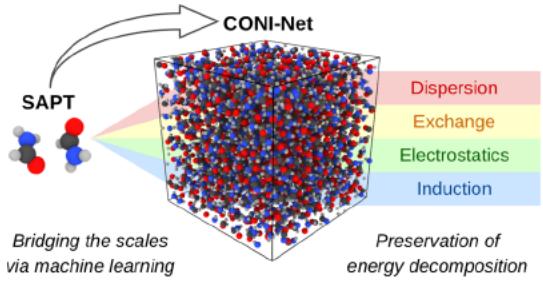
CONI-Net: Machine Learning of Separable Intermolecular Force Fields

Manuel Konrad and Wolfgang Wenzel*

ABSTRACT: Noncovalent interactions (NCIs) play an essential role in soft matter and biomolecular simulations. The ab initio method symmetry adapted perturbation theory allows a precise quantitative analysis of NCIs and offers an inherent energy decomposition, enabling a deeper understanding of the nature of intermolecular interactions. However, this method is limited to small systems, for instance, dimers of molecules. Here, we present a scale bridging approach to systematically derive an intermolecular force field from ab initio data while preserving the energy decomposition of the underlying method. We apply the model in molecular dynamics simulations of several solvents and compare two predicted thermodynamic observables—mass density and enthalpy of vaporization—to experiments and established force fields. For a data set limited to hydrocarbons, we investigate the extrapolation capabilities to molecules absent from the training set. Overall, despite the affordable moderate quality of the reference ab initio data, we find promising results. With the straightforward data set generation procedure and the lack of target data in the fitting process, we have developed a method that enables the rapid development of predictive force fields with an extra dimension of insights into the balance of NCIs.



1. INTRODUCTION

Soft matter and biomolecular simulations cover a wide range of material classes such as solvents, polymers, liquid crystals, and proteins. The common characteristic of these systems is that many of their kinetic and thermodynamic properties are governed by noncovalent interactions (NCIs). The correct modeling of NCIs is, therefore, essential for quantitative and predictive results. On an ab initio level, there are several methods that describe NCIs at different levels of theory. In the supermolecular approach, NCIs are deducted from the difference of monomer energies and the total energy of the complex. Due to the inadequate treatment of electron correlation, many computationally efficient electronic structure methods, such as semiempirical methods and density functional theory (DFT), use dispersion corrections to yield meaningful results for this approach.[1−6] Alternatively, the use of nonlocal density functionals can also enable the modeling of NCIs within the DFT framework.[7−9] Post Hartree−Fock methods that explicitly include electron correlation can provide an accurate description of NCIs from first principles, popular examples are the Møller−Plesset perturbation theory and the coupled cluster method.[10−12] A variant of the latter, CCSD(T)[13−16] extrapolated to the complete basis set limit,[17] is often appointed as the "gold standard" for NCI benchmark data sets.[18,19]

An alternative to the supermolecular approach for the computation of NCIs is the symmetry adapted perturbation theory (SAPT), where the interaction is computed from a perturbative expansion up to a limited order.[20] The resulting separate contributions can be grouped into electrostatics, dispersion, exchange, and induction energy components. Various truncations and flavors of the method offer a broad range of compromises between accuracy and computational cost. While SAPT0 gives qualitative results for systems up to several hundreds of atoms, higher level SAPT methods can approach the gold standard quality.[21−24]

However, the modeling of many soft matter and bioapplications requires simulations of many thousands to millions of atoms.[25−31] In the spirit of multiscale modeling, the traditional solution is the development of analytical force fields to enable simulations at larger scales. For the description of molecular systems, force fields are divided into an intra and intermolecular part. The intramolecular potential is mainly determined by the covalent interactions between the atoms of the molecules. A combination of harmonic contributions and periodic dihedral potentials is a common choice to model these intramolecular interactions. The harmonic parameters can be chosen to reproduce structural data and vibrational frequencies, either from the experiment or quantum mechanics. The parameters of the soft dihedral potentials

can be fitted to torsional angle scans at the ab initio level. The intermolecular potential constitutes a more significant challenge due to the complex interplay of attractive and repulsive regimes and its long range and many body character. Therefore, the parameter fitting process often includes experimental target properties. This allows the choice of simple and computationally cheap functional forms, such as the pairwise additive Lennard−Jones potential combined with atomic partial charge interactions. The top down approach ensures that errors for the cumulative target properties are minimized.[32−37]

An alternative is the bottom up generation of intermolecular force fields from ab initio reference calculations, either for custom applications or as general transferable force fields.[38−49] Using reference methods that offer energy decomposition, such as SAPT, the individual components can even be derived independently or partially grouped to decouple the fit parameters.[43−49] The components are fitted against predefined pairwise functions, often exponential terms or power laws, and as for empirical force fields, can be combined with a baseline model fitted to the electrostatic potential of the monomer such as partial charges in combination with a long range solver.[46−48] In addition to the pairwise additive interactions, many body contributions can explicitly be added to the model.[42,47,48,50] For the transferable approaches, atom types are assigned based on the functional group and chemical intuition. Mixing rules for unlike atom type pairings can further reduce the number of model parameters. Besides strict top down and bottom up approaches, the fitting process of NCIs can also combine ab initio and experimental data.[51,52]

A more recent development is the use of artificial neural networks (ANNs) in surrogate models for the potential energy surface, the so called neural network potentials (NNPs). Here, an important concept is the Behler−Parrinello approach to divide the total model into additive submodels.[53] This concept was successfully applied for the atomization energy from DFT, where the total energy is divided into atomistic contributions, which are evaluated by subnets that share parameters for atoms of the same chemical element.[54] In general, the input for the subnets is a descriptor of the local atomic environment such as symmetry functions, which map the occurrence of distances and angles with neighboring atoms into several fuzzy shells around the considered atom,[53,54] but other representations are also possible.[55−57] One advantage of NNPs is their flexibility to fit complex potential energy surfaces due to the excellent approximation capabilities of their building blocks, the fully connected ANNs. In contrast to conventional analytical methods, there is no need for the manual definition of types, facilitating the development of transferable models.[54]

The division of the model into additive partial contributions for dimensionality reduction that comes with the Behler−Parrinello approach is a general concept and can also be exploited for other quantities than the atomization energy.[58,59] For NCIs, the partitioning scheme, which has proven useful in many analytical models, splits the total interaction energy into atomic pairwise additive contributions. The general applic ability of a pairwise decomposition for NNPs was already shown in a study on atomization energies of methanol, copper clusters, and bulk copper.[60] In the context of NCIs, recent studies applied NNPs based on symmetry function descriptors for a data set of hydrogen bonded complexes, one with atomic and one with pairwise partitioning.[61,62] It could be shown that on the same data set, the pairwise model outperforms the model with atomic contributions.[62] Alternatively, machine learning models can predict partial charges or multipole coefficients that describe electrostatic intermolecular inter actions,[63−66] which combined with an NNP for covalent interactions and a dispersion correction results in a model that can be applied to the dynamic modeling of reactions in molecular systems similar to ReaxFF potentials.[63,67,68] Symmetry functions are powerful descriptors that can result in a good performance for energy predictions. However, the application in molecular simulations requires the implementa tion of an expensive on the fly calculation. More importantly, accurate local energy predictions do not guarantee a smooth distance dependence; consequently, rippled energy curves can cause unphysical forces.[61,62] These are some of the challenges that need to be addressed when developing an NCI model for molecular mechanics applications, and definitive conclusions about the numerical stability and predictive performance for large systems can only be delivered by applying the model in the target simulation method such as molecular dynamics (MD).

Here, we present the component separable noncovalent interaction network (CONI Net), which uses an alternative to the symmetry function descriptor designed for efficient large scale simulations. Like other bottom up force fields, the model is trained on separable dimer interaction energies obtained from SAPT theory, and for each energy component, we train a separate model. The network architecture is based on a Behler−Parrinello network with atomic pairwise energy partitioning.[60,62] The subnets for each pair contribution consist of fully connected neural networks that interpolate between pair fingerprints and a function layer containing several power law terms that model the distance dependence. This approach has two main advantages. Since the distance dependence is treated separately, we can use a pair fingerprint which only depends on equilibrium monomer properties of the involved molecules. Therefore, the neural network results can be precomputed for efficient additive pair interaction evaluations in MD. The second advantage is that with appropriate constraints, the power law terms can ensure monotony and counteract artifacts from overfitting, while the use of multiple terms in the function layer still provides flexible regression capabilities.

In this study, we present two independent applications of the model. First, we train a custom model on a data set of several organic solvents. We acquire thermodynamic properties from MD simulations for all molecules in the data set and compare them to the literature values from experiments and two established force fields. In the second example, we explore the extrapolation capabilities of a model trained on a diverse range of hydrocarbons. Here, the MD prediction is conducted for molecules that do not occur in the training data set.

2. METHODOLOGY

In this work, we developed an NNP for the intermolecular interactions, which can be used as part of a force field for MD simulations. For the intramolecular force field, which is not the scope of this work, we use the established GROMOS force field.[36] In the following sections, we will describe our workflow for the preparation and training of our model, which consists of determining the fingerprint descriptor and partial charges from monomer calculations, the generation of a data set from dimer calculations, and the model training. Finally, the performance of the model to predict thermodynamic properties is tested in

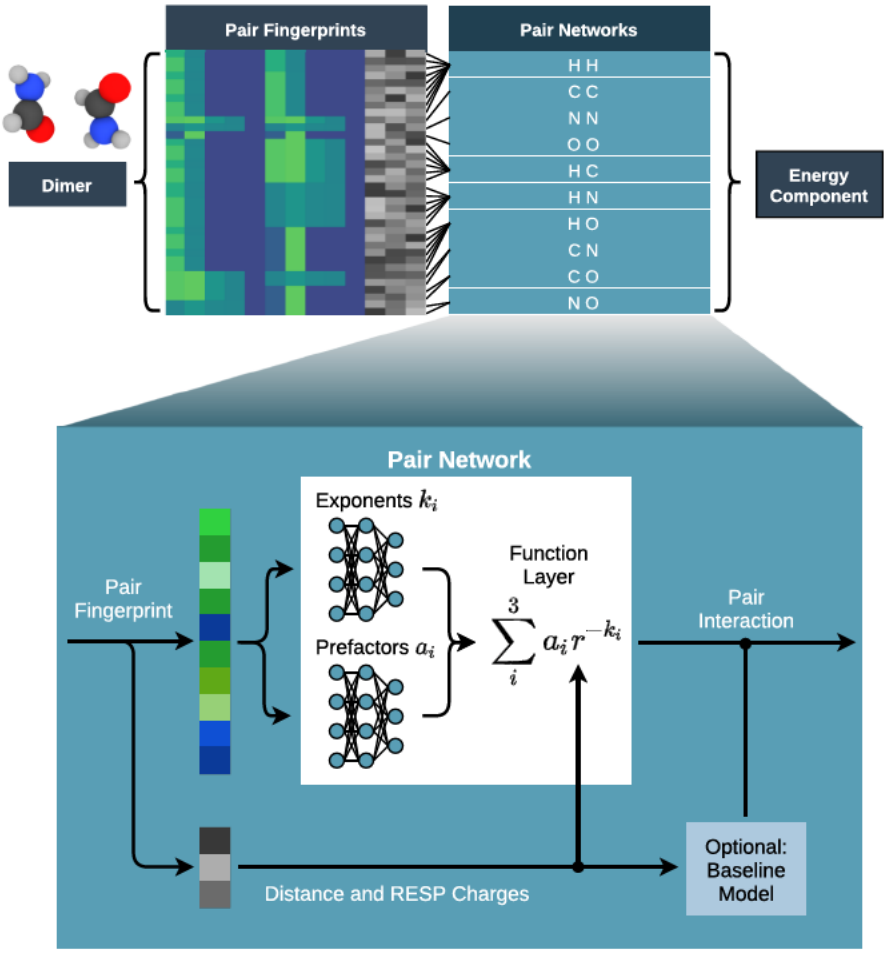


Figure 1. Overview of the CONI Net model for one energy component. For each intermolecular pair of a dimer, the pair fingerprint, the distances, and the RESP charges are fed into a pair network of the corresponding element combination. Parameters are shared for pair network instances of the same type. The pair network consists of two fully connected networks that use the pair fingerprint to evaluate prefactors and exponents, which are used in a function layer together with the distance to compute the pair interaction. For the electrostatic component, an additional baseline energy is computed based on the interaction of the partial charges.

MD simulations. The molecular visualizations in this work were rendered using the OVITO package.[69]

2.1. Fingerprint Descriptor and Partial Charges. For each atom within each molecule, a fingerprint emerging from the local environment inside the molecule is computed. First, the monomer geometry is optimized with the DFT code ORCA[70] using the B3LYP functional[71−73] and the aug cc pVTZ basis set.[74] With the external tool Multiwfn,[75] we compute the partial charges for the baseline electrostatics model in a one stage RESP fitting procedure[76] with additional constraints to obey the symmetry of the MD force field. From an additional single point calculation using the cc pVTZ basis set,[77] Mayer bond orders[78] and Hirshfeld charges[79] are extracted by the postprocessing modules of ORCA. The Hirshfeld charge and the four highest Mayer bond order parameters describe an atomic fingerprint. In order to describe an intermolecular pair, the two atomic fingerprints are combined to a 10 digit pair fingerprint, and the restrained electrostatic potential (RESP) charges and the distance are stored separately. The order of the two atomic fingerprints is fixed for pairs consisting of different chemical elements. For identical elements, the sizes of the individual fingerprint values are compared one by one, and the first differing entries determine the order.

2.2. Data Set Generation. As a reference method for the intermolecular interaction energy, we use SAPT based on monomer wavefunctions from Hartree−Fock theory.[20] The perturbative approach of this method yields separate contributions to the energy and allows grouping into the physically motivated component dispersion, exchange, electro statics, and induction[80]

$$E_{\text{dimer}} = E_{\text{disp}} + E_{\text{exch}} + E_{\text{el}} + E_{\text{ind}} \quad (1)$$

A detailed description of the method is given in ref 80. An overview and benchmarks of the different available truncations can be found in ref 21. For our data set calculations, we chose the SAPT2+3 level with the aug cc pVDZ[74] basis set and density fitting approximation as implemented in Psi4,[12,81−83] which delivers a favorable computational cost to accuracy ratio and moderate disk and memory requirements for the considered molecules. The data set contains various dimer structures for each considered combination. These are prepared the following way:

1. Random rotation of both equilibrium monomer geo metries
2. Determination of dimer distance d_{rep} corresponding to E_{estimate} = 1 kcal/mol (repulsive regime)

3. Choice of dimer distance from interval (d_{rep}, d_{rep} + 5.0 Å), weighed according to $\exp(-\alpha E_{estimate})$

Here, $E_{estimate}$ is an estimate for the dimer energy, which in this work is the sum of the Lennard−Jones energy from an automatically assigned general Amber force field (GAFF)[35] using Ambertools[84] and the electrostatic interaction of the partial charges obtained from the RESP fit, and α is a sampling parameter. With the definition of a sampling temperature T_s by the relation $\alpha = 1/(k_B T_s)$, at room temperature, this corresponds to $\alpha = \frac{1.69}{\text{kcal / mol}}$, which we use for all calculations in this study.

2.3. Model Description and Training Parameters. The total energy is obtained as the combined result of four separate models trained on the dispersion, exchange, electrostatics, and induction components of the SAPT decomposition. The schematic overview of the CONI Net model for one energy component is shown in Figure 1. It is based on the Behler−Parrinello scheme of describing a surrogate model as a sum of contributing submodels.[53] In our model, the submodels are represented by pair networks that describe the interaction of an intermolecular pair of atoms. Parameter sharing is employed for pair networks of the same energy component and element combination. Each pair network consists of the following modules:

- Two fully connected ANNs for the computation of exponents and prefactors that use the pair fingerprint as input
- A function layer that combines the exponents, prefactors, and distances in a sum of power law terms to compute the energy contribution of the pair
- For the electrostatic component, a baseline model based on RESP charges.

The weights and biases of the fully connected ANNs represent the learnable parameters of the model. They contain two hidden layers with four nodes each that use a ReLU activation function. The output layers of both ANNs consist of three nodes with different transformations for exponents and prefactors. The exponents k_i are limited to a fixed range around predefined bias values k_i^{bias}

$$k_i = k_i^{bias} + 2.0 \cdot (\text{sig}(k_i^{out}) - 0.5) \tag{2}$$

Here, the sigmoid function is defined as $\text{sig}(x) = (1 + e^{-x})^{-1}$ and k_i^{out} is the unprocessed value of output node i of the ANN for the exponents. For k_i^{bias}, we choose 8, 10, and 12 for the exchange interaction and 6, 8, and 10 for all other components. For the prefactors, the signs are constrained to positive values for the exchange component and to negative values for the dispersion, induction, and electrostatics components. Furthermore, the prefactors are multiplied by a constant factor of 1000 kcal/mol, which has a similar effect as using different optimizer settings and initializations for the prefactor and exponent networks. Finally, a taper function is applied to all negative pair contributions for distances r below r_{min}, the smallest distance seen during training of the corresponding pair network. This prevents the pair interaction to diverge to negative infinite values, which could cause unstable MD simulations due to rare close encounters. We use a taper factor f as proposed in ref 85

$$r_0 = r_{min} - 0.5 \tag{3}$$

$$x(r) = \frac{r - r_0}{r_{min} - r_0} \tag{4}$$

$$f(x) = (1 - x)^3 (1 + 3x + 6x^2) \tag{5}$$

The taper factor f is applied to the pair interaction $E_p(r)$ in the interval from r_0 to r_{min}, and the final pair interaction $\tilde{E}_p(r)$ is defined as

$$\tilde{E}_p(r) = \begin{cases} 0, & r \le r_0 \\ (1 - f)E_p(r), & r_0 < r < r_{min} \\ E_p(r), & r \ge r_{min} \end{cases} \tag{6}$$

For the implementation of the network model, we use the Python library PyTorch.[86] The weights and biases of the fully connected hidden and output layers are initialized from the uniform distribution

$$\mathcal{U}(-\sqrt{n_f^{-1}}, \sqrt{n_f^{-1}}) \tag{7}$$

with n_f being the number of input features of the layer. For the model training, we apply the Adam optimizer[87] with the recommended settings ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$), a learning rate of 0.001 and a mean squared error loss function. The overall data set is randomly split into a training set, which is used for model training, and a validation set, which is applied for early stopping of the training if the validation loss does not improve for 800 epochs while checking every 100 epochs. The ratio between the number of training and validation data points is 3:1 throughout this work. Furthermore, the pair fingerprints are normalized, and normally distributed noise with a standard deviation of 0.1 is added during training to increase the robustness against overfitting. After early stopping, the noise is switched off, and the model is slightly relaxed by training from the best checkpoint for an additional 100 epochs.

2.4. Thermodynamic Properties from MD. In order to compare the model to other force fields and experimental data, we conduct MD simulations of the liquid bulk phase. The intramolecular GROMOS force field parameters are obtained using the web accessible Automated Topology Builder.[36,88,89] The intermolecular force field is derived from the trained CONI Net models as follows. For each unique pair fingerprint, we sum the four energy components at 500 discrete distances in the range of 0.5−15.0 Å. The resulting energy curves are then derived numerically using the NumPy library[90] to compute the force. Both quantities are stored in a force field table file and can directly be used with the MD package LAMMPS.[91] Furthermore, the partial charges are set to the RESP charges acquired during molecule preparation. For the short range interactions, we use a cutoff of 15.0 Å, and long range electrostatics are computed by Ewald summation as implemented in LAMMPS. The computational efficiency of the tabulated approach is lower compared to Lennard−Jones force fields but of comparable order of magnitude.

The bulk liquid simulation is set up by placing 1000 randomly rotated molecules on a lattice with a sufficiently large lattice constant to avoid initial overlaps. This starting configuration is equilibrated with a series of MD runs using Nosé−Hoover style thermostatting and barostatting. Additionally, a vacuum simulation of a single molecule is conducted to acquire the baseline of the intramolecular potential energy for the following analysis. Here, we apply a Langevin thermostat with a time step of 0.2 fs and a coupling parameter τ_T of 1 ps in an initialization run of 100 ns followed by a production run of 100 ns. An overview of the run parameters for each step of the bulk simulation is given in Table 1, including the simulated

Table 1. MD Parameters of the Preparation Steps 1−3 and the Production Step 4

step	1	2	3	4
time (ps)	200	200	400	2000
time step (fs)	1	1	0.2	0.2
p (atm)	100 → 1	1	1	1
τ_p (ps)	1	1	5	5
τ_T (ps)	0.1	0.1	1	1
kspace rel. err.	1×10^{-4}	1×10^{-4}	1×10^{-5}	1×10^{-5}

time, the time step, the pressure p, the pressure coupling τ_p, the temperature coupling τ_T, and the relative target error in forces of the kspace solver. Since in part we want to compare our values to GAFF[35] and OPLS AA[33] results of ref 92, we use similar simulation parameters in the production runs.

From the production data of the bulk liquid simulation, we extract the average volume and potential energy. Together with the average potential energy from the vacuum simulation, we compute the density and the enthalpy of vaporization of the liquid at the given temperature. We use equivalent expressions to those used in ref 92

$$\rho = \frac{M}{\langle V \rangle} \tag{8}$$

$$\Delta H_{\mathrm{vap}} = E_{\mathrm{pot}}^{\mathrm{vac}} - E_{\mathrm{pot}}^{\mathrm{liq}} + k_{\mathrm{B}}T \tag{9}$$

Here, M is the total mass, k_{B} is Boltzmann's constant, T is the temperature, $\langle V \rangle$ is the average volume during the liquid simulation, and $E_{\mathrm{pot}}^{\mathrm{vac}}$ and $E_{\mathrm{pot}}^{\mathrm{liq}}$ are the average potential energies per molecule determined from the vacuum and liquid simulation, respectively.

3. RESULTS AND DISCUSSION

First, we discuss the learning curve for a single molecule to assess the data efficiency of the model. Then, we present the results of two independent applications of our model. While both models are trained on different data sets, they both use the same set of hyperparameters as described in the methodology section.

3.1. Learning Curve for a Single Molecule. In order to investigate the influence of data set size on the performance of the model to interpolate between different dimer arrangements, we generate a learning curve for the organic molecule methanol. For several total data set sizes, five different random training/validation splits (3:1 ratio) are drawn. The independent test set contains 2000 samples. For each split, five models per energy component are trained. The best model per component and split is selected in two ways. First, by the mean absolute error (MAE) for the validation set of the split, and for comparison, according to the optimum MAE for the test set. The resulting best models for the energy components of a specific split can then be combined to a best total energy model for this split. Hence, we acquire two ensembles of five total energy models for each data set size, corresponding to the two selection criteria and the five unique training/validation splits. In Figure 2, the mean and range of the MAEs for the test set are plotted for different training set sizes and both selections.

As expected, with a growing number of training samples, the average model error for the independent test set decreases. On the one hand, this is caused by the increasing variety of dimer arrangements in the training set. On the other hand, a larger

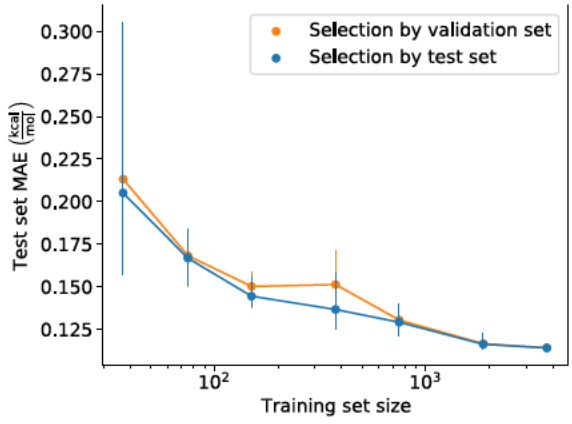


Figure 2. Learning curve for methanol. The points and bars show the arithmetic mean and range of test set errors of the total energy models for the different training/validation splits of the data set. For each split and energy component, the best of five models is used for the total energy model, either selected by the validation set of the split (orange) or by the independent test set (blue).

and more representative validation set size, which scales proportionally to the training set size and controls the early stopping criterion, is less prone to cause under or overfitting and is less sensitive to the random sample of the training/validation split. Therefore, with increasing data set sizes, the range of MAE values decreases, and the test set performances of the models selected by the validation sets converge to the optimally performing models.

3.2. Custom Model for Specific Small Organic Molecules. The first example is the generation of a potential that interpolates between a set of small organic molecules containing the elements carbon, hydrogen, nitrogen, and oxygen. The considered molecules are specified in Table 2.

Table 2. Molecules Included in the CHNO Data Set Used to Train the CONI Net(CHNO) Model[a]

CHNO data set and MD ensemble	
formula	name
C_2H_6O	dimethylether
CH_2O	formaldehyde
C_3H_6O	acetone
C_2H_3N	acetonitrile
CH_4O	methanol
C_2H_6O	ethanol
CH_2O_2	formic acid
CH_3NO	formamide

[a]Here, the same molecules also represent the MD ensemble used to test the prediction of observables.

The data set, in the following referred to as CHNO data set, contains 2000 homodimers per molecule, which gives a total of 12,000 data points in the training and 4000 data points in the validation set. In addition, an independent test set is generated, which contains 4000 data points with the same composition.

We train five models for each energy component while keeping the choice of training and validation set fixed. Finally, we choose the best model for each component according to the MAE for the validation set. In Figure 3, the test set results for the component and total energy models are shown. First, we want to emphasize that the MAEs of the components and

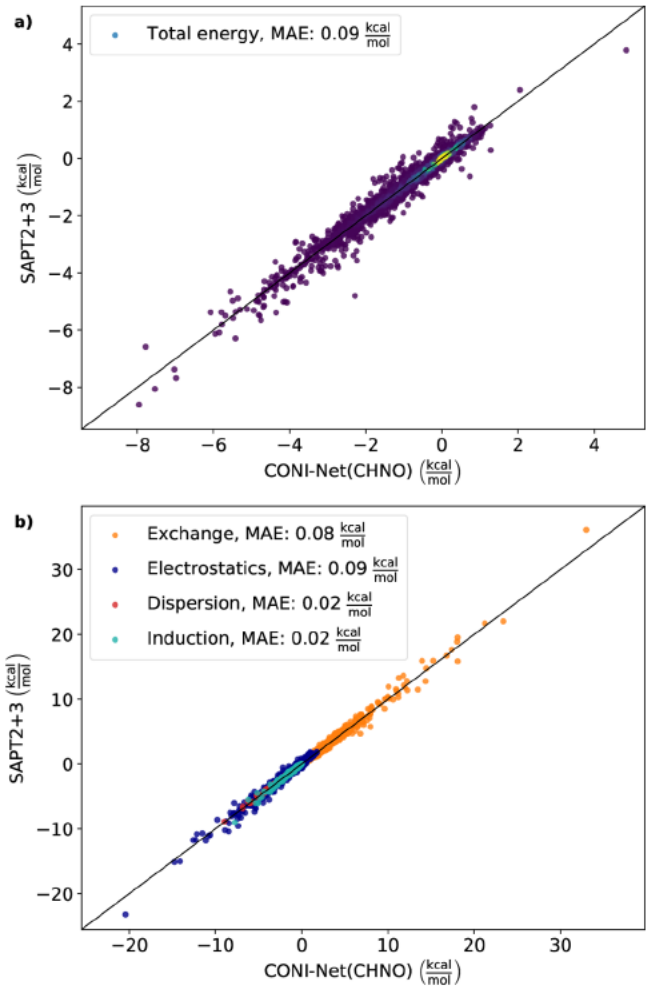


Figure 3. (a) Total energies and (b) energy components obtained with CONI Net(CHNO) versus SAPT2+3 for the CHNO test set. In (a), the color coding represents the local point density in the scatter plot computed via a Gaussian kernel density estimate as implemented in SciPy,[93] violet corresponds to low and yellow to high values.

the total energy are primarily relative indicators to compare models for a specific data set. The absolute magnitude of the MAE is heavily dependent on the composition of the considered data set. The importance of the so called chemical accuracy of 1 kcal/mol, which is a significant sound barrier for the modeling of covalent interactions, should therefore be taken with care for NCI models, especially since the energy ranges are considerably lower.

When comparing the errors of the different energy components, we observe significant differences. One reason is that the different interactions of the model cope differently with the approximations. Charge penetration and Pauli repulsion are emerging from overlapping orbitals. Therefore, the electrostatic and exchange component can have an angle dependence, which is averaged in the isotropic pairwise model. Furthermore, the partial charges of the electrostatic baseline model have to be symmetrized artificially to obey the symmetry of the intramolecular force field, which introduces another obstacle for the model fit. In comparison, the dispersion component with its more long range character copes better with the approximations of the model even for high absolute energies, which is not surprising given the success of analytical dispersion corrections for DFT methods.[1]

Similarly, the induction component also shows a low MAE. Here, the range of values the model has to span is smaller than for the other energy components, a characteristic of the considered molecules in the data set. However, it is important to note that all contributions are fitted into a pairwise additive model to reproduce dimer energies. Especially for the dispersion and induction component, this can constitute a source of error when expanding the system to bulk structures. For instance, in polar or ionic liquids, many body interactions can make up a considerable part of the interaction energy.[94,95]

In order to test the model performance beyond dimer structures, for each molecule (Table 2), we conduct an MD simulation and compare the calculated values for the enthalpy of vaporization ΔH_{vap} and the mass density ρ to the literature values from experiments and for the force fields GAFF and OPLS AA,[92] and the results are shown in Figure 4. From the

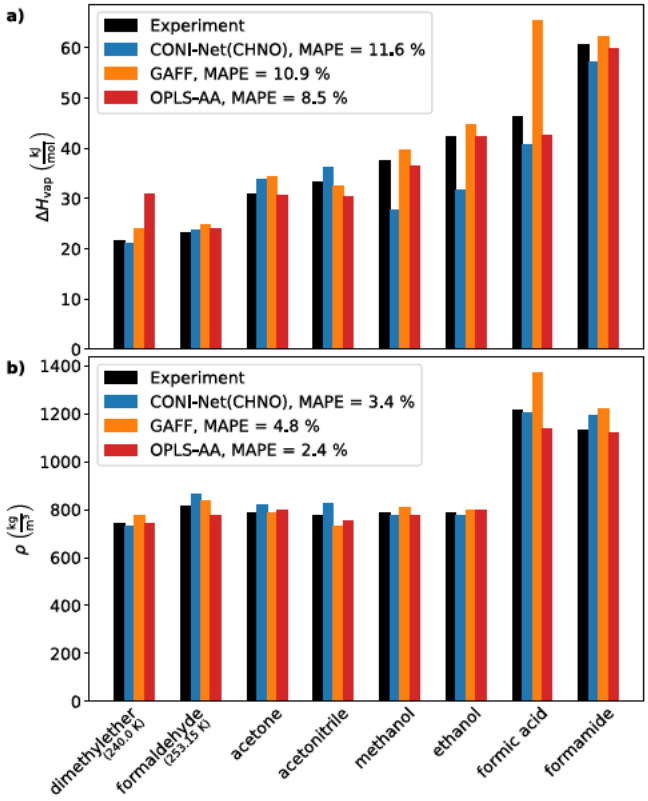


Figure 4. Comparison of observables (a) enthalpy of vaporization ΔH_{vap} and (b) mass density ρ obtained from MD simulations and from experiments. All molecules are included in the CHNO data set used to train the CONI Net(CHNO) model. Values for GAFF and OPLS AA were taken from ref 92. Numerical values and references for the experimental values can be found in Table S1 of Supporting Information.

mean absolute percentage errors (MAPEs), it is evident that for all force fields, the prediction of the enthalpy of vaporization is a more challenging task compared to the mass density. The OPLS AA force field shows excellent performance for both tasks, which is not surprising since it was developed for modeling organic liquids. The GAFF results also show an overall solid performance apart from a large deviation for formic acid. The CONI Net model performs well on predicting the mass density and the MAPE positions between OPLS AA and GAFF for this task. For the enthalpy of vaporization, both under and overbinding values are observed,

resulting in an MAPE slightly larger than the conventional force fields but of a similar order of magnitude. When considering that, in addition to the discussed sources of errors, the limited ab initio accuracy is propagating from the data set to the MD results, the overall performance is promising, particularly since the data set quality still has room for improvement.

In conclusion, the test set results for the CHNO data set (Figure 3) show that the CONI Net model is capable of interpolating the ab initio target energies between both the space of pair fingerprints and different dimer arrangements for a set of specific molecules. The subsequent application in MD simulations of the liquid phase to predict thermodynamic observables (Figure 4) demonstrates the transferability of the model to larger arrangements. However, in order to extrapolate to unknown molecules of the chemical space of organic molecules, more chemical elements and functional groups would have to be considered in the data set generation.

3.3. Transferable Model for Hydrocarbons. For the following application of the model, we limit the chemical space to hydrocarbon molecules. The goal is to generate a prototype of a transferable model. Therefore, the training ensemble should span a variety of characteristic functional groups while keeping the number of carbon atoms low due to the steep scaling of the computational cost of the reference SAPT2+3 calculations. We include molecules with up to three carbon atoms into the training ensemble, with the exception of benzene with six carbon atoms, in order to capture atoms in aromatic bonds. A complete list is given in Table 3.

Table 3. Molecules Included in the CH Data Set Used to Train the CONI Net(CH) Model and Molecules in the MD Ensemble Used to Test the Prediction of Observables

CH data set		MD ensemble	
formula	name	formula	name
CH_4	methane	C_5H_{10}	1-pentene
C_2H_6	ethane	C_5H_{12}	pentane
C_2H_4	ethylene	C_5H_8	cyclopentene
C_2H_2	acetylene	C_5H_8	1-pentyne
C_3H_8	propane	C_6H_{14}	isohexane
C_3H_6	propene	C_6H_{14}	hexane
C_3H_4	propyne	C_6H_{12}	cyclohexane
C_6H_6	benzene	C_7H_8	toluene
		C_8H_{10}	*o*-xylene
		$C_{10}H_8$	naphthalene

In the data set, the **CH data set**, each molecule of the training ensemble is represented by 2000 homodimers, and for each combination, 400 heterodimers are included. Accordingly, the training set contains 20,400 and the validation set contains 6800 data points. Furthermore, a test set with the same composition is created containing 6800 data points. As for the previous application, for each component, we choose the best of five models, all trained with a fixed training/validation split of the data set. The results of the best component and total energy models are shown in Figure 5. Again, we see a component dependent magnitude of the test set error. However, the total energy range and the corresponding MAE are smaller compared to the values of the CHNO test set.

In contrast to the MD simulations for the CONI Net(CHNO) model, here, the molecules of the MD ensemble are not included in the CH data set. Therefore, the

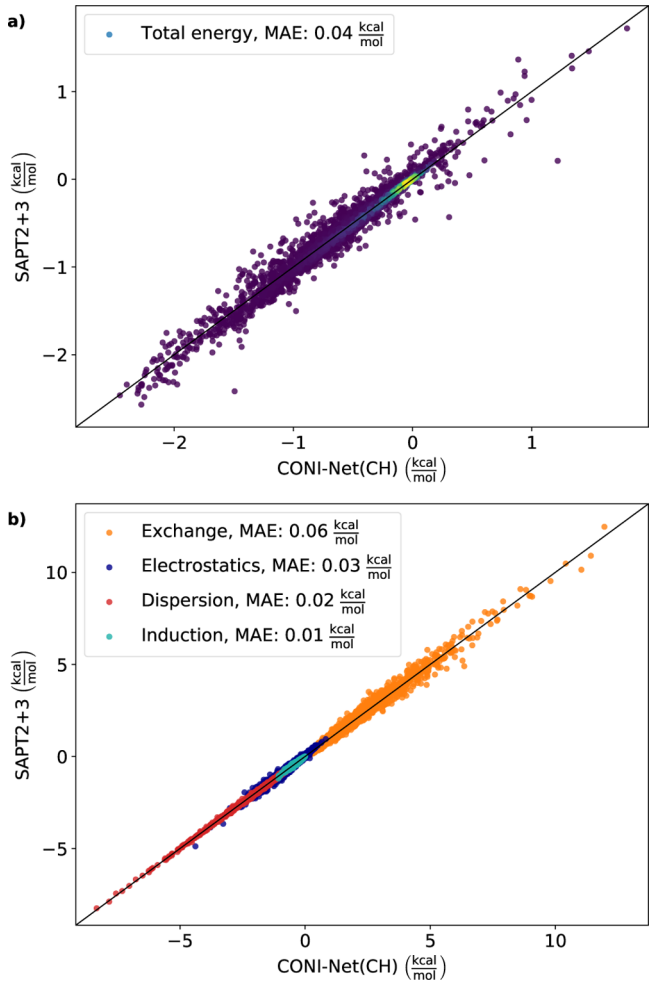


Figure 5. (a) Total energies and (b) energy components obtained with CONI Net(CH) versus SAPT2+3 for the CH test set. In (a), the color coding represents the local point density in the scatter plot computed via a Gaussian kernel density estimate as implemented in SciPy,[93] violet corresponds to low and yellow to high values.

computational limitations are not as strict, and we can use significantly larger molecules (Table 3) since only the fingerprints and partial charges have to be computed in QM calculations once before the MD run. The calculated values for the enthalpy of vaporization ΔH_{vap} and the mass density ρ are shown in Figure 6 and compared to experimental values. The overall more consistent performance for both properties in comparison with the CONI Net(CHNO) model indicates a smaller variation in the systematic errors. This can be explained by the limited set of functional groups of the two considered molecule classes, aliphatic and aromatic compounds. We continue the discussion based on the points of the previous section about the potential sources of errors since they also apply to the CH data set. They include the propagation of the ab initio inaccuracy to the model and the enforced symmetry of RESP charges due to the symmetric intramolecular force field. Additionally, for the molecules of the CH data set, another effect can arise. For the linear alkane molecules, such as pentane and hexane, the RESP charges are fitted to highly symmetric optimized ground state geometries. In the MD simulation, the soft degrees of freedom of the backbone dihedral angles allow the molecule to deviate from this conformation significantly at a low energetic cost. A solution to

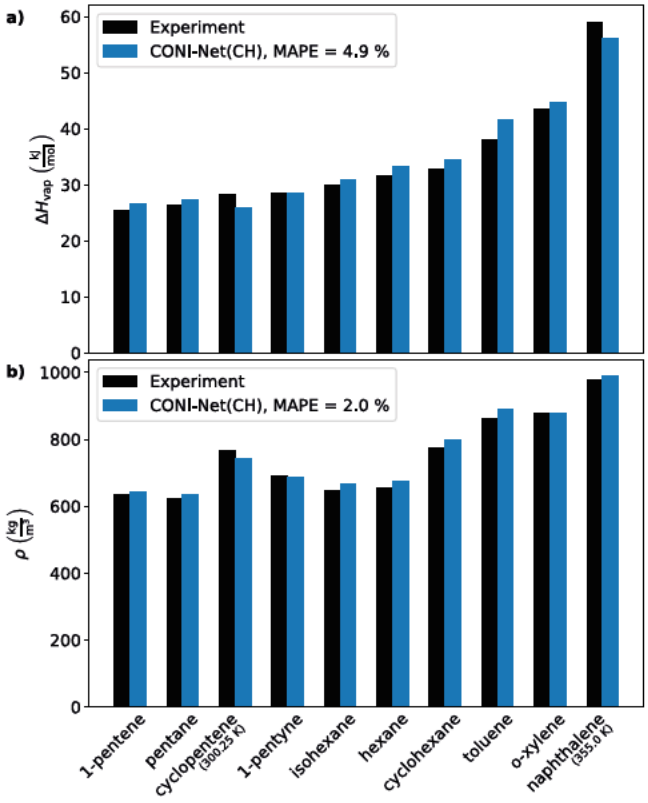


Figure 6. Comparison of observables (a) enthalpy of vaporization ΔH_{vap} and (b) mass density ρ obtained from MD simulations and from experiments. None of the molecules are included in the CH data set used to train the CONI Net(CH) model. Numerical values and references for the experimental values can be found in Table S2 of Supporting Information.

get a more consistent compromise for the charges could be a more complex RESP fitting procedure that includes multiple conformers.

Nevertheless, despite the discussed approximations and the limited accuracy of the reference calculations, the model already shows promising extrapolation capabilities. The MD ensemble includes longer chains, unseen branched and cyclic aliphatic molecules, and new combinations of different functional groups from the training set. In order to go beyond this proof of principle, the minimalistic data set would have to be expanded to include all relevant functional groups and to reveal the nuances of slightly different pair fingerprints, such as from the same functional group in differently sized molecules.

4. CONCLUSIONS

In this study, we presented a systematic bottom up approach for developing an NNP for separable NCIs. It is applicable for custom models comprising one or several specific molecules and transferable models trained on diverse data sets to obtain extrapolation capabilities. The resulting intermolecular force field allows stable MD simulations approaching the predictive and computational performance of established force fields while offering separable interaction contributions and a straightforward parametrization procedure, which does not rely on empirical data. Thus, it represents an alternative for compounds where the existing general force fields do not provide parameters. All steps from data set generation over descriptor preparation to model training are designed to be performed without manual intervention and can be implemented in an automated workflow. Therefore, upscaling and extension of the data set are only limited by computational resources.

The method has the potential for improvement at different levels. One way to improve the predictive performance could be to increase the accuracy of the ab initio method, for example, through larger basis sets. In order to cope with the more demanding scaling of computational cost with system size, an active learning method could decrease the required number of training samples. Here, in a feedback loop, new candidates for data points are chosen based on the current data set. An example is the query by committee selection method that was successfully applied to develop the ANI 1x model, which, despite decreased data set size, has been shown to outperform the original ANI 1 model across several benchmarks.[54,96,97]

In order to improve the transferability of the model to distorted monomers, the dimer sampling, fingerprint calculation, and partial charge fit could be extended to include nonequilibrium monomer geometries. Furthermore, an advanced baseline model could include multipole or many body contributions. The pair fingerprint could be extended by other suitable descriptors that improve the ability of the neural network to capture subtle differences in the atomic bonding environments. Finally, an aspect we have not addressed in this study is the intramolecular force field. A fit of bonded potentials to ab initio vibrational spectra and conformer scans would be the last missing piece to a consistent bottom up force field entirely from first principles.

Regardless of the many possibilities for future enhancements, the present results already represent a step toward elevating separable NCIs from the quantum world to molecular mechanics without relying on prior empirical knowledge. The general procedure allows force field development for many molecular materials, provided a reliable and separable ab initio method is available. Therefore, our model is a potentially useful tool for the molecular design of unknown compounds and opens up new possibilities for analyzing the nature of intermolecular interactions at large scales.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.1c00328.

Numerical values of Figures 4 and 6 (PDF)

Geometries, SAPT2+3 energy components, RESP charges, and fingerprint descriptor elements for all dimer samples in the CHNO and CH data sets (ZIP)

Jupyter notebook demonstrating an implementation of the CONI Net model (ZIP)

AUTHOR INFORMATION

Corresponding Author

Wolfgang Wenzel − *Institute of Nanotechnology, Karlsruhe Institute of Technology, Eggenstein Leopoldshafen 76344, Germany*; Email: wolfgang.wenzel@kit.edu

Author

Manuel Konrad − *Institute of Nanotechnology, Karlsruhe Institute of Technology, Eggenstein Leopoldshafen 76344, Germany*; orcid.org/0000 0001 8196 5071

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors gratefully acknowledge support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy via the Excellence Cluster 3D Matter Made to Order (EXC 2082/1 390761711) and by the GRK 2450.

REFERENCES

(1) Grimme, S. Accurate description of van der Waals complexes by density functional theory including empirical corrections. *J. Comput. Chem.* **2004**, *25*, 1463−1473.
(2) Grimme, S. Semiempirical GGA type density functional constructed with a long range dispersion correction. *J. Comput. Chem.* **2006**, *27*, 1787−1799.
(3) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate *ab initio* parametrization of density functional dispersion correction (DFT D) for the 94 elements H Pu. *J. Chem. Phys.* **2010**, *132*, 154104.
(4) Becke, A. D.; Johnson, E. R. Exchange hole dipole moment and the dispersion interaction revisited. *J. Chem. Phys.* **2007**, *127*, 154108.
(5) Tkatchenko, A.; Scheffler, M. Accurate Molecular Van Der Waals Interactions from Ground State Electron Density and Free Atom Reference Data. *Phys. Rev. Lett.* **2009**, *102*, 073005.
(6) Ambrosetti, A.; Reilly, A. M.; DiStasio, R. A.; Tkatchenko, A. Long range correlation energy calculated from coupled atomic response functions. *J. Chem. Phys.* **2014**, *140*, 18A508.
(7) Dion, M.; Rydberg, H.; Schröder, E.; Langreth, D. C.; Lundqvist, B. I. Van der Waals Density Functional for General Geometries. *Phys. Rev. Lett.* **2004**, *92*, 246401.
(8) Vydrov, O. A.; Van Voorhis, T. Nonlocal van der Waals density functional: The simpler the better. *J. Chem. Phys.* **2010**, *133*, 244103.
(9) Lee, K.; Murray, a. D.; Kong, L.; Lundqvist, B. I.; Langreth, D. C. Higher accuracy van der Waals density functional. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2010**, *82*, 081101.
(10) Møller, C.; Plesset, M. S. Note on an Approximation Treatment for Many Electron Systems. *Phys. Rev.* **1934**, *46*, 618−622.
(11) Čížek, J. On the Correlation Problem in Atomic and Molecular Systems. Calculation of Wavefunction Components in Ursell Type Expansion Using Quantum Field Theoretical Methods. *J. Chem. Phys.* **1966**, *45*, 4256−4266.
(12) Hohenstein, E. G.; Sherrill, C. D. Wavefunction methods for noncovalent interactions: Noncovalent interactions. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2012**, *2*, 304−326.
(13) Bartlett, R. J.; Purvis, G. D. Many body perturbation theory, coupled pair many electron theory, and the importance of quadruple excitations for the correlation problem. *Int. J. Quantum Chem.* **1978**, *14*, 561−581.
(14) Purvis, G. D.; Bartlett, R. J. A full coupled cluster singles and doubles model: The inclusion of disconnected triples. *J. Chem. Phys.* **1982**, *76*, 1910−1918.
(15) Pople, J. A.; Head Gordon, M.; Raghavachari, K. Quadratic configuration interaction. A general technique for determining electron correlation energies. *J. Chem. Phys.* **1987**, *87*, 5968−5975.
(16) Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head Gordon, M. A fifth order perturbation comparison of electron correlation theories. *Chem. Phys. Lett.* **1989**, *157*, 479−483.
(17) Koch, H.; Fernández, B.; Christiansen, O. The benzene−argon complex: A ground and excited state *ab initio* study. *J. Chem. Phys.* **1998**, *108*, 2784−2790.
(18) Jurečka, P.; Šponer, J.; Černý, J.; Hobza, P. Benchmark database of accurate (MP2 and CCSD(T) complete basis set limit) interaction energies of small model complexes, DNA base pairs, and amino acid pairs. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985−1993.
(19) Řezáč, J.; Riley, K. E.; Hobza, P. S66: A Well balanced Database of Benchmark Interaction Energies Relevant to Biomolecular Structures. *J. Chem. Theory Comput.* **2011**, *7*, 2427−2438.
(20) Jeziorski, B.; Moszynski, R.; Szalewicz, K. Perturbation Theory Approach to Intermolecular Potential Energy Surfaces of van der Waals Complexes. *Chem. Rev.* **1994**, *94*, 1887−1930.
(21) Parker, T. M.; Burns, L. A.; Parrish, R. M.; Ryno, A. G.; Sherrill, C. D. Levels of symmetry adapted perturbation theory (SAPT). I. Efficiency and performance for interaction energies. *J. Chem. Phys.* **2014**, *140*, 094106.
(22) Hohenstein, E. G.; Parrish, R. M.; Sherrill, C. D.; Turney, J. M.; Schaefer, H. F. Large scale symmetry adapted perturbation theory computations via density fitting and Laplace transformation techniques: Investigating the fundamental forces of DNA intercalator interactions. *J. Chem. Phys.* **2011**, *135*, 174107.
(23) Williams, H. L.; Chabalowski, C. F. Using Kohn−Sham Orbitals in Symmetry Adapted Perturbation Theory to Investigate Intermolecular Interactions. *J. Phys. Chem. A* **2001**, *105*, 646−659.
(24) Misquitta, A. J.; Szalewicz, K. Intermolecular forces from asymptotically corrected density functional description of monomers. *Chem. Phys. Lett.* **2002**, *357*, 301−306.
(25) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiórkiewicz Kuczera, J.; Yin, D.; Karplus, M. All Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, *102*, 3586−3616.
(26) Takamori, S.; Holt, M.; Stenius, K.; Lemke, E. A.; Grønborg, M.; Riedel, D.; Urlaub, H.; Schenck, S.; Brügger, B.; Ringler, P.; Müller, S. A.; Rammner, B.; Gräter, F.; Hub, J. S.; De Groot, B. L.; Mieskes, G.; Moriyama, Y.; Klingauf, J.; Grubmüller, H.; Heuser, J.; Wieland, F.; Jahn, R. Molecular Anatomy of a Trafficking Organelle. *Cell* **2006**, *127*, 831−846.
(27) Freddolino, P. L.; Arkhipov, A. S.; Larson, S. B.; McPherson, A.; Schulten, K. Molecular Dynamics Simulations of the Complete Satellite Tobacco Mosaic Virus. *Structure* **2006**, *14*, 437−449.
(28) Zink, M.; Grubmüller, H. Mechanical Properties of the Icosahedral Shell of Southern Bean Mosaic Virus: A Molecular Dynamics Study. *Biophys. J.* **2009**, *96*, 1350−1363.
(29) Minoia, A.; Chen, L.; Beljonne, D.; Lazzaroni, R. Molecular modeling study of the structure and stability of polymer/carbon nanotube interfaces. *Polymer* **2012**, *53*, 5480−5490.
(30) Tummala, N. R.; Risko, C.; Bruner, C.; Dauskardt, R. H.; Brédas, J. L. Entanglements in P3HT and their influence on thin film mechanical properties: Insights from molecular dynamics simulations. *J. Polym. Sci., Part B: Polym. Phys.* **2015**, *53*, 934−942.
(31) Roscioni, O. M.; D'Avino, G.; Muccioli, L.; Zannoni, C. Pentacene Crystal Growth on Silica and Layer Dependent Step Edge Barrier from Atomistic Simulations. *J. Phys. Chem. Lett.* **2018**, *9*, 6900−6906.
(32) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179−5197.
(33) Jorgensen, W. L.; Maxwell, D. S.; Tirado Rives, J. Development and Testing of the OPLS All Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225−11236.
(34) Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham, T. E.; DeBolt, S.; Ferguson, D.; Seibel, G.; Kollman, P. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Comput. Phys. Commun.* **1995**, *91*, 1−41.

(35) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **2004**, *25*, 1157−1174.
(36) 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*, 843−856.
(37) 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* **2017**, *14*, 71−73.
(38) McDaniel, J. G.; Schmidt, J. R. Next Generation Force Fields from Symmetry Adapted Perturbation Theory. *Annu. Rev. Phys. Chem.* **2016**, *67*, 467−488.
(39) Stone, A. J.; Misquitta, A. J. Atom−atom potentials from *ab initio* calculations. *Int. Rev. Phys. Chem.* **2007**, *26*, 193−222.
(40) Tafipolsky, M.; Ansorg, K. Toward a Physically Motivated Force Field: Hydrogen Bond Directionality from a Symmetry Adapted Perturbation Theory Perspective. *J. Chem. Theory Comput.* **2016**, *12*, 1267−1279.
(41) Xu, P.; Guidez, E. B.; Bertoni, C.; Gordon, M. S. Perspective: *Ab initio* force field methods derived from quantum mechanics. *J. Chem. Phys.* **2018**, *148*, 090901.
(42) Liu, Y. P.; Kim, K.; Berne, B. J.; Friesner, R. A.; Rick, S. W. Constructing *ab initio* force fields for molecular dynamics simulations. *J. Chem. Phys.* **1998**, *108*, 4739−4755.
(43) Bukowski, R.; Szalewicz, K.; Chabalowski, C. F. Ab Initio Interaction Potentials for Simulations of Dimethylnitramine Solutions in Supercritical Carbon Dioxide with Cosolvents. *J. Phys. Chem. A* **1999**, *103*, 7322−7340.
(44) Hloucha, M.; Sum, A. K.; Sandler, S. I. Computer simulation of acetonitrile and methanol with ab initio based pair potentials. *J. Chem. Phys.* **2000**, *113*, 5401.
(45) Podeszwa, R.; Bukowski, R.; Szalewicz, K. Potential Energy Surface for the Benzene Dimer and Perturbational Analysis of − Interactions. *J. Phys. Chem. A* **2006**, *110*, 10345−10354.
(46) Yu, K.; McDaniel, J. G.; Schmidt, J. R. Physically Motivated, Robust, ab Initio Force Fields for CO_2 and N_2. *J. Phys. Chem. B* **2011**, *115*, 10054−10063.
(47) McDaniel, J. G.; Schmidt, J. R. Physically Motivated Force Fields from Symmetry Adapted Perturbation Theory. *J. Phys. Chem. A* **2013**, *117*, 2053−2066.
(48) McDaniel, J. G.; Schmidt, J. R. First Principles Many Body Force Fields from the Gas Phase to Liquid: A "Universal" Approach. *J. Phys. Chem. B* **2014**, *118*, 8042−8053.
(49) Schmidt, J. R.; Yu, K.; McDaniel, J. G. Transferable Next Generation Force Fields from Simple Liquids to Complex Materials. *Acc. Chem. Res.* **2015**, *48*, 548−556.
(50) Yu, K.; Schmidt, J. R. Many body effects are essential in a physically motivated CO_2 force field. *J. Chem. Phys.* **2012**, *136*, 034503.
(51) Wang, L. P.; Head Gordon, T.; Ponder, J. W.; Ren, P.; Chodera, J. D.; Eastman, P. K.; Martinez, T. J.; Pande, V. S. Systematic Improvement of a Classical Molecular Model of Water. *J. Phys. Chem. B* **2013**, *117*, 9956−9972.
(52) Liu, C.; Piquemal, J. P.; Ren, P. AMOEBA+ Classical Potential for Modeling Molecular Interactions. *J. Chem. Theory Comput.* **2019**, *15*, 4122−4139.
(53) Behler, J.; Parrinello, M. Generalized Neural Network Representation of High Dimensional Potential Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.
(54) 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*, 3192−3203.
(55) Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons. *Phys. Rev. Lett.* **2010**, *104*, 136403.
(56) Schütt, K. T.; Arbabzadah, F.; Chmiela, S.; Müller, K. R.; Tkatchenko, A. Quantum chemical insights from deep tensor neural networks. *Nat. Commun.* **2017**, *8*, 13890.
(57) Schütt, K. T.; Sauceda, H. E.; Kindermans, P. J.; Tkatchenko, A.; Müller, K. R. SchNet − A deep learning architecture for molecules and materials. *J. Chem. Phys.* **2018**, *148*, 241722.
(58) Selvaratnam, B.; Koodali, R. T.; Miró, P. Prediction of optoelectronic properties of Cu_2O using neural network potential. *Phys. Chem. Chem. Phys.* **2020**, *22*, 14910−14917.
(59) Gastegger, M.; Behler, J.; Marquetand, P. Machine learning molecular dynamics for the simulation of infrared spectra. *Chem. Sci.* **2017**, *8*, 6924−6935.
(60) Jose, K. V. J.; Artrith, N.; Behler, J. Construction of high dimensional neural network potentials using environment dependent atom pairs. *J. Chem. Phys.* **2012**, *136*, 194111.
(61) Metcalf, D. P.; Koutsoukas, A.; Spronk, S. A.; Claus, B. L.; Loughney, D. A.; Johnson, S. R.; Cheney, D. L.; Sherrill, C. D. Approaches for machine learning intermolecular interaction energies and application to energy components from symmetry adapted perturbation theory. *J. Chem. Phys.* **2020**, *152*, 074103.
(62) Glick, Z. L.; Metcalf, D. P.; Koutsoukas, A.; Spronk, S. A.; Cheney, D. L.; Sherrill, C. D. AP Net An atomic pairwise neural network for smooth and transferable interaction potentials. *J. Chem. Phys.* **2020**, *153*, 044112.
(63) Yao, K.; Herr, J. E.; Toth, D. W.; Mckintyre, R.; Parkhill, J. The TensorMol 0.1 model chemistry: a neural network augmented with long range physics. *Chem. Sci.* **2018**, *9*, 2261−2269.
(64) Unke, O. T.; Meuwly, M. PhysNet A Neural Network for Predicting Energies, Forces, Dipole Moments, and Partial Charges. *J. Chem. Theory Comput.* **2019**, *15*, 3678−3693.
(65) Zubatyuk, R.; Smith, J. S.; Leszczynski, J.; Isayev, O. Accurate and transferable multitask prediction of chemical properties with an atoms in molecules neural network. *Sci. Adv.* **2019**, *5*, No. eaav6490.
(66) Bereau, T.; DiStasio, R. A.; Tkatchenko, A.; von Lilienfeld, O. A. Non covalent interactions across organic and biological subsets of chemical space: Physics based potentials parametrized from machine learning. *J. Chem. Phys.* **2018**, *148*, 241706.
(67) Vashisth, A.; Ashraf, C.; Zhang, W.; Bakis, C. E.; van Duin, A. C. T. Accelerated ReaxFF Simulations for Describing the Reactive Cross Linking of Polymers. *J. Phys. Chem. A* **2018**, *122*, 6633−6642.
(68) Dasgupta, N.; Yilmaz, D. E.; van Duin, A. Simulations of the Biodegradation of Citrate Based Polymers for Artificial Scaffolds Using Accelerated Reactive Molecular Dynamics. *J. Phys. Chem. B* **2020**, *124*, 5311−5322.
(69) Stukowski, A. Visualization and analysis of atomistic simulation data with OVITO−the Open Visualization Tool. *Modell. Simul. Mater. Sci. Eng.* **2009**, *18*, 015012.
(70) Neese, F. The ORCA program system. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2012**, *2*, 73−78.
(71) Becke, A. D. Density functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **1993**, *98*, 5648−5652.
(72) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle Salvetti correlation energy formula into a functional of the electron density. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1988**, *37*, 785−789.
(73) Vosko, S. H.; Wilk, L.; Nusair, M. Accurate spin dependent electron liquid correlation energies for local spin density calculations: a critical analysis. *Can. J. Phys.* **1980**, *58*, 1200−1211.
(74) Kendall, R. A.; Dunning, T. H.; Harrison, R. J. Electron affinities of the first row atoms revisited. Systematic basis sets and wave functions. *J. Chem. Phys.* **1992**, *96*, 6796−6806.
(75) Lu, T.; Chen, F.F. Multiwfn A multifunctional wavefunction analyzer. *J. Comput. Chem.* **2012**, *33*, 580−592.
(76) Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A. A well behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *J. Phys. Chem.* **1993**, *97*, 10269−10280.
(77) Dunning, T. H. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *J. Chem. Phys.* **1989**, *90*, 1007−1023.

(78) Mayer, I. Charge, bond order and valence in the AB initio SCF theory. *Chem. Phys. Lett.* **1983**, *97*, 270−274.
(79) Hirshfeld, F. L. Bonded atom fragments for describing molecular charge densities. *Theor. Chim. Acta* **1977**, *44*, 129−138.
(80) Patkowski, K. Recent developments in symmetry adapted perturbation theory. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2020**, *10*, No. e1452.
(81) Hohenstein, E. G.; Sherrill, C. D. Density fitting of intramonomer correlation effects in symmetry adapted perturbation theory. *J. Chem. Phys.* **2010**, *133*, 014101.
(82) Hohenstein, E. G.; Sherrill, C. D. Efficient evaluation of triple excitations in symmetry adapted perturbation theory via second order Møller−Plesset perturbation theory natural orbitals. *J. Chem. Phys.* **2010**, *133*, 104107.
(83) Parrish, R. M.; Burns, L. A.; Smith, D. G. A.; Simmonett, A. C.; DePrince, A. E.; Hohenstein, E. G.; Bozkaya, U.; Sokolov, A. Y.; Di Remigio, R.; Richard, R. M.; Gonthier, J. F.; James, A. M.; McAlexander, H. R.; Kumar, A.; Saitow, M.; Wang, X.; Pritchard, B. P.; Verma, P.; Schaefer, H. F.; Patkowski, K.; King, R. A.; Valeev, E. F.; Evangelista, F. A.; Turney, J. M.; Crawford, T. D.; Sherrill, C. D. Psi4 1.1: An Open Source Electronic Structure Program Emphasizing Automation, Advanced Libraries, and Interoperability. *J. Chem. Theory Comput.* **2017**, *13*, 3185−3197.
(84) Case, D. A.; Ben Shalom, I. Y.; Brozell, S. R.; Cerutti, D. S.; Cheatham, T. E., III; Cruzeiro, V. W. D.; Darden, T. A.; Duke, R. E.; Ghoreishi, D.; Giambasu, G.; Giese, T.; Gilson, M. K.; Gohlke, H.; Goetz, A. W.; Greene, D.; Harris, R.; Homeyer, N.; Huang, Y.; Izadi, S.; Kovalenko, A.; Krasny, R.; Kurtzman, T.; Lee, T. S.; LeGrand, S.; Li, P.; Lin, C.; Liu, J.; Luchko, T.; Luo, R.; Man, V.; Mermelstein, D. J.; Merz, K. M.; Miao, Y.; Monard, G.; Nguyen, C.; Nguyen, H.; Onufriev, A.; Pan, F.; Qi, R.; Roe, D. R.; Roitberg, A.; Sagui, C.; Schott Verdugo, S.; Shen, J.; Simmerling, C. L.; Smith, J.; Swails, J.; Walker, R. C.; Wang, J.; Wei, H.; Wilson, L.; Wolf, R. M.; Wu, X.; Xiao, L.; Xiong, Y.; York, D. M.; Kollman, P. A. *AMBER 2019*; University of California: San Francisco, 2019.
(85) Mei, J.; Davenport, J. W.; Fernando, G. W. Analytic embedded atom potentials for fcc metals: Application to liquid and solid copper. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1991**, *43*, 4653−4658.
(86) Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; Chintala, S. PyTorch: An Imperative Style, High Performance Deep Learning Library. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 8026−8037.
(87) Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. **2017**, arXiv:1412.6980.
(88) Malde, A. K.; Zuo, L.; Breeze, M.; Stroet, M.; Poger, D.; Nair, P. C.; Oostenbrink, C.; Mark, A. E. An Automated Force Field Topology Builder (ATB) and Repository: Version 1.0. *J. Chem. Theory Comput.* **2011**, *7*, 4026−4037.
(89) Stroet, M.; Caron, B.; Visscher, K. M.; Geerke, D. P.; Malde, A. K.; Mark, A. E. Automated Topology Builder Version 3.0: Prediction of Solvation Free Enthalpies in Water and Hexane. *J. Chem. Theory Comput.* **2018**, *14*, 5834−5845.
(90) Harris, C. R.; Millman, K. J.; van der Walt, S. J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N. J.; Kern, R.; Picus, M.; Hoyer, S.; van Kerkwijk, M. H.; Brett, M.; Haldane, A.; del Río, J. F.; Wiebe, M.; Peterson, P.; Gérard Marchant, P.; Sheppard, K.; Reddy, T.; Weckesser, W.; Abbasi, H.; Gohlke, C.; Oliphant, T. E. Array programming with NumPy. *Nature* **2020**, *585*, 357−362.
(91) Plimpton, S. Fast Parallel Algorithms for Short Range Molecular Dynamics. *J. Comput. Phys.* **1995**, *117*, 1−19.
(92) Caleman, C.; van Maaren, P. J.; Hong, M.; Hub, J. S.; Costa, L. T.; van der Spoel, D. Force Field Benchmark of Organic Liquids: Density, Enthalpy of Vaporization, Heat Capacities, Surface Tension, Isothermal Compressibility, Volumetric Expansion Coefficient, and Dielectric Constant. *J. Chem. Theory Comput.* **2012**, *8*, 61−74.
(93) Virtanen, P.; Gommers, R.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; van der Walt, S. J.; Brett, M.; Wilson, J.; Millman, K. J.; Mayorov, N.; Nelson, A. R. J.; Jones, E.; Kern, R.; Larson, E.; Carey, C. J.; Polat, İ.; Feng, Y.; Moore, E. W.; VanderPlas, J.; Laxalde, D.; Perktold, J.; Cimrman, R.; Henriksen, I.; Quintero, E. A.; Harris, C. R.; Archibald, A. M.; Ribeiro, A. H.; Pedregosa, F.; van Mulbregt, P.; SciPy 1.0 Contributors. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **2020**, *17*, 261−272.
(94) Heßelmann, A. Correlation effects and many body interactions in water clusters. *Beilstein J. Org. Chem.* **2018**, *14*, 979−991.
(95) Egan, C. K.; Paesani, F. Assessing Many Body Effects of Water Self Ions. II: H3O+(H2O)n Clusters. *J. Chem. Theory Comput.* **2019**, *15*, 4816−4833.
(96) Seung, H. S.; Opper, M.; Sompolinsky, H. Query by committee. *Proceedings of the fifth annual workshop on Computational learning theory COLT '92*: Pittsburgh, Pennsylvania, United States, 1992, pp 287−294. DOI: 10.1145/130385.130417
(97) Smith, J. S.; Nebgen, B.; Lubbers, N.; Isayev, O.; Roitberg, A. E. Less is more: Sampling chemical space with active learning. *J. Chem. Phys.* **2018**, *148*, 241733.