JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

PhyNEO: A Neural-Network-Enhanced Physics-Driven Force Field Development Workflow for Bulk Organic Molecule and Polymer Simulations

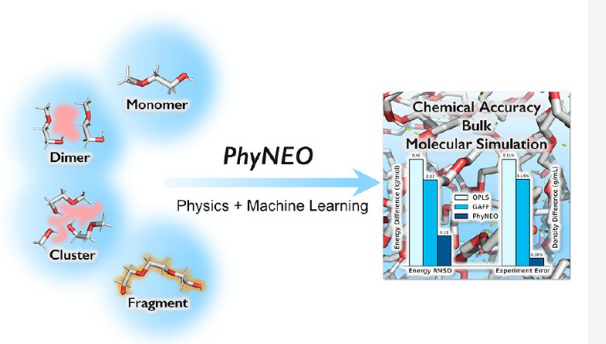
Published as part of Journal of Chemical Theory and Computation virtual special issue "Machine Learning and Statistical Mechanics: Shared Synergies for Next Generation of Chemical Theory and Computation".

Junmin Chen and Kuang Yu*

Cite This: J. Chem. Theory Comput. 2024, 20, 253−265 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: An accurate, generalizable, and transferable force field plays a crucial role in the molecular dynamics simulations of organic polymers and biomolecules. Conventional empirical force fields often fail to capture precise intermolecular interactions due to their negligence of important physics, such as polarization, charge penetration, many-body dispersion, etc. Moreover, the parameterization of these force fields relies heavily on top-down fittings, limiting their transferabilities to new systems where the experimental data are often unavailable. To address these challenges, we introduce a general and fully ab initio force field construction strategy, named PhyNEO. It features a hybrid approach that combines both the physics-driven and the data-driven methods and is able to generate a bulk potential with chemical accuracy using only quantum chemistry data of very small clusters. Careful separations of long-/short-range interactions and nonbonding/bonding interactions are the key to the success of PhyNEO. By such a strategy, we mitigate the limitations of pure data-driven methods in long-range interactions, thus largely increasing the data efficiency and the scalability of machine learning models. The new approach is thoroughly tested on poly(ethylene oxide) and polyethylene glycol systems, giving superior accuracies in both microscopic and bulk properties compared to conventional force fields. This work thus offers a promising framework for the development of advanced force fields in a wide range of organic molecular systems.



1. INTRODUCTION

An accurate and versatile force field brings us closer to the achievement of the Holy Grail of molecular dynamics (MD) simulations with chemical accuracy. The significance of a reliable organic molecular force field cannot be overstated in various research fields, including biomolecules, small molecule drugs, polymers, and organic porous materials. The current workhorses of most industrial MD simulations are empirical force fields such as OPLS-AA[1,2] (OPLS) and GAFF.[3] These force fields employ relatively simple functional forms such as point charges and Lennard-Jones (LJ) potentials with empirically fitted parameters. The force field parameters often need to be refined using macroscopic condensed phase experimental data, and they cannot accurately describe the microscopic details (i.e., polarization, charge penetration, and many-body dispersion) of the potential energy surface (PES). As a result, they fall short of providing a reliable prediction of experimental phenomena at the atomic level, especially when they are applied to new molecules blindly.

Ab initio molecular dynamics (AIMD), which calculates quantum−mechanical atomic forces on-the-fly at each step, circumvents the force field problem. However, most bulk AIMD simulations rely on density functional theory (DFT).[4] With more density functionals being developed nowadays, DFT is becoming more and more reliable. However, choosing the proper density functional for a specific scenario can still be tricky, and thorough benchmarks are often required to ensure sufficient accuracy, especially for the molecular systems. The behaviors of organic molecules (especially polymers) rely on a delicate balance between different interactions (e.g., bonding vs nonbonding, hydrogen bonding vs dispersion, etc.). In many

Received: September 22, 2023
Revised: November 28, 2023
Accepted: November 28, 2023
Published: December 20, 2023

https://doi.org/10.1021/acs.jctc.3c01045
J. Chem. Theory Comput. 2024, 20, 253−265

situations, more precise correlated wave function (CW) methods like Møller−Plesset perturbation (MP) or coupled cluster (CC) are necessary. Unfortunately, CW methods typically scale poorly with respect to system size: they can be quite affordable for small clusters with tens of atoms but prohibitively expensive in a larger box. While DFT-AIMD is already computationally expensive, CW-AIMD becomes virtually impossible for most nontrivial bulk systems.[5]

Recently, machine learning (ML) models, such as BPNN,[6,7] EANN,[8] DeePMD,[9] sGDML,[10] TensorMol,[11] Nequip,[12] etc., have emerged as powerful tools for the construction of PESs, making significant advances toward a better balance between accuracy and simulation speed. However, two challenges remain: (1) a large number of data points are required to train a robust ML model, and data for "bulk-like" geometries are necessary to train a bulk PES. This prevents the application of accurate CW methods in the training of the ML PES: to date, they are still too expensive for massive bulk calculations. (2) It is still a common practice to encode local atomic environments and use them as input for the ML model, which means that the ML PES does not explicitly describe the long-range interactions, which could be important for heterogeneous molecular systems.

In response to these challenges, a plethora of studies have endeavored to address the long-range interaction issue through various methodologies.[13] Following the philosophy of range separation, TensorMol[11] and GEBF-ML[14] choose to separate the conventional Coulomb and LJ terms from the total DFT energies and refine the residual energies using the ML approach. In a similar vein, PhysNet[15] and SpookyNet[16] tackle the issue by incorporating a point charge model and an empirical dispersion correction to describe the long-range interaction. Meanwhile, the SchNet-vdW[17] method delved into segregating long-range van der Waals (vdW) interactions. Building upon this, SCFNN[18] utilizes two self-consistent models to describe long-range polarization: one model is used to predict the short-range Coulomb energy and a separate model is used to capture the linear response of the short-range model caused by the long-range electric field. In a different approach, 4G-HDNNP[19] models the long-range polarization via a global charge equilibration strategy with environment-dependent atomic electronegativities predicted by local neural networks. Additionally, the DPLR[20] model predicts the locations of Wannier centers and uses Wannier centers to compute long-range electrostatic interactions. Remove: It also accommodates the intermediate-range vdW interactions effectively by fitting to the SCAN or SCAN0 DFT data.

Nonetheless, in the majority of existing studies, the details of electrostatic and dispersion interactions are described at the point charge and C6 level, with higher-order multipole contributions being almost universally neglected. They are usually derived from bulk-like DFT training data, which are often not accurate enough and not easy to obtain for large heterogeneous molecular systems. The ML methods still face great challenges in the description of weak nonbonding interactions and the generation of high-quality training data. Therefore, although tremendous progress has been made in the development of ML potential, a scalable and chemically accurate force field for general bulk organic molecular simulations is still missing. In this work, we aim to build a workflow for the construction of such a force field that exploits the full advantages of both physical and ML models.

Besides the ML methods, another approach also uses the range separation strategy but relies on many-body expansion (MBE) to describe the short-range interactions. The representative model is MB-pol, which achieves remarkable success in both water clusters and bulk systems.[21,22] More recently, it has been transferred to simple polymers such as alkanes[23] and polyhydrocarbons.[23,24] However, further validation is needed to show the capability of MBE in more complex bulk organic molecules. When dealing with a heterogeneous molecular system, how to determine the interacting "bodies" in a way that promotes faster convergence of MBE is still an open question.

Previously, some of us showed that accurate physics-driven intermolecular potentials can be built based on the perturbation theory.[25−29] The long-range atomic parameters (i.e., multipoles, dispersion coefficients, polarizabilities, etc.) can be obtained using monomer DFT and time-dependent DFT (TD-DFT) calculations and localization techniques such as iterative stockholder analysis (ISA).[30,31] The medium- and short-range parameters can be then fitted term-by-term, according to the physically meaningful energy decomposition given by the symmetry-adapted perturbation theory (SAPT).[32−37] The procedure was then further refined by van Vleet et al. with more advanced short-range terms,[38] which gives a better description of charge penetration effects. This methodology was shown to be highly effective in small molecule fluids,[25] ionic liquids,[39] and rigid organic porous materials.[40] Compared to other studies, this methodology possesses great advantages in computational efficiency and transferability as all parameters are physically meaningful and most of them are directly computed instead of fitted. However, two problems remain before it can be used as a general organic force field: 1. Its accuracy in strong interacting systems (e.g., hydrogen bonds) is still limited and 2. it does not tackle the intramolecular interaction, which is important for large molecules and polymers. Recently, using different ML techniques, we have addressed these two issues separately: using a range-separated ML model,[41] we show that small molecules with strong hydrogen bonds (i.e., water) can be described with extraordinary accuracy. Meanwhile, a subgraph neural network (sGNN)[42] model has been shown to give a highly scalable description of the intramolecular interaction of polymers with weak nonbonding interactions. In this work, two techniques are combined: by exploiting a clean separation of both long- and short-range interactions and nonbonding−bonding interactions, we formulate a general workflow for organic molecule force field construction, named PhyNEO (physics-driven force field with neural network enhancement for organic molecules). In PhyNEO, we use ab initio data of small clusters with 10−25 heavy atoms, which only include the closest neighbors of an atom, to train the potential. Without any further empirical refinement, the PhyNEO potential accurately predicts the bulk properties of the polymer systems with and without hydrogen bonds. It is thus demonstrated that PhyNEO warrants a promising solution to solve the problems that remain in molecular force field development.

2. METHODS AND COMPUTATIONAL DETAILS

2.1. PhyNEO Workflow. In this work, we focus on the development of PhyNEO and demonstrate its robustness, effectiveness, and transferability in handling both non-hydrogen-bonding and hydrogen-bonding systems. We will use polyethylene glycol (PEG) and poly(ethylene oxide) (PEO)

https://doi.org/10.1021/acs.jctc.3c01045
J. Chem. Theory Comput. 2024, 20, 253−265

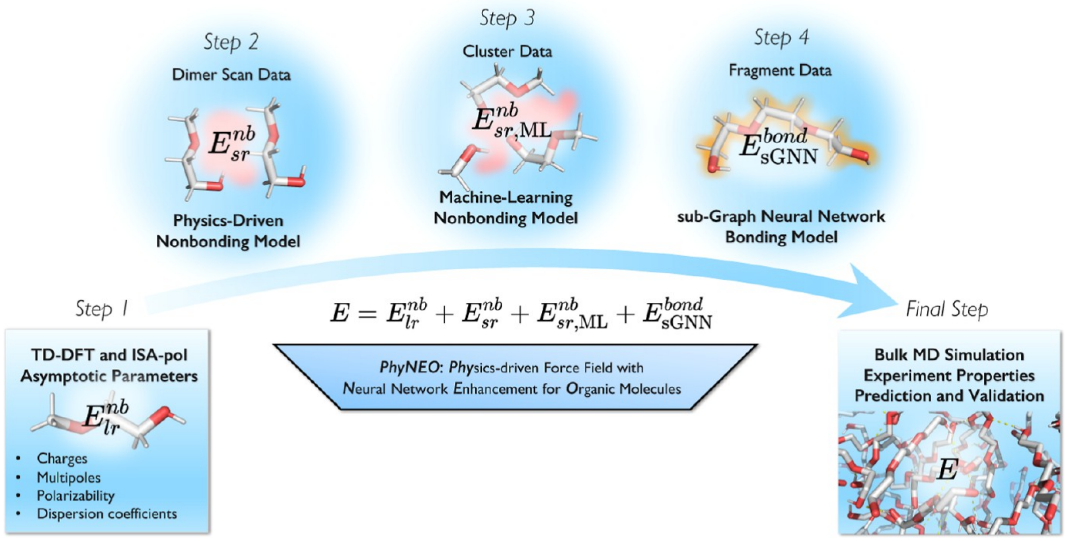


Figure 1. Schematic depiction of the PhyNEO force field development workflow. The whole procedure is composed of four major steps: (1) long-range asymptotic parameter calculations; (2) physics-driven short-range nonbonding model development; (3) short-range ML nonbonding model fitting; and (4) sGNN bonding model fitting. Once the PhyNEO force field is constructed, it can be validated against experimental properties through bulk MD simulations. The data needed in each step are also labeled in the scheme.

polymer chains as proof-of-concept examples. Throughout this paper, PEO and PEG denote polymer chains with chemical formulas $H_3C{-}O{-}(CH_2{-}CH_2{-}O)_n{-}CH_3$ and $HO{-}(CH_2{-}CH_2{-}O)_n{-}H$, respectively. PEG features terminal hydroxyl groups (−OH) with hydrogen bonds, while PEO does not. To simplify the notation, the PEG and PEO chains of different lengths are denoted as PEG[n] and PEO[n], where n represents the number of repeating "$CH_2{-}CH_2{-}O$" units.

The workflow to build a PhyNEO potential is schematically depicted in Figure 1. The workflow involves the utilization of three distinct models: the physics-motivated nonbonding model ($E^{nb}_{lr} + E^{nb}_{sr}$), ML short-range nonbonding correction ($E^{nb}_{sr,ML}$), and the sGNN bonding model (E^{bond}_{sGNN}). The total energy could be computed using the following equation

$$E = E^{nb} + E^{bond}_{sGNN} \tag{1}$$

$$E^{nb} = E^{nb}_{lr} + E^{nb}_{sr} + E^{nb}_{sr,ML} \tag{2}$$

The physics-driven nonbonding terms of PhyNEO are further separated into long-range E^{nb}_{lr} and short-range E^{nb}_{sr} contributions, as done in our and others' previous studies.[27,38,41,42] The four parameterization steps of PhyNEO are described below:

1. As shown in Figure 1, the first step of PhyNEO is to obtain the long-range terms E^{nb}_{lr}, which include the electrostatic (es), polarization (pol), and dispersion (disp) interactions

$$E^{nb}_{lr} = E^{lr}_{es} + E^{lr}_{pol} + E^{lr}_{disp} \tag{3}$$

$$E^{lr}_{es} = \sum_{i<j} \frac{q_i q_j}{r_{ij}} + \sum_{i<j} \sum_{tu} Q_t^i T_{tu} Q_u^j \tag{4}$$

$$E^{lr}_{pol} = E_{ind-dip}(\alpha_i) \tag{5}$$

$$E^{lr}_{disp} = -\sum_{i<j} \sum_{n=6,8,10} \frac{C^n_{ij}}{r^n_{ij}} \tag{6}$$

$$C^n_{ij} = \sqrt{C^n_i C^n_j} \tag{7}$$

The atomic charges and multipoles (q_i and Q^i_t) are computed at the DFT level and are distributed to atoms via ISA. The charge density susceptibility matrices at zero and imaginary frequencies are computed at the TD-DFT level, and then distributed using the ISA-pol[31] method. Atomic polarizabilities (α_i) and dispersion coefficients (C^{2n}_i) can then be obtained via multipole expansion and the Casimir−Polder relation. All multipoles are truncated at the quadrupole level, dispersions are truncated at the C10 level, and isotropic induced point dipoles [$E_{ind-dip}(\alpha_i)$] are used to describe the polarization energy. For polarization [$E_{ind-dip}(\alpha_i)$], we use a thole damping algorithm that is identical to the MPID[43] model used in polarizable CHARMM.

For typical closed-shell organic molecules, these asymptotic parameters can be computed using relatively small fragments [i.e., 2-methoxyethanol (SMILES: COCCO) and dimethoxyethane (SMILES: COCCOC) for PEG and PEO, respectively] and is highly transferable to larger molecules. It is noted that all of these "asymptotic parameters" are in principle conformation-dependent. But in this work, we use values averaged over different conformations for simplicity, which perform reasonably well for PEG/PEO. The procedure to obtain the asymptotic nonbonding parameters is well-established in numerous previous studies.[25−29,31,40]

2. Once the asymptotic parameters are determined, we move forward to train the short-range charge penetration terms and the damping functions (E^{nb}_{sr}). For this part, we adopt the physically decomposed Slate-type functions proposed by van Vleet et al.[38]

$$E^{nb}_{sr} = E_{ex} + E^{sr}_{es} + E^{sr}_{disp} + E^{sr}_{pol} + E_{dhf} \tag{8}$$

$$E_{ex} = \sum_{i<j} A^{ex}_{ij} P(B_{ij}, r_{ij}) \exp(-B_{ij} r_{ij}) \tag{9}$$

https://doi.org/10.1021/acs.jctc.3c01045
J. Chem. Theory Comput. 2024, 20, 253−265

$$E_{es}^{sr} = \sum_{i<j} \left\{ -A_{ij}^{es} P(B_{ij}, r_{ij}) \exp(-B_{ij} r_{ij}) + [f_1(B_{ij} r_{ij}) - 1] \frac{q_i q_j}{r_{ij}} \right\} \tag{10}$$

$$E_{disp}^{sr} = \sum_{i<j} \left\{ -A_{ij}^{disp} P(B_{ij}, r_{ij}) \exp(-B_{ij} r_{ij}) + \sum_{n=6,8,10} [1 - f_n(x)] \frac{C_{ij}^n}{r_{ij}^n} \right\} \tag{11}$$

$$E_{pol}^{sr} = \sum_{i<j} -A_{ij}^{pol} P(B_{ij}, r_{ij}) \exp(-B_{ij} r_{ij}) \tag{12}$$

$$E_{dhf} = \sum_{i<j} -A_{ij}^{dhf} P(B_{ij}, r_{ij}) \exp(-B_{ij} r_{ij}) \tag{13}$$

$$f_n(x) = 1 - e^{-x} \sum_{i=0}^{n} \frac{x^k}{k!} \tag{14}$$

$$P(B_{ij}, r_{ij}) = \frac{1}{3}(B_{ij} r_{ij})^2 + B_{ij} r_{ij} + 1 \tag{15}$$

$$x_{ij} = B_{ij} r_{ij} - \frac{2B_{ij}^2 r_{ij} + 3B_{ij}}{B_{ij}^2 r_{ij}^2 + 3B_{ij} r_{ij} + 3} r_{ij} \tag{16}$$

$$A_{ij} = A_i A_j \tag{17}$$

$$B_{ij} = \sqrt{B_i B_j} \tag{18}$$

The damping exponents (B_i) are shared among all of the energy components. Instead of deducing B_i from the tails of the ISA weighting functions, we train all short-range parameters (i.e., A_i and B_i) directly to dimer interaction energy data computed using DFT-SAPT (symmetry-adapted perturbation theory based on DFT). For typical organic molecules, DFT-SAPT provides nonbonding interaction energies at the CCSD(T) level of accuracy and also offers a physically meaningful energy decomposition. Exploiting this advantage, the target function is designed to include the losses of all components, which greatly enhance the transferability of the final parameters

$$\begin{aligned} L_1 = {} & \lambda_{tot} \sum (E_{tot} - E_{tot}^{ref})^2 + \lambda_{ex} \sum (E_{ex} - E_{ex}^{ref})^2 \\ & + \lambda_{es} \sum (E_{es}^{lr} + E_{es}^{sr} - E_{es}^{ref})^2 \\ & + \lambda_{ind} \sum (E_{pol}^{lr} + E_{pol}^{sr} - E_{pol}^{ref})^2 \\ & + \lambda_{disp} \sum (E_{disp}^{lr} + E_{disp}^{sr} - E_{disp}^{ref})^2 \\ & + \lambda_{dhf} \sum (E_{dhf} - E_{dhf}^{ref})^2 \end{aligned} \tag{19}$$

where λ is the weight of each component. To prioritize the fidelity of the total energy, we set λ_{tot} to 1.0 while assigning a value of 0.1 to all other components. Empirically, such weight assignment leads to satisfactory fitting performances for all components (see the details in the Supporting Information).

The data required in this step are generated via fragment dimer scans. Dimer configurations are first extracted from bulk OPLS-AA MD simulations and scanned along the direction that connects the centers of mass of the two molecules. The separation between the two molecules spans a range of 1.4−6.2 Å, effectively conducting a comprehensive sampling over both short- and medium-range distances. We found that stochastic optimizers developed to train neural networks, such as ADAM,[44] are also highly efficient to train the nonlinearly coupled short-range parameters, where conventional optimizers often fail due to overfitting. Therefore, we adopt the ADAM optimizer in this work, and each dimer scan, which is typically composed of 4−12 single point calculations at different separations, is fed in as a minibatch. In the previous work, we often trained the model using only homodimer (dimers composed of the same types of molecules) data and completely rely on the combination rule to describe heterodimers (dimers composed of different types of molecules). However, here, we find that the addition of heterodimer data is important to generate transferable parameters, possibly due to the fact that we are simply fitting, instead of computing these parameters; thus, more data are needed. In principle, the former approach features a computational cost of $O(N)$, with N being the number of molecule types, while the latter approach scales as $O(N^2)$. In reality, one does not need to go over all possible heterodimers, so the $O(N^2)$ scale can be largely reduced. Furthermore, the error introduced in this step can be suppressed in the next step, where an ML model is introduced to fit the residual error. Nevertheless, for such small training molecules, the DFT-SAPT calculation is relatively cheap and can be done on a massive scale; thus, we adopt the $O(N^2)$ approach here for simplicity.

3. The pairwise additive isotropic short-range terms trained in the last step provide a basic approximation for the repulsive wall between atoms. It greatly enhances the numerical stability of the simulation by preventing the atoms from entering unphysically short distances. However, the short-range nonbonding interaction can be complicated by other physics such as charge transfer and many-body exchange/dispersion.[45] This leads to significant anisotropy[46] and many-body characters to the PES, which is difficult to capture using conventional approaches. Such effects could be particularly important in hydrogen bonds,[29] which are ubiquitous in biological and chemical systems. Following the protocol we developed to describe bulk water,[41] we introduce an additional short-range ML correction ($E_{sr,ML}^{nb}$) fitted to small cluster intermolecular energies computed at the MP2 level. The embedded atom neural network (EANN)[8] is chosen to fit $E_{sr,ML}^{nb}$ due to its excellent computational efficiency and its ability to avoid severe overfitting with limited data. We note that due to the carefully refined physical potential, the additional ML correction is extremely localized compared to the naïve ML models. Therefore, we can train this part using tiny clusters with less than 25 heavy atoms, allowing the use of high-quality training data generated by accurate CW methods. For simplicity, MP2 is used in this demo (which is accurate enough for PEG/PEO), but higher-level methods such as DLPNO−CCSD(T)[47] can certainly be used too. The target function to be optimized in this step is written as

$$L_2 = \sum (E_{lr}^{nb} + E_{sr}^{nb} + \Delta E_{sr,ML}^{nb} - \Delta E_{MP2}^{nb})^2 \tag{20}$$

and

$$\Delta E_{MP2}^{nb} = E_{MP2}(\text{cluster}) - \sum_A E_{MP2}(A) \tag{21}$$

https://doi.org/10.1021/acs.jctc.3c01045
J. Chem. Theory Comput. 2024, 20, 253−265

$$\Delta E^{nb}_{sr,ML} = E^{nb}_{sr,ML}(\text{cluster}) - \sum_A E^{nb}_{sr,ML}(A) \quad (22)$$

where A runs over all fragments in the cluster. In this work, we deal with flexible molecules with significant conformation fluctuations. Following the logic of our previous work,[42] we intend to separate the nonbonding and bonding interactions and tackle the latter one using an internal-coordinate-based GNN model, which is much more efficient. Therefore, in this step, the training target is the intermolecular MP2 energy (ΔEMP2nb) instead of the total MP2 energy (E_{MP2}).

Note that in previous studies,[41] one often trains a ML model ($E^{nb}_{sr,ML}$) to fit ΔEMP2nb directly, corresponding to the following target function

$$L_2' = \sum (E^{nb}_{lr} + E^{nb}_{sr} + E^{nb}_{sr,ML} - \Delta E^{nb}_{MP2})^2 \quad (23)$$

The key difference between L_2 and L_2' is the cancellation of intramolecular nonbonding energy, which is not presented in ΔEMP2nb and thus should not be included in $\Delta E^{nb}_{sr,ML}$ either. For small rigid molecules with relatively fixed intramolecular nonbonding interactions, the difference between L_2 and L_2' is trivial and often ignored, but it can be important for larger flexible molecules. To generate the cluster training data, we first sample the bulk COCCOC and COCCO molecules using the OPLS-AA force field and then carve out small clusters from the bulk. Each cluster is based on one central atom and all atoms within a 4 Å cutoff around the central atom. The dangling covalent bonds are then capped to form complete molecules, and the detailed capping scheme is described in the Methods and Computational Details section. In general, the clusters are formed by either full COCCOC and COCCO molecules or smaller fragments including C, COC, CO, and COC. The long-range parameters of all the small fragments are also determined using the same approach as that described above, and a short-range ML correction $E^{nb}_{sr,ML}$ is trained to fit the residual error.

4. Once we determine the nonbonding terms ($E^{nb}_{lr} + E^{nb}_{sr} + E^{nb}_{sr,ML}$), we adopt the procedure described in our previous paper[42] to train the sGNN bonding energy (E^{bond}_{sGNN}). The training data are generated using small molecules including PEG[3] and PEO[3], with conformation variations sampled using MD at the OPLS-AA level. The total energies of these short-chain molecules are computed at the MP2/aug-cc-pVTZ level of theory, and the nonbonding terms are computed and separated from the total energy. The residual conformation energies are considered to be pure bonding interactions and are tackled using sGNN. The sGNN method provides a quite scalable framework for the development of bonding energies, given that an accurate nonbonding potential is provided. Unlike other neural network models, sGNN takes internal coordinates (IC) as input, which is a more natural coordinate for bonding energy than Cartesian distances. The topological distance cutoff for the sGNN model was set to be two bonds, which effectively incorporates all interactions within 5 bonds. Correspondingly, all nonbonding interactions are turned off within 5 bonds. Furthermore, in addition to our previous work,[42] we incorporate layer normalization in each layer of sGNN in this work, enhancing the network's fitting efficiency and capability.

Once the last step is finished, we obtain a complete PhyNEO potential ready to be used in bulk simulations of molecules of any size. In the entire process, PhyNEO requires only high-quality ab initio data from small clusters, rather than bulk structures. The computational costs for the first and fourth steps are trivial as they only need single molecule data, while the third step is the most computationally intensive step that is likely to be the bottleneck of a general workflow. Nevertheless, since only small cluster data are required, the PhyNEO workflow already represents a significant advancement compared to the current state-of-the-art ML potential workflow.

2.2. Computational Details. *2.2.1. Reference Data Generation.* In different steps of the PhyNEO workflow, we use different types of ab initio data, which we introduce in more detail below, and the data set composition (Table S1) is provided as part of the Supporting Information. In general, the structures of the training data were all generated using bulk MD simulations with the OPLS-AA[1,2] force field, and the OPLS-AA parameters were generated by LigParGen.[48] The sampling simulations were executed using a Langevin thermostat and a Monte Carlo barostat, utilizing a time step of 1 fs at conditions of (300 K, 1 bar) or (600 K, 1 bar). The cutoff distance implemented was 10 Å, and all sampling simulations were performed using the OpenMM-7.7[49] program. Once the structure is sampled, the ab initio reference energies were calculated using the MOLPRO[50] program at either the DFT-SAPT or MP2 level, depending on the type of data needed in each step. We introduce the technical details of the ab initio calculation in each step below.

2.2.2. Parameterization of the Physics-Driven Nonbonding Terms. The long-range asymptotic parameters, including charges, multipoles (up to quadrupole), and dipole polarizabilities, were obtained from TD-DFT calculation and ISA (ISA-pol), and the dispersion coefficients can be computed using the Casimir−Polder relation. The procedure is well-documented in existing studies[31] and was performed using CamCASP 6.1[31,51,52] interfacing with NWChem 6.8.[53] The atomic parameters were computed using the asymptotically corrected PBE0 functional and the aug-cc-pVTZ basis set at geometries optimized at the MP2/aug-cc-pVDZ level of theory. For molecules with rotating dihedrals, multiple conformations exist, and the asymptotic parameters may vary depending on the conformations. For these molecules, we first ran OPLS MD simulations to identify the distribution peaks of each dihedral angle, which can be mapped to several representative conformations that need to be examined. For example, for COCCOC, 48 conformations can be identified according to the joint dihedral distribution, and for COCCO, there are 24. These conformations were then optimized using MP2, and the final atomic parameters were taken as averages over the results of all of these conformations.

In order to train the pairwise short-range interactions (E^{nb}_{sr}), we created a training set consisting of 600 dimers. Density fitting DFT-SAPT calculations were conducted using the aug-cc-pVTZ basis set for each dimer to obtain the decomposed interaction energies. Due to the truncation and capping scheme necessary in the ML model ($E^{nb}_{sr,ML}$) training (vide infra), smaller fragments including C and COC from COCCOC and CO and COC from COCCO also need to be included. In total, six types of dimers, including 3 homodimers and 3 heterodimers, are included for dimer scan calculations.

After the dimer scan results are obtained, the parameter for the E^{nb}_{sr} term can be trained using the ADAM optimizer, employing a learning rate of 0.01. The optimization process should be conducted for a maximum of 2000 epochs with a

https://doi.org/10.1021/acs.jctc.3c01045
J. Chem. Theory Comput. 2024, 20, 253−265

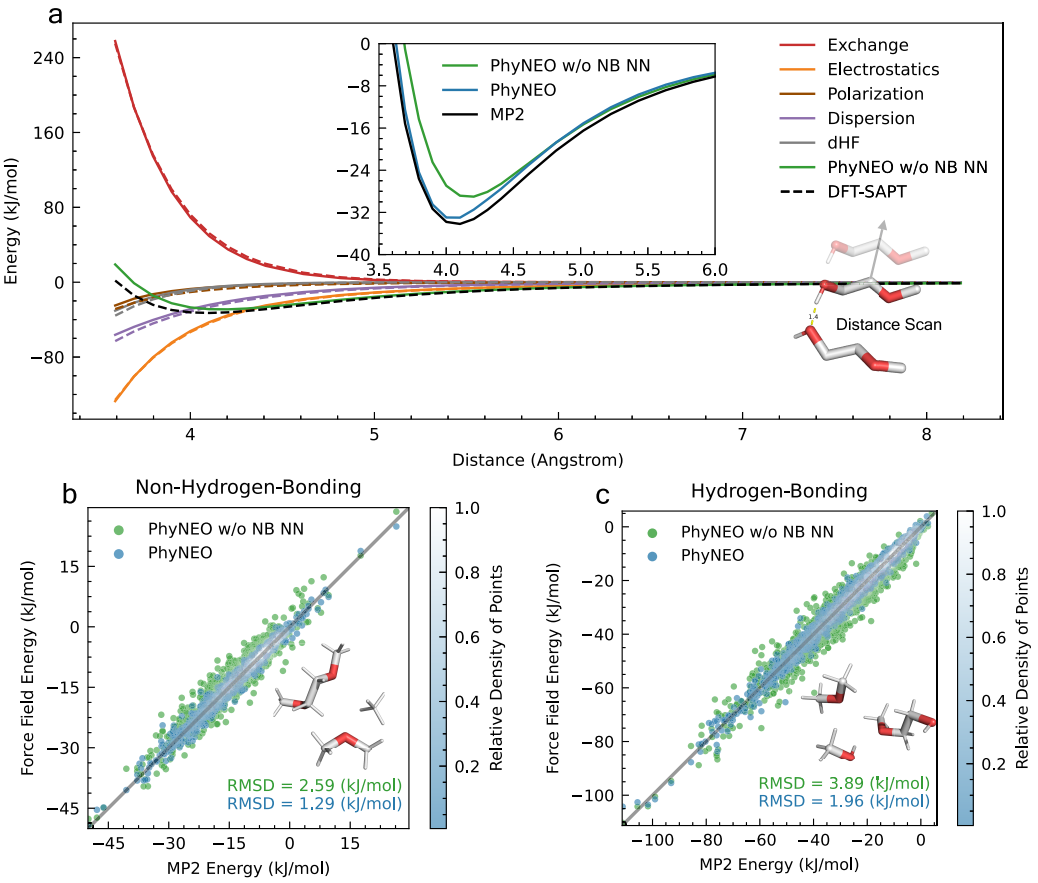


Figure 2. Validation of PhyNEO on intermolecular interactions: panel (a) illustrates the PES for a hydrogen-bonded 2-methoxyethanol (COCCO) dimer scan. The total energy is decomposed into physically meaningful components and is compared with that obtained from DFT-SAPT. The dashed lines represent the energy components calculated from DFT-SAPT, while the solid lines are the components given by PhyNEO. The total PhyNEO energy is also compared with the MP2 results and is shown in the inset of panel (a). Panel (b) shows the independent test of small cluster intermolecular energies of the PEO system, PhyNEO versus the MP2 reference. Both full PhyNEO and PhyNEO without the nonbonding ML enhancement (labeled as "PhyNEO w/o NB NN") are shown. Panel (c) presents the same comparison for the PEG systems.

batch size of 12. Training should be halted when the performance of the model ceases to improve on a held-out validation data set.

2.2.3. Parameterization of the ML Nonbonding Term. To generate the data needed in the training of $E^{nb}_{sr,ML}$, we first needed to build small clusters from bulk COCCOC and COCCO OPLS simulations. We first equilibrated both systems at a pressure of 1 bar and temperatures of 300 and 600 K, followed by 30 ns *NVT* sampling runs. Subsequently, we randomly selected a central atom and built the corresponding cluster by including all molecules within a 4 Å cutoff distance from the central atom. To further limit the size of the cluster, if part of a neighboring molecule is outside of the cutoff distance, the molecule is truncated and capped as follows: first of all, both COCCO and COCCOC molecules were divided into COC and CO fragments. Then, a fragment would be fully removed if all its heavy (non-hydrogen) atoms are outside of the cutoff, and the resulting dangling bonds are capped with H. For the COCCOC system, one additional rule was applied: if only the terminal C of the COCCOC molecule is within the cutoff, it is replaced by a methane (C) molecule. Following this scheme, we built 10384 clusters for the PEO system and 15100 clusters for the PEG system in total (see details in the Supporting Information). The intermolecular energies of these many-body clusters were then computed at MP2/aug-cc-pVTZ with counterpoise corrections, composing the many-body data set. In all cases, 90% of data sets were used for training, while the rest were used for testing. We then used the embedded atom neural network (EANN)[8] model to fit the $E^{nb}_{sr,ML}$ term. All NN parameters and hyperparameters were optimized using the ADAMW optimizer, and the learning rate decays from 0.001 to 1×10^{-6} with a 0.5 decay factor. The network contained two hidden layers with 64 neurons for each input and hidden layer. In each fitting, the maximal number of iterations was set to be 1000.

2.2.4. Parameterization of the sGNN Bonding Model. The sGNN[42] training and testing data were generated by *NVT* MD simulations using the OPLS-AA force field. PEG[3] and PEO[3] were used as training molecules, and we drew 10000 conformations at 300 K and 10000 conformations at 1000 K, leading to an ensemble of 20000 structures for each molecule. The single molecule conformation energies were then computed using the MP2/aug-cc-pVTZ method, forming the single-molecule data set. Among them, 1000 conformations from the 300 K ensemble were randomly selected to form the small molecule testing set and others were used as training data. To validate the size transferability of PhyNEO, we further generated 1000 conformations for larger molecules (i.e., PEG[7] and PEO[7]) from 300 K *NVT* simulations. The sGNN model contained two hidden layers with 20 neurons in each input and hidden layer. The ADAM optimizer was employed to train the sGNN model, and the learning rate was

https://doi.org/10.1021/acs.jctc.3c01045
J. Chem. Theory Comput. 2024, 20, 253−265

set to be 0.0001. We ran the optimizations for 4000 epochs at maximum. Before training, the target energies were shifted and scaled to a standard normal distribution.

2.2.5. Molecular Dynamics Simulation. All fittings in this work were conducted using the DMFF differentiable molecular force field platform,[54] and we also used it to perform force and virial calculations in MD simulations. A highly integrated interface of DMFF has been implemented in the i-PI[55] package to run PhyNEO MD. We computed the virial matrix under periodic boundary conditions based on Thompson's paper[56] and validated the correctness of the *NPT* implementation by comparing the GAFF simulation results with OpenMM-7.7[49] (see details in the Supporting Information). All PhyNEO MD simulations were performed using the Bussi−Zykova−Parrinello barostat and Langevin thermostat with a time step of 0.5 fs. The total simulation time was 200 ps to converge the density and radial distribution function (RDF) results. The simulation temperature was set to be 293.15 K for PEO and 300 K for PEG, consistent with the experimental conditions. OPLS and GAFF simulations were performed under the same conditions.

3. RESULTS AND DISCUSSION

3.1. Accuracy in a Microscopic PES. We first examine the accuracy of PhyNEO in the microscopic details of the PES, in which empirical force fields often perform quite poorly.

First, we investigate the intermolecular interactions of the dimer and small cluster, the results of which are shown in Figure 2, with the corresponding RMSD (root-mean-square deviation) given in each panel.

We pick a hydrogen-bonded dimer for the dimer scan test, in which the physics-driven parts E^{nb}_{lr} + E^{nb}_{sr} of energies are decomposed into ex, es, pol, disp, and dhf components, with each component including both the corresponding long- and short-range contributions. The results are compared to the corresponding DFT-SAPT results term by term and are shown in Figure 2a. The overall agreement is excellent in the long range beyond 4−5 Å but gradually deteriorates in the short range as the charge penetration effects start to manifest. The short-range error primarily comes from the pol, disp, and dHF components, while the Slater-type function captures the short-range ex and es interactions with good accuracy. The total physics-driven energies are shown in the inset of Figure 2a and are compared with MP2 results. Apparently, pure physics-driven potential (labeled as "PhyNEO w/o NB NN" in the figure) is still inaccurate in the hydrogen-bonded region. Meanwhile, a localized ML nonbonding correction $\Delta E^{nb}_{sr,ML}$ fitted to small cluster data largely improves the performance, demonstrating its outstanding fitting capability in the short range. A similar trend can also be seen in the cluster tests (Figure 2b for the PEO system and Figure 2c for the PEG system): the ML nonbonding correction reduces the error of the intermolecular interaction by 50%, in both hydrogen-bonded and non-hydrogen-bonded cases. Meanwhile, we also note that the physics-driven short-range interaction E^{nb}_{sr} is indispensable, which offers a basic repulsion wall that greatly enhances the numerical stability of the MD simulation. Also, the accurate long-range physical potential largely lowers the training cost and enhances the transferability of the ML model. Therefore, the robustness and the transferability of the physics-driven model and the fitting capability of the ML model are combined in PhyNEO, due to a proper range separation scheme.

The introduction of $E^{nb}_{sr,ML}$ improves the quality of the nonbonding potential and thus further ensures a cleaner separation between the bonding and nonbonding energies. Therefore, it also leads to a more transferable intramolecular sGNN potential than that in our previous work.[42] To test the validity of PhyNEO in intramolecular interactions, we train the sGNN model using small fragments (PEG[3] and PEO[3]), and we examine its accuracy in longer and more flexible molecules (PEG[7] and PEO[7]). Figure 3 demonstrates that

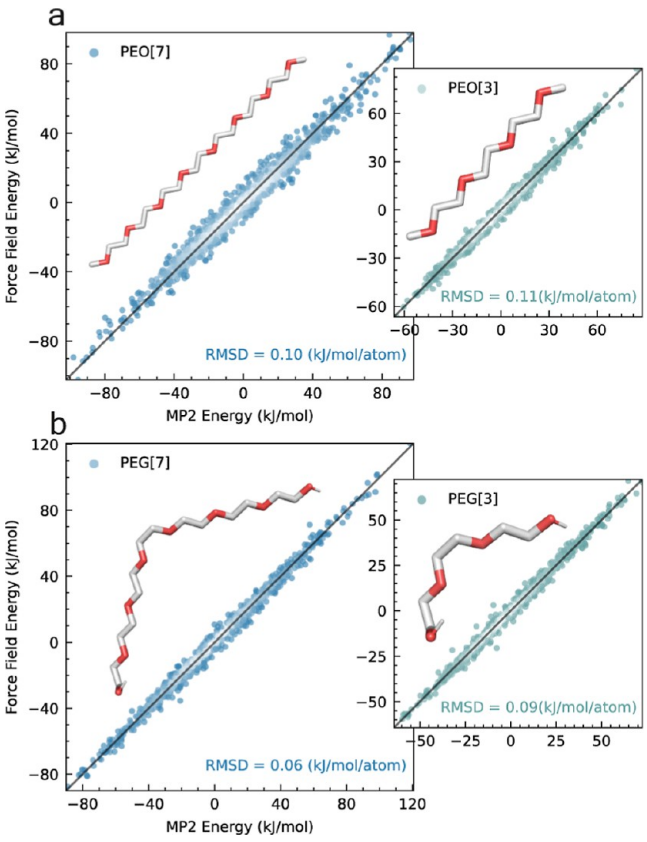


Figure 3. Validation of PhyNEO on intramolecular interactions: panel (a) shows a comparison of the conformation energies of PEO, computed by both PhyNEO and MP2. The comparison is done for both PEO[3] and PEO[7] to examine the transferability of the model in molecules of different sizes. Panel (b) displays the same comparison for PEG.

in both PEG (0.09 kJ/mol/atom for PEG[3] versus 0.06 kJ/mol/atom for PEG[7]) and PEO (0.09 kJ/mol/atom for PEO[3] versus 0.09 kJ/mol/atom for PEO[7]) systems, the error of sGNN remains stable with respect to increasing chain length, showing excellent scalability between molecules of different sizes. Again, this observation highlights the extrapolation capability of sGNN when nonbonding and bonding interactions are well-separated.

While we have shown that PhyNEO performs well in both inter- and intramolecular interactions, it is the delicate balance between these two terms that determines the behaviors of polymers. To examine the accuracy of the final PhyNEO potential, we further investigate the total energies of hydrogen-bonded complexes of PEG[3] and PEG[7], including both inter- and intramolecular contributions. In total, 1000 PEG[3] and 1000 PEG[7] complexes were extracted from 300 K NVT MD simulations using the OPLS-AA force field, and the total

https://doi.org/10.1021/acs.jctc.3c01045
J. Chem. Theory Comput. 2024, 20, 253−265

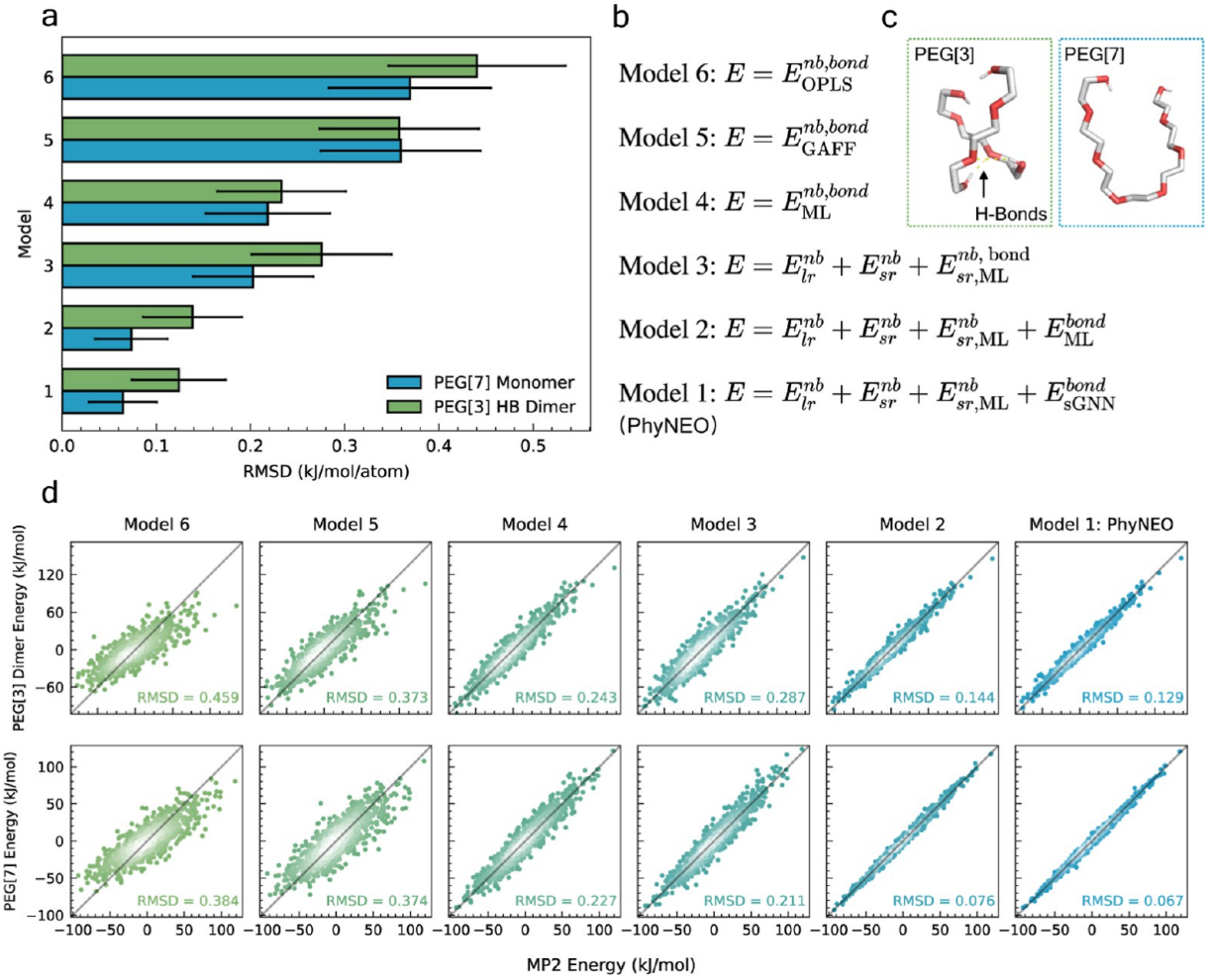


Figure 4. Validation on the total energies of (1) hydrogen-bonded PEG[3] clusters and (2) PEG[7] monomers. Total energies computed by using MP2 are compared to those calculated using different models with different levels of accuracy. Panel (a) shows the performances of different models in the two test sets; panel (b) shows the definition equations of different models. Panel (c) shows the typical cluster structures in the two test sets. The energy comparison and the RMSD of each model are plotted and labeled in the corresponding subgraph in panel (d).

energies were computed at the MP2/aug-cc-pVTZ level of theory for reference. These PEG[3] hydrogen-bonded configurations involve both inter- and intramolecular hydrogen bonds, forming a quite challenging test set for conventional force fields. The results are shown in Figure 4. Comparing to the conventional force field including both OPLS-AA (model 6) and GAFF (model 5), the error of PhyNEO decreases by 67.5−73.9% respectively, showing its great superiority.

To further demonstrate the importance of short-/long-range and bonding/nonbonding separations in PhyNEO, we build three more models, labeled as "model 2−5" in Figure 4. In model 2, we use EANN, instead of sGNN, to fit the bonding energies (i.e., using two EANN models to fit the bonding and nonbonding energies separately). Comparing to PhyNEO (model 1), model 2 shows slightly worse fitting quality, which can be more significant in more distorted geometries sampled at higher temperatures (see Figure S5). The advantages of sGNN over EANN in describing bonding energies are possibly due to two reasons: (1) the nonbonding interactions between bonded atoms are excluded based on topological relations, instead of Cartesian distances (nonbonding interactions within 5 bonds are excluded in here). Correspondingly, the bonding-topology-based sGNN model is more compatible with such nonbonding exclusion scheme, comparing to the Cartesian-coordinate-based EANN model. (2) The sGNN model uses ICs as inputs, which is probably a more efficient descriptor for bonding energies than Cartesian coordinates, as similar phenomena were also found in a previous work.[57] In model 3, we combine the short-range bonding and nonbonding energies and fit them using a single EANN model. The results are worse than those of both model 2 and PhyNEO, clearly demonstrating the importance of bonding/nonbonding separation. In model 4, we further abandon the range separation scheme and fit the total energies using a single ML model. The resulting pure ML model, trained using only small cluster data, is not only less accurate in both training and testing data sets but also unstable in MD simulations.

In addition, we compare the computational efficiency between model 1−4, as detailed in the Supporting Information. In the current implementation, the computational bottlenecks in all four models are neighbor list construction and the multipolar polarizable particle mesh Ewald (PME) calculation, while the ML parts are relatively cheap. However, we do note that this result can be attributed to the inefficient implementation of the neighbor list and PME algorithm in DMFF and may change after future code optimization. Nevertheless, for the ML parts, the "single-engine" model 3 runs faster than the "dual-engine" model 1/2, as expected.

https://doi.org/10.1021/acs.jctc.3c01045
J. Chem. Theory Comput. 2024, 20, 253−265

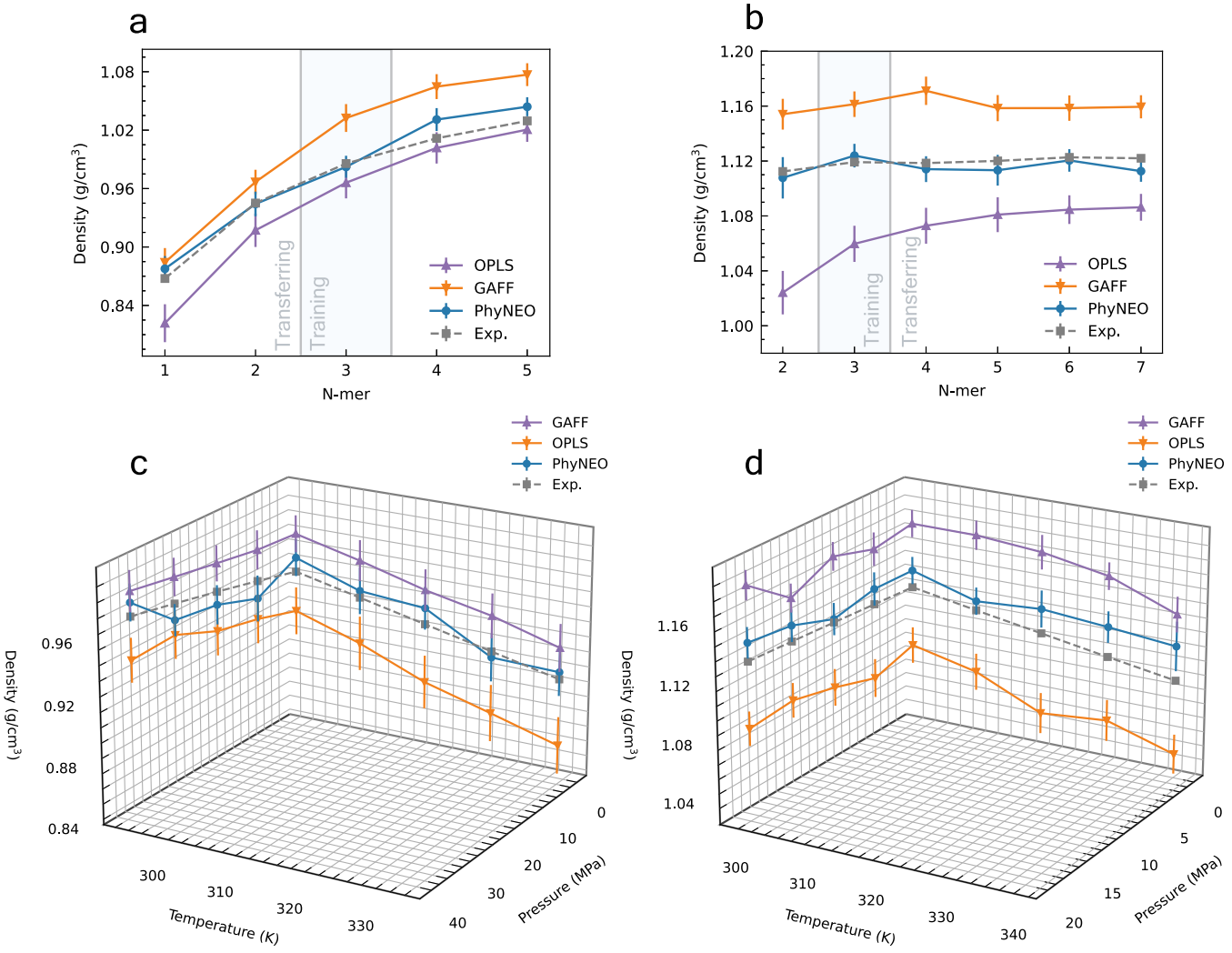


Figure 5. Densities predicted by OPLS, GAFF, and PhyNEO, in comparison with experimental results. Panel (a) shows the density comparison of non-hydrogen-bonded systems, ranging from PEO[1] to PEO[5]. Panel (b) shows the density comparison of hydrogen-bonded systems, ranging from PEG[2] to PEG[7]. In both panels, the shaded areas represent PEG[3] and PEO[3], which are used to train PhyNEO. The other areas are labeled as "transferring" areas as no training data are generated from these areas. Panels (c) and (d) show the temperature/pressure dependencies of PEO[3] and PEG[4] (PEG200) densities, respectively.

Meanwhile, the sGNN-based PhyNEO potential, with slightly higher accuracy, runs faster than that of the EANN-based model 2. This again indicates the higher efficiency of the sGNN model in the description of the bonding energy.

Through these comparisons, we clearly show the keys to the success of PhyNEO: the physically meaningful separations of different energy components and the usage of different physical-driven and ML models to tackle them separately. Such a comparison directly proves the importance of incorporating physical understandings in ML models, which can improve the performance of the final force field.

Also, as we show in the next section, these results are much more profound than just a better fitting: due to the clean separation of short-/long-range and bonding/nonbonding interactions, the improvement brought by EANN and sGNN can be reliably transferred to larger molecules and bulk systems, from which no training data are required. Such transferability is the true strength of PhyNEO, which warrants it as a force field with strong predictive power for bulk organic molecules.

3.2. Performance in Bulk Simulations. It is well-known that a better fitting to cluster energies does not necessarily indicate a better performance in bulk simulations. Conventional force fields often adopt a top-down protocol, fine-tuning the LJ parameters to reproduce experimental properties such as densities and evaporation enthalpies. The top-down protocol can deteriorate the microscopic accuracy of the potential, leading to the question of how trustworthy the MD simulation is at the atomic level. Here, we show that PhyNEO not only outperforms conventional force fields in energy fitting but also gives better predictions for bulk properties. We choose the densities of PEG and PEO oligomers of different lengths and the temperature/pressure dependencies of PEO[3] and PEG[4] (PEG200) densities as our benchmark, the experimental data of which are well-documented in the literature.[58−62] The densities are computed using both conventional force fields and PhyNEO, and the results are presented in Figure 5. Since density is typically included as a top-down fitting target, it is reasonable to see that both OPLS-AA and GAFF force fields can capture the PEO and PEG densities within 3−6% of error. Meanwhile, the densities

https://doi.org/10.1021/acs.jctc.3c01045
J. Chem. Theory Comput. 2024, 20, 253−265

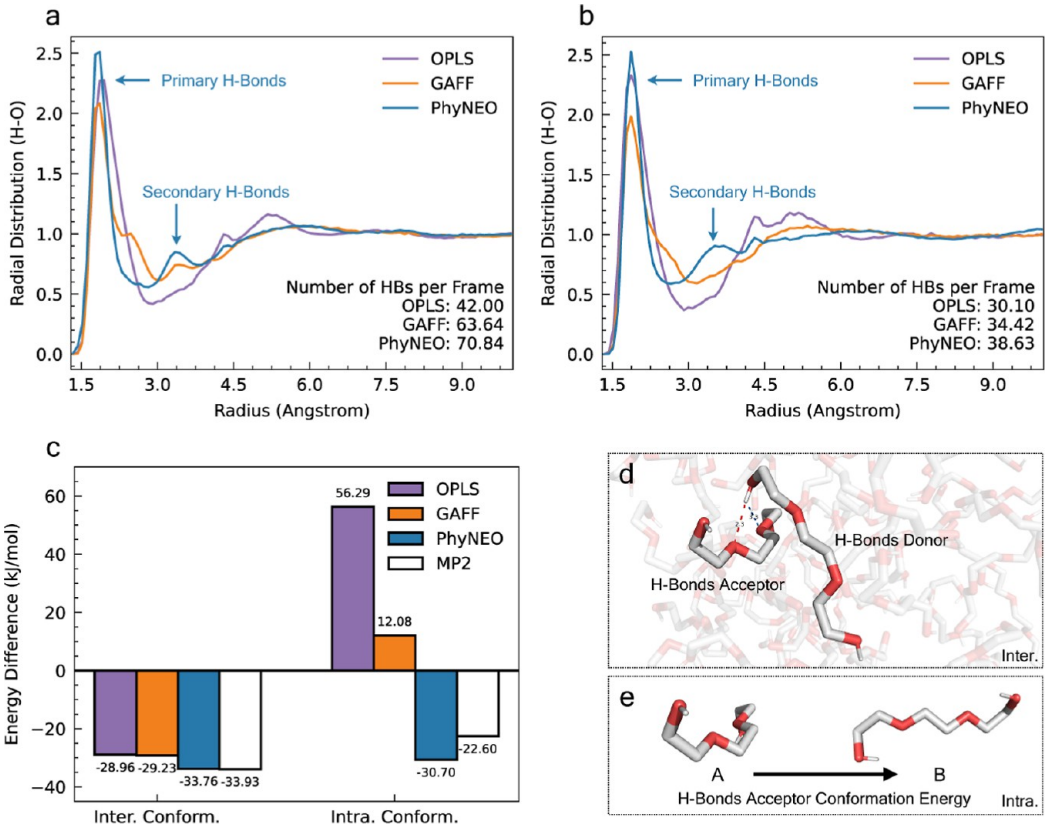


Figure 6. H−O RDFs computed by the OPLS, GAFF, and PhyNEO models. Panel (a) and (b) show the RDFs of PEG[3] and PEG[5] systems, respectively. A typical dimer structure involving both primary and secondary hydrogen bonds is selected, and its energies (including both inter- and intramolecular components) are shown in panel (c), in comparison with MP2. Panel (d) shows the structure of this dimer. Panel (c) shows how the intramolecular conformation energy is defined (using the H-bond acceptor as an example) and the reference conformation used in intramolecular energy calculation.

predicted by PhyNEO show even more superior agreement with the experimental results, with errors below and around 1% for all PEG and PEO systems. Even though PhyNEO (especially its intramolecular part) was primarily trained using PEG[3] and PEO[3], it is remarkable to see that it performs well in both longer chains (PEG[7] and PEO[5]) and shorter chains (PEG[2] and PEO[1]). Comparing PEG with PEO, the effects of hydrogen bonds can be clearly observed: not only does PEG feature higher densities but also its density remains constant with the increasing chain length, while PEO shows larger density with longer chains. All of these critical features are accurately captured by PhyNEO, while conventional force fields such as OPLS-AA fail to describe the density profile of PEG qualitatively. It is worth noting again that PhyNEO is a purely ab initio model that does not rely on any experimental inputs. Therefore, PhyNEO provides a powerful prediction tool for the bulk properties of organic systems.

To illustrate the reliability of a potential, it is important to investigate the liquid microstructures, which could exhibit notable differences despite similar densities. To understand the structure of hydrogen bonds, we examined the RDF between hydroxyl H and O in PEG, the results of which are plotted in Figure 6a for PEG[3] and Figure 6b for PEG[4]. The first RDF peak around 1.8 Å, which is related to the primary hydrogen bond, is already quite different among the three models: the higher peak intensity and the narrower peak width of PhyNEO indicate a much stronger hydrogen bond than that from GAFF and OPLS-AA. A similar trend can be observed in the total number of hydrogen bonds, OPLS-AA and GAFF underestimate the number of hydrogen bonds by 40 and 10%, respectively. It is interesting to see that even though GAFF underestimates the hydrogen bond strength, it overestimates the density, possibly due to extra error cancellations. An even more interesting observation can be made for the secondary peak located at around 3.3 Å, especially in short molecules such as PEG[3]. While PhyNEO predicts a prominent secondary O−H peak at this position, GAFF predicts a much weaker interaction and OPLS-AA misses it completely. Deeper investigation reveals that this secondary hydrogen bond is associated with the interaction between a hydroxyl H with an O atom while being hydrogen-bonded to its neighboring O. A typical "double hydrogen-bonded" structure is plotted in Figure 6d. It can be seen that in such a structure, the H bond acceptor must adopt a cis conformation, the energy of which might be overestimated by OPLS and GAFF.

To further validate our findings, we conduct additional calculations on the formation energy of one representative "double hydrogen-bonded" structure and compare the results with MP2/aug-cc-pVTZ calculations. We decompose the total formation energy into intermolecular and intramolecular parts and show the results in Figure 6c. For the intermolecular part, all three models give similar results to those by MP2, with conventional force fields slightly underestimating the binding energy compared to those of PhyNEO and MP2. A larger error can be seen in the intramolecular part: while PhyNEO

https://doi.org/10.1021/acs.jctc.3c01045
J. Chem. Theory Comput. 2024, 20, 253−265

underestimates the energy by 8 kJ/mol, both OPLS-AA and GAFF overestimate the energy by over 30 kJ/mol. Such a large deviation clearly explains the inaccurate behavior of OPLS-AA and GAFF in the secondary RDF peak.

This comparison highlights the importance of a force field with high microscopic fidelity and the superior performance of PhyNEO. Due to its extraordinary prediction capability, the fully ab initio PhyNEO serves as a highly competent methodology for the construction of the next-generation general organic force field.

4. CONCLUSIONS AND FUTURE PERSPECTIVE

In this work, we present a versatile force field development workflow, PhyNEO, for the atomistic simulations of organic molecular systems. PhyNEO combines an ab initio physics-driven multipolar polarizable force field with an ML short-range enhancement. The key idea of the workflow is the clean separation of the long-/short-range interactions and the bonding/nonbonding energies. Physical models with strong transferability and low training cost are used to give a refined description of the long-range interactions, and ML models with strong fitting capabilities are used to describe the short-range correction and local bonding terms. The PhyNEO force field features strong transferability and scalability and consequently high data efficiency. For proof-of-concept purposes, PEO and PEG are selected as test examples, representing typical polymeric systems without and with hydrogen bonds. Using ab initio data of small clusters with fewer than 25 heavy atoms as the training set, we are able to build a force field that applies to both small molecule clusters and large bulk simulations. The microscopic details of the PES are accurately captured by PhyNEO and also show a much stronger capability for bulk property predictions than conventional force fields. The potential also shows excellent transferability among molecules of different sizes and thus is easily applicable to polymers with longer lengths. In general, PhyNEO provides a promising theoretical framework for the development of the next-generation general-purpose organic force field. Meanwhile, we note that in order to obtain a truly general force field that covers the wide chemical space of organic molecules, there are still a few issues that need to be solved. For example, more systematic, automatic, and efficient approaches are needed for molecule fragmentation and nonequilibrium geometry sampling, especially for many-body clusters. It is also interesting to develop an ML model to predict atomic parameters directly from the local chemical environment. Such a model should account for the geometry-dependent fluctuations of the atomic parameters, leading to a better description of the long-range interactions. Also, in current multipolar force fields such as AMOEBA, one often needs to define a local frame for each atom using its neighboring atoms, within which atomic multipoles can be defined. Such a local frame definition is often subject to discontinuities caused by atom index permutations and structural changes. An equivariant neural network with intrinsically built-in permutation symmetry can potentially avoid such problems. The incorporation of these techniques would significantly enhance the capability of PhyNEO, and it is a subject of our future research.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.3c01045.

Data set composition, validation of the implementation of NPT MD simulation, and comparison of PhyNEO with ab initio calculations (PDF)

AUTHOR INFORMATION

Corresponding Author

Kuang Yu − *Tsinghua-Berkeley Shenzhen Institute and Institute of Materials Research (iMR), Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, Guangdong 518055, P. R. China;* orcid.org/0000-0001-9142-5263; Email: yu.kuang@sz.tsinghua.edu.cn

Author

Junmin Chen − *Tsinghua-Berkeley Shenzhen Institute, Tsinghua University, Shenzhen, Guangdong 518055, P. R. China;* orcid.org/0000-0002-6069-9162

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.3c01045

Notes

The authors declare no competing financial interest.

A sample of the PhyNEO force field and an example of a PhyNEO MD simulation is provided at https://github.com/Jeremydream/PhyNEO. The underlying DMFF package is available at https://github.com/deepmodeling/DMFF. Any additional codes not listed here are available from the authors upon request.

ACKNOWLEDGMENTS

We thank the National Natural Science Foundation of China (22103048) and Shenzhen Science and Technology Innovation Committee (WDZC20200819115243002) for their financial support of this work.

REFERENCES

(1) 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.

(2) Jorgensen, W. L.; Tirado-Rives, J. Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6665−6670.

(3) 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.

(4) Tuckerman, M. *Statistical Mechanics: Theory and Molecular Simulation*; Oxford University Press: USA, 2010.

(5) Partridge, H.; Schwenke, D. W. The determination of an accurate isotope dependent potential energy surface for water from extensive ab initio calculations and experimental data. *J. Chem. Phys.* **1997**, *106*, 4618−4639.

(6) Behler, J.; Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.

(7) Behler, J. Constructing high-dimensional neural network potentials: A tutorial review. *Int. J. Quantum Chem.* **2015**, *115*, 1032−1050.

(8) Zhang, Y.; Hu, C.; Jiang, B. Embedded Atom Neural Network Potentials: Efficient and Accurate Machine Learning with a Physically Inspired Representation. *J. Phys. Chem. Lett.* **2019**, *10*, 4962−4967.

(9) Zhang, L.; Han, J.; Wang, H.; Car, R.; Weinan, E. Deep Potential Molecular Dynamics: A Scalable Model with the Accuracy of Quantum Mechanics. *Phys. Rev. Lett.* **2018**, *120*, 143001.

(10) Chmiela, S.; Sauceda, H. E.; Müller, K. R.; Tkatchenko, A. Towards exact molecular dynamics simulations with machine-learned force fields. *Nat. Commun.* **2018**, *9*, 3887.
(11) Yao, K.; Herr, J.; Toth, D.; Mckintyre, R.; Parkhill, J. The TensorMol-0.1 model chemistry: a neural network augmented with long-range physics. *Chem. Sci.* **2018**, *9*, 2261−2269.
(12) Batzner, S.; Musaelian, A.; Sun, L.; Geiger, M.; Mailoa, J. P.; Kornbluth, M.; Molinari, N.; Smidt, T. E.; Kozinsky, B. E. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat. Commun.* **2022**, *13*, 2453.
(13) Behler, J.; Csányi, G. Machine Learning Potentials for Extended Systems: A Perspective. *Eur. Phys. J. B* **2021**, *94*, 142.
(14) Cheng, Z.; Zhao, D.; Ma, J.; Li, W.; Li, S. An on-the-fly approach to construct generalized energy-based fragmentation machine learning force fields of complex systems. *J. Phys. Chem. A* **2020**, *124*, 5007−5014.
(15) 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.
(16) Unke, O. T.; Chmiela, S.; Gastegger, M.; Schütt, K. T.; Sauceda, H. E.; Müller, K. R. SpookyNet: Learning force fields with electronic degrees of freedom and nonlocal effects. *Nat. Commun.* **2021**, *12*, 7273.
(17) Westermayr, J.; Chaudhuri, S.; Jeindl, A.; Hofmann, O. T.; Maurer, R. J. Long-Range Dispersion-Inclusive Machine Learning Potentials for Structure Search and Optimization of Hybrid Organic−Inorganic Interfaces. *Digit. Discov.* **2022**, *1*, 463−475.
(18) Gao, A.; Remsing, R. C. Self-consistent determination of long-range electrostatics in neural network potentials. *Nat. Commun.* **2022**, *13*, 1572.
(19) Ko, T. W.; Finkler, J. A.; Goedecker, S.; Behler, J. A fourth-generation high-dimensional neural network potential with accurate electrostatics including non-local charge transfer. *Nat. Commun.* **2021**, *12*, 398.
(20) Zhang, L.; Wang, H.; Muniz, M. C.; Panagiotopoulos, A. Z.; Car, R. A deep potential model with long-range electrostatic interactions. *J. Chem. Phys.* **2022**, *156*, 124107.
(21) Babin, V.; Leforestier, C.; Paesani, F. Development of a "first principles" water potential with flexible monomers: Dimer potential energy surface, VRT spectrum, and second virial coefficient. *J. Chem. Theory Comput.* **2013**, *9*, 5395−5403.
(22) Reddy, S. K.; Straight, S. C.; Bajaj, P.; Huy Pham, C.; Riera, M.; Moberg, D. R.; Morales, M. A.; Knight, C.; Götz, A. W.; Paesani, F. On the accuracy of the MB-pol many-body potential for water: Interaction energies, vibrational frequencies, and classical thermodynamic and dynamical properties from clusters to liquid water and ice. *J. Chem. Phys.* **2016**, *145*, 194504.
(23) Bull-Vulpe, E. F.; Riera, M.; Bore, S. L.; Paesani, F. Data-Driven Many-Body Potential Energy Functions for Generic Molecules: Linear Alkanes as a Proof-of-Concept Application. *J. Chem. Theory Comput.* **2023**, *19*, 4494−4509.
(24) Hajibabaei, A.; Ha, M.; Pourasad, S.; Kim, J.; Kim, K. S. Machine learning of first-principles force-fields for alkane and polyene hydrocarbons. *J. Phys. Chem. A* **2021**, *125*, 9414−9420.
(25) 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.
(26) Yu, K.; Kiesling, K.; Schmidt, J. R. Trace Flue Gas Contaminants Poison Coordinatively Unsaturated Metal−Organic Frameworks: Implications for CO_2 Adsorption and Separation. *J. Phys. Chem. C* **2012**, *116*, 20480−20488.
(27) 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.
(28) McDaniel, J. G.; Schmidt, J. Next-generation force fields from symmetry-adapted perturbation theory. *Annu. Rev. Phys. Chem.* **2016**, *67*, 467−488.
(29) McDaniel, J. G.; Schmidt, J. Physically-Motivated Force Fields from Symmetry-Adapted Perturbation Theory. *J. Phys. Chem. A* **2013**, *117*, 2053−2066.
(30) Stone, A. J. *The Theory of Intermolecular Forces*, 2nd ed.; Oxford Univ. Press: Oxford, 2013.
(31) Misquitta, A. J.; Stone, A. J. ISA-Pol distributed polarizabilities and dispersion models from a basis-space implementation of the iterated stockholder atoms procedure. *Theor. Chem. Acc.* **2018**, *137*, 153.
(32) Jansen, G.; Hesselmann, A. Comment on "Using Kohn-Sham orbitals in symmetry-adapted perturbation theory to investigate intermolecular interactions. *J. Phys. Chem. A* **2001**, *105*, 11156−11157.
(33) Heßelmann, A.; Jansen, G. First-order intermolecular interaction energies from Kohn-Sham orbitals. *Chem. Phys. Lett.* **2002**, *357*, 464−470.
(34) Heßelmann, A.; Jansen, G. Intermolecular induction and exchange-induction energies from coupled-perturbed Kohn-Sham density functional theory. *Chem. Phys. Lett.* **2002**, *362*, 319−325.
(35) Heßelmann, A.; Jansen, G. Intermolecular dispersion energies from time-dependent density functional theory. *Chem. Phys. Lett.* **2003**, *367*, 778−784.
(36) Heßelmann, A.; Jansen, G. The helium dimer potential from a combined density functional theory and symmetry-adapted perturbation theory approach using an exact exchange-correlation potential. *Phys. Chem. Chem. Phys.* **2003**, *5*, 5010−5014.
(37) Heßelmann, A.; Jansen, G.; Schutz, M. Density-functional theory-symmetry-adapted intermolecular perturbation theory with density fitting: A new efficient method to study intermolecular interaction energies. *J. Chem. Phys.* **2005**, *122*, 14103.
(38) Van Vleet, M. J.; Misquitta, A. J.; Stone, A. J.; Schmidt, J. R. Beyond Born−Mayer: Improved Models for Short-Range Repulsion in Ab Initio Force Fields. *J. Chem. Theory Comput.* **2016**, *12*, 3851−3870.
(39) Son, C.; McDaniel, J.; Schmidt, J.; Cui, Q.; Yethiraj, A. First-Principles United Atom Force Field for the Ionic Liquid BMIM +BF4-: An Alternative to Charge Scaling. *J. Phys. Chem. B* **2016**, *120*, 3560−3568.
(40) McDaniel, J. G.; Yu, K.; Schmidt, J. Ab initio, physically motivated force fields for CO_2 adsorption in zeolitic imidazolate frameworks. *J. Phys. Chem. C* **2012**, *116*, 1892−1903.
(41) Yang, L.; Li, J.; Chen, F.; Yu, K. A transferrable range-separated force field for water: Combining the power of both physically-motivated models and machine learning techniques. *J. Chem. Phys.* **2022**, *157*, 214108.
(42) Wang, X.; Xu, Y.; Zheng, H.; Yu, K. A Scalable Graph Neural Network Method for Developing an Accurate Force Field of Large Flexible Organic Molecules. *J. Phys. Chem. Lett.* **2021**, *12*, 7982−7987.
(43) Huang, J.; Simmonett, A. C.; Pickard, F. C.; MacKerell, A. D.; Brooks, B. R. Mapping the Drude polarizable force field onto a multipole and induced dipole model. *J. Chem. Phys.* **2017**, *147*, 147.
(44) Kingma, D. P.; Ba Adam, J. A method for stochastic optimization. *arXiv* **2014**, 1412.6980 preprint.
(45) Yu, K.; Schmidt, J. Many-body effects are essential in a physically motivated CO2 force field. *J. Chem. Phys.* **2012**, *136*, 034503.
(46) Van Vleet, M. J.; Misquitta, A. J.; Schmidt, J. New angles on standard force fields: Toward a general approach for treating atomic-level anisotropy. *J. Chem. Theory Comput.* **2018**, *14*, 739−758.
(47) Guo, Y.; Riplinger, C.; Becker, U.; Liakos, D. G.; Minenkov, Y.; Cavallo, L.; Neese, F. Communication: An improved linear scaling perturbative triples correction for the domain based local pair-natural orbital based singles and doubles coupled cluster method DLPNO-CCSD(T). *J. Chem. Phys.* **2018**, *148*, 011101.
(48) Dodda, L. S.; Cabeza de Vaca, I.; Tirado-Rives, J.; Jorgensen, W. L. LigParGen web server: an automatic OPLS-AA parameter generator for organic ligands. *Nucleic Acids Res.* **2017**, *45*, W331−W336.

https://doi.org/10.1021/acs.jctc.3c01045
J. Chem. Theory Comput. 2024, 20, 253−265

(49) Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L.-P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D.; et al. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput. Biol.* **2017**, *13*, No. e1005659.
(50) Werner, H.-J.; Knowles, P. J.; Knizia, G.; Manby, F. R.; Schütz, M. Molpro: a general-purpose quantum chemistry program package. *WIREs Comput. Mol. Sci.* **2012**, *2*, 242−253.
(51) Misquitta, A. J.; Stone, A. J. Ab Initio Atom−Atom Potentials Using CamCASP: Theory and Application to Many-Body Models for the Pyridine Dimer. *J. Chem. Theory Comput.* **2016**, *12*, 4184−4208.
(52) Misquitta, A. J.; Stone, A. J.; Fazeli, F. Distributed Multipoles from a Robust Basis-Space Implementation of the Iterated Stockholder Atoms Procedure. *J. Chem. Theory Comput.* **2014**, *10*, 5405−5418.
(53) Valiev, M.; Bylaska, E.; Govind, N.; Kowalski, K.; Straatsma, T.; Van Dam, H.; Wang, D.; Nieplocha, J.; Apra, E.; Windus, T.; de Jong, W. NWChem: A comprehensive and scalable open-source solution for large scale molecular simulations. *Comput. Phys. Commun.* **2010**, *181*, 1477−1489.
(54) Wang, X.; Li, J.; Yang, L.; Chen, F.; Wang, Y.; Chang, J.; Chen, J.; Feng, W.; Zhang, L.; Yu, K. DMFF: An Open-Source Automatic Differentiable Platform for Molecular Force Field Development and Molecular Dynamics Simulation. *J. Chem. Theory Comput.* **2023**, *19*, 5897−5909.
(55) Ceriotti, M.; More, J.; Manolopoulos, D. E. i-PI: A Python interface for ab initio path integral molecular dynamics simulations. *Comput. Phys. Commun.* **2014**, *185*, 1019−1026.
(56) Thompson, A. P.; Plimpton, S. J.; Mattson, W. General formulation of pressure and stress tensor for arbitrary many-body interaction potentials under periodic boundary conditions. *J. Chem. Phys.* **2009**, *131*, 154107.
(57) Vassilev-Galindo, V.; Fonseca, G.; Poltavsky, I.; Tkatchenko, A. Challenges for Machine Learning Force Fields in Reproducing Potential Energy Surfaces of Flexible Molecules. *J. Chem. Phys.* **2021**, *154*, 094119.
(58) Hoffmann, M. M.; Horowitz, R. H.; Gutmann, T.; Buntkowsky, G. Densities, Viscosities, and Self-Diffusion Coefficients of Ethylene Glycol Oligomers. *J. Chem. Eng. Data* **2021**, *66*, 2480−2500.
(59) Hoffmann, M. M.; Kealy, J. D.; Gutmann, T.; Buntkowsky, G. Densities, Viscosities, and Self-Diffusion Coefficients of Several Polyethylene Glycols. *J. Chem. Eng. Data* **2022**, *67*, 88−103.
(60) Conesa, A.; Shen, S.; Coronas, A. Liquid Densities, Kinematic Viscosities, and Heat Capacities of Some Ethylene Glycol Dimethyl Ethers at Temperatures from 283.15 to 423.15 K. *Int. J. Thermophys.* **1998**, *19*, 1343−1358.
(61) Chang, J.-S.; Lee, M.-J.; Lin, H.-M. PVT of Fractionation Cuts of Poly(ethylene glycol) and Poly(propylene glycol) from 298 K to 338 K and Pressures up to 30 MPa. *J. Chem. Eng. Jpn.* **1999**, *32*, 611−618.
(62) López, E. R.; Daridon, J. L.; Plantier, F.; Boned, C.; Fernández, J. Temperature and Pressure Dependences of Thermophysical Properties of Some Ethylene Glycol Dimethyl Ethers from Ultrasonic Measurements. *Int. J. Thermophys.* **2006**, *27*, 1354−1372.

https://doi.org/10.1021/acs.jctc.3c01045
J. Chem. Theory Comput. 2024, 20, 253−265