Description of Potential Energy Surfaces of Molecules using FFLUX Machine Learning Models

Zak E. Hughes[1,2,†], *Joseph C. R. Thacker*[1,2], *Alex L. Wilson*[1,2] *and Paul L. A. Popelier*[1,2,*]

[1] Manchester Institute of Biotechnology, The University of Manchester, Manchester, M1 7DN, UK

[2] School of Chemistry, The University of Manchester, Manchester, M13 9PL, UK

ABSTRACT. A new type of model, FFLUX, to describe the interaction between atoms has been developed as an alternative to traditional force fields. FFLUX models are constructed from applying the kriging machine learning method to the topological energy partitioning method, Interacting Quantum Atoms (IQA). The effect of varying parameters in the construction of the FFLUX models is analyzed, with the most dominant effects found to be the structure of the molecule and the number of conformations used to build the model. Using these models the optimization of a variety of small organic molecules is performed, with sub kJ mol[-1] accuracy in the energy of the optimized molecules. The FFLUX models are also evaluated in terms of their performance in describing the potential energy surfaces (PESs) associated with specific degrees of freedoms within molecules. While the accurate description of PESs presents greater challenges than individual minima, FFLUX models are able to achieve errors of <2.5 kJ mol[-1] across the full C-C-C-C dihedral PES of *n*-butane, indicating the future possibilities of the technique.

1

1. INTRODUCTION

With the huge advances made in computer technology made in the last two decades, molecular modelling and simulation have become key techniques in chemistry and physics providing vital insights into the interaction of systems. However, one of the key challenges in computational chemistry remains obtaining a description of a system's potential energy surface (PES) that is accurate, transferable, computationally efficient and as simple as possible.[1] The most rigorous method of describing the interaction between particles is provided by a quantum mechanical (QM) description of some kind. However, the high computational cost of QM models means that, even with the advances in computer power/speed, they are, and will be for the foreseeable future, unsuitable to use in the study of many systems. The major alternative method of describing a PES of a system is to use a classical force field (FF), which consists of a series of potentials describing the different interactions between sites in the system. Ensuring that the PES described by a FF is one that accurately reproduces reality (or at least the most relevant or important attributes of it) is a complex task. Even modern biomolecular FFs, which are the result of many decades of work, are still being revised and developed in the light of new data.[2-4]

While the traditional biomolecular FFs, (e.g. CHARMM[2] or AMBER[3]) have many advantages they also suffer from a number of limitations, such as a point charge description of electrostatics, the artificial distinction of bonding and non-bonding interactions, a lack of reactivity, limited transferability, etc. Because of these limitations the last decade has seen an increased focus on developing a new generation of FFs that both address these limitations and provide a more accurate and deeper description of the PESs of (bio)molecular systems.[5,6] A *non-exhaustive* list of such next-generation force fields includes AMOEBA[7], NEMO[8], SIBFA[9], XED[10], EFP[11] and DMACRYS[12], which all share the important advance of multipolar electrostatics as justified in a recent perspective[13], as do the Gaussian Multipole Model (GMM)[14], the Exact Potential Multipole Method (EPMM)[15], and the water potential[16] family ASP-Wn[17] and DPP2[18].

Our strategy to develop such a next-generation FF is based around combining machine learning (which is also in this context by several other groups[19-28]) with the quantum chemical topological energy partitioning scheme called Interacting Quantum Atoms (IQA).[29] Under this approach, originally called Quantum Chemical Topology Force Field (QCTFF),[30] now renamed FFLUX,[31,32] the resulting FF has interacting atoms that are endowed with quantum mechanical

information. Unlike the other next-generation force fields, our approach is not based on long-range Rayleigh Schrödinger perturbation, which has far reaching consequences for its design.

In the development of a FFLUX model the properties of atoms, whether multipole moments or energy contributions, are assigned from QM calculations using the IQA partitioning scheme. The machine learning method kriging (also known as Gaussian process regression[33,34]) is then used to construct potentials that link the properties of atom i with its position relative to the position of other atoms in the system (either all the atoms in a system or a subset of them). Kriged potentials have been used to model the variation of atomic multipole moments and of kinetic, exchange and correlation energies as a function of configurational changes in the system.[32] Recent work has indicated that FF models developed from the kriging of a single physical property, the atomic energy term E_{IQA}^{A} (which consists of a sum of all the different types of energy contributions),[32] are able to provide a description of the PES of small molecules.

It has been shown that the description of PESs by FFLUX is of such quality that the energy and geometry of molecules at an energy minimum can be reproduced to considerable accuracy (<1 kJ mol[1] and root-mean-square-deviation < 0.2 Å in the case of glycine).[35] However, to reproduce a PES a FF must not only be able to describe a minimum but also attributes such as the difference in energy between minima and the height of energy barriers separating them. To date FFLUX has not been systematically tested to see how well it reproduces specific attributes on PESs away from the global minimum. In addition, there are still a number of open questions regarding the optimal methods of developing FFLUX potentials. Therefore, the present study has two aims: (1) to refine the procedure for the generation of FFLUX FFs by comparing results of a number different molecular species, representative of the organic groups found in biomolecules, (2) to show that FFLUX is able to reproduce PESs associated with bond stretches, angle bends and dihedral rotations within such molecules.

2. METHODS

2.1 Procedure for Generation of FFLUX Potentials

In this work FFLUX models are developed and tested for five small organic molecules, which are representative of the groups found in amino acids: methanol, propane, *n*-butane, isobutane and N-methylacetamide (NMA). As outlined in previous work,[32,35] the development of FFLUX models for the atoms in a molecule involves a number of distinct steps: (1) generation of a conformational ensemble used for training the machine learning models (and for validation purposes), (2) calculation of the wavefunction of each conformation, (3) calculation of the IQA atomic energies for each conformation, and (4) mapping of the IQA atomic energies to geometric features via the kriging machine learning method. Finally, the resulting FF is tested by calculating the energy of different conformations and by optimizing molecular geometries. The computational details associated with each step are described below in Section 2.2.

2.2 Computational Details

2.2.1 Generation of Conformational Ensembles

During the development of FFLUX as a FF method, a number of different methods of generating an ensemble of conformers, to be used to construct FFLUX models, have been tried. In the present work the standard method of conformation generation was based on distortion according to the normal modes of molecules via the in-house code TYCHE.[36] Normal mode sampling is an established method to obtain a physically relevant sampling of PESs, also used recently[37] in the training of a deep neural network on DFT energies of organic molecules. Normal mode sampling involves taking "seed" geometries, typically conformations corresponding to energy minima, of the molecule and then putting energy (corresponding to an input temperature) into the normal modes leading to an ensemble of randomly distorted temperature-weighted geometries of the molecule. For propane, *n*-butane, isobutane and NMA, the B3LYP/6-31+G(d,p) minimum energy conformation was used as the single input "seed" and the distortions carried out at a temperature of 1750 K (preliminary studies have shown that this temperature gives a good level of sampling around bonds/angles). This method was also used for methanol, but with three input seeds, corresponding to the three equivalent minimum energy conformations around the H-C-O-H dihedral angle, used to construct the ensemble. In the case of methanol, two alternative methods of generating conformational ensembles were used in addition to the ensemble constructed via TYCHE in order to allow comparison. From this point on ensemble E1

refs to TYCHE's output. Ensemble E2 was generated using the in-house code EROS. This method also involved distortion of the molecule along the normal modes but the distortions were not made by putting energy into the system but instead by randomly distorting the normal modes, to a maximum of 20% of the minimum energy conformation. For example, the C-O bond, with a bond length of 1.425 Å, could take any value between 1.140 and 1.710 Å. The final method for conformer generation was to perform an in vacuo MD simulation of the molecule and randomly extract conformations. These MD simulations were performed using the MMFF94x FF[38] using the Molecular Operating Environment program.[39] These in vacuo simulations were performed at 300 K for 1 ns using a timestep of 1 fs and with conformations dumped every 1 ps. The conformations obtained from these simulations make up ensemble E3.

2.2.2 DFT Calculations

After generation of the ensemble the wavefunctions of all the conformations in the ensemble were obtained from DFT calculations using the B3LYP[40] functional and the 6-31+G(d,p) basis set with 6d orbitals included. All DFT calculations were performed using GAUSSIAN09.[41]

2.2.3 Atomic Property Calculations

The IQA partitioning scheme derives from Quantum Chemical Topology (QCT),[42-44] a parameter-free approach to partitioning the wavefunction using only the gradient vector of the electron density. Under IQA the energy of a molecule, E_{IQA}^{mol}, is partitioned into the sum of atomic energies, E_{IQA}^{A}, which are composed of intra-atomic, E_{intra}^{A} (for atom A), and inter-atomic, components $V^{AA'}$, where A' is any atom but atom A. It is possible to the break down the intra- and inter-atomic energies further into kinetic, exchange-correlation and electrostatic contributions and construct FFLUX models for these various different components. However, because a previous study showed that a FFLUX model built at the E_{IQA}^{A} level performed at least as well as the combination of models built from the separate contributions,[32] all the models in the present work are built at the E_{IQA}^{A} level.

The calculations of the IQA energies from the wavefunctions were performed using the program AIMAll,[45] with the default parameter options and with the original implementation for the calculation of the two-electron integrals (i.e. not using the "TWOe implementation"[46]).

2.2.4 Construction of Models via Kriging Machine Learning

Before building the kriging models any conformations where the atomic integration error of AIMAll, $L(\Omega)$, was above a threshold value for one or more atoms were removed from the ensemble of sample conformations. For propane, models were constructed using $L(\Omega) = 0.001$, 0.0005 and 0.00025 Ha, for methanol $L(\Omega) = 0.00025$ Ha and for all other molecules $L(\Omega) = 0.0005$ Ha. A number, N_{trn}, of the remaining conformations in the sample were then randomly selected as the training set used to build the kriging model, (some of) the remaining conformations in the sample are then used as a validation set, allowing the quality of kriging models to be tested.

Full methodological details of how a property of an atom can be linked to the geometrical features through kriging machine learning were given in previous work.[35,47] In this work the kriging machine learning generates potentials that express the atomic energies E_{IQA}^A as a function of positions of all the atoms in the molecule. The predicted molecular energy of a conformation is the sum of all the predicted atomic energies, $\hat{E}_{IQA}^A$, and where each atomic energy is given by

$$\hat{E}_{IQA}^A = \mu^A + \sum_{j=1}^{N_{\text{trn}}} a_j^A \exp\left[-\sum_{h=1}^{N_{\text{feat}}} \theta_h^A \left|f_{h,j}^A - f_h^A\right|^{p_h^A}\right] \qquad (1)$$

where μ^A is the mean value of all the training data points, a_j^A is the kriging weight of training point j, θ_h^A is the activity of feature-space described by the summation index h, $f_{h,j}^A$ is the known feature value from training point j, f_h^A is the current feature for which a prediction must be made and p_h^A is the smoothness of the feature space. Kriging models can be constructed by optimizing θ_h^A and/or p_h^A. In the present study p_h^A was fixed at 2 as initial tests (see Section S1, Figure S1 and Table S1 in the Supporting Information) showed that for molecules the size of those in the present study optimizing p_h^A (as well as θ_h^A) resulted in only limited improvement, if any, in the performance of the resulting kriging models. In addition, there is a significant computational cost associated with optimizing p_h^A.

The kriging models were calculated using the in-house developed program FEREBUS,[48] with the values of the Kriging parameters optimized using particle swarm optimization to maximize the concentrated log-likelihood, as outlined previously.[48,49] FEREBUS uses the radial basis function (RBF) kernel but alternative kernels, such as the Matérn one, can also be used. Preliminary work shows that the latter can outperform the former but a systematic study will be published in the future.

2.2.5 Geometry Optimizations

Geometry optimizations were performed using an in-house modified version of the DL_POLY_4.08 program.[50] The simulations were performed using the 0 K optimizer, where, at each timestep, the system is relaxed before the integration of the equations of motion. The equations of motion were integrated using the velocity Verlet algorithm. The results of tests with different timesteps are described in Section S2 and Table S2, with models giving consistent, and accurate, results for timesteps of 0.5, 1.0 and 2.0 fs. Ultimately, a timestep of 1 fs was used for all production simulations.

3. RESULTS AND DISCUSSION

3.1 Determination of the Minimum Energy Conformations of Molecules Using FFLUX Models

Before investigating the ability of FFLUX models to reproduce the PES associated with different degrees of freedom within molecules, (vida infra) FFLUX models are compared and tested across the range of molecules - methanol, propane, *n*-butane, isobutane and NMA - in order to assess their performance at describing energy and geometry of the systems, and determine which input parameters, for example N_{trn} or $L(\Omega)$, have the greatest effect on the quality of a model.

An initial test of the quality of the generated FFLUX model is its ability to reproduce the energy of conformations in the validation set. The prediction error of a conformation is the absolute difference in the sum of the atomic energies calculated from AIMAll with those calculated using FFLUX. A plot of prediction error against the cumulative percentage of conformations in the validation set results in an S-curve. Figure 1 shows S-curves for a variety of different models, constructed using different parameters. S-curves give an indication of the quality of kriging models from the position of the curve on the x-axis (the lower the values the better the model), their gradient (the steeper the curve the better the model) and the shortness of the tail near 100% (the shorter the better). Figure 1(a) compares S-Curves for models containing 1000 conformations in the training set for all five molecules (1500 to 2500 conformations are in each validation set). The S-curves shown in Figure 1(a) suggest that the models for all five molecules are of good quality, with more than 95% of the energies of the validation conformations within 10 kJ mol[1] of their B3LYP energy in all cases. Figure 1(b) shows S-curves of

models built for propane with different number of conformations in the training set, while Figure S2 shows equivalent S-curves for n-butane, isobutane and NMA. Figure S3 presents the same error distributions as in Figure 1 but in an alternative way, which is non-cumulative (unlike an S-curve). Table 1 gives the mean absolute errors (MAE) and maximum absolute errors for the different models. In general, the greater the number of atoms in the system the more challenging it is to generate a model of an equivalent quality due to increase in the dimensionality of the features in the model. However, the particular geometric attributes of a molecule also play a role in model quality, with the MAEs for n-butane being lower than those of isobutane at the same number of training conformations. Increasing N_{tm} results in S-curves that are shifted to lower errors and lower MAE, again suggesting better quality models. Overall, it appears that for the size of molecules tested here a training set of 1000 conformations will typically result in FFLUX models, which give MAE < 4.0 kJ mol^{-1} when tested against the validation set. In contrast to increasing the number of conformations, reducing the atomic integration error threshold has a negligible effect (at least over the range tested) as seen from the S-curves for propane models with different $L(\Omega)$ ($N_{tm} = 500$ for all three models), see Figure 1(c). As changing $L(\Omega)$ resulted in no improvement all further experiments and analysis of propane models were done with models build using $L(\Omega) = 0.0005$ Ha.
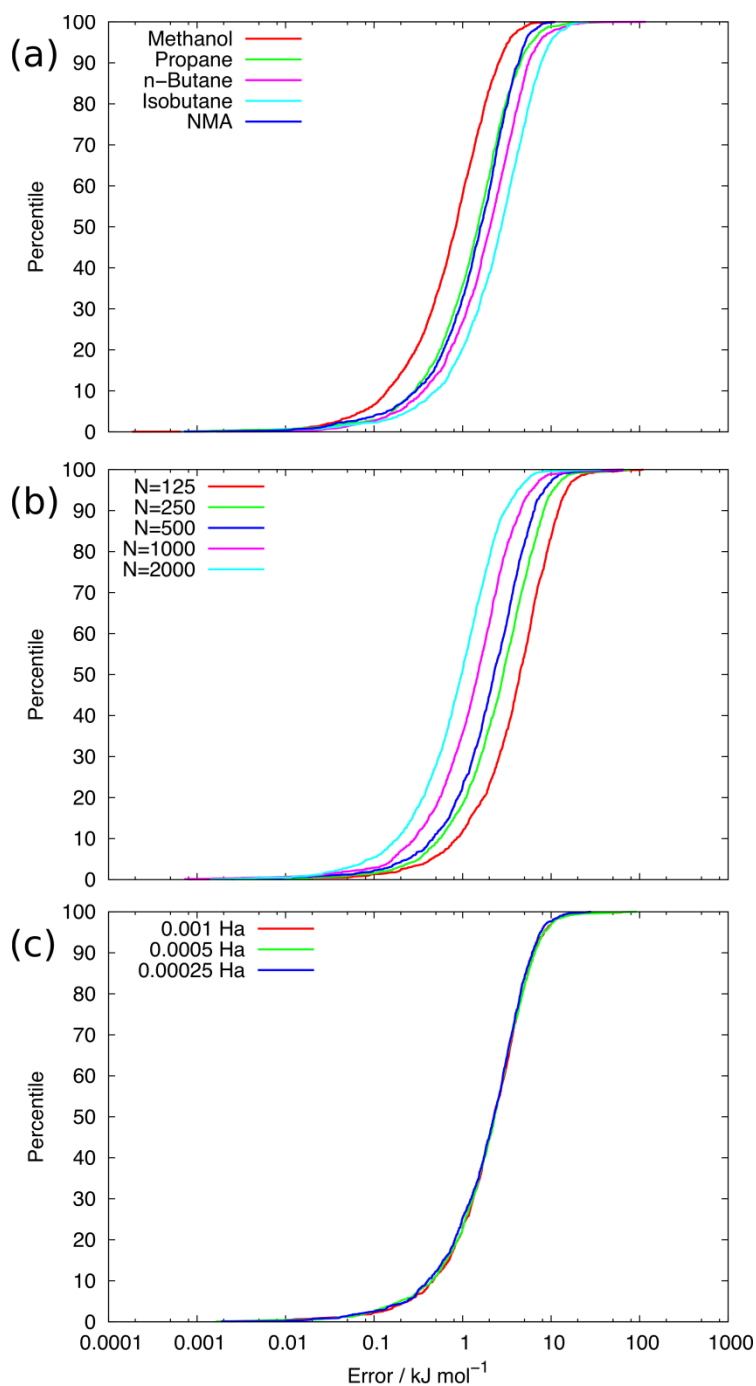
8

Figure 1. S-Curves showing the prediction error of the sum of the atomic E_{IQA}^A energies: (a) Models for the different molecules, each model built from 1000 conformations. (b) Models of propane built from varying numbers of conformations ($L(\Omega)$ constant at 0.0005 Ha). (c) Models of propane built with varying $L(\Omega)$ (N_{tm} constant at 500).

Table 1. Mean absolute errors, kJ mol^{-1}, (maximum absolute errors in parentheses) of the conformations in the validation sets for FFLUX models generated with varying numbers of conformations in the training set, N_{trn}, for propane, n-butane, isobutane and NMA.

Molecule	N_{trn}				
	125	250	500	1000	2000
Methanol	3.03 (17.71)	1.95 (18.17)	1.15 (10.68)	0.72 (7.02)	0.52 (5.02)
Propane	5.97 (110.53)	4.05 (77.94)	3.21 (65.30)	2.15 (66.21)	1.43 (36.84)
n-Butane	7.75 (105.95)	5.74 (107.66)	4.13 (74.85)	2.84 (116.78)	2.11 (30.07)
Isobutane	8.51 (214.81)	6.63 (182.17)	5.04 (129.22)	3.61 (26.40)	2.70 (161.07)
NMA	7.62 (35.21)	4.45 (38.20)	2.75 (19.07)	2.00 (11.17)	1.48 (10.55)

While S-curves provide an indication of the quality of a model they only compare the energy difference of conformations, and do not provide information about *where* the FFLUX model deviates from the PES it has been built from. To further gauge the accuracy of the different FFLUX models geometry optimizations of all conformations in the TYCHE generated samples were performed. In general, the FFLUX minimum energy conformation was an excellent match to the DFT minimum both in terms of energetics, Table 2, and geometry, Table 3. Each model optimized the geometry of more than 99% of the 4000 conformations it was tested against to the same minimum on the PES. For all the molecules except NMA, a model generated using a training set of 500 conformations was able to obtain a minimum energy within 1 kJ mol^{-1} of the value of the DFT minimum. Even in the case of NMA sub kJ mol^{-1} accuracies were obtained for models with 1000 conformations in the training set. Likewise the difference between the RMSD over all the atoms in the system, i.e. including the hydrogen atoms, of the B3LYP minimum energy conformation and the FFLUX optimized minima was excellent for methanol, propane, n-butane and isobutane while reasonable for NMA. When calculated over all atoms, the RMSD calculated for NMA is significantly higher than those for the other molecules. This worse fit is due to the fact that in a few of the FFLUX optimized conformations the hydrogens on the methyl carbons are displaced, as evidenced by the much lower values of the RMSDs calculated for the heavy atoms only. A similar issue with the optimization of methyl hydrogens was observed in our previous study on glycine[35]. Our hypothesis is that the large energy penalty that arises

10

from the distortion of the H-C-N-O dihedral "drowns out" any energy penalty that might arise from poor optimization of the methyl hydrogens, i.e. it is favorable for FFLUX to ensure the planarity of the central amide group, possibly at the expense of methyl hydrogen geometry optimization

As with the S-curves increasing the number of conformations in the training set improves the performance of the FFLUX model at capturing the energy and geometry of a global minimum. However, the results summarized in Tables 2 and 3 indicate that a "good" S-curve for the validation set does not necessarily correspond to a FFLUX model that gives minimum energy conformation in good agreement with the B3LYP value. This assertion is most obvious in the case of NMA, which yields S-curves with a lower MAE than n-butane and isobutane, yet the geometry-optimized minimum-energy conformations obtained using FFLUX are "better" in the case of n-butane and isobutane than in that of NMA. As such, while S-curves provide a reasonable initial test of a FFLUX model they do not show the full picture in regards to the quality of a model.

Table 2. Difference in energy (kJ mol^{-1}) of the minimum energy conformation obtained from DFT calculations with those determined using the FFLUX models, with varying numbers of conformations in the training set, N_{trn}.

Molecule	N_{trn}				
	125	250	500	1000	2000
Methanol	-0.758	-0.167	-0.052	-0.032	-0.016
Propane	-1.787	-0.775	-0.410	-0.167	-0.041
n-Butane	-5.224	-1.716	-0.720	-0.269	-0.129
Isobutane	-2.577	-1.234	-0.733	-0.612	-0.220
NMA	-9.876	-3.424	-3.692	-0.753	-0.088

Table 3. Root mean square difference (RMSD), Å, between the minimum energy conformation obtained from DFT calculations with those determined using the FFLUX models, with varying numbers of conformations in the training set, N_{trn}. RMSD are calculated over all atoms and over all the non-hydrogen atoms.

Molecule	Atoms included in RMSD calculation	N_{trn}				
		125	250	500	1000	2000
Methanol	All	0.013±0.001	0.013±0.001	0.009±0.001	0.008±0.001	0.002±0.001
	Non-H	0.001±0.001	0.001±0.001	0.001±0.001	0.001±0.001	0.001±0.001
Propane	All	0.071±0.002	0.011±0.001	0.043±0.001	0.014±0.001	0.004±0.001
	Non-H	0.042±0.001	0.010±0.004	0.010±0.004	0.008±0.005	0.006±0.005
n-Butane	All	0.060±0.001	0.037±0.001	0.026±0.001	0.016±0.001	0.026±0.002
	Non-H	0.028±0.001	0.024±0.001	0.007±0.001	0.009±0.001	0.007±0.001
Isobutane	All	0.116±0.006	0.022±0.001	0.023±0.001	0.017±0.001	0.012±0.001
	Non-H	0.012±0.001	0.007±0.001	0.006±0.001	0.008±0.001	0.004±0.001
NMA	All	0.105±0.077	0.199±0.084	0.179±0.147	0.147±0.006	0.039±0.021
	Non-H	0.040±0.029	0.034±0.004	0.018±0.002	0.013±0.001	0.028±0.016

Overall these results show that FFLUX models are able to reproduce the energy and geometry of the minimum energy conformations to a very high level. As might be expected, increasing the number of conformations used for training the FFLUX model usually leads to an improvement in the agreement between the model and the target PES. For all the molecules tested except NMA, a training set of 500 conformations (generated via a temperature-based normal mode distortion method) is sufficient to result in a difference in energy of less than 1 kJ mol[-1] between the FFLUX and DFT minimum energy conformations.

3.2 Evaluation of the Performance of Different Training Ensembles for Methanol

It is clear from both the above results and previous studies[32] that one of the most important elements in constructing an accurate FFLUX model is the ensemble of conformations used in the training of the machine learned potential. Increasing the number of conformations in the training set (generally) leads to improvement in the accuracy of the resulting model, but is also more costly both in terms of constructing the model and in its evaluation at each MD timestep. Thus, it is highly desirable to use a method of generating training conformations that can give a model with high degree of accuracy constructed from only a small number of conformations. To this end, ensembles of conformations were generated for methanol using three different procedures and the resulting models compared. Ensemble E1 was generated using a temperature-based distortion of normal modes, ensemble E2 was generated from randomly sampled distortion of normal modes within limits, and ensemble E3 generated from extracting random snapshots of an in vacuo MD simulation of a methanol molecule at 300 K.

Figure 2 shows the S-curves generated with varying number of conformations in the training set for each ensemble, the mean and maximum absolute errors for each model against the validation sets are summarized in Table 4. It is immediately obvious that E1 and E3, which are broadly comparable with each other, result in much better S-curves than E2, and have significantly lower mean and maximum absolute errors. Geometry optimizations were performed in order to test the molecules further, with the difference in energies of the B3LYP and FFLUX optimized conformations, ΔE_{min}, (Table 4) greatest for the models generated using E2. However, even in the case of E2, a model with 500 conformations in the training set sufficed to ensure $\Delta E_{min} < 1$ kJ mol^{-1}. Comparing E1 and E3, the latter shows a small improvement over the former both regarding the errors measured from the validation set and ΔE_{min}. However, both models return excellent agreement with the B3LYP description of the molecule. The greater accuracy of E1 and E3 indicates that ensembles of conformations used to construct the FFLUX model should contain a reasonable percentage of conformations close to the global minimum. Thus, ensemble E2, which is not weighted by any temperature dependence, provides the kriging machine learning models with as much information about the system in high energy states as in low energy states. We note that a FFLUX model might result in seemingly larger errors when tested against the validation set because it samples more "difficult" parts of conformational space, which may result in a worse S-curve. Population distributions of the energies of the conformations that are used in the training sets for ensembles E1, E2 and E3 do suggest that the poor performance of E2

13

is partly due its use of high energy conformations in the training set (see Figure S4). Moreover, as highlighted in recent work, the level of error observed against a validation set of conformations may be higher than the level of error observed in the calculation of other properties, e.g. the vibrational spectrum[51]. In other words, just because it results in a poor S-Curve, E2 may still be able to reproduce other properties of the system with a reasonable degree of accuracy. However, for the present study the key motivation is to reproduce the PES of the system, at which E2 is clearly less successful than the other two ensembles.

While the above results indicate the need for training sets to provide detailed information of the system at the minimum, an accurate description of the minimum is necessary but not sufficient for a FF to reproduce the PESs of a system. The next step (see Section 3.3) in testing FFLUX is determining where the FFLUX models deviate most strongly from the PES that they are attempting to reproduce.
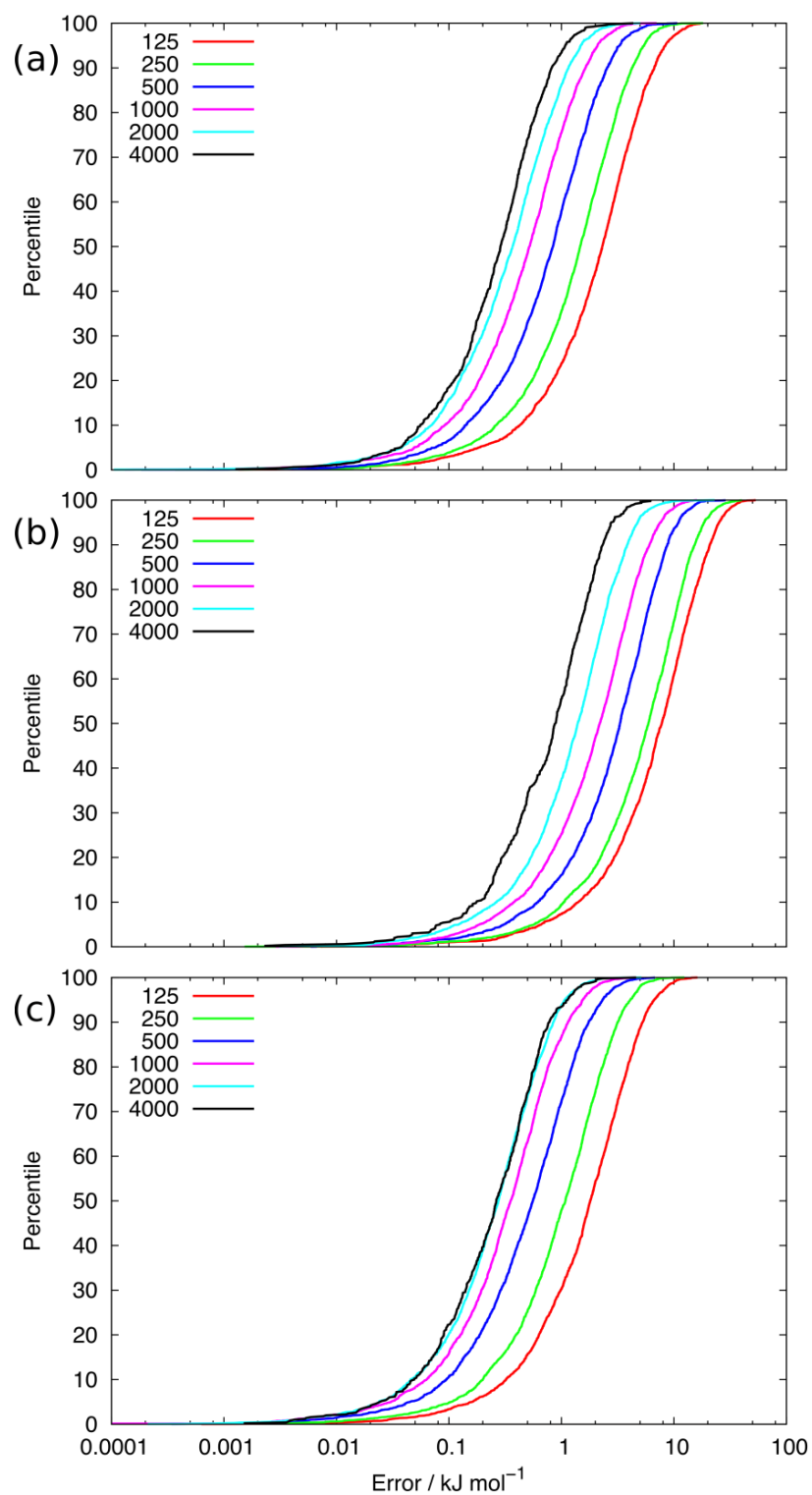
Figure 2. S-Curves showing the prediction error of the sum of the atomic E_{IQA}^{A} energies for different models of methanol, made using (a) ensemble E1 (TYCHE), (b) ensemble E2 (EROS) and (c) ensemble E3 (MD).

15

Table 4. Mean absolute errors (MAE) and maximum absolute errors of the conformations in the validation sets and the difference in energy between the minimum energy conformations predicted by the DFT calculations and by FFLUX, ΔE_{min}, for models generated from the three different ensembles of methanol conformations and with varying numbers of conformations in the training sets, N_{trn}.

N_{trn}	MAE / kJ mol^{-1}			Maximum / kJ mol^{-1}			ΔE_{min} / kJ mol^{-1}		
	E1	E2	E3	E1	E2	E3	E1	E2	E3
125	3.03	9.72	2.42	17.71	53.39	16.18	-0.758	-3.530	-0.620
250	1.95	7.32	1.45	18.17	40.12	12.39	-0.167	-2.839	-0.136
500	1.15	4.16	0.79	10.68	28.73	6.78	-0.052	-0.923	-0.059
1000	0.72	2.80	0.51	7.02	16.01	4.50	-0.032	-0.462	-0.022
2000	0.52	1.81	0.38	5.02	23.02	5.14	-0.016	-0.089	-0.014
4000	0.39	1.11	0.38	4.36	6.21	4.61	-0.013	-0.042	-0.014

3.3 Ability of FFLUX to Describe Potential Energy Surfaces

The ability of a FF to reproduce the energy of minimum energy conformations alone is insufficient for a FF to model molecular systems accurately. Ideally a FF needs to be able to reproduce the full PES of a system and, in the best case scenario, with transferability across a wide range of conditions. In practice few FFs meet this ideal model but at the very least a FF needs to be able to reproduce the PES(s) around minima, and the energy difference between different minima with a reasonable degree of accuracy. To see how FFLUX performs in this regard, the PES of the C-O bond stretch, C-O-H angle bend and H-C-O-H dihedral angle rotation in methanol have been mapped using the B3LYP functional, and the energy difference between the DFT calculations and different FFLUX models compared.

The PES along the bond/angle/dihedral was obtained by performing a relaxed scan along the degree of freedom in question (i.e. performing a geometry optimization for all the other degrees of freedom in the molecule) at the B3LYP/6-31+G(d,p) level. The scan along the C-O bond was performed over the range 1.285-1.585 Å in 0.01 Å steps, the scan of the C-O-H angle from 93-125° in 1° steps, and the scan over the H-C-O-H dihedral over the full 360° in 5° steps (giving a total of 31, 33 and 72 conformations for the bond, angle and dihedral scans, respectively). The energy of conformations obtained from the DFT calculations was then compared with their energy within different FFLUX models. As in the case of the S-curves and ΔE_{min} the performance of models derived from E2 was much worse than models derived from the other two ensembles, as such the discussion below concentrates on models constructed from E1 and E3.
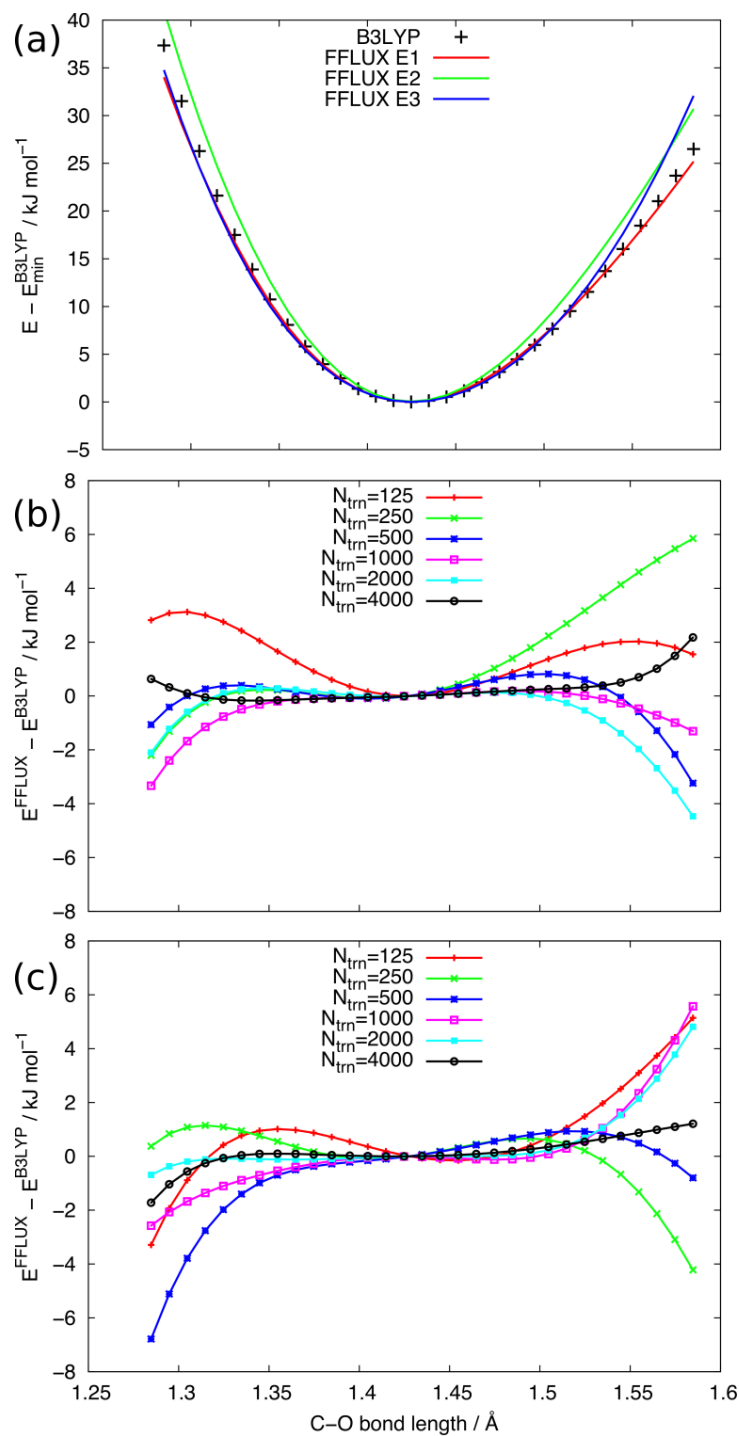
Figure 3. Comparison of the B3LYP and FFLUX models potential energy surfaces of the C-O bond length in methanol: (a) energy of conformations relative to the B3LYP minimum energy conformation for FFLUX models with 1000 conformations in the training set, (b) and (c) the difference in energy between FFLUX and DFT for models constructed from (b) E1 and (c) E3.
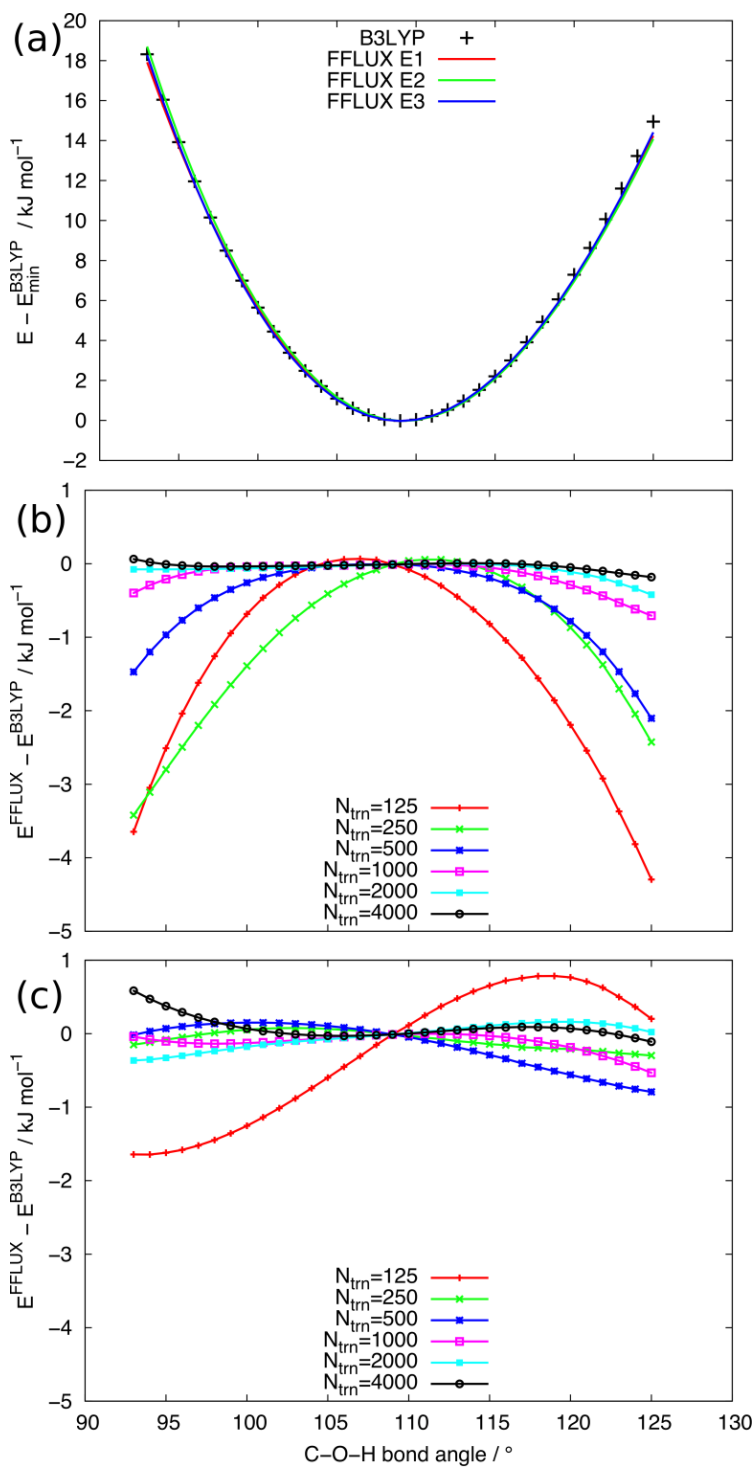
Figure 4. Comparison of the B3LYP and FFLUX models potential energy surfaces of the C-O-H bond angle in methanol: (a) energy of conformations relative to the B3LYP minimum energy conformation for FFLUX models with 1000 conformations in the training set. (b) and (c) the difference in energy between FFLUX and DFT for models constructed from (b) E1 and (c) E3.

The performance of the FFLUX models in describing the PES of the relaxed scan of the C-O bond length and C-O-H bond angle are shown in Figures 3 and 4, respectively (see also Figure S5). There are a number of points to note from these results. First, as in the case of the S-curves and ΔE_{min}, the performance of models derived from E2 was much worse than models derived from the other two ensembles (see Figure S5), and with the models constructed from ensemble E3 slightly better than those constructed from E1. Second, the FFLUX models typically show very good agreement to B3LYP at the center of the PES (close to the optimized geometry) but become less accurate the further the bond length/angle is from the optimal value. Third, the C-O-H bond angle is reproduced by the FFLUX models to a greater level of accuracy than the the C-O bond stretch, due to the greater pliability of the bond angle bend. Fourth, whereas FFLUX models built from only 125 conformations are able to reproduce the minimum energy with a high degree of accuracy (< 1 kJ mol^{-1}), achieving the same level of accuracy across the full range of the PESs tested requires a greater number of training conformations, especially in the case of the C-O bond stretch.

In the case of the bond stretch and angle bend the PES have a single minimum and, on the range sampled, no maxima. In contrast the PES of the methanol H-C-O-H dihedral, being periodic, has three energetically equivalent minima and maxima. It should be noted that in the present formulation of FFLUX each atom is distinct, so that the three hydrogens bonded to the carbon atom are not equivalent within the FFLUX model. Previous work has also indicated that while normal mode sampling methods tend to sample bond lengths and angle quite well, they have a harder time sampling dihedral angles.[36]
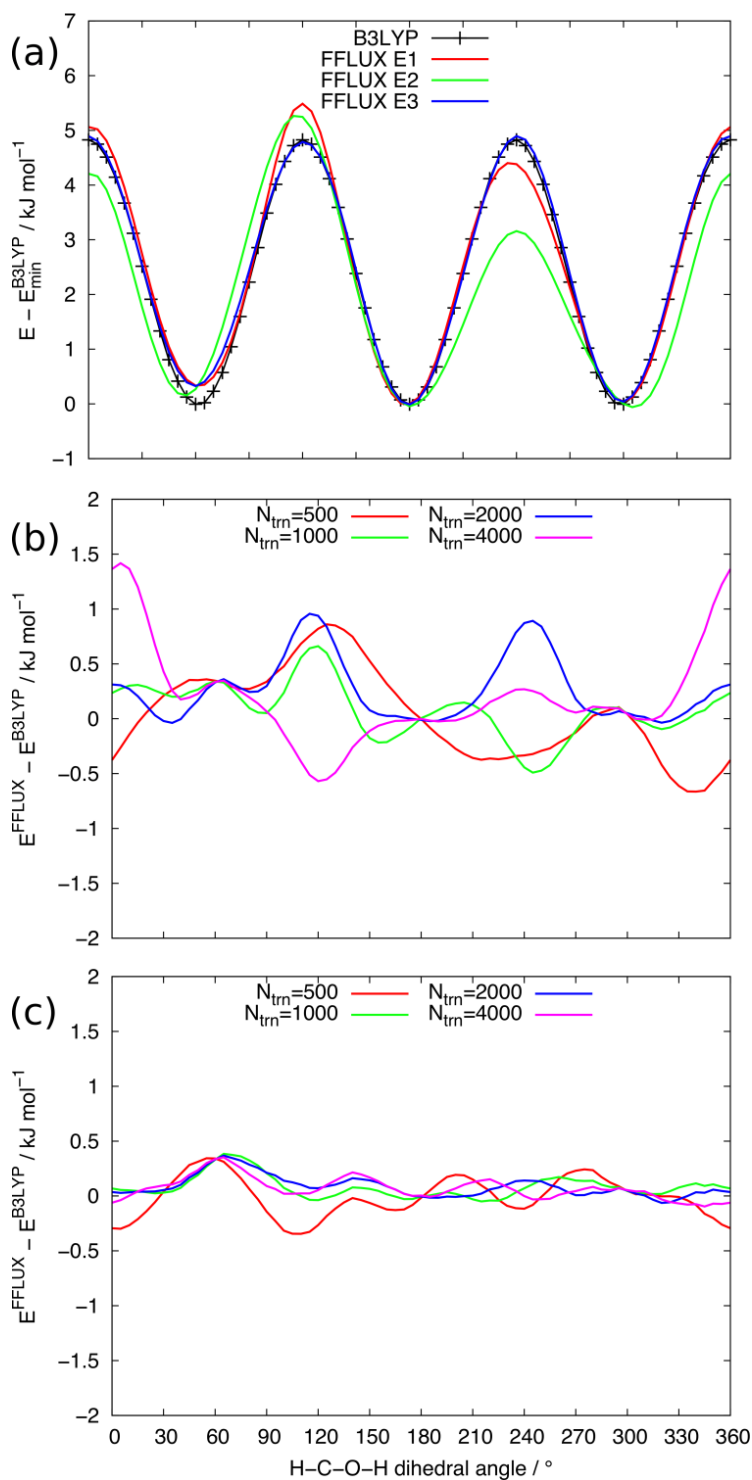
Figure 5. Comparison of the B3LYP and FFLUX models potential energy surfaces of the H-C-O-H dihedral angle of methanol: (a) energy of conformations relative to the B3LYP minimum energy conformation for FFLUX models with 1000 conformations in the training set, (b) and (c) the difference in energy between FFLUX and DFT for models constructed from (b) E1 and (c) E3.

Thus, accurately reproducing the PES of the H-C-O-H dihedral is a tough challenge for the FFLUX models. The PESs of the dihedral angle at the B3LYP level and for FFLUX models containing 1000 conformations in their training set is shown in Figure 5(a). Yet again the models built from the E2 ensemble perform significantly worse than those constructed from other two ensembles (Figure S5(c)), with even a model trained from 1000 conformations predicting the location of some of the minima/maxima wrongly. For the E1 models 500+ conformations in the training set is enough to ensure the location of the stationary points in the PES are captured accurately, and with a maximum error in the energy of <1.5 kJ mol⁻¹. The models based on E3 perform even better, 500 conformations is enough to ensure deviations in the energy of <0.5 kJ mol⁻¹. Unfortunately even the best model (E3 N_{tm}=4000) fails in one respect, the three minima are not energetically equivalent, with the minimum at 60° being 0.33 kJ mol⁻¹ higher in energy than the other two minima. This is not a direct failing of FFLUX but, in this case, arises from the fact that in the calculation of the E_{IQA}^A energies, AIMALL introduces atomic integration errors[52] due to the numerical imprecision of the quadrature over complex volumes, resulting in the three minima becoming slightly non-equivalent energetically.

In consequence of the successful reproduction of the PES of the different normal modes of methanol by FFLUX models, it was decided to test if the FFLUX models could reproduce a complex PES, such as that defined by the C-C-C-C dihedral rotation in *n*-butane. Not only is *n*-butane a larger molecule than methanol but the C-C-C-C dihedral PES has energy barriers 3 or 5 times the size of those in the H-C-O-H dihedral rotation of methanol and also has non-equivalent minima (and maxima). The ensemble of configurations that were to make up the training set for the new model of *n*-butane was constructed as follows. First, a relaxed scan of the C-C-C-C dihedral of *n*-butane, in 5° steps, was performed at the B3LYP level, resulting in 73 configurations (0/360° was included twice). Second, each of these 73 conformations was used as seed geometries for a temperature based normal mode distortion, which resulted in the generation of 100 distorted conformations from each seed. Finally, a random 4000 of the 7300 sample conformations were used to construct the training set of the model. Geometry optimizations using the resulting FFLUX model were then performed starting from each of the 73 seed conformations. Figure 6(a) shows the PES of the C-C-C-C dihedral for B3LYP DFT calculations and for the FFLUX model, as well of each of the 73 seed conformations as determined from AIMAll calculations (it should be noted that these 73 seed conformations *do not* form part of the training set for the FFLUX model). Figure 6(b) shows the difference between the energies calculated with AIMAll/FFLUX and the B3LYP energies. FFLUX is able to reproduce large parts of the PES

with a high degree of accuracy: even the worst energy difference is less than 2.5 kJ mol[-1]. However, there are parts of the PES that the FFLUX model does not capture accurately. Interestingly, while the global minimum and maximum are reproduced well by the FFLUX model, the description of the PES at the local minima/maxima is poor. In fact, the description of the PES at ~120/240° is so poor that false minima have been introduced to the PES, geometry optimizations performed from these starting conformations result in optimized conformations with a dihedral angle of ~120/240°, i.e. the FFLUX models have local minima at these dihedral angles.

The reason for the inaccuracies in the FFLUX model may arise from (1) deviations from the B3LYP PES introduced during the partitioning of the wavefunction into the atomic E_{IQA}^{A} by AIMAll and/or (2) the kriging model constructed not accurately reproducing the AIMAll PES. Comparison of the AIMAll and B3LYP energies suggests that both of the areas probably need improvement. Certainly the AIMAll calculations do introduce some level of error, see Figure 6(b). In fact, the difference between the AIMAll and B3LYP energies (MAE = 0.38 kJ mol[-1]) is not that much less than the difference between the FFLUX and B3LYP energies (MAE = 0.67 kJ mol[-1]).
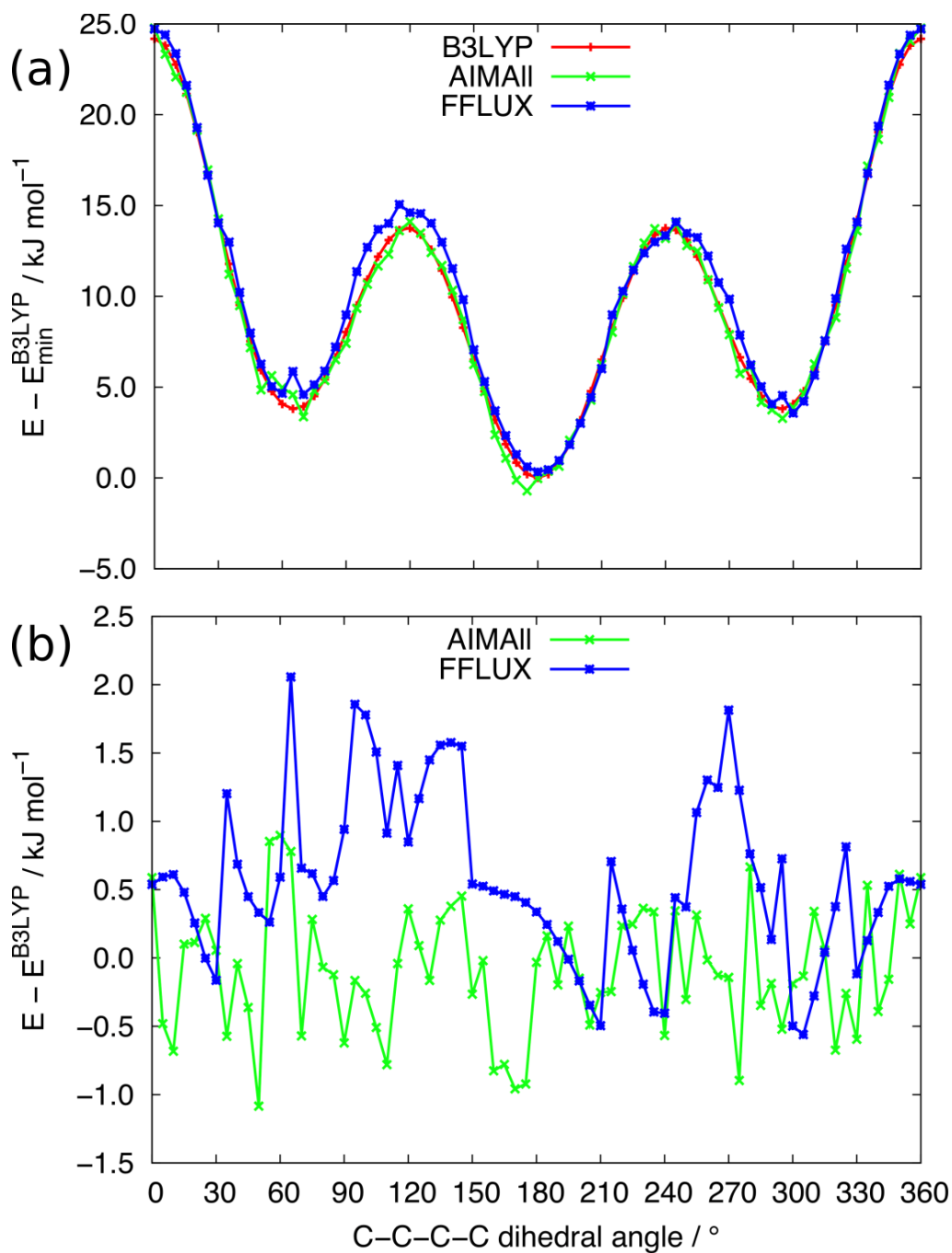
Figure 6. Comparison of the B3LYP and FFLUX models potential energy surfaces of the C-C-C-C dihedral of *n*-butane: (a) energy of conformations relative to the B3LYP minimum energy conformation for B3LYP calculations, AIMAll calculations and the FFLUX model. (b) The difference in energy between the AIMAll calculations/FFLUX model and the B3LYP calculations.

However, the false local maxima that occur in the FFLUX model at 60 and 330° suggest that there are also deficiencies in the kriging model itself (although the fact that the AIMAll PES is rather "rugged" around some of the stationary points may be partly responsible). These deficiencies of the kriging models can arise from a number of factors such poorly optimized θ_h^A parameters or a poor distribution of training points.

3.4 Assessment of FFLUX Models and Future Developments

The above results show that, in principle, kriging machine learning methods can generate a model that is capable of reproducing a given PES (in this case the PES described by the B3LYP functional), with a considerable degree of accuracy. The method of FFLUX model construction detailed in the present work is able to reproduce the energies and geometries of the global minimum to < 0.5 kJ mol⁻¹ and < 0.05 Å, respectively. In addition, FFLUX models are able to provide a good description of the PES of specific normal modes within molecules, reproducing energy barriers and stationary points other than the global minimum. However, the agreement between FFLUX and the target PES can be improved further, both by revisions within the current methodological framework as well as by longer term future developments, including adaptive sampling[53], biasing[54] with metadynamics and the incorporation of symmetry[19] into the models. Finally, we may consider Bayesian optimization[55] as an appealing global optimization method to be used as an alternative to particle swarm optimization in order to optimize the hyperparameters. This method can also be applied to determine the ideal sampling of the PES for kriging, as was done before.[56]

Regarding the improvements to models that can be made with revisions to the current procedure, the first modification could be to try and reduce the errors between the AIMAll and B3LYP (or other target) PESs. These AIMAll integration errors do not appear to flow on into the FFLUX model in a manner that leads to a poor description of the PES close to the global minimum. However, such integration errors can become significant when looking at the description of specific PESs. Refinements to AIMAll parameters, such as increasing the quality of the integration grid, may bring the AIMAll PES closer to that of the target. The second area for revision is regarding the conformations that are used to construct the models. Increasing the number of conformations that a model is constructed from is one route to increasing model performance (generally), though with diminishing returns and increasing computational cost. A more effective approach is to use an alternative method of conformation generation that results in a more efficient ensemble. The results of the current work

have shown that an ensemble of conformations generating from non-Boltzmann-weighted sampling of normal modes is significantly less efficient (for molecules of 6 to 14 atoms) than one generated from Boltzmann- weighted sampling of normal modes. Moreover, an ensemble generated from in vacuo MD simulations seems more efficient still. A more through-going analysis of the performance of ensembles generated from MD simulations is an area of future work. Finally, very recent preliminary work in our group on adaptive sampling[53] proved that the previously applied training sets used to model the water monomer were much larger than necessary. A reduction in training set size is possible by a few dozen times, without loss of accuracy. Such reduction is important because the evaluation of a kriging model is of the order of the number of its training points, and the ultimate use of FFLUX will be in the MD simulation of condensed matter.

4. CONCLUSIONS

The development of alternatives to the traditional (bio)molecular force fields is an important future step for computational chemistry. By combining the kriging machine learning method with the Interacting Quantum Atoms quantum chemical topology partitioning scheme it has been possible to develop potentials that contain quantum mechanical information in their description of atomic interactions. The resulting FFLUX models are able to reproduce minimum energy conformations of molecules to <1 kJ mol^{-1} and <0.05 Å (in terms of the root-mean square deviation) on the DFT reference values. In addition, the ability of the FFLUX models to map out PESs associated with specific degrees of freedom in methanol and n-butane has been investigated, with the best performing models able to achieve a <2.5 kJ mol^{-1} accuracy across the full range of the PESs tested. The performance of the FFLUX models depends on both the size and method of generation of the ensemble of conformations used to construct the model. While increasing the number of conformations in the training ensemble typically leads to an improvement in the model there is diminishing return and an increasing computational cost associated with simply adding more and more conformations to the training ensemble. A more fruitful area for model construction to concentrate on is to refine/develop new techniques for the generation of an ensemble. FFLUX models need a high density of conformations at/near minima to give high accuracy for these features but also need higher energy conformations sampled so that the model can describe the full PES of a system. The present work has shown that

ensembles that contain a broad population distribution result in less effective FFLUX models than ensembles made with a Boltzmann distribution. The next step in the development of FFLUX is to explore ensemble generation techniques that are able to reduce to number of conformations needed to model systems accurately.

Supporting Information

The Supporting Information is free of charge on the ACS publications website.

Analysis of the effect of optimizing p_h^A; analysis of the effect of molecular dynamics timestep used; S-curves for n-butane, isobutane and NMA for varying N_{tm}; energy difference between the PESs of methanol constructed from ensemble E2 and the DFT calculations (PDF).

AUTHOR INFORMATION

Corresponding Author

*Phone: +44 161 3064511. E-mail: pla@manchester.ac.uk

Present Addresses

†Present Address: School of Chemistry and Biosciences, University of Bradford, Bradford, BD7 1DP, UK.

REFERENCES

(1) Houk, K. N.; Lui, F. Holy Grails for Computational Organic Chemistry and Biochemistry *Acc*. *Res*. *Chem*. **2017**, *50*, 539.

(2) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; De Groot, B. L.; Grubmüller, H.; MacKerell, A. D. CHARMM36m: An Improved Force Field for Folded and Intrinsically Disordered Proteins. *Nat*. *Methods* **2016**, *14*, 71.

(3) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. Ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from Ff99SB. *J*. *Chem*. *Theory Comput*. **2015**, *11*, 3696.

(4) Schmid, N.; Eichenberger, A. P.; Choutko, A.; Riniker, S.; Winger, M.; Mark, A. E.; Van Gunsteren, W. F. Definition and Testing of the GROMOS Force-Field Versions 54A7 and 54B7 *Eur*. *Biophys*. *J*. **2011**, *40*, 843.

(5) Lemkul, J. A.; Huang, J.; Roux, B.; MacKerell, A. D. J. An Empirical Polarizable Force Field Based on the Classical Drude Oscillator Model: Development History and Recent Applications *Chem*.*Rev*. **2016**, *116*, 4983−5013.

(6) Albaugh, A.; Boateng, H. A.; Bradshaw, R. T.; Demerdash, O. N.; Dziedzic, J.; Mao, Y.; Margul, D. T.; Swails, J.; Zeng, Q.; Case, D. A.; Eastman, P.; Wang, L.-P.; Essex, J. W.; Head-Gordon, M.; Pande, V. S.; Ponder, J. W.; Shao, Y.; Skylaris, C.-K.; Todorov, I. T.; Tuckerman, M. E.; Head-Gordon, T. Advanced Potential Energy Surfaces for Molecular Simulation. *J*.*Phys*.*Chem*.*B* **2016**, *120*, 9811.

(7) Ren, P. Y.; Wu, C. J.; Ponder, J. W. Polarizable Atomic Multipole-Based Molecular Mechanics for Organic Molecules *J*.*Chem*.*Theory Comput*. **2011**, *7*, 3143.

(8) Holt, A.; Boström, J.; Karlström, G.; Lindh, R. A NEMO potential that includes the dipole–quadrupole and quadrupole–quadrupole polarizability *J*.*Comput*.*Chem*. **2010**, *31*, 1583.

(9) Chaudret, R.; Gresh, N.; Parisel, O.; Piquemal, J. P. Many-Body Exchange-Repulsion in Polarizable Molecular Mechanics. I. Orbital-Based Approximations and Applications to Hydrated Metal Cation Complexes *J*.*Comput*.*Chem*. **2011**, *32*, 2949.

(10) Vinter, J. G. Extended Electron Distributions Applied to the Molecular Mechanics of some Intermolecular Interactions *J*.*Comput*.*Aided Mol*. *Des*. **1994**, *8*, 653.

(11) Ghosh, D.; Kosenkov, D.; Vanovschi, V.; Williams, C. F.; Herbert, J. M.; Gordon, M. S.; Schmidt, M. S.; Slipchenko, L. V. Noncovalent Interactions in Extended Systems Described by the Effective Fragment Potential Method: Theory and Application to Nucleobase Oligomers *J*.*Phys*.*Chem*.*A*. **2010**, *114*, 12739.

(12) Bardwell, D. A.; Adjiman, C. S.; Arnautova, Y. A.; Bartashevich, E.; Boerrigter, S. X. M.; Braun, D. E.; Cruz-Cabeza, A. J.; Day, G. M.; Della Valle, R. G.; Desiraju, G. R.; van Eijck, B. P.; Facelli, J. C.; Ferraro, M. B.; Grillo, D.; Habgood, M.; Hofmann, D. W. M.; Hofmann, F.; Jose, K. V. J.; Karamertzanis, P. G.; Kazantsev, A. V.; Kendrick, J.; Kuleshova, L. N.; Leusen, F. J. J.; Maleev, A. V.; Misquitta, A. J.; Mohamed, S.; Needs, R. J.; Neumann, M. A.; Nikylov, D.; Orendt, A. M.; Pal, R.; Pantelides, C. C.; Pickard, C. J.; Price, L. S.; Price, S. L.; Scheraga, H. A.; van de Streek, J.; Thakur, T. S.; Tiwari, S.; Venuti, E.; Zhitkov, I. K. Towards crystal structure prediction of complex organic compounds - a report on the fifth blind test *Acta Cryst*.*B* **2011**, *67*, 535.

(13) Cardamone, S.; Hughes, T. J.; Popelier, P. L. A. Multipolar Electrostatics *Phys*.*Chem*.*Chem*.*Phys*. **2014**, *16*, 10367.

(14) Wheatley, R. J.; Mitchell, J. B. O. Gaussian Multipoles in Practice: Electrostatic Energies for Intermolecular Potentials *J*.*Comput*.*Chem*. **1994**, *15*, 1187.

(15) Volkov, A.; King, H. F.; Coppens, P.; Farrugia, L. On the calculation of the electrostatic potential, electric field and electric field gradient from the aspherical pseudoatom model. *Acta Cryst*. **2006**, *A62*, 400–408.

(16) Ren, P.; Ponder, J. W. Polarizable atomic multipole water model for molecular mechanics simulation *J*.*Phys*.*Chem*.*B* **2003**, *107*, 5933.

(17) Hodges, M. P.; Stone, A. J.; Xantheas, S. S. Contribution of Many-Body Terms to the Energy for Small Water Clusters: A Comparison of ab Initio Calculations and Accurate Model Potentials *J*.*Phys*.*Chem*.*A* **1997**, *101*, 9163.

(18) Kumar, R.; Wang, F.-F.; Jenness, G.; Jordan, K. A. A second generation distributed point polarizable water model *J*.*Chem*.*Phys*. **2010**, *132*, 014309.

(19) Uteva, E.; Graham, R. S.; Wilkinson, R. D.; Wheatley, R. J. Interpolation of intermolecular potentials using Gaussian processes *J*.*Chem*.*Phys*. **2017**, *147*, 161706.

(20) Behler, J. Representing potential energy surfaces by high-dimensional neural network potentials *J*.*Phys*.*Condens*.*Matter* **2014**, *26*, 183001.

(21) Behler, J. Perspective: Machine learning potentials for atomistic simulations. *J*.*Chem*.*Phys*. **2016**, *145*, 170901

(22) Bartok, A. P.; Gillan, M. J.; Manby, F. R.; Csanyi, G. Machine-learning approach for one- and two-body corrections to density functional theory: Applications to molecular and condensed water *Phys*.*Rev*.*B* **2013**, *88*, 054104.

(23) Delhommelle, J. Editorial *Molecular Simulation* **2018**, *44*, 1032.

(24) Sosso, G. C.; Deringer, V. L.; Elliott, S. R.; Csanyi, G. Understanding the thermal properties of amorphous solids using machine-learning-based interatomic potentials *Molecular Simulation* **2018**, *44*, 866.

(25) Mittal, S.; Shukla, D. Recruiting machine learning methods for molecular simulations of proteins *Molecular Simulation* **2018**, *44*, 891.

(26) Swann, E.; Sun, B.; Cleland, D. M.; Barnard, A. S. Representing molecular and materials data for unsupervised machine learning *Molecular Simulation* **2018**, *44*, 905.

(27) Hachmann, J.; Afzal, M. A. F.; Haghighatlari, M.; Pal, Y. Building and deploying a cyberinfrastructure for the data-driven design of chemical systems and the exploration of chemical space *Molecular Simulation* **2018**, *44*, 921.

(28) Thurston, B. A.; Ferguson, A. L. Machine learning and molecular design of self-assembling pi-conjugated oligopeptides *Molecular Simulation* **2018**, *44*, 930.

(29) Blanco, M. A.; Martín Pendás, A.; Francisco, E. Interacting quantum atoms: a correlated energy decomposition scheme based on the quantum theory of atoms in molecules. *J.Chem.Theor.Comput.* **2005**, *1*, 1096.

(30) Popelier, P. L. A. QCTFF: On the Construction of a Novel Protein Force Field *Int.J.Quant.Chem.* **2015**, *115*, 1005.

(31) Popelier, P. L. A. Molecular Simulation by Knowledgeable Quantum Atoms *Phys.Scr.* **2016**, *91*, 033007.

(32) Zielinski, F.; Maxwell, P. I.; Fletcher, T. L.; Davie, S. J.; Di Pasquale, N.; Cardamone, S.; Mills, M. J. L.; Popelier, P. L. A. Geometry Optimization with Machine Trained Topological Atoms *Scientific Reports* **2017**, *7*, 12817.

(33) Cressie, N. *Statistics for Spatial Data*; Wiley, New York, USA, 1993.

(34) Rasmussen, C. E.; Williams, C. K. I. *Gaussian Processes for Machine Learning.*; The MIT Press: Cambridge, USA, 2006.

(35) Thacker, J. C. R.; Wilson, A. L.; Hughes, Z. E.; Burn, M. J.; Maxwell, P. I.; Popelier, P. L. A. Towards the simulation of biomolecules: optimisation of peptide-capped glycine using FFLUX *Molec.Simulation* **2018**, *under review*.

(36) Hughes, T. J.; Cardamone, S.; Popelier, P. L. A. Realistic Sampling of Amino Acid Geometries for a Multipolar Polarizable Force Field *J.Comput.Chem.* **2015**, *36*, 1844.

(37) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost *Chem.Sci.* **2017**, *8*, 3192.

(38) Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490.

(39) MOE,2013.08 ed.; Chemical Computing Group ULC, 1010 Sherbrooke St. West, Suite #910, Montreal, QC, H3A 2R7, CANADA: 2018.

(40) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. AB-INITIO CALCULATION OF VIBRATIONAL ABSORPTION AND CIRCULAR-DICHROISM SPECTRA USING DENSITY-FUNCTIONAL FORCE-FIELDS *J.Phys.Chem.* **1994**, *98*, 11623.

(41) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J.; Gaussian, Inc.: Wallingford, CT, USA, 2009.

(42) Popelier, P. L. A. In *Structure and Bonding. Intermolecular Forces and Clusters, Ed, D.J.Wales*; Springer: Heidelberg, Germany, 2005; Vol. 115, p 1.

(43) Popelier, P. L. A. In *Challenges and Advances in Computational Chemistry and Physics dedicated to "Applications of Topological Methods in Molecular Chemistry"*; Chauvin, R., Lepetit, C., Alikhani, E., Silvi, B., Eds.; Springer: Switzerland, 2016, p 23.

(44) Ayers, P. L.; Boyd, R. J.; Bultinck, P.; Caffarel, M.; Carbó-Dorca, R.; Causá, M.; Cioslowski, J.; Contreras-Garcia, J.; Cooper, D. L.; Coppens, P.; Gatti, C.; Grabowsky, S.; Lazzeretti, P.; Macchi, P.; Pendás, A. M.; Popelier, P. L. A.; Ruedenberg, K.; Rzepa, H.; Savin, A.; Sax, A.; Schwarz, W. E. H.; Shahbazian, S.; Silvi, S.; Solà, M.; Tsirelson, V. Six questions on topology in theoretical chemistry *Comput.Theor.Chem.* **2015**, *1053*, 2.

(45) AIMALL, Keith, T. A. Overland Park KS, USA, 2016.

(46) Polestshuk, P. M. Accurate Integration Over Atomic Regions Bounded by Zero-Flux Surfaces *J.Comp.Chem.* **2013**, *34*, 206.

(47) Mills, M. J. L.; Popelier, P. L. A. Electrostatic Forces: formulae for the first derivatives of a polarisable, anisotropic electrostatic potential energy function based on machine learning. *J.Chem.Theory Comput*. **2014**, *10*, 3840–3856.

(48) Di Pasquale, N.; Bane, M.; Davie , S. J.; Popelier, P. L. A. FEREBUS: Highly Parallelized Engine for Kriging Training *J.Comput.Chem*. **2016**, *37*, 2606.

(49) Di Pasquale, N.; Davie, S. J.; Popelier, P. L. A. Optimization Algorithms in Optimal Predictions of Atomistic Properties by Kriging *J.Chem.Theor.Comp*. **2016**, *12*, 1499.

(50) Todorov, I. T.; Smith, W.; Trachenko, K.; Dove, M. T. DL_POLY_3: new dimensions in molecular dynamics simulations via massive parallelism *J.Mater.Chem*. **2006**, *16*, 1911.

(51) Kamath, A.; Vargas-Hernandez, R. A.; Krems, R. V.; Carrington, J. T.; Manzhos, S. Neural networks vs Gaussian process regression for representing potential energy surfaces: A comparative study of fit quality and vibrational spectrum accuracy *J.Chem.Phys*. **2018**, *148*, 241702.

(52) Aicken, F. M.; Popelier, P. L. A. Atomic properties of selected biomolecules. Part 1. The interpretation of atomic integration errors *Can.J.Chem*. **2000**, *78*, 415.

(53) Liu, H.; Cai, J.; Ong, Y.-S. An adaptive sampling approach for Kriging metamodeling by maximizing expected prediction error. *Comp.Chem.Eng*. **2017**, *106*, 171.

(54) Mones, L.; Bernstein, N.; Csányi, G. Exploration, Sampling, And Reconstruction of Free Energy Surfaces with Gaussian Process Regression *J.Chem.Theory Comput*. **2016**, *12*, 5100–5110.

(55) Snoek, J.; Larochelle, H.; Adams, R. P. Practical Bayesian Optimization of Machine Learning Algorithms *Adv.Neur. Inf.Process.Sys*. **2012**, *25*, 2951.

(56) Vargas-Hernandez, R. A.; Guan, Y.; Zhang, D. H.; Krems, R. V. A machine-learning approach for the inverse scattering problem in quantum reaction dynamics. *https://arxiv.org/abs/1711.06376* **2018**.

(57) Rafat, M.; Popelier, P. L. A. Visualisation and integration of quantum topological atoms by spatial discretisation into finite elements. *J.Comput.Chem*. **2007**, *28*, 2602.

Table of Contents Image