Less is more: Sampling chemical space with active learning

Justin S. Smith,[1] Ben Nebgen,[2] Nicholas Lubbers,[2] Olexandr Isayev,[3,a)]
and Adrian E. Roitberg[1,a)]
[1]*Department of Chemistry, University of Florida, Gainesville, Florida 32611, USA*
[2]*Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA*
[3]*UNC Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA*

(Received 28 January 2018; accepted 2 May 2018; published online 22 May 2018)

The development of accurate and transferable machine learning (ML) potentials for predicting molecular energetics is a challenging task. The process of data generation to train such ML potentials is a task neither well understood nor researched in detail. In this work, we present a fully automated approach for the generation of datasets with the intent of training universal ML potentials. It is based on the concept of active learning (AL) via Query by Committee (QBC), which uses the disagreement between an ensemble of ML potentials to infer the reliability of the ensemble's prediction. QBC allows the presented AL algorithm to automatically sample regions of chemical space where the ML potential fails to accurately predict the potential energy. AL improves the overall fitness of ANAKIN-ME (ANI) deep learning potentials in rigorous test cases by mitigating human biases in deciding what new training data to use. AL also reduces the training set size to a fraction of the data required when using naive random sampling techniques. To provide validation of our AL approach, we develop the COmprehensive Machine-learning Potential (COMP6) benchmark (publicly available on GitHub) which contains a diverse set of organic molecules. Active learning-based ANI potentials outperform the original random sampled ANI-1 potential with only 10% of the data, while the final active learning-based model vastly outperforms ANI-1 on the COMP6 benchmark after training to only 25% of the data. Finally, we show that our proposed AL technique develops a universal ANI potential (ANI-1x) that provides accurate energy and force predictions on the entire COMP6 benchmark. This universal ML potential achieves a level of accuracy on par with the best ML potentials for single molecules or materials, while remaining applicable to the general class of organic molecules composed of the elements CHNO. *Published by AIP Publishing.* https://doi.org/10.1063/1.5023802

I. INTRODUCTION

The development of accurate force fields[1–3] for the efficient simulation of large and small molecular systems has been a cornerstone of modern computational chemistry. The popularity of force fields is driven by low computational cost relative to more accurate and transferable quantum mechanical (QM) methods, such as density functional theory[4] (DFT) or post-Hartree-Fock[5–7] methods. However, parametrizing *universal* force fields—applicable to any chemical system in any chemical environment—has remained an elusive goal due to the restrictive functional form and tedious atom typing of classical force fields. For this reason, a "zoo" of force fields has been developed over the last 30 years with applications in various regions of chemistry and physics, such as materials, proteins, carbohydrates, and small drug-like molecules.[8–11] Drawing a line between where these system-specific force fields work and where they fail is a challenging task.

In recent years, machine learning (ML) methods have been successfully applied in many areas of chemistry and physics research.[12–19] Specifically, ML approaches for the prediction of interatomic potential energy surfaces (referred to as ML potentials) have exhibited chemical accuracy compared to QM models at roughly the computational cost of classical force fields.[20–31] ML potentials promise to bridge the speed vs. accuracy gap between force fields and QM methods. Many recent studies rely on a philosophy of parametrization to one chemical system at a time,[22,25] single component bulk systems[27,28] or many equilibrium structures, i.e., QM7 and QM9 datasets.[32,33] While parametrization to one system at a time can achieve high accuracy with relatively small amounts of QM, it has the downside that one must generate additional QM data and train a new ML model for every new chemical system. Using this approach in any study requires extra parametrization time due to the non-universality of the potentials. Additionally, parametrization to only equilibrium geometries does not attempt to describe the range of conformations visited in atomistic simulations. For these reasons, single system and equilibrium dataset ML potentials do not aim to build an extensible and transferable (universal) ML potential.

Our work on the ANAKIN-ME (ANI) method for developing the ANI-1 potential[34] is one example of a universal ML atomistic potential for organic molecules. The methodology

[a)]Authors to whom correspondence should be addressed: olexandr@olexandrisayev.com and roitberg@ufl.edu is built upon the concept of an atomic environment descriptor first developed by Behler and Parrinello[35] and refined to perform significantly better on large and diverse datasets of organic molecules. A key aspect of the ANI methodology was the focus on dataset diversity, which promotes the learning of low level interactions (by utilizing localized descriptors) for better transferability. For training the ANI-1 model, we calculated over 22 × 10^6 structural conformations from 57 000 distinct small organic molecules using DFT.[36] The ANI-1 dataset was built through an exhaustive sampling of molecules containing between one and eight C, N, and O atoms from the GDB-11 database, with H atoms added to saturate the configurations. The ANI-1 dataset is built on a philosophy of dataset construction that samples small molecule *conformational and configurational space* at the same time. The ANI-1 potential was shown to be chemically accurate for systems of 50 atoms and more, demonstrating extensibility and transferability to much larger molecules than those in the training set. This phenomenon, whereby an ML model is trained on small systems (which could be thought of as fragments of large systems), then demonstrated to be extensible to large systems has also been confirmed in other recent studies.[29,37,38] Other recent work had success in developing universal ML property predictors for organic based chemical systems away from their local minima.[29]

When it comes to developing or optimizing ML model training datasets, human intuition currently drives the experiment design. The resulting datasets tend to be clustered, sparse, and incomplete; recent work finds that people tend to favor inclusion of "successful" experiments and tend to forget "failed" experiments.[39] The comprehensive incorporation of all data is the strength of ML approaches to artificial intelligence (AI). With sufficient data, an AI-driven machine can more effectively choose the next step in experiments or simulations than humans, speeding up the optimization of a given dataset, while also reducing the overall amount of data required. As robotics transforms chemical synthesis,[40] manufacturing, and transportation, constituting a modern industrial revolution,[41,42] achieving the analogous revolution in computational methods will require AI and, in particular, the emulation of scientific intuition, reasoning, and decision making. Such an ambitious program will not be accomplished all at once and will instead require incremental progress as AI algorithms are developed.

In this work, we present a fully automated approach of dataset generation for training universal ML potentials. It is based on the concept of active learning (AL) which has been successfully applied to develop single system ML potentials[37,43–46] and in other areas[47,48] of chemical sciences. We develop a two-component technique for training universal ML potentials. The first component is a dataset reduction algorithm for eliminating redundancy in an existing training set. The second is an active learning algorithm based on the Query by Committee[49] (QBC) approach for selecting new training data. For a complete and rigorous validation of universal potentials, we also develop the COmprehensive Machine-learning Potential (COMP6) benchmark suite for organic molecules and bio-molecules. The COMP6 benchmark samples the chemical space (for molecules containing C, H, N, and O) of molecules larger than those included in the training set, as well as non-covalent interactions via the S66x8 benchmark.[50] The COMP6 benchmark is publicly available on GitHub [https://github.com/isayev/COMP6]. Using the active learning scheme, a potential can be trained to the accuracy of ANI-1 using 90% less data, even while sampling from smaller molecules. After further exploration of chemical space, our potential (dubbed ANI-1x) strongly out-performs ANI-1, while being trained on a dataset that is only 25% of the size.

II. METHODS

In the context of this work, the goal of active learning is to infer an accurate predictor from labeled training data. These labeled data are input-output pairs (X, y), where the output y represents the correct answer to a question associated with the input X. In the problem of ML potential training, the label y may be the "yes"/"no" answer to whether the potential correctly describes a molecule X. As part of the active learning process, this question may be answered empirically for a given substance. The Query by Committee (QBC) approach uses the disagreement between models trained to similar data to experimentally infer the correctness of an ensemble's prediction. This is by the following reasoning: if an ensemble of predictors has a high variance, then some models in the ensemble must have a relatively high error from the ground truth. Therefore, selection of compounds that have a high variance of ensemble predictions in search of new molecules and conformation can be employed to sample high error regions' chemical space automatically, minimizing the need for redundant QM calculations. Several studies provided empirical evidence that this method of sampling indeed improves the overall fitness of ML potentials for single systems.[37,51] In this work, we apply this concept in a massive search of chemical space to develop a superior training set for universal ML ANI[34] potentials. These ANI potentials are applicable to organic molecules containing C, H, N, and O. With minimal modification, the same approach could be used for other areas of chemical sciences, e.g., materials.

A. Sample selection via Query by Committee

We show how, in a rigorous statistical way, one can obtain *a priori* information about what new samples should be included in subsequent generations of an ML potential training set. The *a priori* information is obtained by the QBC[49] algorithm. QBC measures the disagreement between students (models) of a committee (ensemble); then, the algorithm selects new examples where the students disagree by a preset inclusion criterion. Finally, new reference data for selected examples are obtained and included in the next committee training iteration. As a test of agreement, we choose to include new data point i only for test cases which generate a value ρ_i greater than an inclusion criterion $\hat{\rho}$, where ρ_i is defined as

$$\rho_i = \frac{\sigma_i}{\sqrt{N_i}}. \tag{1}$$

In Eq. (1), σ_i is the standard deviation of predictions from an ensemble (see Sec. II E for details) of ANI potentials and N_i is the number of atoms in the given test system. The square root is applied to N_i since the potentials are atomistic, and the total energy error is assumed to be a random distribution, centered around zero, per atom. That is, cancellation of error on a per atom basis can lead to artificially low per atom errors (and standard deviations in this case) on larger molecules when a square root is not applied. This is necessary when using a single value of $\hat{\rho}$ to test across molecules with varying numbers of atoms as is done in this work.

Figure 1 provides an example of how the inclusion criterion $\hat{\rho}$ is determined. In this 2-dimensional density plot, $\varepsilon_i = \left|MAX\left(\left\{E_T^{ANI}\right\}_i^{ens} - E_{T,i}^{REF}\right)\right|/\sqrt{N_i}$, where N_i is the number of atoms in the i-th molecule. Therefore, ε_i is the largest per atom prediction error of any model in the ensemble of ANI models for test molecule i. The test data used in this example is the GDB07to09 test set which is described in Sec. II C. The ANI model used to determine $\hat{\rho}$ in this example is the ANI model which initialized the AL process (Sec. II B). The value $\hat{\rho}$ is determined from the choice of what value of ε is considered too large and what percentage of epsilon over that should be considered as fail cases. Therefore, $\hat{\rho} = 0.23$ was empirically selected as it is the value which allows selection of 98% of all $\varepsilon_i > 1.5$ kcal/mol.

The example from Fig. 1 shows that $\hat{\rho} = 0.23$ kcal/mol selects 58% of all test data as fail cases. As evidence that the chosen definition of ε_i allows for the statistical determination of poorly fit data, it is shown that before selecting any data (i.e., for all ρ_i), 26% of the complete test set ε_i are greater than 1.5. However, this is 44% when considering all $\varepsilon_i > 1.5$ kcal/mol which correspond to $\rho_i > \hat{\rho}$. This shows that the determined $\hat{\rho}$ leads to a selection of data with a greater number of $\varepsilon_i > 1.5$ kcal/mol within its population. As further validation of the approach, the application of the concept is shown to choose "bad" data by calculating the root mean squared error (RMSE) of the potential energy (E) for the mean prediction of the ensemble of ANI models vs. reference DFT calculations. For all i molecular structures corresponding to $\rho_i > \hat{\rho}$, the E RMSE is 7.4 kcal/mol. On the other hand, for all i molecular structures corresponding to $\rho_i \leq \hat{\rho}$, the E RMSE is 1.5 kcal/mol. Therefore, in a statistical way, the method chooses new data which are significantly higher in error compared to GDB07to09 which are randomly generated data.

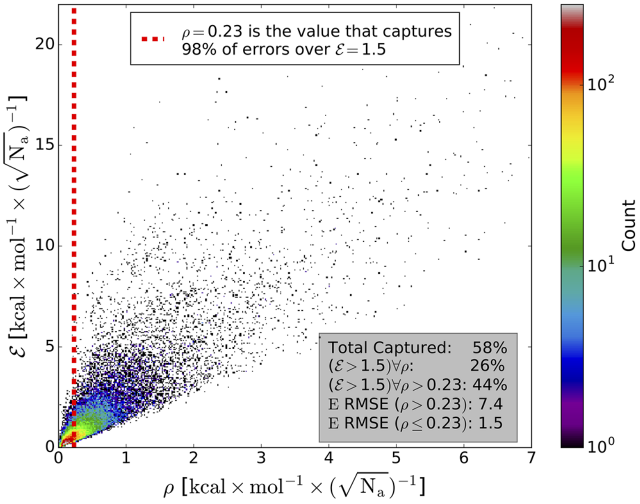

FIG. 1. Example of choosing a value $\hat{\rho}$ which captures 98% of all errors (ε) over 1.5 kcal/mol on the GDB07to09 benchmark set using the initial (before using active learning) ANI model ensemble. The value which accomplished this is found to be $\hat{\rho} = 0.23$. This value of $\hat{\rho}$ used in Query by Committee results in the selection of 58% of all test data. Initially 26% of all ε are greater than 1.5. 44% of ε corresponding to $\rho > \hat{\rho}$ are greater than 1.5. Splitting the dataset along $\rho = \hat{\rho}$ results in a total energy RMSE of the ANI ensemble prediction vs. reference DFT of 7.4 kcal/mol for all values $\rho > \hat{\rho}$ and 1.5 kcal/mol for all values $\rho \leq \hat{\rho}$.

With enough processing time on HPC resources, the rate-limiting step of a QBC data selection cycle using ANI potentials is the training of a new ensemble of ANI models. Complete training of a single network takes 40 min per one million data points on a single NVIDIA Tesla V100 GPU. To reduce the number of models trained, QBC is used in batches, searching configurational and conformational (chemical) space for tens of thousands of new reference data points that fail the agreement test. Finally, labels (reference potential energies, E^{REF}) are computed for all molecules in the selected batch. This process may lead to some redundant data. However, the alternative, retraining a new model ensemble after the addition of every new data point, will be impractically slow.

B. Automatic chemical space sampling via active learning

Figure 2 shows the overall workflow of the iterative AL algorithm. The algorithm is initialized from an existing random sampling generated dataset which may contain some amount of redundant data. This initial dataset (ANI-1 in this work) is then reduced through an iterative approach with the goal of minimizing the overall dataset size, while not impacting predictive performance. The reduction algorithm is provided in detail in Fig. 2(a). Figure 2(a) is initialized with a random subsampled 2% of the original ANI-1 dataset. Then, iteratively, the remaining data are tested, and 2% subsets of the fail cases are added to the training set. Here, a fail case is defined as $|E_{ANI} - E_{DFT}|/\sqrt{N} > 0.04$ kcal/mol, where N is the number of atoms in the molecule. The algorithm is terminated when less than 5% of the data not yet added to the training set are considered as fail cases. The remaining <5% high error data are added to the final dataset. Hyper-parameters for the reduction algorithm can be tuned to further reduce redundancies in the data, at the cost of more cycles, and therefore, longer run time. The final reduced dataset is used to bootstrap the remaining cycles of the active learning algorithm. If a dataset such as ANI-1 is not available, this step can be replaced with the generation of a small amount randomly sampled data across many small, one to five C, N, O atoms, molecules. However, this will lead to more active learning cycles before achieving the desired result.

With the reduced dataset, the configurational search [Fig. 2(b)] is initialized. The configurational search is carried out by randomly sampling an external database of small molecules [e.g., GDB-11,[52,53] ChEMBL,[54–56] algorithmically generated dipeptides using RDKit (www.rdkit.org), automatically generated dimers], embedding the molecule in 3D space

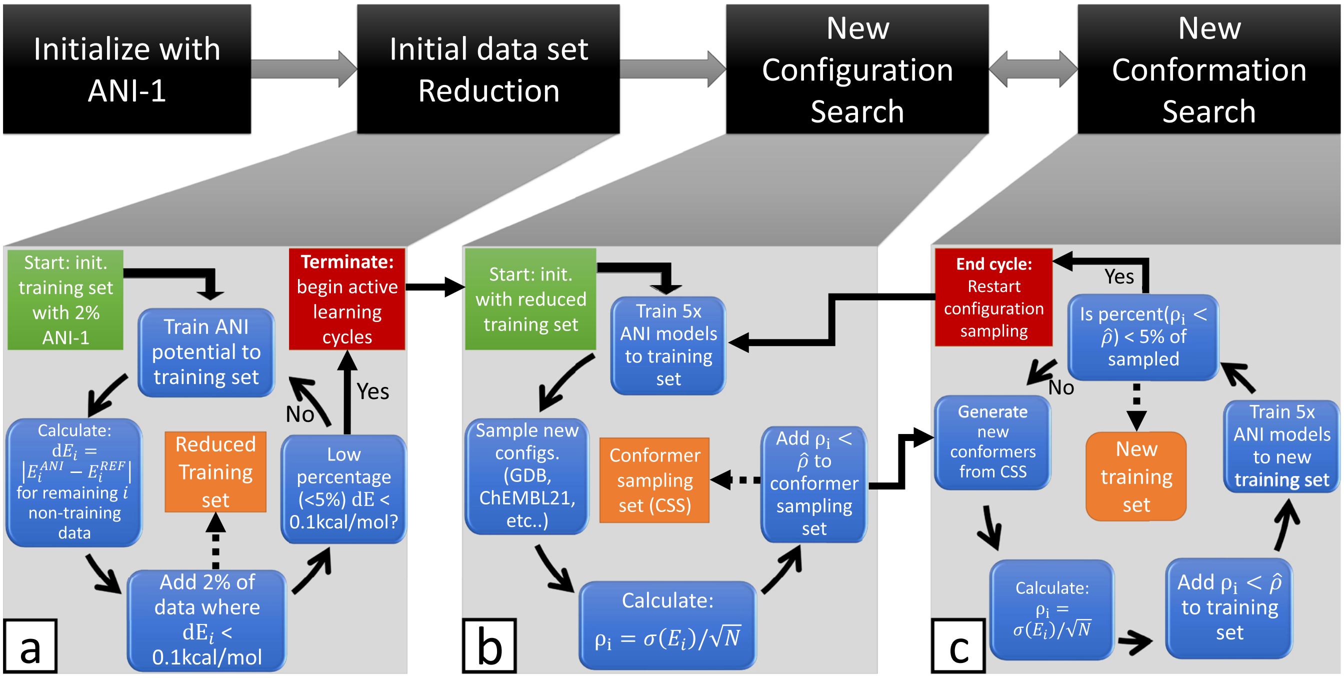
Initialize with ANI-1
Initial data set Reduction
New Configuration Search
New Conformation Search
Start: init. training set with 2% ANI-1
Train ANI potential to training set
Terminate: begin active learning cycles
Calculate: $dE_i = |E_i^{ANI} - E_i^{REF}|$ for remaining i non-training data
Reduced Training set
No
Yes
Low percentage (<5%) dE < 0.1kcal/mol?
Add 2% of data where $dE_i <$ 0.1kcal/mol
a
Start: init. with reduced training set
Train 5x ANI models to training set
Sample new configs. (GDB, ChEMBL21, etc..)
Conformer sampling set (CSS)
Add $\rho_i < \hat{\rho}$ to conformer sampling set
Calculate: $\rho_i = \sigma(E_i)/\sqrt{N}$
b
End cycle: Restart configuration sampling
Yes
Is percent($\rho_i < \hat{\rho}$) < 5% of sampled
No
Generate new conformers from CSS
New training set
Train 5x ANI models to new training set
Calculate: $\rho_i = \sigma(E_i)/\sqrt{N}$
Add $\rho_i < \hat{\rho}$ to training set
c

FIG. 2. Fully automated AL workflow for data generation. The algorithm contains 3 steps: (a) an existing dataset reduction, (b) a configurational search, and (c) a conformational search.

with RDKit, and then optimizing the initial structure with the UFF[57] force field. See Sec. S1.2.3 of the supplementary material for details on dimer generation. Next, ANI energies are computed using an ensemble of five ANI models trained to the current AL dataset (see Sec. II D for details on ensemble prediction and training). Finally, $\rho_i = \sigma_i/\sqrt{N}$ is computed. Here, σ_i is the standard deviation of the ensemble's energy predictions for molecule i and N is the number of atoms in the molecule. The test of whether to include the molecule corresponding to a given ρ_i is $\rho_i > \hat{\rho}$. The selection of $\hat{\rho}$ is explained in Sec. II A. All molecules that fail this test are included in the new conformer sampling set. Any molecules added to the conformer sampling set are geometry optimized with the correct reference QM level of theory using tight SCF and optimization convergence criteria.

With the configurational search complete, a conformational search cycle [Fig. 2(c)] is initialized, whereby the conformer sampling set (a set of equilibrium molecules generated in the configuration sampling step) is used to generate a set of new non-equilibrium molecules ($\hat{X}$). The conformers in $\hat{X}$ are generated via one of the three techniques which are designed to sample various regions of chemical space. These sampling techniques are listed as follows:

- Diverse normal mode sampling (DNMS). A version of normal mode sampling (NMS) as presented in our previous work,[36] but with diversity selection used to reduce redundant data and a bias toward near equilibrium structures. A detailed description of DNMS is provided in Sec. S1.2.1 of the supplementary material.
- K random trajectory sampling (RTS). We run short (4 ps) molecular dynamics simulations, with an ensemble of ANI networks, starting with random velocities equal to 300 K and heated slowly to 1000 K over the simulation time. During the simulation, each step QBC is used to check whether the current structure fails the agreement test. Once the simulation reaches a conformation that fails the test, dynamics is terminated, and new QM properties (e.g. energies) are generated and included in the next AL cycles training set. This is repeated to generate multiple new samples. A detailed description of RTS is provided in Sec. S1.2.2 of the supplementary material.
- Molecular dynamics generated dimer sampling. Dimers are generated by randomly placing and orienting molecules from the conformer sampling set into a box with periodic boundary conditions. A molecular dynamics simulation for 5 ps is then carried out on the box. Every 50 steps, the box is fragmented into only dimer pairs within the desired cutoff radius. Each new dimer pair is tested using the QBC approach, failed tests are kept as new data, and QM properties are generated for inclusion in the training set. A detailed description of the dimer sampling approach used here is provided in Sec. S1.2.3 of the supplementary material.

After new data are selected, labels are computed and included in the training set, and a new ensemble of ANI potentials is trained. The conformational search cycles are repeated until the model stops improving within the COMP6 benchmarks (see details in Sec. II C). Finally, the entire cycle is restarted from the configurational sampling step. This process is carried out to produce a total of 37 cycles including many configurational and conformational searching cycles. Throughout this work, we will refer to various intermediate active learned ANI models as AL1 through AL6. The AL6 potential is the final potential reached in this work and is referred to as the ANI-1x potential which is provided for free in a python package integrated with the atomic simulation environment (ASE) package[58] [https://github.com/isayev/ASE_ANI]. The first row in Table I provides information about the final dataset from this work, labeled as ANI-1x. Notably, the size of the ANI-1x dataset, at 5.5 × 10^6 structures, is 25% the size of the dataset used in training the original ANI-1 potential (22 M).

C. Development of the COMP6 benchmark suite

To validate that the active learning process generates an ANI potential which outperforms the original ANI-1 potential and that each cycle's resulting AL ANI potentials consistently outperforms previous versions of AL ANI potentials, we develop the COmprehensive Machine-learning Potential (COMP6) benchmark. COMP6 is a benchmark suite composed of five rigorous benchmarks that cover broad regions of organic and bio-chemical space (for molecules containing C, N, O, and H atoms) and a sixth one built from the existing S66x8[50] noncovalent interaction benchmark. The five new benchmark sets are referred to as GDB7to9, GDB10to13, Tripeptides, DrugBank, and ANI-MD. See Table I for a detailed description. The benchmarks range from a mean molecule size of 17 atoms to 75 atoms, with the largest molecule being 312 atoms. Following is a description of the methods used to develop each benchmark. Energies and forces for all non-equilibrium molecular conformations presented have been calculated using the ωB97x[59] density functional with the 6-31G(d) basis set[60] as implemented in the Gaussian 09[61] electronic structure software. Hirshfeld charges and molecular dipoles are also included in the benchmark. An analysis of these properties will be carried out in future work.

- S66x8 benchmark. This dataset is built from the original S66x8[50] benchmark for comparing accuracy between different methods in describing noncovalent interactions common in biological molecules. S66x8 is developed from 66 dimeric systems involving hydrogen bonding, pi-pi stacking, London interactions, and mixed influence interactions. While the keen reader might question the use of this benchmark without dispersion corrections, since dispersion corrections such as the D3[62] correction by Grimme *et al.* are *a posteriori* additions to the produced energy, then a comparison without the correction is equivalent to a comparison with the same dispersion corrections applied to both models.
- ANI Molecular Dynamics (ANI-MD) benchmark. Forces from the ANI-1x potential are applied to run 1 ns of vacuum molecular dynamics with a 0.25 fs time step at 300 K using the Langevin thermostat on 14 well-known drug molecules and two small proteins. System sizes range from 20 to 312 atoms. A random subsample of 128 frames from each 1 ns trajectory is selected, and reference DFT single point calculations are performed to obtain QM energies and forces.
- GDB7to9 benchmark. The GDB-11 subsets containing 7 to 9 heavy atoms (C, N, and O) are subsampled and randomly embedded in 3D space using RDKit [www.rdkit.org]. A total of 1500 molecule SMILES [opensmiles.org] strings are selected: 500 per 7, 8, and 9 heavy-atom sets. The resulting structures are optimized with tight convergence criteria, and normal modes/force constants are computed using the reference DFT model. Finally, diverse normal mode sampling (DNMS) is carried out to generate non-equilibrium conformations.
- GDB10to13 benchmark. Subsamples of 500 SMILES strings each from the 10 and 11 heavy-atom subsets of GDB-11[52,53] and 1000 SMILES strings from the 12 and 13 heavy-atom subsets of the GDB-13[63] database are randomly selected. DNMS is utilized to generate random non-equilibrium conformations.
- Tripeptide benchmark. 248 random tripeptides containing H, C, N, and O are generated using FASTA strings and randomly embedded in 3D space using RDKit. As with GDB7to9, the molecules are optimized, and normal modes are computed. DNMS is utilized to generate random non-equilibrium conformations.
- DrugBank benchmark. This benchmark is developed through a subsampling of the DrugBank[64] database of real drug molecules. 837 SMILES strings containing C, N, and O are randomly selected. Like the GDB7to9 benchmark, the molecules are embedded in 3D space, structurally optimized, and normal modes

TABLE I. Description of the final active learning generated training dataset (ANI-1x) and all six COMP6 benchmark datasets. Mean relative energy range is the average range of relative energies for each set of conformers. Energy prediction range is the real prediction range in the benchmark; this is the range that the ANI model predicts energies in, i.e. energies with all per atom shifts removed. All energies are given in kcal/mol.

Purpose	Dataset	Molecule source	Configurations (conformations)	Atoms/molecule mean (std. dev.)	Mean relative energy range	Energy prediction range
Training	ANI-1x	ANI-1 + AL	63 865 (5 496 771)	15 (5)	97.6	6 400
Testing	S66x8	S66x8	66 (528)	20 (7)	6.00	2 800
	ANI-MD	PDB	14 (1 791)	75 (72)	35.0	31 000
	GDB7to9	GDB-11	1500 (36 000)	17 (3)	78.0	1 900
	GDB10to13	GDB-13	2996 (47 670)	25 (4)	214.0	2 300
	Tripeptides	RDKit	248 (1 984)	53 (7)	102.0	4 200
	DrugBank	DrugBank	837 (13 379)	44 (20)	167.0	14 000

are computed. DNMS is utilized to generate random non-equilibrium conformations.

D. Error metrics for validation on the COMP6 benchmark suite

This work uses three error metrics for comparing different versions of ANI potentials: potential energy (E), conformer energy difference (ΔE), and atomic force component errors (F).

- Potential energy (E) error is a comparison of E_i^{M1}, the potential energies produced by model M1 for molecule i, to E_i^{M2}, the potential energies produced by model M2 for molecule i.
- The conformer energy difference (ΔE) error is calculated per set of conformers. In the benchmark dataset, K sets of conformers are supplied, one per molecular configuration. For a given set of conformers k, the conformer energy difference between conformers i and j where for a given model M is obtained by computing $\Delta E_{ij}^{M,k} = E_i^{M,k} - E_j^{M,k}$. Finally, error is calculated between $\Delta E_{ij}^{M1,k}$ and $\Delta E_{ij}^{M2,k}$ for all k, i, and $j > i + 1$ for models M1 and M2.
- The atomic force (F) error metric is a comparison between the individual components (x, y, z) of each atom's force vector for all conformations included in the given benchmark.

Comparisons are given in mean absolute error (MAE), and root mean squared error (RMSE) throughout this article. The comparison of MAE along with RMSE can give information about outliers in a model's predictions. For example, two models can have the same MAE for a prediction on a given benchmark, while the RMSE can be much higher for one than the other. For this reason, it is good practice to provide both MAE and RMSE when comparing two methods on some benchmark.

E. Property prediction with an ensemble of ANI models

For energy and force predictions, we use the mean prediction of an ensemble of ANI potentials. The concept of using an ensemble mean for ML model prediction is common practice in the ML community. Recently, it has been adopted in the area of ML molecular property prediction.[31,37,65] All potentials used to generate results in this work utilize the mean prediction for an ensemble of $L = 5$ ANI potentials trained to a 5-fold cross validation split of the training dataset. The potential energy (E) is represented by

$$E = \frac{1}{L}\sum_{i=1}^{L} E_i,$$

where E_i is the potential energy prediction from each of an ensemble's L ANI models. Since the models are independent, atomic forces for the ensemble can be derived as the component-wise mean of the forces from the L individual ANI models. The use of an ensemble as described above decreases ANI vs. DFT E RMSE by 0.67 kcal/mol, ΔE RMSE by 0.68 kcal/mol, and F RMSE by 2.1 kcal/mol × Å^{-1} over the entire COMP6 benchmark, with an error reduction of 17%, 19%, and 28%, respectively.

III. RESULTS AND DISCUSSIONS

The supplementary material provided with this work contains various tables detailing the results obtained on the COMP6 benchmark by the ANI potentials discussed in this work. Tables S1–S7 of the supplementary material provide an analysis of the ΔE, E, and F errors obtained for six subsequent active learned ANI potentials, AL1 through AL6, and the original ANI-1 potential. Note that the publicly released ANI-1x potential is the AL6 ANI potential. Tables S8–S10 of the supplementary material provide an analysis of the individual ANI-MD trajectory results for the ANI-1x potential. Table S9 of the supplementary material provides per atom energy errors for the ANI-1x potential vs. DFT and shows that the mean energy prediction RMSE per atom for all trajectories is 0.05 kcal/mol per atom. This level of accuracy is on par with single molecule or bulk metal ML potentials as described in recent work by Behler.[25] Table S11 of the supplementary material provides details on the ANI models introduced in this work. Finally, Tables S12–S17 of the supplementary material provide errors for COMP6 considering conformers within select energy ranges for the ANI-1x potential. These tables show much lower errors for conformations which are thermally accessible to room temperature molecular dynamics simulations. As shown in Table S17 of the supplementary material, thermally accessible conformations (within 50 kcal/mol) have an E MAE/RMSE of 0.064/0.105 kcal/mol per atom and ΔE MAE/RMSE of 0.049/0.070 kcal/mol per atom over the complete COMP6 benchmark.

Figure 3 provides evidence of the ANI-1x force prediction capabilities. Also, most tables in the supplementary material further establish the accuracy of ANI potential force prediction on the COMP6 benchmark suite. By construction, ANI potentials provide analytic and energy-conservative forces, a requirement for molecular dynamics simulations. It is noteworthy that force training, which can be computationally expensive, **is not required** to achieve these force prediction results. The forces compared in the DFT correlation density plots in Fig. 3 are from all trajectories combined in the COMP6 ANI-MD benchmark. We compare the same figures for ANI-1x (left), DFTB (center), and PM6 (right). DFTB and PM6 are included as a baseline for the comparison. The ANI-MD benchmark is a rigorous test case for any ML potential's force prediction because the molecules supplied in the dataset range from 20 to 312 atoms, with an average size of 75 atoms. A breakdown of the errors for each trajectory in the ANI-MD benchmark is supplied in Tables S8–S10 of the supplementary material.

The closest comparison in the literature can be found in recent work on a system specific ML potential for an alanine tripeptide where a force RMSE of 3.4 kcal/mol × Å^{-1} was achieved with test data from a 350 K MD trajectory.[37] The force error from this work was obtained by training directly to energies and analytic forces from fragments of the molecule being tested. In the case of the ANI-1x potential,

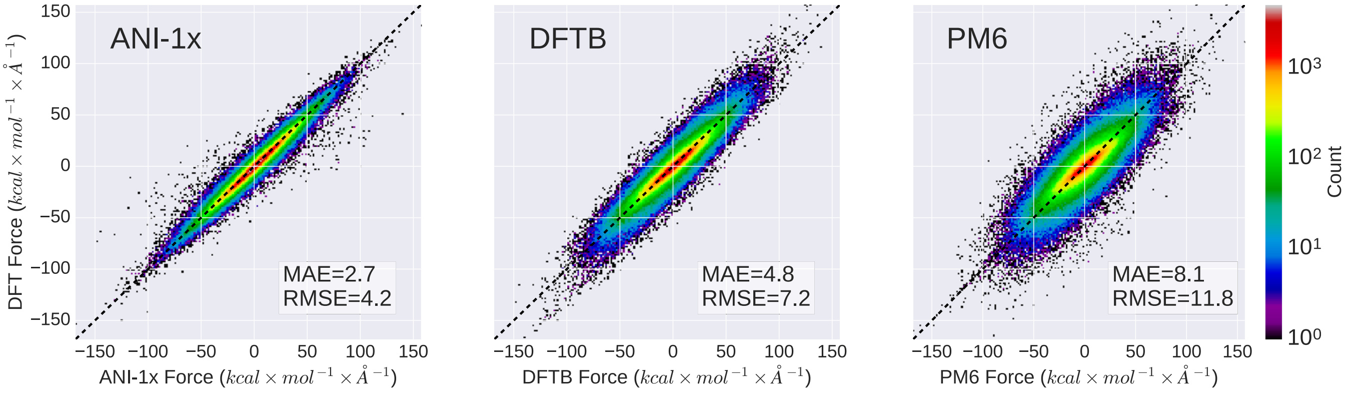

FIG. 3. Force correlation plots comparing ANI-1x, DFTB (3ob-3-1 parameter set for bio-molecules), and PM6 to DFT reference calculations are provided from left to right, respectively, for the complete ANI-MD benchmark. Molecules in the ANI-MD benchmark are composed of a mean of 75 atoms with the largest being Trp-cage (1L2Y), a 20-residue (312-atom) protein. DFTB and PM6 are provided as a baseline of comparison. Mean absolute errors (MAEs) and root mean squared errors (RMSEs) are provided in the bottom right of each figure. The color bar scale is the same for all figures allowing a proper density comparison.

which was used to predict the forces for the creation of the ANI-MD benchmark, a MAE/RMSE of 2.7/4.2 kcal/mol × Å^{-1} is obtained vs. *a posteriori* DFT calculations on 128 random frames from each of the 14 molecules' 1 ns molecular dynamics trajectories. More impressive, F MAE/RMSE for the neutralized 20-residue Trp-cage (1L2Y) and 10-residue chignolin (1UAO) proteins are 3.1/4.6 kcal/mol × Å^{-1} and 3.3/4.7 kcal/mol × Å^{-1}, respectively. ANI-1x also exhibits a force MAE/RMSE of 2.3/3.3 kcal/mol × Å^{-1} within the energy range of 50 kcal/mol on the tripeptide benchmark (non-equilibrium conformations from 248 randomly generated tripeptides) from COMP6 (see Table S14 of the supplementary material). 50 kcal/mol is roughly the accessible energy range of 350 K molecular dynamics simulations. Finally, considering the ANI-1x potential was utilized to generate 1 ns of stable 300 K molecular dynamics trajectories (for building the ANI-MD benchmark) shows the applicability of ANI predicted forces in molecular dynamics simulation. All the previously mentioned results from the ANI-1x potential were obtained **without** the need of direct force training.

Figure 4 provides a plot of E RMSE achieved on COMP6 vs. dataset size for various active learned datasets and the original ANI-1 dataset. With only 2 × 10^6 data points, the active learned ANI potentials already outperform the original ANI-1 potential across the entire COMP6 benchmark. Once the active learned ANI potential reaches 5.5 × 10^6 data points, it five times outperforms ANI-1 and is approaching chemical accuracy from the reference DFT calculations. In the new COMP6 benchmark, diversity selection in the normal mode sampling helps ensure a more uniform sampling of energy states within the energy range being fit to and tested within. Therefore, general errors on COMP6 vs. the ANI-1 potential's original results are expected to be much higher on this complex benchmark than the results published on the less rigorous test sets from the original ANI-1 work. Table I provides the average energy ranges for each benchmark in COMP6 and the final training set (ANI-1x), as well as the energy prediction (atomization energy) range.

Most benchmarks in COMP6 (all but the ANI-MD benchmark) were used during the active learning process to validate the improvement in accuracy and universality of new active learned ANI models. Figure 5 provides the learning curves for six intermediate active learned ANI potentials on each benchmark in COMP6. Table S11 of the supplementary material provides information of the chemical space sampled in each of these datasets. The horizontal dashed lines in Fig. 5 represent the original ANI-1 ensemble predictions on each of the benchmarks for the property corresponding to its color. AL1 is the ANI potential used to initialize the active learning process. It was trained to a reduced [Fig. 2(a)] version of the one through six heavy atom subsets of the ANI-1 dataset. AL2 through AL6 are successive versions of the active learned ANI potentials. More details for each active learning cycle

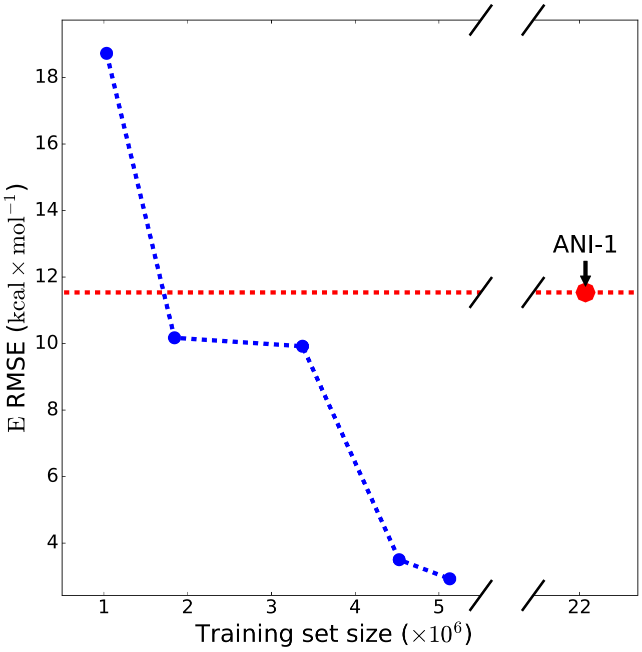

FIG. 4. Comparison of potential energy (E) RMSE obtained on the entire COMP6 benchmark vs. training set size (total molecular conformation included in the training set). The x-axis represents the progression of the active learning process. Plot points are obtained by ANI potentials (blue) trained to various versions of the active learned dataset and an ANI potential (red) trained on the original ANI-1 dataset.

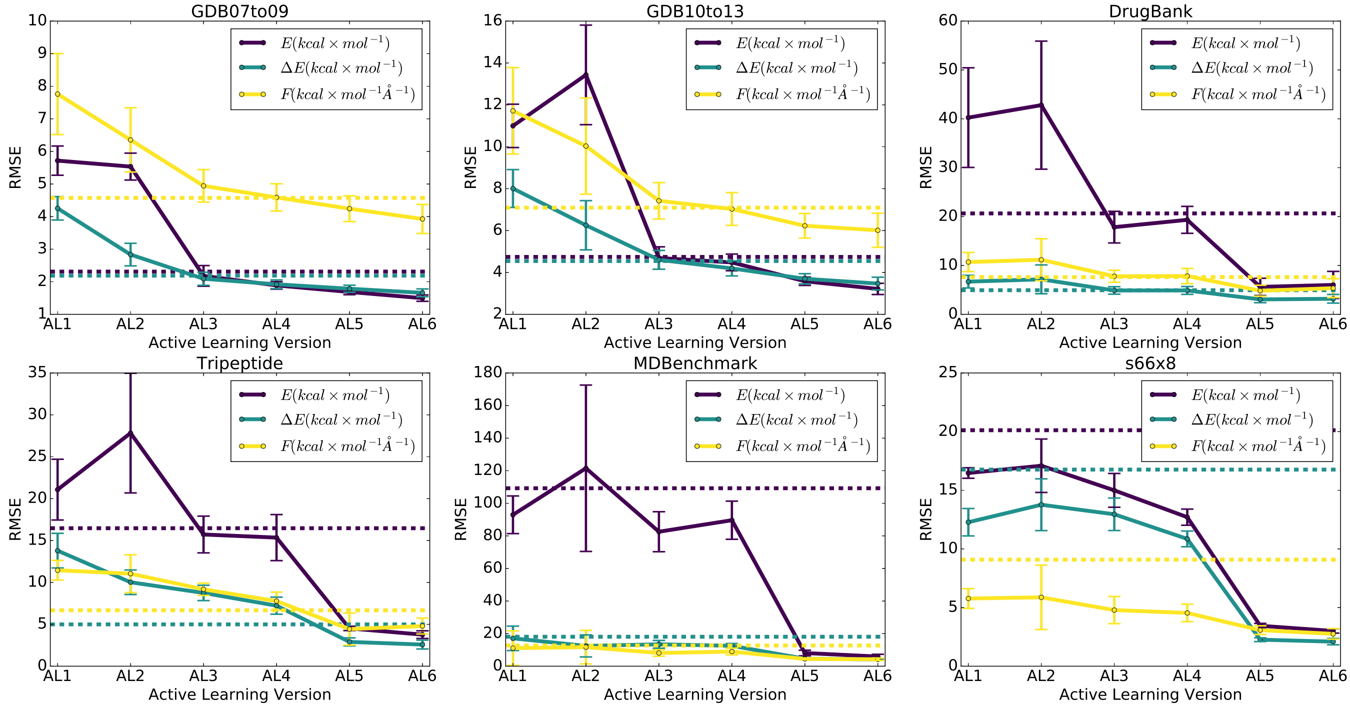

FIG. 5. Individual COMP6 benchmark learning curves for successive versions of the active learned potentials. RMSE is provided for three properties: potential energy (E), conformer energy differences (ΔE), and force components (F). The error bars on the solid lines represent one standard deviation of each of the five ANI models in the ensemble used to make the mean prediction. The horizontal dashed lines represent the mean prediction of ANI-1.

shown in Fig. 5 is provided in Table S11 of the supplementary material. During the active learning process, small molecules (one to six C, N, and O atoms) were initially sampled, with the size of the molecules sampled gradually increased as the active learning process continued. AL3 is where the AL models begin to statistically match or outperform the original ANI-1 model in most metrics. It is notable that AL3 accomplished this feat, while only having sampled 1.8×10^6 conformations from molecules with up to 7-heavy atoms from GDB-11. This shows that the active learning techniques employed in this work sample chemical space far better than random sampling techniques. Especially considering the ANI-1 dataset includes 22×10^6 conformations from larger, up to 8-heavy atom, molecules.

Eventually, between the AL4 and AL5 steps, amino acids, generated dipeptides, generated small molecule dimers, and small ChEMBL molecules were added to the sampling set. This is apparent from the large drop in error between AL4 and AL5 for the DrugBank, Tripeptides, and S66x8 benchmarks. Active learning sampling was also driven into the GDB-11's 9-heavy atom subset for sampling during the production of AL6. Tables S2–S7 of the supplementary material provide all data shown in Fig. 5 along with Table S1 of the supplementary material which describes the obtained errors across all benchmarks combined. The latest ANI potential, ANI-1x (shown as AL6), achieves remarkable property prediction on the complete benchmark with errors (MAE/RMSE) of 1.9/3.4 kcal/mol (E), 1.8/3.0 kcal/mol (ΔE), and 3.1/5.3 kcal/mol × Å^{-1} (F) within the full energy range of the benchmark.

In general, as each ANI potential's fitness improves in Fig. 5, the standard deviation (shown as vertical error bars) of each property prediction for a given ensemble decreases as well. This is a sign that each model in the ensemble is obtaining enough chemical interaction information through active learning that the models begin agreeing on their predictions for these larger systems. By the final iteration of the active learning cycles, an active learned dataset of 5.5 M data points is used in training the ANI-1x potential. The ANI-1x potential outperforms the ANI-1 potential on all properties across all benchmarks. Furthermore, the ANI-1x dataset is 25% the size of the original ANI-1 dataset which contains a total of 22 M data points.

There has been recent discussion in the literature (Herr *et al.*[66]) questioning the validity of using data generated from a single sampling technique to successfully extrapolate to out-of-sample data. We believe such a critique is well placed and has a particular impact when defining system-specific potentials or other models of limited scope. In the present work, we combined several sampling techniques, attempting to cover all relevant regions of conformational and configurational (chemical) space. We test model performance on a separate and very diverse set of systems, showing not only accuracy, but extensibility to molecules and conformations much larger than the training set. Accuracies on benchmarks generated with different sampling techniques are comparably accurate: On the ANI-MD benchmark (mean energy range of 35 kcal/mol), we achieve force MAE of 2.49 kcal/mol × Å^{-1} (Table S10 of the supplementary material). On the overall COMP6 benchmark restricted to the energy range of 50 kcal/mol, we achieve a force MAE of 2.48 kcal/mol × Å^{-1} (Table S12 of the supplementary material). The fact that MD-sampled test points show quite similar error to the mostly DNM-sampled benchmark is evidence that the active learning procedure with hybrid sampling methods produces a model that is robust.

IV. CONCLUSIONS

In pursuit of automated dataset generation for the development of universal machine learned potentials, we introduce automatic active learning techniques for sampling sparsely explored regions of chemical space. The algorithm begins with the reduction of an existing dataset to remove redundant data without loss of accuracy. New conformations of molecules are generated through normal mode sampling, molecular dynamics sampling, and random dimer sampling. Periodically the algorithm samples new molecular configurations from a variety of sources to diversify its exploration of chemical space. The result is a new potential (ANI-1x) developed though successive generations of the active learning process. The ANI-1x potential is packaged in a user-friendly *Python* library, which is publicly available on GitHub [https://github.com/isayev/ASE_ANI]. We also introduce the COMP6 benchmark for monitoring the progress of active learning cycles and for comparison to future universal potentials. The ANI-1x potential achieves errors (MAE/RMSE) of 1.6/3.0 kcal/mol (E), 1.4/2.3 kcal/mol (ΔE), and 2.7/4.5 kcal/mol × Å^{-1} (F) when testing on points within 100 kcal/mol of the energy minima for the complete COMP6 benchmark.

The COMP6 benchmark suite consists of six diverse benchmark test sets. The COMP6 benchmark suite is made publicly available for comparing future ML potentials [https://github.com/isayev/COMP6]. As provided, properties are calculated using the ωB97x density functional with the 6-31G(d) basis set; however, it could be recomputed using the desired quantum level of theory. For complete transparency, we provide the exact error metrics used to measure accuracy on the COMP6 benchmark suite. It is our hope that the COMP6 benchmark will provide the universal ML potential development community with a rigorous benchmark for comparison of ML potential methods on organic molecules in the extrapolative regime. The COMP6 benchmark suite constitutes a first benchmark of its kind for the comparison of universal ML potentials in this rapidly changing and ever-growing field.

The ANI-1x potential was trained to less than 100 conformations per molecular configuration in its training set, compared to 400 for the ANI-1 dataset. The accuracy of the ANI-1x potential is on par with the best single molecule or material ML potentials, while most single molecule parametrized ML potentials require many hundreds to thousands of conformations to parametrize a single system. This further validates the configurational and conformational big data sampling philosophy introduced in the original ANI-1 work. Since the mean molecule size in the ANI-1x active learning training set is only 15 total atoms (8 heavy atoms), the generation of more accurate post-Hartree-Fock datasets is now feasible.

The high-level of universal accuracy achieved by the ANI-1x potential can be attributed to the capacity of neural networks to learn low level interactions from properly developed descriptors. We hypothesize the use of spatially localized descriptors (i.e., the atomic environment vector[34] with modified angular symmetry function) within the cutoff to contribute greatly to this ability. This contrasts with descriptor sets that represent the entire chemical environment at once, and thus, interactions must be inferred through the entire set of non-local descriptors by the ML model.

Given the prospects of high-throughput experiments, robotic synthesis, and intelligent software, we are currently witnessing a transformation of science into a more data-driven automated discovery. The envisioned chemical AI imitates human decision making by transferring responsibility to an objective machine learning system. If successful overall, the approach will revolutionize the way computational methods are developed. As one possible building block to construct such AI, we introduced a fully automated workflow to select and calculate QM training data for accurate, transferable, and extensible ML potentials. These techniques can aid in the generation of universal potentials for a wide variety of current and future ML models.

SUPPLEMENTARY MATERIAL

See supplementary material for complete technical details about training of ensemble of neural networks (Sec. S1.1) and sampling methods (Sec. S1.2). Tables S1–S10 list individual and complete COMP6 benchmarks. Table S11 lists details about ANI potential at various AL cycles. Tables S12–S17 list COMP6 benchmarks for ANI-1x within select energy ranges.

ACKNOWLEDGMENTS

J.S.S. thanks the University of Florida for the graduate student fellowship and the Los Alamos National Laboratory Center for Non-linear Studies for resources and hospitality. This work was performed, in part, at the Center for Integrated Nanotechnologies, an Office of Science User Facility operated for the U.S. Department of Energy (DOE) Office of Science. Los Alamos National Laboratory, an affirmative action equal opportunity employer, is operated by Los Alamos National Security, LLC, for the National Nuclear Security Administration of the U.S. Department of Energy under Contract No. DE-AC52-06NA25396. O.I. acknowledges support from DOD-ONR No. N00014-16-1-2311 and Eshelman Institute for Innovation award. The authors acknowledge Extreme Science and Engineering Discovery Environment (XSEDE) Award No. DMR110088, which is supported by National Science Foundation Grant No. ACI-1053575. This research was done in part using resources provided by the Open Science Grid[67,68] which is supported by the National Science Foundation Award No. 1148698 and the U.S. Department of Energy's Office of Science. We gratefully acknowledge the support of the U.S. Department of Energy through the LANL/LDRD Program for this work. The authors thank Roman Zubatyuk and Kipton Barros for invaluable discussions on the topics presented in this work.

[1] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling, Proteins Struct. Funct. Genet. **65**, 712 (2006).

[2] K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov, and A. D. Mackerell, J. Comput. Chem. **31**, 671 (2010).

[3] T. A. Halgren, J. Comput. Chem. **17**, 490 (1996).

[4] K. S. Thanthiriwatte, E. G. Hohenstein, L. A. Burns, and C. D. Sherrill, J. Chem. Theory Comput. **7**, 88 (2011).
[5] H. J. Monkhorst, Int. J. Quantum Chem. **12**, 421 (1977).
[6] G. D. Purvis and R. J. Bartlett, J. Chem. Phys. **76**, 1910 (1982).
[7] D. Cremer, Wiley Interdiscip. Rev.: Comput. Mol. Sci. **1**, 509 (2011).
[8] J. Huang and A. D. Mackerell, J. Comput. Chem. **34**, 2135 (2013).
[9] H. Sun, J. Phys. Chem. B **102**, 7338 (1998).
[10] K. N. Kirschner, A. B. Yongye, S. M. Tschampel, J. González-Outeiriño, C. R. Daniels, B. L. Foley, and R. J. Woods, J. Comput. Chem. **29**, 622 (2008).
[11] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling, J. Chem. Theory Comput. **11**, 3696 (2015).
[12] T. Moot, O. Isayev, R. W. Call, S. M. McCullough, M. Zemaitis, R. Lopez, J. F. Cahoon, and A. Tropsha, Mater. Discovery **6**, 9 (2016).
[13] M. Ragoza, J. Hochuli, E. Idrobo, J. Sunseri, and D. R. Koes, J. Chem. Inf. Model. **57**, 942 (2017).
[14] B. Liu, B. Ramsundar, P. Kawthekar, J. Shi, J. Gomes, Q. Luu Nguyen, S. Ho, J. Sloane, P. Wender, and V. Pande, ACS Cent. Sci. **3**, 1103 (2017).
[15] J. N. Wei, D. Duvenaud, and A. Aspuru-Guzik, ACS Cent. Sci. **2**, 725 (2016).
[16] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. Von Lilienfeld, J. Chem. Theory Comput. **11**, 2087 (2015).
[17] A. Lavecchia, Drug Discovery Today **20**, 318 (2015).
[18] O. Isayev, C. Oses, C. Toher, E. Gossett, S. Curtarolo, and A. Tropsha, Nat. Commun. **8**, 15679 (2017).
[19] E. Kim, K. Huang, A. Saunders, A. McCallum, G. Ceder, and E. Olivetti, Chem. Mater. **29**, 9436 (2017).
[20] B. Kolb, B. Zhao, J. Li, B. Jiang, and H. Guo, J. Chem. Phys. **144**, 224103 (2016).
[21] M. Hellström and J. Behler, Phys. Chem. Chem. Phys. **19**, 82 (2017).
[22] T. H. Ho, N.-N. Pham-Tran, Y. Kawazoe, and H. M. Le, J. Phys. Chem. A **120**, 346 (2016).
[23] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, and A. Tkatchenko, Nat. Commun. **8**, 13890 (2017).
[24] K. Yao, J. E. Herr, S. N. Brown, and J. Parkhill, J. Phys. Chem. Lett. **8**, 2689 (2017).
[25] J. Behler, Angew. Chem., Int. Ed. **56**, 12828 (2017).
[26] V. Botu, R. Batra, J. Chapman, and R. Ramprasad, J. Phys. Chem. C **121**, 511 (2017).
[27] S. Kondati Natarajan, T. Morawietz, and J. Behler, Phys. Chem. Chem. Phys. **17**, 8356 (2015).
[28] J. Behler, R. Martoňák, D. Donadio, and M. Parrinello, Phys. Status Solidi B **245**, 2618–2629 (2008).
[29] K. Yao, J. E. Herr, D. W. Toth, R. Mcintyre, and J. Parkhill, Chem. Sci. **9**, 2261 (2018).
[30] K. Schütt, P.-J. Kindermans, H. E. Sauceda Felix, S. Chmiela, A. Tkatchenko, and K.-R. Müller, in *Advances in Neural Information Processing Systems*, edited by I. Guyon, U. V Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc., 2017), Vol. 30, pp. 992–1002.
[31] N. Lubbers, J. S. Smith, and K. Barros, J. Chem. Phys. **148**, 241715 (2018).
[32] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, Sci. Data **1**, 140022 (2014).
[33] M. Rupp, A. Tkatchenko, K.-R. Muller, and O. A. von Lilienfeld, Phys. Rev. Lett. **108**, 058301 (2012).
[34] J. S. Smith, O. Isayev, and A. E. Roitberg, Chem. Sci. **8**, 3192 (2017).
[35] J. Behler and M. Parrinello, Phys. Rev. Lett. **98**, 146401 (2007).
[36] J. S. Smith, O. Isayev, and A. E. Roitberg, Sci. Data **4**, 170193 (2017).
[37] M. Gastegger, J. Behler, and P. Marquetand, Chem. Sci. **8**, 6924 (2017).
[38] B. Huang and O. A. von Lilienfeld, preprint arXiv:1707.04146 (2017).
[39] P. Raccuglia, K. C. Elbert, P. D. F. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier, and A. J. Norquist, Nature **533**, 73 (2016).
[40] P. J. Kitson, G. Marie, J.-P. Francoia, S. S. Zalesskiy, R. C. Sigerson, J. S. Mathieson, and L. Cronin, Science **359**, 314 (2018).
[41] T. Chapman, Nature **421**, 661 (2003).
[42] R. D. King, J. Rowland, S. G. Oliver, M. Young, W. Aubrey, E. Byrne, M. Liakata, M. Markham, P. Pir, L. N. Soldatova, A. Sparkes, K. E. Whelan, and A. Clare, Science **324**, 85 (2009).
[43] E. V. Podryabinkin and A. V. Shapeev, Comput. Mater. Sci. **140**, 171 (2017).
[44] N. J. Browning, R. Ramakrishnan, O. A. von Lilienfeld, and U. Roethlisberger, J. Phys. Chem. Lett. **8**, 1351 (2017).
[45] P. O. Dral, A. Owens, S. N. Yurchenko, and W. Thiel, J. Chem. Phys. **146**, 244108 (2017).
[46] A. A. Peterson, R. Christensen, and A. Khorshidi, Phys. Chem. Chem. Phys. **19**, 10978–10985 (2017).
[47] D. Reker and G. Schneider, Drug Discovery Today **20**, 458 (2015).
[48] J. M. Hernández-Lobato, J. Requeima, E. O. Pyzer-Knapp, and A. Aspuru-Guzik, preprint arXiv:1706.01825 (2017).
[49] H. S. Seung, M. Opper, and H. Sompolinsky, in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory–COLT 1992* (ACM Press, New York, USA, 1992), pp. 287–294.
[50] B. Brauer, M. K. Kesharwani, S. Kozuch, and J. M. L. Martin, Phys. Chem. Chem. Phys. **18**, 20905 (2016).
[51] I. Kruglov, O. Sergeev, A. Yanilkin, and A. R. Oganov, Sci. Rep. **7**, 8512 (2017).
[52] T. Fink, H. Bruggesser, and J. L. Reymond, Angew. Chem., Int. Ed. **44**, 1504 (2005).
[53] T. Fink and J. L. Raymond, J. Chem. Inf. Model. **47**, 342 (2007).
[54] S. Jupp, J. Malone, J. Bolleman, M. Brandizi, M. Davies, L. Garcia, A. Gaulton, S. Gehant, C. Laibe, N. Redaschi, S. M. Wimalaratne, M. Martin, N. Le Novre, H. Parkinson, E. Birney, and A. M. Jenkinson, Bioinformatics **30**, 1338 (2014).
[55] A. P. Bento, A. Gaulton, A. Hersey, L. J. Bellis, J. Chambers, M. Davies, F. A. Krüger, Y. Light, L. Mak, S. McGlinchey, M. Nowotka, G. Papadatos, R. Santos, and J. P. Overington, Nucleic Acids Res. **42**, D1083 (2014).
[56] M. Davies, M. Nowotka, G. Papadatos, F. Atkinson, G. van Westen, N. Dedman, R. Ochoa, and J. Overington, Challenges **5**, 334 (2014).
[57] A. K. Rappe, C. J. Casewit, W. A. Goddard III, K. S. Colwell, and W. M. Skiff, J. Am. Chem. Soc. **114**, 10024 (1992).
[58] A. Hjorth Larsen, J. Jørgen Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, E. D. Hermes, P. C. Jennings, P. Bjerre Jensen, J. Kermode, J. R. Kitchin, E. Leonhard Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. Bergmann Maronsson, T. Maxson, T. Olsen, L. Pastewka, A. Peterson, C. Rostgaard, J. Schiøtz, O. Schütt, M. Strange, K. S. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng, and K. W. Jacobsen, J. Phys.: Condens. Matter **29**, 273002 (2017).
[59] J.-D. Chai and M. Head-Gordon, J. Chem. Phys. **128**, 084106 (2008).
[60] R. Ditchfield, W. J. Hehre, and J. A. Pople, J. Chem. Phys. **54**, 724 (1971).
[61] G. M. J. Frisch, W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, and J. L. Sonnenberg, GAUSSIAN 09, Revision c.01, Gaussian, Inc., Wallingford, CT, 2009.
[62] S. Grimme, J. Antony, S. Ehrlich, and H. Krieg, J. Chem. Phys. **132**, 154104 (2010).
[63] L. C. Blum and J.-L. Reymond, J. Am. Chem. Soc. **131**, 8732 (2009).
[64] V. Law, C. Knox, Y. Djoumbou, T. Jewison, A. C. Guo, Y. Liu, A. Maciejewski, D. Arndt, M. Wilson, V. Neveu, A. Tang, G. Gabriel, C. Ly, S. Adamjee, Z. T. Dame, B. Han, Y. Zhou, and D. S. Wishart, Nucl. Acids Res. **42**, D1091 (2014).
[65] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, preprint arXiv:1704.01212 (2017).
[66] J. E. Herr, K. Yao, R. McIntyre, D. Toth, and J. Parkhill, J. Chem. Phys. **148**, 241710 (2018).
[67] R. Pordes, D. Petravick, B. Kramer, D. Olson, M. Livny, A. Roy, P. Avery, K. Blackburn, T. Wenaus, F. Würthwein, I. Foster, R. Gardner, M. Wilde, A. Blatecky, J. McGee, and R. Quick, J. Phys.: Conf. Ser. **78**, 012057 (2007).
[68] I. Sfiligoi, D. C. Bradley, B. Holzman, P. Mhashilkar, S. Padhi, and F. Würthwein, in *2009 WRI World Congress on Computer Science and Information Engineering CSIE 2009* (IEEE, 2009), pp. 428–432.