

# Learning from the density to correct total energy and forces in first principle simulations

Cite as: J. Chem. Phys. **151**, 144102 (2019); <https://doi.org/10.1063/1.5114618>

Submitted: 10 June 2019 . Accepted: 19 September 2019 . Published Online: 08 October 2019

Sebastian Dick , and Marivi Fernandez-Serra 



View Online



Export Citation



CrossMark

## ARTICLES YOU MAY BE INTERESTED IN

### [Adventures in DFT by a wavefunction theorist](#)

The Journal of Chemical Physics **151**, 160901 (2019); <https://doi.org/10.1063/1.5116338>

### [CrystaLattE: Automated computation of lattice energies of organic crystals exploiting the many-body expansion to achieve dual-level parallelism](#)

The Journal of Chemical Physics **151**, 144103 (2019); <https://doi.org/10.1063/1.5120520>

### [The simulation of X-ray absorption spectra from ground and excited electronic states using core-valence separated DFT/MRCI](#)

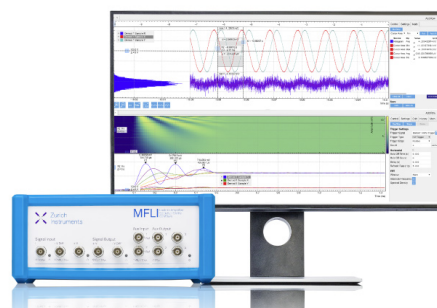
The Journal of Chemical Physics **151**, 144104 (2019); <https://doi.org/10.1063/1.5110418>

## Challenge us.

What are your needs for periodic signal detection?



Zurich  
Instruments



# Learning from the density to correct total energy and forces in first principle simulations

Cite as: J. Chem. Phys. 151, 144102 (2019); doi: 10.1063/1.5114618

Submitted: 10 June 2019 • Accepted: 19 September 2019 •

Published Online: 8 October 2019



View Online



Export Citation



CrossMark

Sebastian Dick<sup>a)</sup>  and Marivi Fernandez-Serra<sup>a)</sup> 

## AFFILIATIONS

Physics and Astronomy Department, Stony Brook University, Stony Brook, New York 11794-3800, USA  
and Institute for Advanced Computational Science, Stony Brook University, Stony Brook, New York 11794-3800, USA

<sup>a)</sup>sebastian.dick@stonybrook.edu and maria.fernandez-serra@stonybrook.edu

## ABSTRACT

We propose a new molecular simulation framework that combines the transferability, robustness, and chemical flexibility of an *ab initio* method with the accuracy and efficiency of a machine learning model. The key to achieve this mix is to use a standard density functional theory (DFT) simulation as a preprocessor for the atomic and molecular information, obtaining a good quality electronic density. General, symmetry preserving, atom-centered electronic descriptors are then built from this density to train a neural network to correct the baseline DFT energies and forces. These electronic descriptors encode much more information than local atomic environments, allowing a simple neural network to reach the accuracy required for the problem of study at a negligible additional cost. The balance between accuracy and efficiency is determined by the baseline simulation. This is shown in results where high level quantum chemical accuracy is obtained for simulations of liquid water at standard DFT cost or where high level DFT-accuracy is achieved in simulations with a low-level baseline DFT calculation at a significantly reduced cost.

Published under license by AIP Publishing. <https://doi.org/10.1063/1.5114618>

## I. INTRODUCTION

The development of atomistic simulation methods that make use of machine learning (ML) techniques to achieve close or beyond *ab initio* level accuracy at a highly reduced computational cost is proving to be an efficient and robust alternative to traditional force field approaches.<sup>1-7</sup> This high level of activity is facilitated and supported by the availability and continuous generation of larger and improved databases of highly accurate calculations or experiments both for molecular<sup>8-10</sup> and solid systems.<sup>11-13</sup> While in many cases, newly proposed methodologies and force-fields are only at the proof of concept stage, some of the most promising methods have already produced results that provide new physical insight into challenging problems.<sup>14</sup> Beyond predicting atomic structures of complex materials,<sup>15</sup> they have been used to address problems previously inaccessible to *ab initio* level calculations such as the structure of the surface of water or its temperature dependent dynamical properties using the MB-pol model.<sup>16-20</sup>

However, even if all these new force fields are facilitating the simulation of much larger systems and achieving longer dynamic

scales with near density functional theory (DFT) accuracy<sup>3,7,21,22</sup> or beyond,<sup>4,5,23</sup> this does not imply that DFT is becoming obsolete. There is still a need to obtain accurate energies, densities, and other physical observables that can be derived from first-principles methods such as DFT. Hence, developing approaches that aim to improve DFT by searching for more accurate exchange and correlation potentials using information from correlated wave function methods is a worthwhile endeavor. It is known that DFT-based methods within their lower rungs of approximation to the exchange and correlation (XC) energy struggle in simulating systems like water or ionic solutions.<sup>24-27</sup> The use of more expensive XC approximations, such as hybrid functionals,<sup>28</sup> may deliver the required accuracy; however, the size of systems for which these simulations are feasible is limited. This trade-off between accuracy, flexibility, and computational cost seems to lie at the heart of many problems the molecular science community is presently facing.

Some progress has been made in the development of machine-learned density functionals. Apart from promising efficient orbital-free DFT calculations by learning the kinetic energy functional,<sup>29-31</sup>

these methods have been shown to allow to skip self-consistent calculations altogether by directly learning the map from atomic potential to electron density.<sup>32,33</sup> While these applications still appear to be limited to one-dimensional systems or small molecules, other approaches have shown promising results in resolving this size extensivity problem.<sup>34–37</sup> The approach of Chandrasekaran *et al.*<sup>37</sup> seems to be very promising in that they also fit the local density of states in addition to the density. They can achieve highly accurate densities, and preliminary results suggest that these densities are good enough for accurate energy predictions.

In this work, we propose an alternative approach called Machine Learned Correcting Functionals (MLCFs) which, rather than replacing DFT, complements it by adding a machine learned functional of the electron density, expanded into a set of localized basis functions. Even though not explicitly shown here, this machine-learned density functional, based on local electronic descriptors, can in principle be used in self-consistent calculations. Exploring this path will be the topic of future work. The principal concept of this new methodology could be termed “Informed Machine Learning.” We argue that all the available physical insight into a problem should be used to help the ML-model make its predictions and minimize its generalization error. In MLCFs, this translates to DFT assuming the role of a data-preprocessor that maps the atomic coordinates to a self-consistent electron density which is input into an artificial neural network (ANN). The ANN is then fitted to the difference in energy and forces between the baseline method which was used to obtain the electron density and a higher-level reference method of choice. Thus, MLCFs draw upon and optimize the strengths of DFT, which is effective in calculating mean field electrostatic interactions, but often fails to accurately describe nonlocal interactions rooted in quantum mechanics. At the same time, DFT provides a solid foundation which the ML-model can fall back on when faced with unseen scenarios in which it will likely fail to make reliable predictions.

All results presented in this work were obtained with our MLCF implementation, available at Ref. 38.

## II. METHODS

### A. Charge density representation

As in any machine learning application, the representation of the data used for input (aka, the features) should encode as much information in as few variables as possible to avoid redundancies and make sure that the ANN model is quick to train and not prone to overfitting. Furthermore, we require the number of descriptors to scale linearly with the number of atoms in the system to guarantee size extensivity of the energy.

To ensure that the model respects rotational symmetry, we further impose the constraint that descriptors need to transform covariantly under  $SO(3)$  rotations.

One way to create such a descriptor is to project the electron density in real space onto a set of atom-centered orbitals, which were inspired by the work of Bartok *et al.* on representing chemical environments.<sup>39</sup> While completing this manuscript, the authors became aware of the work by Grisafi *et al.*<sup>33</sup> which uses a very similar approach.

The angular and radial basis functions are given by spherical harmonics  $Y_l^m(\theta, \phi)$  and

$$\tilde{\zeta}_n(r) = \begin{cases} \frac{1}{N}(r-r_i)^2(r_o-r)^{n+2}, & \text{for } r > r_i \text{ and } r < r_o, \\ 0, & \text{else} \end{cases}, \quad (1)$$

respectively. Here, we have used a radius  $r_o$  and a normalization factor  $N$ , the latter being determined by numerical integration. The radial functions defined above correspond to the ones used by Bartok *et al.*<sup>39</sup> except for the addition of an inner cutoff radius  $r_i$  which we included to disregard the core area. Moreover, the exponents were chosen so that the basis has vanishing first and second derivatives at both cutoff radii to ensure smoothness at these points. We found that this somewhat mitigates artifacts arising from the discrete nature of the euclidean grid used to represent the electron density.

After orthogonalizing the radial functions with the transformation

$$\zeta_n(r) = \sum_{n'} S_{nn'}^{-1/2} \tilde{\zeta}_n(r), \quad \text{where } S_{nn'} = \int dr \tilde{\zeta}_n(r) \tilde{\zeta}_{n'}(r), \quad (2)$$

the full basis is given as  $\psi_{nlm}(\vec{r}) = Y_l^m(\theta, \phi) \zeta_n(r)$ .

The descriptors  $c_{nlm}^{\alpha,I}$  for atom  $I$  of species  $\alpha$  at position  $\vec{r}_I$  can be obtained by projecting the electron density  $\rho$  onto the corresponding basis functions  $\psi_{nlm}^{\alpha}$ ,

$$\tilde{c}_{nlm}^{\alpha,I} = \int_{\vec{r}} \rho(\vec{r} - \vec{r}_{\alpha,I}) \psi_{nlm}^{\alpha*}(\vec{r}). \quad (3)$$

Introduction of an atomic species label  $\alpha$  has made it possible to have different bases for distinct atomic species.

To ensure rotational invariance, the basis should be aligned with a uniquely defined local coordinate system (LCS). Knowing the Euler angles  $\{\alpha, \beta, \gamma\}$  that relate the global coordinate system (GCS) to the LCS, the rotated descriptors are given as

$$c_{nlm} = \sum_{m'} (D^l(\alpha, \beta, \gamma))_{mm'}^{-1} \tilde{c}_{nlm'}, \quad (4)$$

with  $D^l$  being the Wigner  $D$ -matrix for angular momentum.<sup>40–42</sup> The LCS of a given atom can be defined through the position of atoms in its immediate environment.

Even though this approach has been proven to work for machine learned force fields,<sup>3</sup> it would be desirable to have a definition of the LCS that is completely independent of the atomic coordinates and only relies on electronic information.

Recognizing that the descriptors associated with the p-orbitals  $\tilde{c}_{n1m}$  transform like vectors under  $SO(3)$  rotations, we can use these vectors to define a local coordinate system that only depends on the electron density (see Appendixes A–E for more details).

Aligning the descriptors independently of structural variables has the major advantage that conformers that are close in their electron densities are described by similar features. This is particularly important for MLCFs that only correct the exchange correlation energy  $E_{xc}$  as these conformers will necessarily exhibit the same error in  $E_{xc}$ .

The ideal number of radial and angular basis functions as well as the inner and outer cutoff radii can be determined by cross-validation (CV). Doing so, the hyperparameters listed in Table I were found to provide a good compromise between accuracy and computational efficiency.

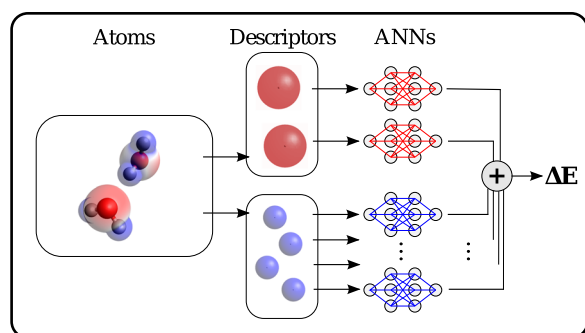
**TABLE I.** Hyperparameters used to create the charge density descriptors from top to bottom: Number of radial functions, maximum angular momentum, inner radial cutoff, and outer radial cutoff. Radii are given in units of Angstrom.

	Oxygen	Hydrogen
$n_{rad}$	2	2
$l_{max}$	2	1
$r_i$	0.05	0.0
$r_o$	1.0	1.5

## B. Machine learning models

Inspired by the well-established Behler-Parrinello networks (BPNs),<sup>21</sup> the energy functional for a given number of molecules was represented as an artificial neural network (ANN) consisting of a sum of smaller atomic networks, each of which only sees the local charge density around the atom it is associated with (see Fig. 1). Furthermore, networks belonging to atoms of the same species are identical to guarantee permutational invariance. Separate models were trained directly on the force corrections to be used in molecular dynamics (MD) simulations. Details about the force models are provided in the [Appendixes A–E](#).

Both models were trained using the Adam<sup>43</sup> optimizer with training rate  $\alpha = 0.001$  and decay rates  $\beta_1 = 0.9$  and  $\beta_2 = 0.999$ . The validation set was created by splitting off 20% from the original training set, and the number and size of layers as well as the training rate was determined using cross-validation. No mini-batches were used for training, i.e., the entire training set was passed through the neural network for every parameter update. The maximum number of epochs was set to 40 000, but due to early stopping, this value was never reached. Typical numbers of training epochs ranged between 10 000 and 30 000 and training times were of the order of minutes. A network with three layers of 8 nodes and one with three layers of 16 nodes were determined to have the optimal accuracy for energy and force predictions, respectively; however, satisfying results were obtained with a wide variety of model sizes.

**FIG. 1.** Architecture of a Behler-Parrinello type neural network. Starting with a given configuration, the self-consistent electron density is computed with the baseline DFT. The density is then encoded in descriptors that are fed into artificial neural networks (ANNs) which calculate atomic contributions to the total energy correction  $\Delta E$ .

The sigmoid function was chosen as activation,

$$f(x) = \frac{1}{1 + e^{-x}}, \quad (5)$$

as it was determined through cross-validation to be less prone to over-fitting than other possible choices such as the hyperbolic tangent or the rectified linear unit. One possible problem that practitioners should be aware of when using sigmoid or related functions as activations is what is commonly known as the vanishing gradient problem. The vanishing gradient problem stems from the fact that for very large and small input values, the derivative of the sigmoid function vanishes, which makes training exceedingly difficult for networks with many layers. This is usually mitigated by using rectified linear units (ReLU) as activation in deep neural networks.<sup>44</sup> We have not encountered any problems during the training process, which is most likely due to the limited depth of our neural network.

## III. RESULTS

### A. Gas phase water

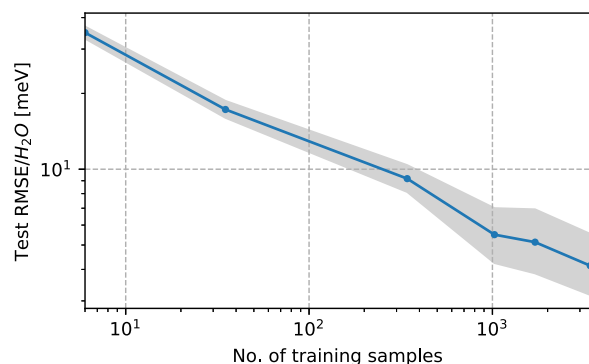
We tested our method on water clusters of varying size. The baseline calculations were conducted using SIESTA<sup>45</sup> with a quadruple zeta doubly polarized basis set and a van der Waals density functional of Dion *et al.*<sup>46</sup> with GGA exchange modified by Berland and Hyldgaard (vdW-cx).<sup>47</sup> This XC functional has been shown to be as good as other vdW-type functionals in describing liquid water.<sup>48</sup>

For the reference energies, we use the MB-pol force field<sup>16–18</sup> which is fitted to highly accurate coupled cluster calculations of water monomers and clusters. MB-pol has been shown to accurately reproduce the structural and thermodynamic properties of condensed phases of water; hence, it is a superior model to any vdW or GGA XC functional. It is important to emphasize that our method is of course not limited to the use of force-fields as a reference method. As only moderately sized data sets of relatively small systems are needed, high level quantum chemistry calculations to produce these data sets are expensive but feasible.

The MLCF was trained to interpolate between the baseline and reference method, i.e., the target values used for fitting are given as the differences in energy  $\Delta E = E_{ref} - E_{base}$ .

The data set comprised 2000 dimers and 1500 trimer configurations that were all sampled from the data that was used to fit MB-pol<sup>16–18</sup> and an additional 682 monomer configurations. Of these monomers, 300 had bond lengths and angles uniformly distributed on a grid between 0.90 and 1.07 Å and 90° and 118°. The remaining 382 monomers were sampled from the dimer configurations, only keeping structures that were not within the previously defined parameter grid.

The data were split into 80% training and 20% hold-out set. [Figure 2](#) shows how the generalization error (i.e., the error on the test set) changes with training set size. The reduced training sets were sampled from the original set while maintaining the relative contributions of monomers, dimers, and trimers. However, a minimum amount of two structures per subset was enforced. As expected,<sup>49</sup> the training curve follows a power law.



**FIG. 2.** Root mean square error (RMSE) on test set of fixed size as the number of training samples is varied. Reduced sets were sampled from the full training set described in the main text while keeping the relative contributions of monomers, dimers, and trimers constant. The shaded region indicates the 95% confidence interval which was obtained by resampling from the test set 10 000 times.

We further created samples of larger water clusters with  $n = 4, 5, 8, 16$  molecules (50 samples each) for testing the size extensivity of the network.

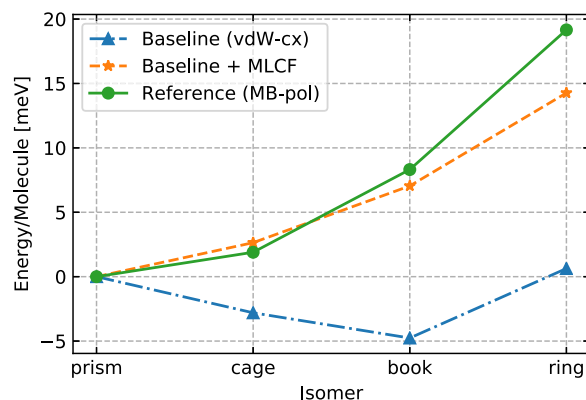
In [Table II](#), the errors in relative energies of isomers are split up into root mean square (rms) error, mean absolute error (MAE), and maximum (absolute) error.

As can be seen in [Table II](#), despite only being trained on systems with up to three molecules, our model performs well for all cluster sizes and significantly decreases errors in relative energies of isomers, all rms errors being below 5 meV/Molecule and mean absolute errors below 3 meV/Molecule. Moreover, [Fig. 3](#) shows that the overall energy ordering of hexamer structures is correctly reproduced by the MLCF, a task that has so far proven largely elusive to GGA-level DFT calculations.<sup>48</sup>

To gain further insights, we can follow the approach of previous works<sup>48,50</sup> and split up the total energy of water clusters into their  $n$ -body contributions. This was done on the same data sets that were used in [Ref. 48](#) to optimize the GGA exchange enhancement factor for water systems.

**TABLE II.** Model performance on test sets containing clusters of different sizes. Reported are the root mean square error (RMSE), mean absolute error (MAE) and, maximum absolute error (max. Error), given in meV/Molecule. The baseline method errors are given in parentheses.

No. Molecules	RMSE	MAE	Max. Error
1	2.29(53.19)	1.25(43.13)	14.71(151.19)
2	4.33(40.17)	2.91(31.28)	31.03(136.91)
3	2.89(28.25)	2.16(22.29)	12.27(75.20)
4	2.79(9.69)	2.15(7.93)	7.70(24.95)
5	2.64(11.24)	2.19(8.97)	6.23(36.69)
8	3.43(9.26)	2.72(7.34)	8.05(22.67)
16	2.75(6.28)	2.15(5.03)	6.19(17.15)



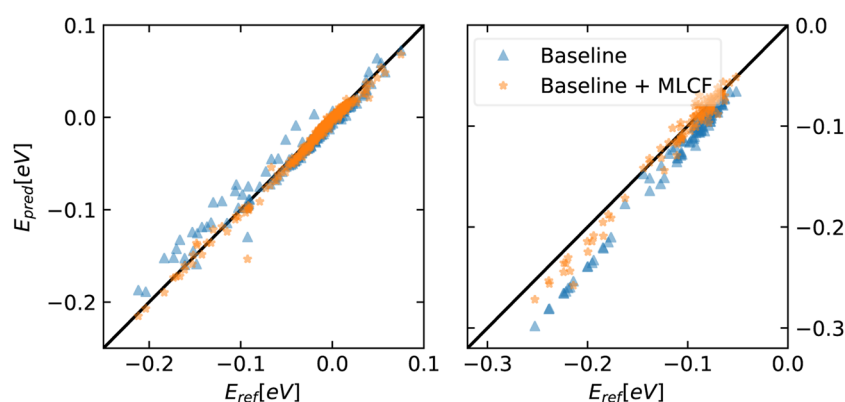
**FIG. 3.** Energies of water hexamers with respect to the prism isomer, which is correctly determined to be the most stable structure by the MLCF.

[Figure 4](#) shows that our model not only corrects the predominant errors in the one body energy ([Table II](#)) but also improves two and three body energies. This might seem surprising as the MLCF is only provided with the local charge density around each atom. However, the model seems to learn how to infer higher order interaction energies from correlations in the local charge density. This indicates that the long range and possibly nonlocal interactions are encoded within local density spatial correlations, enabling us to use radial cutoffs for the descriptors that are very short (see [Table I](#)), comparable to the nearest neighbor distance, which is about 1 Å in water molecules. In order to obtain similar accuracy using only structural information, much larger cutoffs are needed.<sup>39,50</sup>

## B. Comparison with machine learned force fields

Rather than competing with state of the art machine learned force fields, our work is aimed at investigating possible paths toward machine learned functionals that use the electron density as input. Nevertheless, it is instructive to compare the performance of our model to ML methods that make predictions using only geometric information. In particular, we want to investigate their capability to extrapolate, which restricts our possible choices to methods that learn a size extensive representation of the total energy. This disqualifies models that assume a fixed input size such as the sGDM approach by [Chmiela et al.](#)<sup>4,23,51,52</sup> Out of the plethora of publicly available ML force-fields, we chose to compare our method to SchNet<sup>5,53</sup> and a Behler-Parrinello network (BPN)<sup>21</sup> with weighted atom-centered symmetry functions (WACSFs)<sup>54</sup> both implemented in the open-source library SchNetPack.<sup>55</sup>

For a fair comparison, we require these force fields to merely learn the difference between the reference (MB-pol) and baseline (vdW-cx) energies and we will therefore refer to them as  $\Delta$ -SchNet and  $\Delta$ -WACSF from here on. In analogy to our analysis for MLCF, we want to test how a model trained on monomers, dimers, and trimers extrapolates to larger water clusters. The performance of non- $\Delta$  models, i.e., models that are trained on total energies, is shown in [Fig. 9](#) in the [Appendixes A–E](#). All non- $\Delta$  models trained show generalization errors above 10 meV/H<sub>2</sub>O and can



**FIG. 4.** Predicted vs reference two (left) and three (right) body energies for the data sets taken from Ref. 48. For the two body energies, the RMSE is 4.20 for the baseline method and 2.80 for the MLCF corrected method. For the three body energies, the errors are 7.10 and 3.56, respectively. Errors are given in meV/Molecule.

therefore not compete in accuracy with MLCF or their  $\Delta$ -learning counterparts.

It is not clear, *a priori*, which is the appropriate cutoff distance for the force fields. For non- $\Delta$  models, it is obvious that the cutoff should be chosen equal to the largest atomic distance in the training set, which is about 8 Å. Intuitively we would expect that for  $\Delta$ -models, a smaller cutoff will be sufficient and maybe even beneficial for the model's capability to extrapolate. Instead of fixing the radial cutoff, we opted to train a collection of models with cutoff ranging from 2 to 8 Å. For every choice of cutoff, the optimal set of hyperparameters was determined by a grid-search and cross-validation (see Appendixes A–E for details).

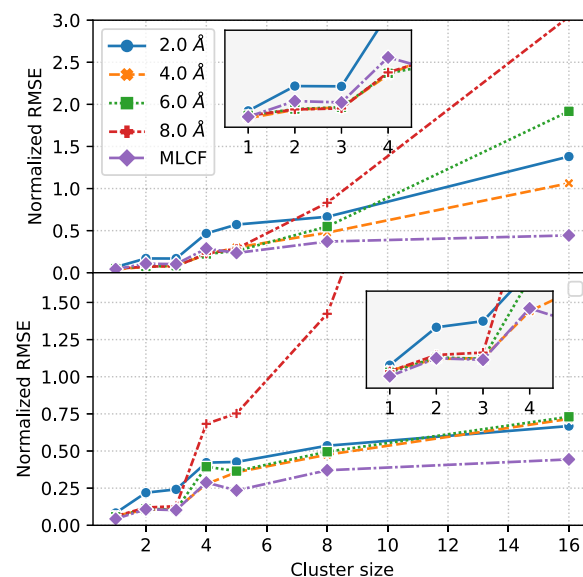
As can be seen in Fig. 5, for both  $\Delta$ -SchNet and  $\Delta$ -WACSF, a cutoff of 4 Å results in the best extrapolation to larger clusters.  $\Delta$ -SchNet outperforms  $\Delta$ -WACSF for small clusters but performs worse in terms of extrapolation.  $\Delta$ -SchNet is slightly more accurate than MLCF for cluster sizes below 5 molecules, but both force-fields are less capable of extrapolating to large clusters than MLCF. It is instructive to see that using the electron density as input, smaller cutoff radii (1.0 Å–1.5 Å) are sufficient to reach competitive accuracy. This suggests that the network learns to infer how nonlocal information is encoded in the local charge density.

To go one step further, we would like to compare the models' performance in systems with other molecules present. For this, we used the publicly available data set S66x8, which contains dissociation curves for 66 noncovalent complexes relevant to biomolecular structures.<sup>56</sup> We limited our calculations to a subset with complexes that contain at least one water molecule. For  $\Delta$ -SchNet and  $\Delta$ -WACSF, we used the models with cutoff 4 Å from above, the MLCF remained unchanged to the one used for water clusters.

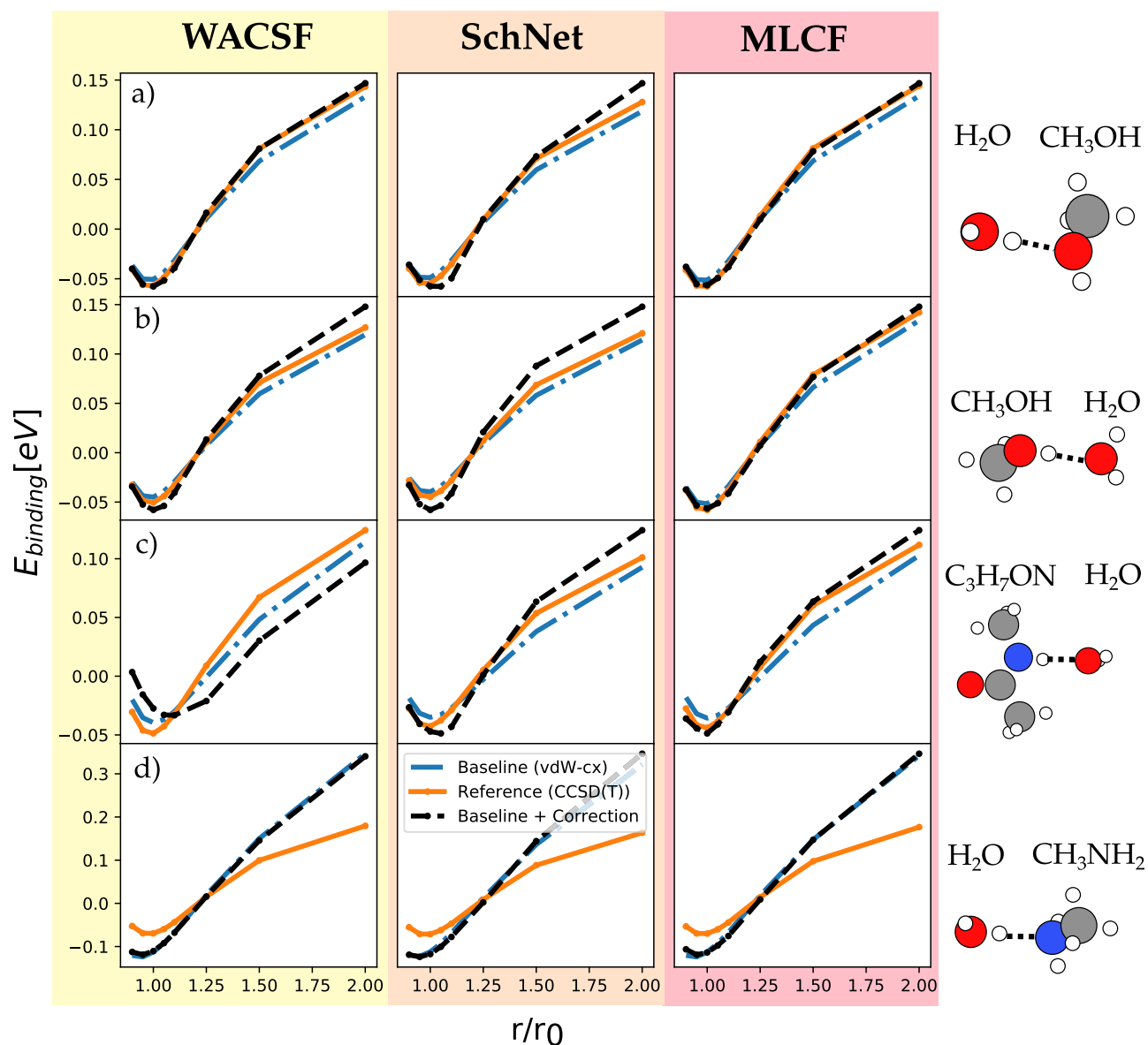
By design, the MLCF is only ever used for atomic species that it was trained on, i.e., oxygen and hydrogen. For the force-fields, best results were achieved if corrections were restricted to these elements as well. Therefore, when using these force-fields, we masked out all other atoms before creating the input features. A comparison of all three methods for a selected subset can be seen in Fig. 6. None of the methods correctly predicts the dissociation curve for water and methylamine [Fig. 6(d)]. This is due to the fact that the H–N hydrogen bond was not contained in the training set. In fact, the

nitrogen atom is “invisible” to the machine learning method in all cases. For the systems containing methanol,  $\Delta$ -WACSF and MLCF both make reasonable predictions, the latter being more accurate than the former in the case where methanol acts as the donor molecule. In all cases,  $\Delta$ -SchNet shows deviations for large distances and around the equilibrium distance.

The most instructive example is given by dissociation curve for *n*-methylacetamine and water. The crucial difference to the systems above is that now the hydrogen atom participating in the bond is covalently bonded to a nitrogen atom. In terms of the electron density, the changes to OH are small, which explains why MLCF still performs well. However, making use of only structural information,



**FIG. 5.** Root mean squared errors (RMSEs) of  $\Delta$ -SchNet (top) and  $\Delta$ -WACSF (bottom) models trained for different cutoff radii (see legend) on water monomers, dimers, and trimers and tested on random clusters of increasing size. The errors are normalized by dividing them through the RMSE of the underlying baseline method (vdW-cx) and compared to those of MLCF (purple line) which can also be found in Table II.



**FIG. 6.** Comparison of, from left to right,  $\Delta$ -WACSF,  $\Delta$ -SchNet, and MLCF on a subset of dissociation curves taken from the s66x8<sup>56</sup> data set. From top to bottom, the figures describe hydrogen bonds between (a) water and methanol, (b) methanol and water, (c) n-methylacetamine (C<sub>3</sub>H<sub>7</sub>ON) and water, and (d) water and methylamine (CH<sub>3</sub>NH<sub>2</sub>), where the molecule listed first is understood to be the hydrogen donor. The energies depicted correspond to binding energies for the reference calculations (orange) and total energies for the baseline (blue) and corrected (black) calculations. As the molecules stay rigid during the relative displacement, binding and total energies merely differ by a constant. We therefore shifted every curve by its mean energy to align them and draw better comparisons between the methods. The depicted distances were normalized by dividing them through their respective equilibrium values, which were determined with MP2.

both force-fields perform worse than the baseline method in this case.

We do want to stress that this comparison is not intended to disqualify any of the above tested force-fields. These methods are extremely accurate and efficient at the purposes they are developed for. We merely use the above example to highlight tasks where the MLCF method outperforms these approaches.

### C. Liquid water

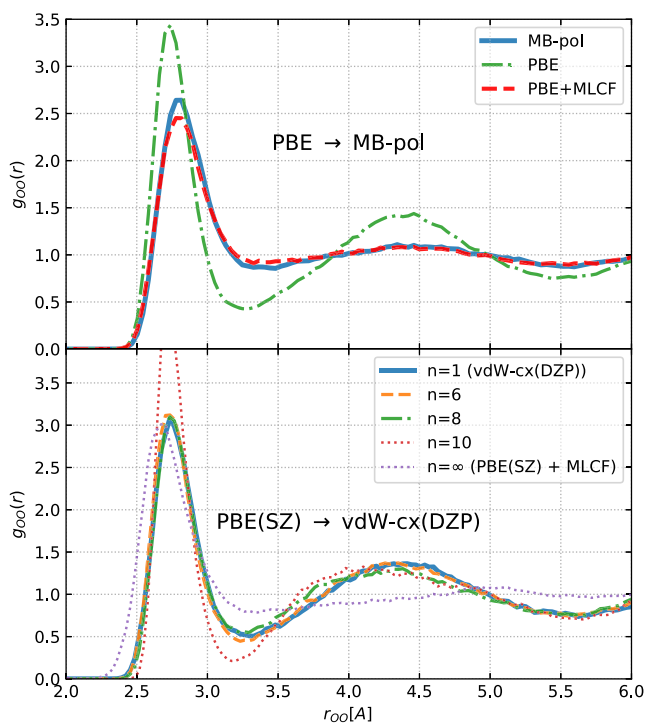
Having proven successful for gas-phase calculations, we went on to assess the method's performance in molecular dynamics simulations on liquid water. For reasons of practicality, we opted to replace the previously employed van der Waals functional with the faster Perdew-Burke-Ernzerhof (PBE)<sup>57</sup> functional. We trained a

MCLF to interpolate between this baseline method and MB-pol, which has been shown to produce pair correlation functions close to experimental results. Further details about the training data set used are given in the [Appendixes A–E](#).

We conducted a DFT-based molecular dynamics (MD) simulation of 128 water molecules at experimental density in a periodic box using a Nose-Hoover thermostat<sup>58</sup> at a temperature of 300 K. The simulation length was 10 ps with a time step of 0.5 fs using 4 ps for equilibration and 6 ps to create the correlation functions. As can be seen in the top section of [Fig. 7](#), our MCLF provides a significant improvement over PBE, which is known to over-structure liquid water. The MCLF approach also corrects the dynamical properties of the baseline calculation, bringing them to the level of the reference method. This can be seen in the vibrational density of states (VDOS) plots shown in the [Appendixes A–E](#).

#### D. Basis set correction

Beyond correcting the density functional approximation, MLCFs can also be used to extrapolate between two given basis sets. Thus, extremely fast calculations can be performed with minimal [single zeta (SZ)] basis-sets that with the help of MLCFs



**FIG. 7.** Oxygen-oxygen radial distribution functions  $g_{OO}(r)$  for two different applications of MLCFs. *Top:* Correcting for errors in the XC functional by interpolating between PBE and MB-pol. *Bottom:* Accelerating DFT calculations by correcting for both functional and basis-set related convergence errors, combining MLCF with the time step mixing method proposed by Anglada *et al.*<sup>59</sup> Interpolation between a quick and dirty PBE(SZ) simulation and a slow and more accurate vdW-cx(DZP) simulations.  $n$  indicates the number of fast steps needed before correcting with a well converged step.

are then brought to the accuracy of more conservative basis sizes.

Here, it should be noted that MLCFs rely upon the quality of the electron density used as input. Following an approach by Kim *et al.*,<sup>60</sup> the error that the MLCF is trained to correct can be decomposed as

$$\Delta E = E_{ref}[n_{ref}] - E_{base}[n_{base}] = \Delta E_F + \Delta E_D, \quad (6)$$

with  $E_F = E_{ref}[n_{base}] - E_{base}[n_{base}]$  and  $E_D = E_{ref}[n_{ref}] - E_{ref}[n_{base}]$  being the functional and density driven error, respectively. By design, MLCFs are only effective at correcting functional driven errors as deviations in the density induce long-range effects through electrostatic interactions that are hard to correct with our semilocal approach. This generally does not pose any serious problems as most density functional approximations yield a good quality electron density. Using very sparse basis sets, however, imposes an upper limit on the accuracy of the density, making density driven errors become dominant.

It turns out that despite the inaccuracies produced by the small basis set, we can still obtain satisfactory results in MD simulations by using a method developed by Anglada *et al.*<sup>59</sup> In their paper, they propose to alternate fast steps calculated with a cheaper, less accurate method with slow steps obtained with a well converged method. The forces used to integrate the equations of motion with a time step  $\Delta t$  are given as  $\vec{F}(t) = \vec{F}_{fast}(t) + \Delta\vec{F}(t)$ , where the correction term  $\Delta\vec{F}(t)$  is defined as

$$\Delta\vec{F} = \begin{cases} n\delta\vec{F}, & \text{if } (t/\Delta t \bmod n = 0), \\ 0, & \text{else,} \end{cases} \quad (7)$$

with  $\delta\vec{F} = \vec{F}_{slow} - \vec{F}_{fast}$ . The motivation behind this idea is that the error  $\delta\vec{F}$  is small compared to the absolute forces and relatively smooth over time. Therefore, it is sufficient to only correct for this error every  $n$  steps, the limiting value of  $n$  being dependent on both the system and the accuracy of the fast method.

To test this, we ran AIMD simulations of 64 water molecules, at 300 K and experimental density for 10 ps.

The fast steps were calculated using the PBE exchange-correlation functional with a single zeta basis set and subsequently corrected with an adequately trained MLCF. For the slow, well-converged steps, we used the vdW-cx functional and a double zeta polarized (DZP) basis set. The MLCF was trained to lift the accuracy of the fast method to that of the slow one. For that, we used the same procedure and data set as for the gas-phase MLCF discussed above. However, we re-did all the DFT calculations replacing vdW-cx by PBE with single zeta basis set and MB-pol by vdW-cx with a polarized double zeta basis set, adjusting baseline and reference methods to fit the problem at hand.

As can be seen in the bottom part of [Fig. 7](#), the radial distribution function remains unchanged for  $n \leq 6$  and reasonable results can be achieved up to  $n \leq 8$ .

It is important to point out that these results could only be obtained by correcting the fast steps with an MLCF. In its original form, even though proven successful for systems like liquid silica,<sup>59</sup> the mixing method failed for water. Furthermore, the upper limit for the number of subsequent fast steps  $n_{lim}$  is strongly



system-dependent and rigorous testing should be conducted before using this mixing scheme. However, if carefully used, this method promises significant speed-ups that are roughly equal to  $n_{lim}$  for large system sizes (see [Appendixes A–E](#) for detailed analysis).

#### IV. CONCLUSION

In conclusion, we have presented and tested a new molecular simulation framework in which the electronic density is used to train a neural network to correct baseline DFT energies and forces to the accuracy provided by a higher level method. Our results indicate that real space semilocal correlations of the electronic density as encoded in our localized density descriptors contain information about longer range effects. These effects can be corrected through the use of a neural network trained on purely local data, an observation that might support the potential for developing machine learned density functionals using local density descriptors<sup>33</sup> and total energy targets.

The scope of this method is not the same as that of machine learned force fields, given that it aims to stay within the realm of DFT. However, we have been able to show that indeed the use of local descriptors built from the electronic density outperforms those methods in terms of transferability. This suggests that electronic descriptors such as the ones proposed here could be explored to improve the accuracy of force fields.

After completion of this manuscript, the authors became aware of work by Bogojeski *et al.*<sup>61</sup> that is very closely related to the work presented here. Similar to our approach, Bogojeski *et al.* use density-based inputs and  $\Delta$ -learning to achieve quantum chemical accuracy. The main difference between the two approaches lies in the choice of basis functions and the way symmetries are encoded. In the method of Bogojeski *et al.*, the molecule is first aligned with a global coordinate system which is defined through some molecular axes. The electron density is subsequently expanded in a Fourier basis. These design choices seem to restrict their method to systems of fixed size, and results presented are limited to small and medium sized molecules. In contrast, as we have shown above, by using semilocal, atom-centered basis functions, our method is size-extensive and can therefore extrapolate to systems it was not explicitly trained on, in particular, condensed phases.

#### ACKNOWLEDGMENTS

We acknowledge funding from DOE Awards Nos. DE-SC0001137 and DE-SC0019394. Sebastian Dick was supported by a fellowship from The Molecular Sciences Software Institute under NSF Grant No. ACI-1547580. We would like to thank Stony Brook Research Computing and Cyberinfrastructure and the Institute for Advanced Computational Science at Stony Brook University for access to the high-performance SeaWulf computing system, which was made possible by a \$1.4M National Science Foundation Grant (No. 1531492). Finally, we want to express our thanks to Jose Soler for valuable advice concerning this project.

#### APPENDIX A: COORDINATE SYSTEMS

In this work, several definitions of a local coordinate system (LCS) are used.

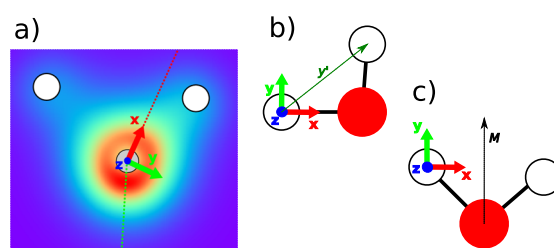


FIG. 8. Definition of the (a) electronic, (b) nearest-neighbor, and (c) molecular local coordinate system.

#### 1. Electronic

Given a set of descriptors  $\{\tilde{c}_{nlm}\}$  for a given atom, the descriptors associated with the p-orbitals  $\tilde{c}_{nlm}$  transform like vectors under  $SO(3)$  rotations. Starting with  $n = 1$  and  $l = 1$ , the corresponding descriptor vector  $\tilde{c}_{11m}$ , after being brought into its real form and normalized, is used as the first local axis. Increasing  $n$  allows one to find a second vector of  $l = 1$  descriptors that, using Gram-Schmidt orthogonalization, defines the second local axis.

In the example given in [Fig. 8\(a\)](#), the first vector  $\tilde{c}_{11m}$  (dashed red line) picks up on the asymmetry of the covalent bonds and is bent toward the closer hydrogen atom, whereas the second vector  $\tilde{c}_{21m}$  (dashed green line) points toward the lone pair. Problems may arise if all p-descriptors are collinear as is the case for highly symmetric systems such as a water monomer with covalent bonds of equal length. In cases like this, one has to revert to using the nearest neighbor rule as described below to determine the second axis. However, extended systems are rarely this symmetrical and an increase in the number of radial functions is usually enough to resolve issues of linear dependence and find a set of p-descriptors that span a coordinate system.

#### 2. Nearest neighbor

To determine the LCS around a given atom, one can also use structural information about the atom's immediate environment. The first axis is defined as the direction to the atom's nearest neighbor. The second axis is simply given by the direction to the next nearest neighbor, orthogonalized with respect to axis number one. If all three atoms happen to lie on a line, more distant atoms are taken into consideration until noncollinearity is achieved.

#### 3. Molecular

For nonreactive models, the molecular axes can be used as a local coordinate system. The axis assignment is hereby arbitrary, but following convention, we define the y-axis as the bisector (M) of the HOH-angle, the x-axis as the one parallel to the molecular plane, and the z-axis orthogonal to it [see [Fig. 8\(c\)](#)].

#### APPENDIX B: DETAILS ON COMPARISON WITH FORCE FIELDS

We determined the optimal values for a selected subset of hyperparameters for both SchNet and WACSF by cross-validation (CV), where the validation set was identical to the one used for

**TABLE III.** Optimal hyperparameters determined through cross-validation for SchNet. The hyperparameters that were optimized were the dimension of the embedding space denoted as “features” and the number of interaction layers. An “x” in the column named  $\Delta$  denotes that the model was trained to correct vdW-cx.

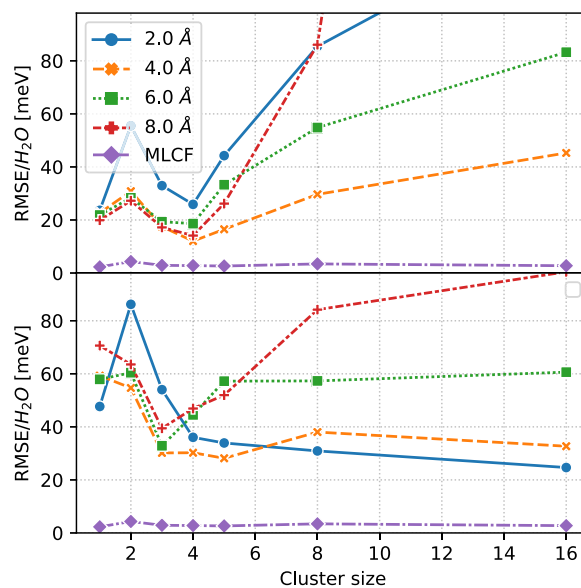
Cutoff	$\Delta$	Features	Interaction layers
2.0	x	128	4
4.0	x	256	4
6.0	x	256	3
8.0	x	256	2
2.0		256	3
4.0		256	3
6.0		256	4
8.0		256	4

MLCF. For SchNet, we fixed the number of Gaussians used to expand the interatomic distance to 25, the dimension of the embedding space (features), as well as the number of interaction layers were determined by CV. For WACSF, the number of fully connected layers in the neural network that acts as an estimator was fixed to three, whereas the number of nodes in each layer as well as the number of radial and angular basis functions was determined by CV. For both SchNet and WACSF, a batch size of 100 together with a maximum number of training epochs of 300 was employed. Adam<sup>45</sup> together with a learning rate decay of 0.5, and early stopping was used for optimization. The optimal set of hyperparameters is shown in Tables III and IV for SchNet and WACSF, respectively.

Figure 9 shows the performance of non- $\Delta$  models compared to that of MLCF. Even for cluster sizes contained in the training set, the force-fields cannot reach competitive accuracy. This is due to the limited size of the training set. It should be noted, however, that even though an increase in training set size will most likely lead to a better performance for cluster sizes 1–3, a significant improvement in accuracy for larger clusters is not to be expected as the model will not have enough information to extrapolate to these systems. Better results can be achieved if the force fields

**TABLE IV.** Optimal hyperparameters determined through cross-validation for WACSF. The hyperparameters that were optimized were the number of angular and radial basis functions and the number of nodes in the three layer fully connected neural network. An “x” in the column named  $\Delta$  denotes that the model was trained to correct vdW-cx.

Cutoff	$\Delta$	Angular	Nodes	Radial
2.0	x	8	128	22
4.0	x	8	128	22
6.0	x	4	128	66
8.0	x	8	128	66
2.0		8	128	66
4.0		8	128	44
6.0		4	128	66
8.0		6	128	66



**FIG. 9.** Root mean squared errors (RMSE) in meV per molecule of SchNet (top) and WACSF (bottom) models trained for different cutoff radii (see legend) on water monomers, dimers, and trimers and tested on random clusters of increasing size. The errors are compared to those of MLCF (purple line) which can also be found in Table II.

are trained to correct the baseline method as shown in the main text.

## APPENDIX C: FORCE MODEL

Given that the mapping from atomic configuration to electron density is obtained by conducting a full self-consistent DFT calculation, the forces cannot be derived from the energy functional as the derivatives  $\partial\Delta E/\partial r_{i,\alpha} = \Delta F_{i,\alpha}$  are unknown.

We therefore have to train a model to directly predict the force corrections  $\Delta\vec{F}$ , where again each atomic species is treated with a separate model. Taking derivatives of a machine learned energy model is known to amplify its errors. By using a model directly trained on forces, we can avoid this problem. However, as there is no coupling between energy and force models, the resulting method does not exactly preserve energy and momentum. These effects are due to and thus comparable in magnitude to the fitting error of the force MLCF which in our case is of the order of 0.01 eV/Å. As an ad-hoc solution for our molecular dynamics simulations, we manually set the mean force acting on the system to zero at every time step, making sure that the relative acceleration between oxygen and hydrogen atoms remained unaltered. Ongoing research is being conducted into how energy and momentum conservation can be included in the fitting procedure.

## APPENDIX D: MOLECULAR DYNAMICS

To test our method's performance in molecular dynamics simulations, we trained an MLCF to interpolate between PBE<sup>37</sup>

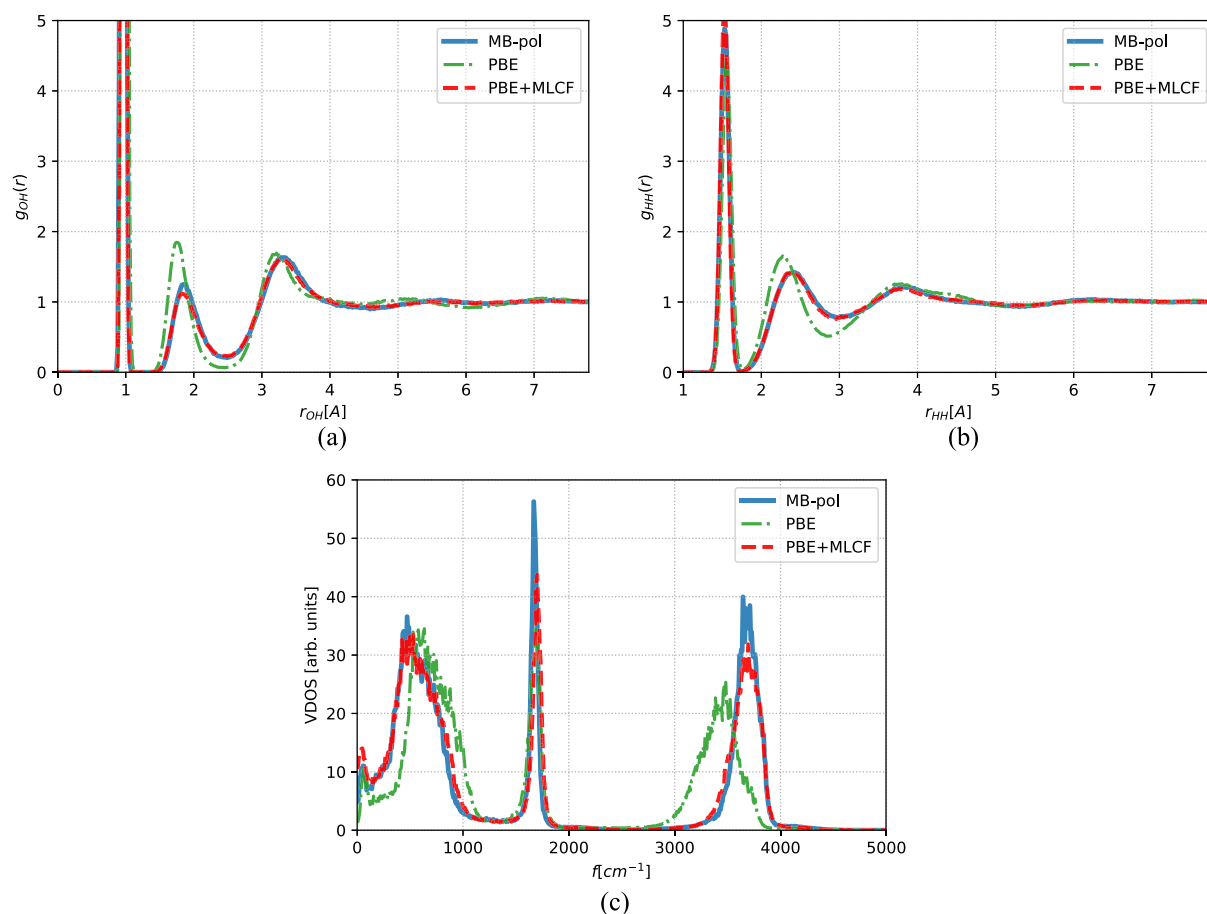
(baseline) and MB-pol<sup>16-18</sup> (reference). The previously employed quadruple zeta doubly polarized was replaced by a faster double zeta polarized (DZP) basis set, and the real space mesh cutoff energy was reduced from 300 Ry to 200 Ry. This real space cutoff can be interpreted as the upper limit for the kinetic energy of plane waves that can be represented on this grid without aliasing.

Moving to condensed phase systems, the electron density and thereby the descriptors undergo significant changes, making an MLCF trained on only small clusters inaccurate. The most obvious remedy is to simply train the MLCF on condensed phase data. This, however, has the drawback that reference calculations become increasingly costly. We avoided this issue by training the model on electron densities of 16(H<sub>2</sub>O) clusters that were electrostatically embedded in liquid water modeled by TIP4P/2005,<sup>62</sup> opening the possibility of using more expensive, wave function based methods to perform reference calculations in the future. In detail, this means that a data set was created by sampling a random molecule and its 15 nearest neighbors from snapshots of an MD simulation and treating all remaining water molecules in the snapshot with TIP4P/2005. Doing so, the number of hydrogen bonds that the central molecule

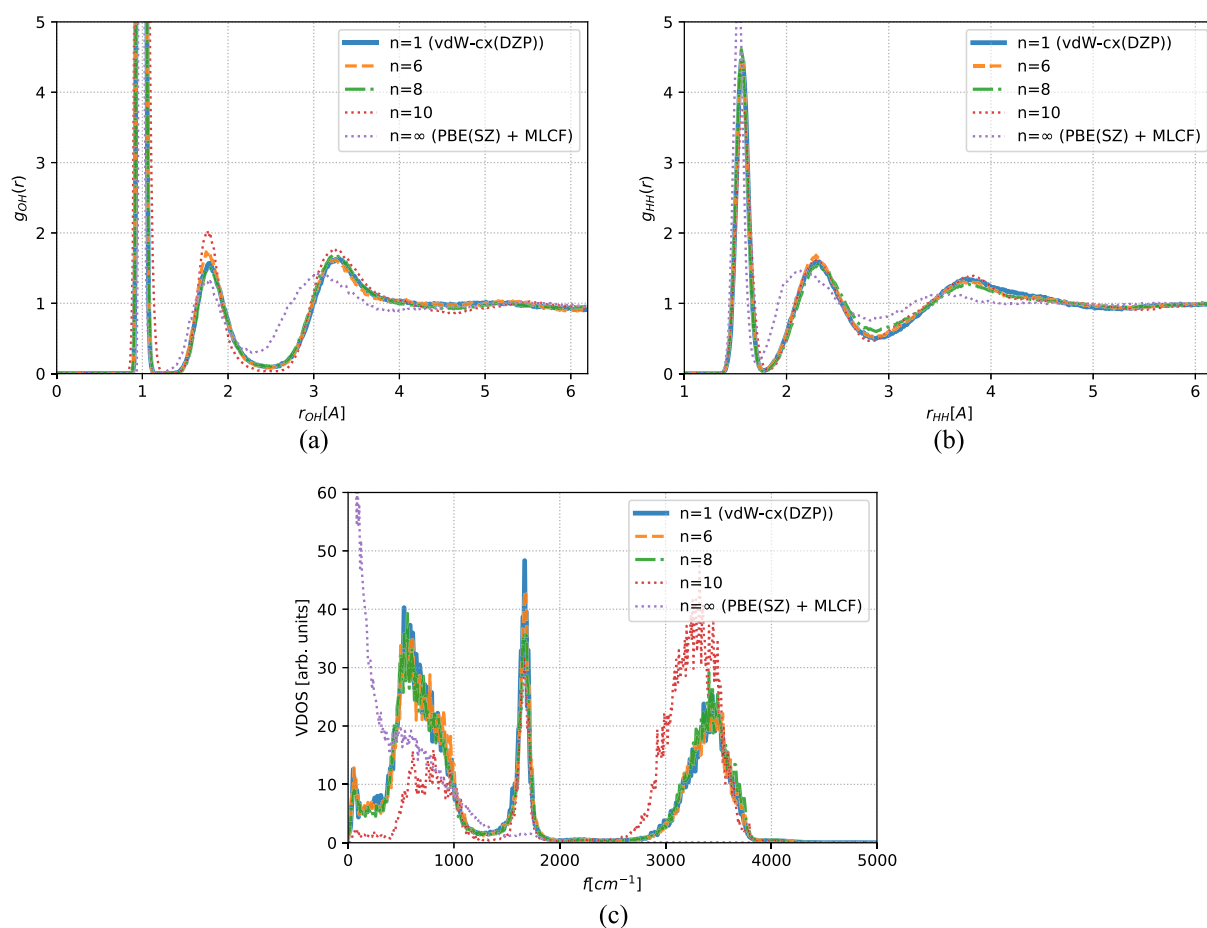
forms with its neighbors was used as a stratifying parameter during the sampling. For the descriptor alignment, we chose to employ the aforementioned nearest-neighbor rule. To obtain the target energies and forces, the embedding was removed and calculations were performed on clusters only, thereby restricting the correction to short-range effects. Subsequently, only the central molecule was used in fitting the force corrections.

For the MD simulations that used a single zeta (SZ) basis as baseline method, we found that a model trained on smaller clusters (monomers, dimers, and trimers) exhibited sufficient accuracy. Including larger clusters or using embedding did not lead to any significant improvement. We accredit this to the “stiffness” of the SZ basis set which is not capable of describing bonding and polarization effects, making electron densities of small clusters and condensed phases practically indistinguishable. We found, however, that using molecular local coordinate systems (as opposed to nearest neighbor or electronic ones) lead to a slight improvement in model performance.

In addition to the oxygen-oxygen pair correlation function which can be found in the main text, Figs. 10 and 11 show the OH



**FIG. 10.** (a) and (b) Radial distribution functions and (c) vibrational density of states for an MLCF that is trained to correct errors in the XC functional by interpolating between PBE and MB-pol.



**FIG. 11.** (a) and (b) Radial distribution functions and (c) vibrational density of states. MLCF is used to accelerate DFT calculations by correcting for both functional and basis-set related convergence errors, combining MLCF with the time step mixing method proposed by Anglada *et al.*<sup>59</sup> Interpolation between a quick and dirty PBE(SZ) simulation and a slow and more accurate vdW-cx(DZP) simulations.  $n$  indicates the number of fast steps needed before correcting with a well converged step.

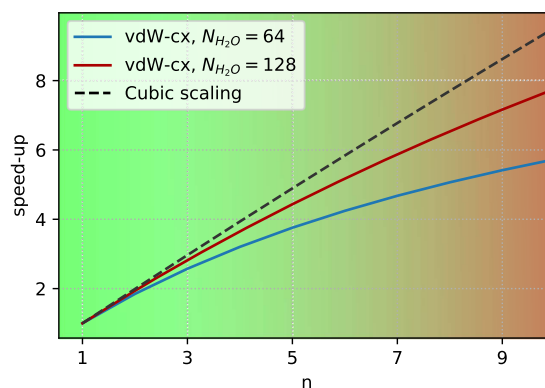
and HH correlation functions and the vibrational density of states (VDOS) which was obtained by taking the Fourier-transform of the velocity autocorrelation function.

## APPENDIX E: SPEED-UP

The speed-up that can be achieved with the mixing method by Anglada *et al.*<sup>59</sup> depends crucially on both the mixing parameter  $n$  and the ratio between the time it takes to propagate the system with the fast method  $t_f$  and the time it takes with the slower, accurate method  $t_s$ . Alternating  $(n-1)$  fast steps with one slow step, the speed-up  $\eta$  is given by

$$\eta = \frac{nt_s}{(n-1)t_f + t_s}. \quad (\text{E1})$$

Figure 12 shows  $\eta$  obtained if PBE with SZ basis and vdW-cx<sup>47</sup> with DZP are mixed for which  $t_f \approx 35$  s and  $t_s \approx 423$  s ( $N = 64$ ) and  $t_f \approx 89$  s and  $t_s \approx 2779$  s ( $N = 128$ ). The values are obtained for calculations on



**FIG. 12.** Speed-up obtained by mixing PBE(SZ) with vdW-cx(DZP) for finite size systems compared to the speed-up that would be achieved for an algorithm that scales strictly cubic in the number of orbitals. The color gradient indicates the method's reliability to reproduce the reference method for a given value of  $n$ .

a single Intel Xeon E5-2683v3 CPU and compared to the theoretical speed-up for an algorithm that is strictly cubic scaling in the number of orbitals. We see that for large enough system sizes  $\eta \approx n$ .

## REFERENCES

- <sup>1</sup>J. Behler, *J. Chem. Phys.* **145**, 170901 (2016).
- <sup>2</sup>T. T. Nguyen, E. Székely, G. Imbalzano, J. Behler, G. Csányi, M. Ceriotti, A. W. Götz, and F. Paesani, *J. Chem. Phys.* **148**, 241725 (2018).
- <sup>3</sup>L. Zhang, J. Han, H. Wang, R. Car, and E. Weinan, *Phys. Rev. Lett.* **120**, 143001 (2018).
- <sup>4</sup>S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, and K.-R. Müller, *Sci. Adv.* **3**, e1603015 (2017); e-print [arXiv:1611.04678](https://arxiv.org/abs/1611.04678).
- <sup>5</sup>K. Schütt, P.-J. Kindermans, H. E. S. Felix, S. Chmiela, A. Tkatchenko, and K.-R. Müller, *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2017), pp. 991–1001.
- <sup>6</sup>K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. Von Lilienfeld, K.-R. Müller, and A. Tkatchenko, *J. Phys. Chem. Lett.* **6**, 2326 (2015).
- <sup>7</sup>J. S. Smith, O. Isayev, and A. E. Roitberg, *Chem. Sci.* **8**, 3192 (2017).
- <sup>8</sup>R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, and A. Aspuru-Guzik, *ACS Cent. Sci.* **4**, 268 (2018).
- <sup>9</sup>M. M. Ghahremanpour, P. J. Van Maaren, and D. Van Der Spoel, *Sci. Data* **5**, 180062 (2018).
- <sup>10</sup>R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. Von Lilienfeld, *Sci. Data* **1**, 140022 (2014).
- <sup>11</sup>S. Kirklín, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, and C. Wolverton, *npj Comput. Mater.* **1**, 15010 (2015).
- <sup>12</sup>K. Winther, M. J. Hoffmann, O. Mamun, J. R. Boes, J. K. Nørskov, M. Bajdich, T. Bligaard, M. Bajdich *et al.*, *Sci. Data* **6**(1), 75 (2019).
- <sup>13</sup>A. Zakutayev, J. Perkins, M. Schwarting, R. White, K. Munch, W. Tumas, N. Wunder, and C. Phillips, “High throughput experimental materials database,” Technical Report 75, National Renewable Energy Laboratory-Data (NREL-DATA), Golden, CO, USA, 2017.
- <sup>14</sup>M. Hellström, V. Quaranta, and J. Behler, *Chem. Sci.* **10**, 1232 (2019).
- <sup>15</sup>V. L. Deringer, N. Bernstein, A. Bartók, M. J. Cliffe, R. N. Kerber, L. E. Marbella, C. P. Grey, S. R. Elliott, and G. Csányi, *J. Phys. Chem. Lett.* **9**, 2879 (2018).
- <sup>16</sup>V. Babin, C. Leforestier, and F. Paesani, *J. Chem. Theory Comput.* **9**, 5395 (2013).
- <sup>17</sup>V. Babin, G. R. Medders, and F. Paesani, *J. Chem. Theory Comput.* **10**, 1599 (2014); e-print [arXiv:1210.7022v1](https://arxiv.org/abs/1210.7022v1).
- <sup>18</sup>G. R. Medders, V. Babin, and F. Paesani, *J. Chem. Theory Comput.* **9**, 1103 (2014).
- <sup>19</sup>S. Sun, F. Tang, S. Imoto, D. R. Moberg, T. Ohto, F. Paesani, M. Bonn, E. H. G. Backus, and Y. Nagata, *Phys. Rev. Lett.* **121**, 246101 (2018).
- <sup>20</sup>S. K. Reddy, D. R. Moberg, S. C. Straight, and F. Paesani, *J. Chem. Phys.* **147**, 244504 (2017).
- <sup>21</sup>J. Behler and M. Parrinello, *Phys. Rev. Lett.* **98**, 146401 (2007).
- <sup>22</sup>L. Zhang, J. Han, H. Wang, W. Saidi, R. Car, and E. Weinan, *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2018), pp. 4436–4446.
- <sup>23</sup>S. Chmiela, H. E. Sauceda, K.-R. Müller, and A. Tkatchenko, *Nat. Commun.* **9**, 3887 (2018).
- <sup>24</sup>R. A. DiStasio, Jr., B. Santra, Z. Li, X. Wu, and R. Car, *J. Chem. Phys.* **141**, 084502 (2014).
- <sup>25</sup>M. J. Gillan, D. Alfè, and A. Michaelides, *J. Chem. Phys.* **144**, 130901 (2016).
- <sup>26</sup>Y. Yao and Y. Kanai, *J. Chem. Theory Comput.* **14**, 884 (2018).
- <sup>27</sup>M. Chen, H.-Y. Ko, R. C. Remsing, M. F. C. Andrade, B. Santra, Z. Sun, A. Selloni, R. Car, M. L. Klein, J. P. Perdew *et al.*, *Proc. Natl. Acad. Sci. U. S. A.* **114**, 10846 (2017).
- <sup>28</sup>A. D. Becke, *J. Chem. Phys.* **98**, 5648 (1993).
- <sup>29</sup>J. Seino, R. Kageyama, M. Fujinami, Y. Ikabata, and H. Nakai, *J. Chem. Phys.* **148**, 241705 (2018).
- <sup>30</sup>J. C. Snyder, M. Rupp, K. Hansen, K.-R. Müller, and K. Burke, *Phys. Rev. Lett.* **108**, 253002 (2012).
- <sup>31</sup>K. Yao and J. Parkhill, *J. Chem. Theory Comput.* **12**, 1139 (2016).
- <sup>32</sup>F. Brockherde, L. Vogt, L. Li, M. E. Tuckerman, K. Burke, and K.-R. Müller, *Nat. Commun.* **8**, 872 (2017).
- <sup>33</sup>A. Grisafi, A. Fabrizio, B. Meyer, D. M. Wilkins, C. Corminboeuf, and M. Ceriotti, *ACS Cent. Sci.* **5**(1), 57–64 (2018).
- <sup>34</sup>X. Lei and A. J. Medford, preprint [arXiv:1901.10822](https://arxiv.org/abs/1901.10822) (2019).
- <sup>35</sup>R. Nagai, R. Akashi, S. Sasaki, and S. Tsuneyuki, *J. Chem. Phys.* **148**, 241737 (2018).
- <sup>36</sup>R. Nagai, R. Akashi, and O. Sugino, preprint [arXiv:1903.00238](https://arxiv.org/abs/1903.00238) (2019).
- <sup>37</sup>A. Chandrasekaran, D. Kamal, R. Batra, C. Kim, L. Chen, and R. Ramprasad, *npj Comput. Mater.* **5**, 22 (2019).
- <sup>38</sup>D. Sebastian, *semodi/mlcf: First production release (Version v1.0)*, Zenodo (16 September 2019). [http://doi.org/10.5281/zenodo.3416639](https://doi.org/10.5281/zenodo.3416639)
- <sup>39</sup>A. P. Bartók, R. Kondor, and G. Csányi, *Phys. Rev. B* **87**, 184115 (2013).
- <sup>40</sup>E. Wigner, *Ann. Math.* **40**, 149 (1939).
- <sup>41</sup>N. Thomas, T. Smidt, S. Kearnes, L. Yang, L. Li, K. Kohlhoff, and P. Riley, preprint [arXiv:1802.08219](https://arxiv.org/abs/1802.08219) (2018).
- <sup>42</sup>T. Cohen and M. Welling, in *International Conference on Machine Learning* (Microtome Publishing, 2016), pp. 2990–2999.
- <sup>43</sup>D. P. Kingma and J. Ba, preprint [arXiv:1412.6980](https://arxiv.org/abs/1412.6980) (2014).
- <sup>44</sup>G. B. Goh, N. O. Hodas, and A. Vishnu, *J. Comput. Chem.* **38**, 1291 (2017).
- <sup>45</sup>J. M. Soler, E. Artacho, J. D. Gale, A. Garcia, J. Junquera, P. Ordejón, and D. Sánchez-Portal, *J. Phys.: Condens. Matter* **14**, 2745 (2002).
- <sup>46</sup>M. Dion, H. Rydberg, E. Schröder, D. C. Langreth, and B. I. Lundqvist, *Phys. Rev. Lett.* **92**, 246401 (2004).
- <sup>47</sup>K. Berland and P. Hylgaard, *Phys. Rev. B* **89**, 035412 (2014).
- <sup>48</sup>M. Fritz, M. Fernández-Serra, and J. M. Soler, *J. Chem. Phys.* **144**, 224101 (2016).
- <sup>49</sup>C. Cortes, L. D. Jackel, S. A. Solla, V. Vapnik, and J. S. Denker, *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 1994), pp. 327–334.
- <sup>50</sup>A. P. Bartók, M. J. Gillan, F. R. Manby, and G. Csányi, *Phys. Rev. B* **88**, 054104 (2013).
- <sup>51</sup>S. Chmiela, H. E. Sauceda, I. Poltavsky, K.-R. Müller, and A. Tkatchenko, *Comput. Phys. Commun.* **240**, 38–45 (2019).
- <sup>52</sup>H. E. Sauceda, S. Chmiela, I. Poltavsky, K. R. Müller, and A. Tkatchenko, *J. Chem. Phys.* **150**, 114102 (2019).
- <sup>53</sup>K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, *J. Chem. Phys.* **148**, 241722 (2018).
- <sup>54</sup>M. Gastegger, L. Schwiedrzik, M. Bittermann, F. Berzsenyi, and P. Marquetand, *J. Chem. Phys.* **148**, 241709 (2018).
- <sup>55</sup>K. Schütt, P. Kessel, M. Gastegger, K. Nicoli, A. Tkatchenko, and K.-R. Müller, *J. Chem. Theory Comput.* **15**, 448 (2018).
- <sup>56</sup>J. Rezac, K. E. Riley, and P. Hobza, *J. Chem. Theory Comput.* **7**, 2427 (2011).
- <sup>57</sup>J. P. Perdew, K. Burke, and M. Ernzerhof, *Phys. Rev. Lett.* **77**, 3865 (1996).
- <sup>58</sup>G. J. Martyna, M. L. Klein, and M. Tuckerman, *J. Chem. Phys.* **97**, 2635 (1992).
- <sup>59</sup>E. Anglada, J. Junquera, and J. M. Soler, *Phys. Rev. E* **68**, 055701 (2003).
- <sup>60</sup>M.-C. Kim, E. Sim, and K. Burke, *Phys. Rev. Lett.* **111**, 073003 (2013).
- <sup>61</sup>M. Bogojeski, L. Vogt-Maranto, M. E. Tuckerman, K.-R. Müller, and K. Burke, “Density functionals with quantum chemical accuracy: From machine learning to molecular dynamics” (to be published).
- <sup>62</sup>J. L. Abascal and C. Vega, *J. Chem. Phys.* **123**, 234505 (2005).