V. Shapeev, and A. R. Oganov, Phys. Rev. B **99**, 064114 (2019).

[11] S. Hajinazar, J. Shao, and A. N. Kolmogorov, Phys. Rev. B **95** (2017).

[12] S. Chmiela, H. Sauceda, K. Müller, and A. Tkatchenko, Nature Communications **9** (2018).

[13] K. Lee, D. Yoo, W. Jeong, and S. Han, Computer Physics Communications **242**, 95 (2019).

[14] H. Chan, B. Narayanan, M. Cherukara, F. Sen, K. Sasikumar, S. Gray, M. Chan, and S. Sankaranarayanan, The Journal of Physical Chemistry C **123** (2019), doi: 10.1021/acs.jpcc.8b09917.

[15] J. S. Smith, O. Isayev, and A. E. Roitberg, Chem. Sci. **8**, 3192 (2017).

[16] W. Jeong, K. Lee, D. Yoo, D. Lee, and S. Han, The Journal of Physical Chemistry C **122** (2018), doi: 10.1021/acs.jpcc.8b08063.

[17] V. Botu, R. Batra, J. Chapman, and R. Ramprasad, The Journal of Physical Chemistry C **121**, 511 (2017), https://doi.org/10.1021/acs.jpcc.6b10908.

[18] T. Huan, R. Batra, J. Chapman, S. Krishnan, L. Chen, and R. Ramprasad, npj Computational Materials **3** (2017).

[19] J. Behler, Journal of Physics: Condensed Matter **26**, 183001 (2014).

[20] S. L. Frederiksen, K. W. Jacobsen, K. S. Brown, and J. P. Sethna, Phys. Rev. Lett. **93**, 165501 (2004).

[21] E. Podryabinkin and A. V. Shapeev, Computational Materials Science **140**, 171 (2017).

[22] K. Gubaev, E. V. Podryabinkin, G. L. Hart, and A. V. Shapeev, Computational Materials Science

156, 148 (2019).

[23] R. Jinnouchi, F. Karsai, and G. Kresse, Phys. Rev. B **100**, 014105 (2019).

[24] J. Vandermause, S. B. Torrisi, S. Batzner, Y. Xie, L. Sun, A. M. Kolpak, and B. Kozinsky, npj Computational Materials **6**, 1 (2020).

[25] G. Csányi, T. Albaret, M. Payne, and A. De Vita, Physical review letters **93**, 175503 (2004).

[26] W. Pronobis, A. Tkatchenko, and K.-R. Müller, Journal of Chemical Theory and Computation **14**, 2991 (2018), pMID: 29750522, https://doi.org/10.1021/acs.jctc.8b00110.

[27] J. Beirlant, E. Dudewicz, L. Györfi, and I. Dénes, International Journal of Mathematical and Statistical Sciences **6**, 17 (1997).

[28] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, Phys. Rev. Lett. **104**, 136403 (2010).

[29] A. P. Thompson, L. P. Swiler, C. R. Trott, S. M. Foiles, and G. J. Tucker, Journal of Computational Physics **285**, 316 (2015), arXiv:1409.3880 [cond-mat.mtrl-sci].

[30] Clarke and Smith, J Chem Phys **84**, 2290 (1986).

[31] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning* (MIT Press, 2006).

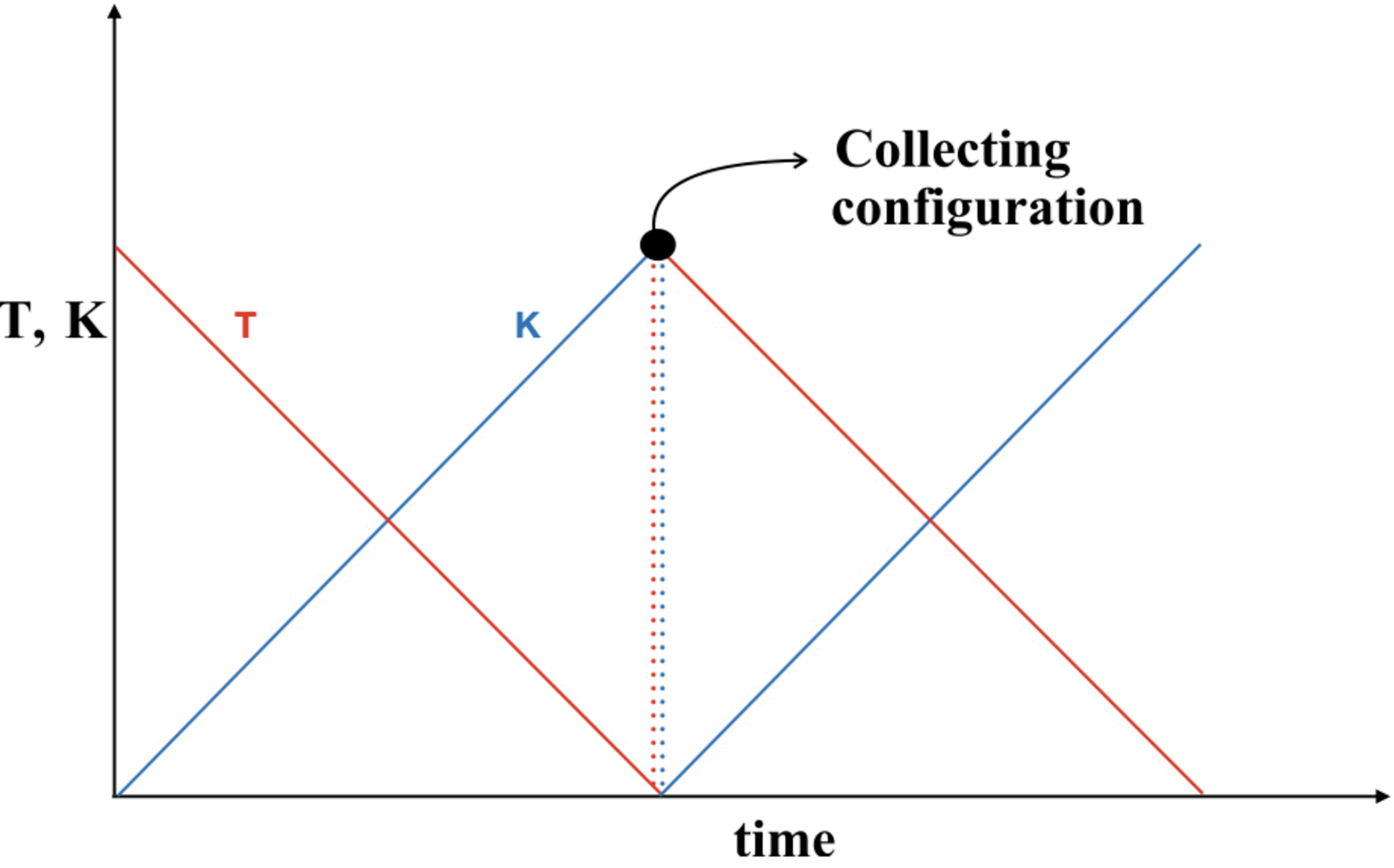
[32] C. Handley and J. Behler, The European Physical Journal B **87** (2014).

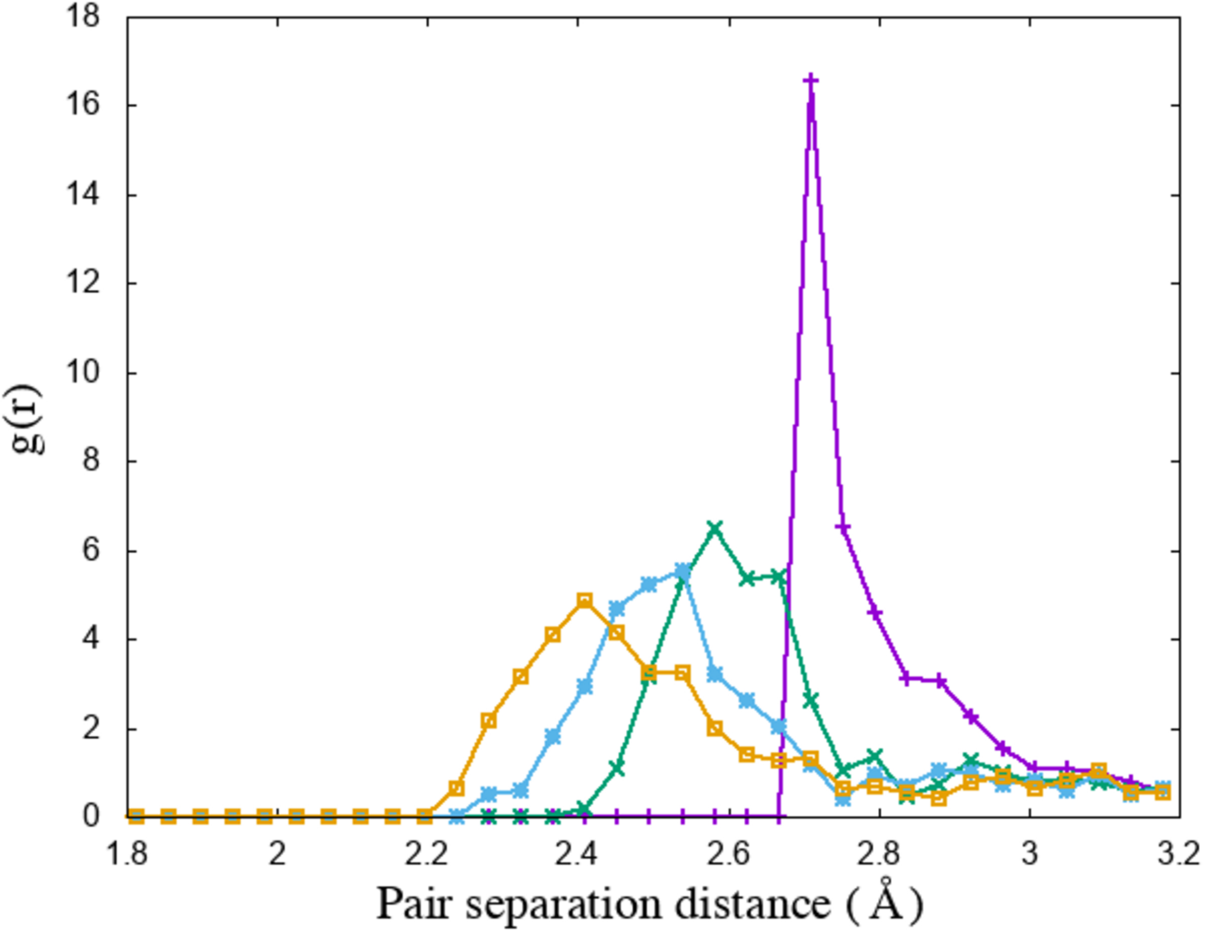
[33] W. J. Szlachta, A. P. Bartók, and G. Csányi, Phys. Rev. B **90**, 104108 (2014).

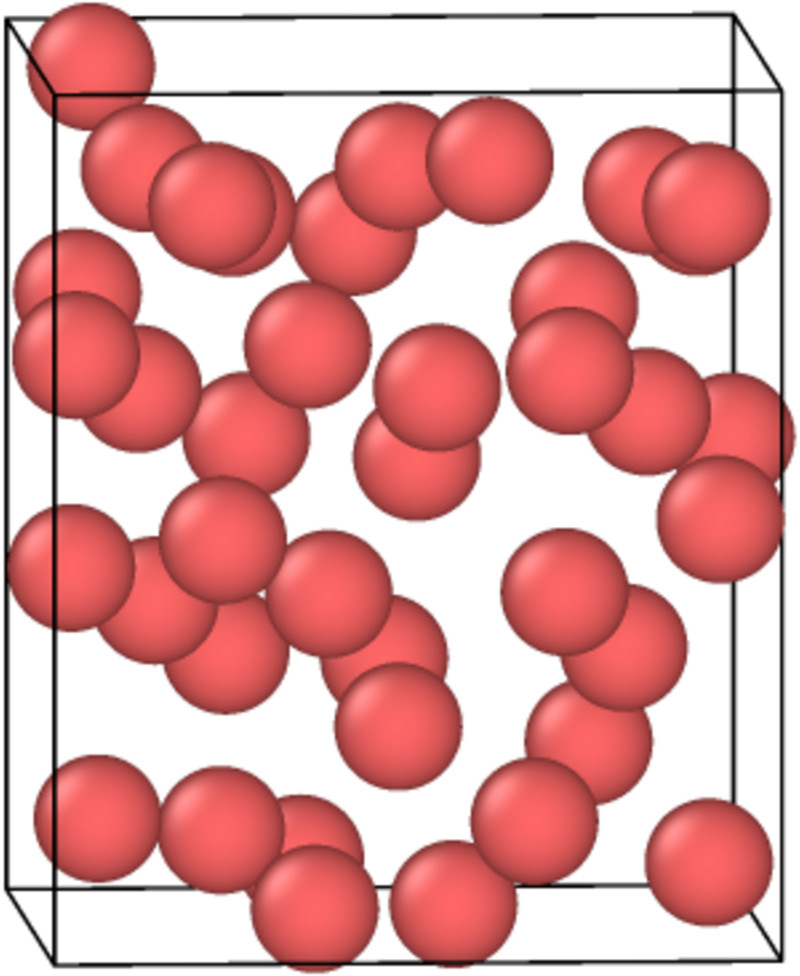
[34] J. Byggmästar, A. Hamedani, K. Nordlund, and F. Djurabekova, Phys. Rev. B **100**, 144105 (2019).

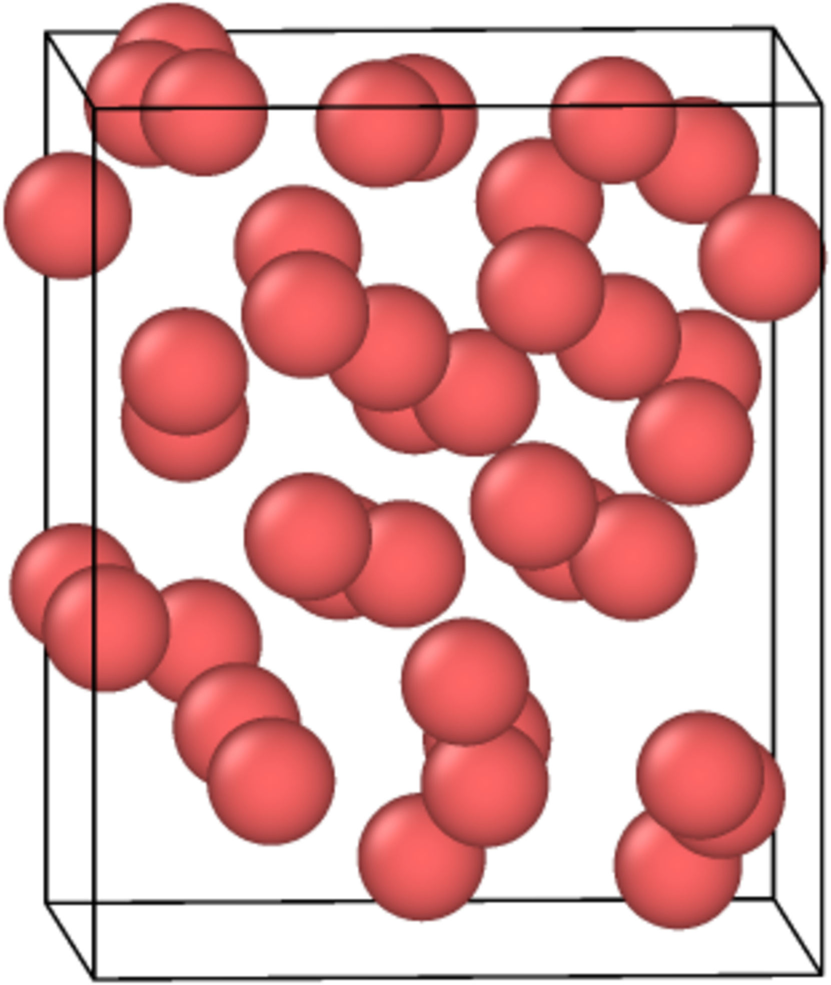
[35] M. A. Cusentino, M. A. Wood, and A. P. Thompson, The Journal of Physical Chemistry A **124**, 5456 (2020), pMID: 32432859, https://doi.org/10.1021/acs.jpca.0c02450.

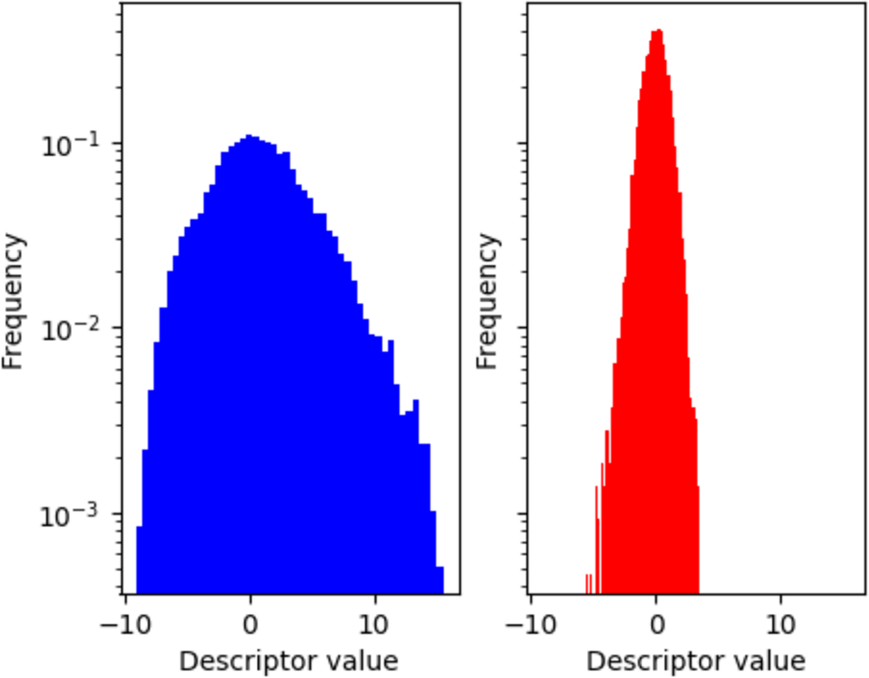
[36] E. V. Podryabinkin, E. V. Tikhonov, A. V. Shapeev, and A. R. Oganov, Physical Review B **99**, 064114 (2019).

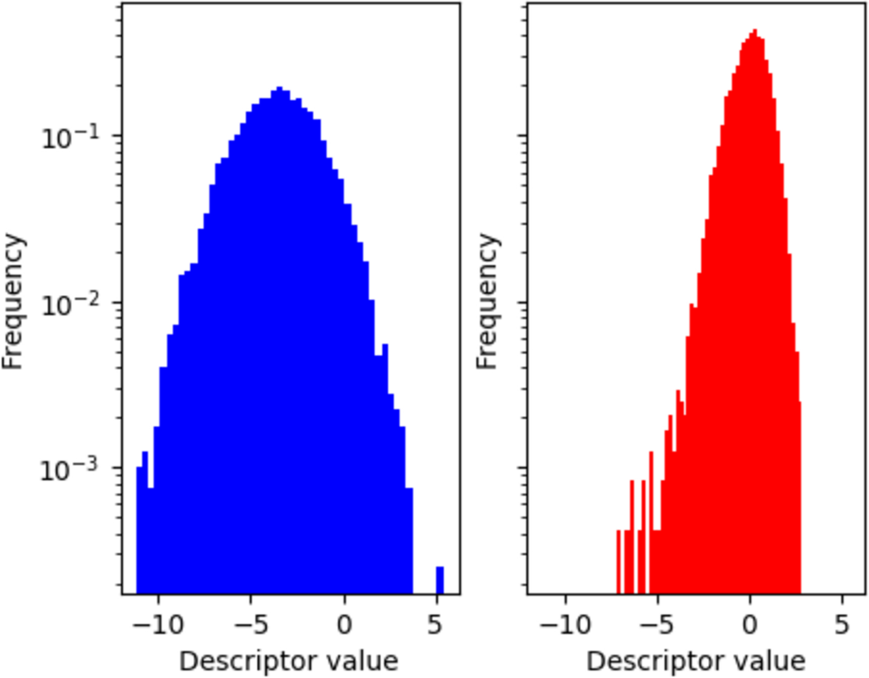

Collecting
configuration
T, K
T
K
time


18
16
14
12
10
8
6
4
2
0
g(r)
1.8
2
2.2
2.4
2.6
2.8
3
3.2
Pair separation distance (Å)

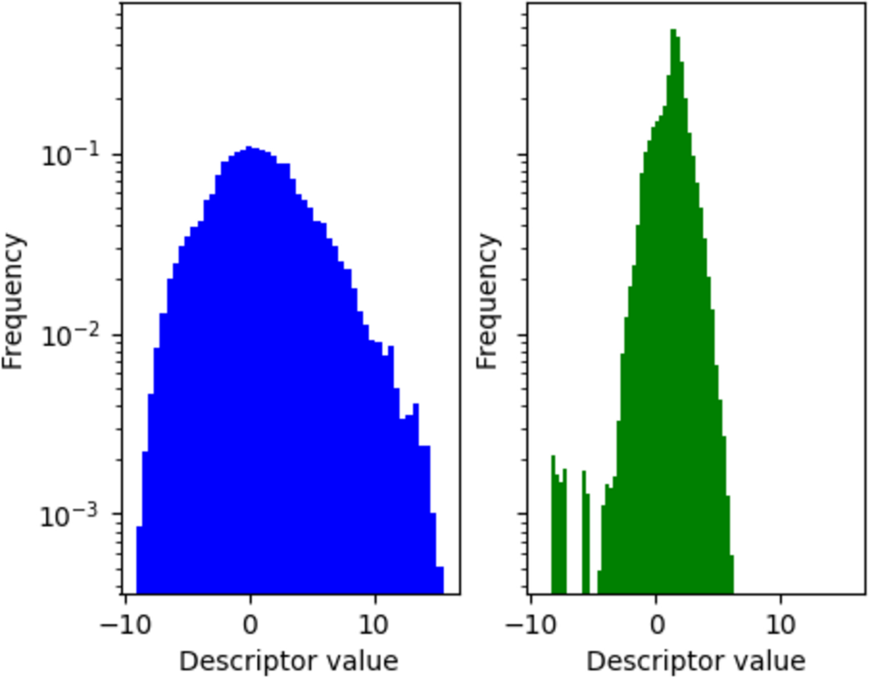




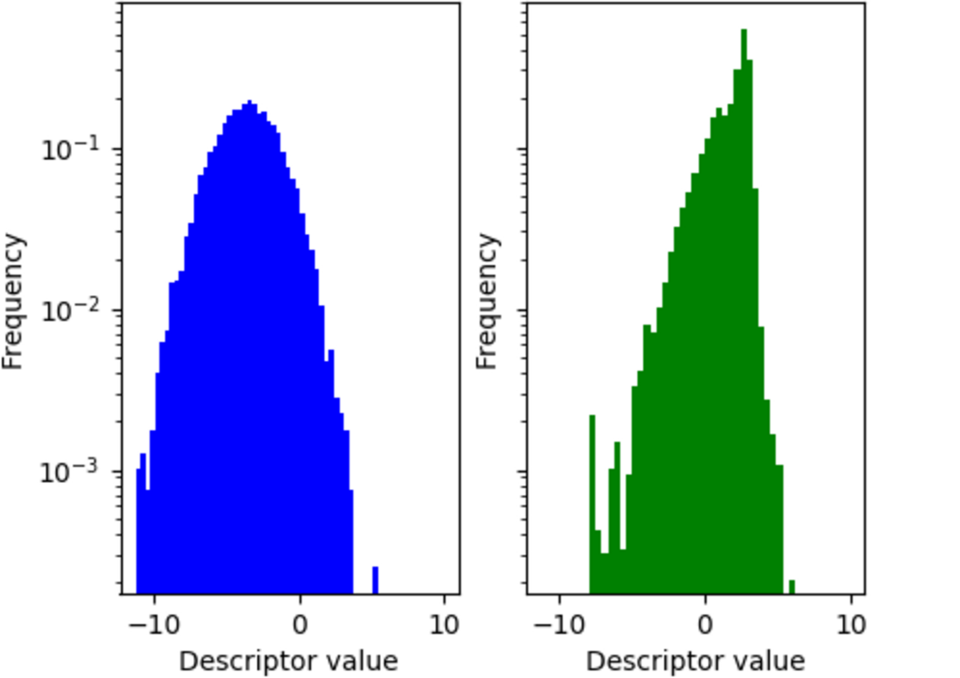

Frequency
10^{-1}
10^{-2}
10^{-3}
−10
0
10
Descriptor value
Frequency
−10
0
10
Descriptor value

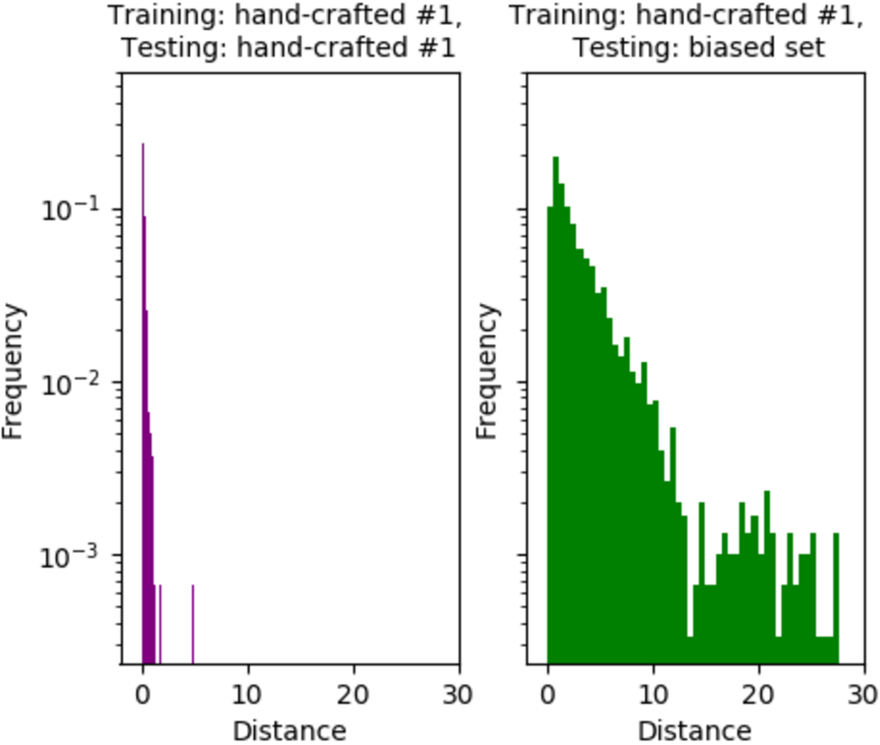


Frequency
10^{-1}
10^{-2}
10^{-3}
−10
−5
0
5
Descriptor value
Frequency
−10
−5
0
5
Descriptor value



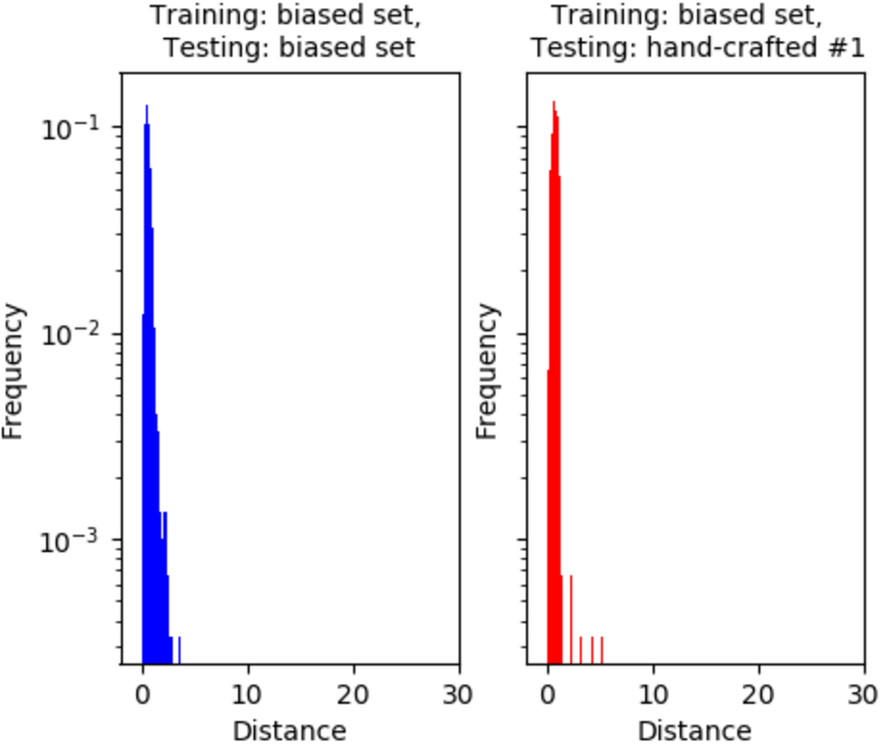
10^{-1}
10^{-2}
10^{-3}
Frequency
−10
0
10
Descriptor value
Frequency
−10
0
10
Descriptor value


Frequency
10^{-1}
10^{-2}
10^{-3}
−10
0
10
Descriptor value
Frequency
−10
0
10
Descriptor value



Training: hand-crafted #1, Testing: hand-crafted #1

Training: hand-crafted #1, Testing: biased set


Training: biased set,
Testing: biased set
Training: biased set,
Testing: hand-crafted #1
Frequency
10^{-1}
10^{-2}
10^{-3}
0
10
20
30
Distance
Frequency
0
10
20
30
Distance