JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

Developing a Differentiable Long-Range Force Field for Proteins with E(3) Neural Network-Predicted Asymptotic Parameters

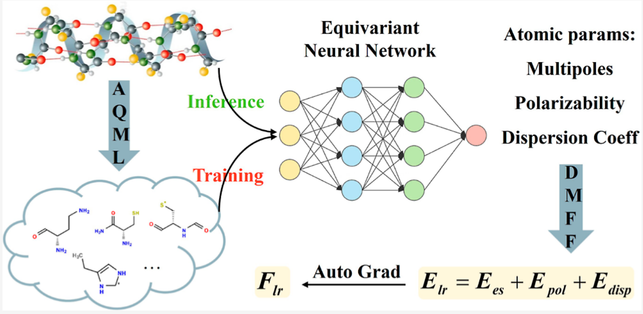
Published as part of Journal of Chemical Theory and Computation virtual special issue "Machine Learning and Statistical Mechanics: Shared Synergies for Next Generation of Chemical Theory and Computation".

Zheng Cheng, Hangrui Bi, Siyuan Liu, Junmin Chen, Alston J. Misquitta, and Kuang Yu*

Cite This: *J. Chem. Theory Comput.* 2024, 20, 5598−5608 Read Online

ACCESS | Metrics & More | Article Recommendations | SI Supporting Information

ABSTRACT: Accurately describing long-range interactions is a significant challenge in molecular dynamics (MD) simulations of proteins. High-quality long-range potential is also an important component of the range-separated machine learning force field. This study introduces a comprehensive asymptotic parameter database encompassing atomic multipole moments, polarizabilities, and dispersion coefficients. Leveraging active learning, our database comprehensively represents protein fragments with up to 8 heavy atoms, capturing their conformational diversity with merely 78,000 data points. Additionally, the E(3) neural network (E3NN) is employed to predict the asymptotic parameters directly from the local geometry. The E3NN models demonstrate exceptional accuracy and transferability across all asymptotic parameters, achieving an R^2 of 0.999 for both protein fragments and 20 amino acid dipeptide test sets. The long-range electrostatic and dispersion energies can be obtained using the E3NN-predicted parameters, with an error of 0.07 and 0.02 kcal/mol, respectively, when compared to symmetry-adapted perturbation theory (SAPT). Therefore, our force fields demonstrate the capability to accurately describe long-range interactions in proteins, paving the way for next-generation protein force fields.



1. INTRODUCTION

Molecular dynamics (MD) is widely used to study the structures and the functions of proteins, including protein folding and protease catalysis.[1−4] The insights provided by MD are instrumental in advancing drug design and unraveling the pathophysiology of diseases. Underlying all MD simulations is the force field, which describes the intra- and intermolecular interactions and determines the accuracy and reliability of the simulation results.[5−11] Currently, the most commonly used force fields[12−14] adopt simple functional forms with empirical parameters. These empirical force fields feature high computational efficiency, making them the primary models for industrial applications in the study of large molecular systems. These force fields are typically fitted to reproduce the experimental macroscopic properties of interest through top-down approaches but do not represent the microscopic details of the potential energy surface (PES) faithfully. Therefore, there are great challenges for these force fields to predict macroscopic properties outside of their calibration set. Meanwhile, another approach that is becoming increasingly popular is generating a force field purely based on ab initio data. Ideally, to accurately describe the intra- and intermolecular interactions, one would employ high-level correlated wave (CW) function methods such as MP2 or even CCSD(T). However, the computational demands of such methods scale unfavorably with system size, rendering direct energy evaluations at the CW level impractical for large proteins. Hence, the real challenge lies in how to learn the low-dimensional calculations of protein fragments and extrapolate them to larger proteins, and apparently such extrapolation is highly nontrivial.

The groundbreaking quantum mechanic/molecular mechanics (QM/MM) approach,[15,16] as initially proposed by Warshel and Levitt, describe the long-range interactions with the MM method, while handling the short-range interactions using the QM method. This long-/short-range separation approach provides an efficient solution of extending low-dimensional ab initio calculations to large organic molecules, and it has been extensively utilized to elucidate chemical processes

Received: March 16, 2024
Revised: June 3, 2024
Accepted: June 4, 2024
Published: June 18, 2024

https://doi.org/10.1021/acs.jctc.4c00337

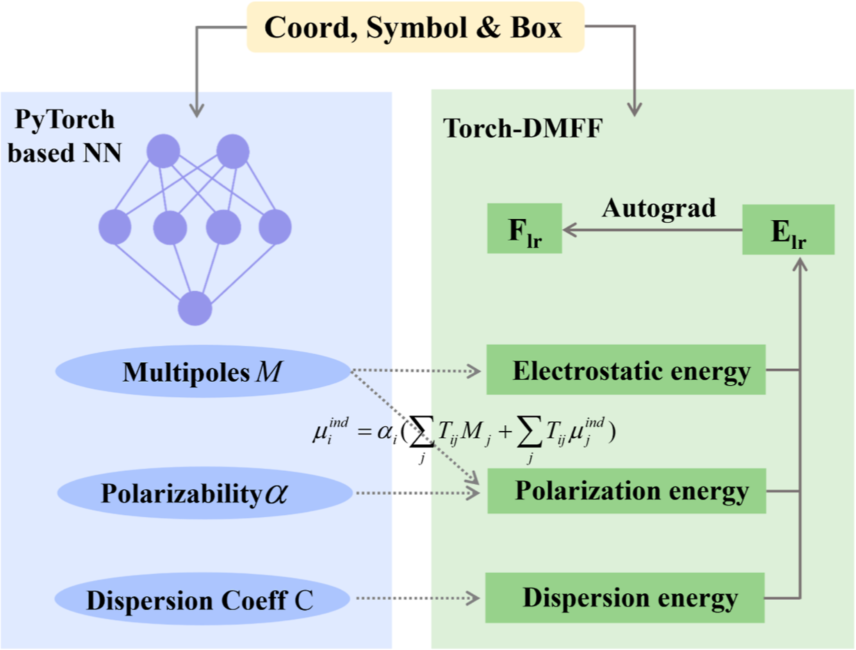


Figure 1. Schematic long-range interaction calculations with NN-based asymptotic parameters and Torch-DMFF.

occurring both in solution and within proteins.[17−22] Despite its wide applications, the QM/MM approach still relies on resource-intensive ab initio calculations to resolve short-range interactions, and the empirical-force-field-like long-range interactions can hardly achieve a CW level of accuracy. In the past decade, there has been tremendous progress toward a physical-driven intermolecular (nonbonding) potential for organic molecules. Employing time-dependent DFT (TD-DFT) coupled with appropriate population analysis,[23] we can now directly compute asymptotic atomic parameters, encompassing atomic multipole moments, polarizabilities, and dispersion coefficient.[24−29] These parameters enable the description of long-range interactions at the CW level, while retaining the computational efficiency of classical force fields. More recently, the descriptor-based or the graph-based machine learning models developed by different groups have been ingeniously applied to a wide range of inorganic materials and small organic molecules, showing superior fitting capabilities.[30−42] These data-driven methods typically express the PES of the system as a summation of atomic energies. The local environment of the atom is encoded into a feature vector, and a neural network or Gaussian process is then employed to predict the atomic energies. This approach provides a general scheme for fitting ab initio short-range interactions. Some models also employ conventional coulomb and LJ terms to address the long-range interactions outside of the cutoff radius, but the high order permanent multipole moments and induction terms are often ignored.[35,43−46] It is therefore natural to combine the local ML models with the CW-level asymptotic physical-driven potentials to build a new protein PES.

Our previous work demonstrated that CW-level PES can be built based on the range-separation scheme for simple polymers and bulk water.[47−49] In such a methodology, only monomer-based atomic asymptotic parameters (i.e., multipole moments, polarizabilities, and dispersion coefficients) are needed to describe the long-range interactions. In principle, for complex biomolecules such as proteins, the same protocol can be applied to construct a universal CW-level force field, given that accurate asymptotic atomic parameters are available. However, such efforts are hindered by the large chemical and conformational space of large biomolecules. The asymptotic parameters may vary subject to local chemical and structural changes, and capturing such variation is critical for further development of the complete range-separation force field. However, most existing data sets for organic molecules are limited to energy and force data,[38,50−54] and the most extensive data set to date, SPICE, only provides atomic multipole moments.[50] Important atomic property data, such as dispersion coefficients and polarizabilities, are missing, so an accurate long-range protein force field is yet to be developed.

In the current work, we aim to fill this gap by systematically sampling the chemical and geometrical space of proteins fragments. A database is built containing all important asymptotic data including polarizabilities and high-order dispersion coefficients. Based on this database, we employ the architecture of state-of-the-art(SOAT) E3NN model[39,40,55] to learn the mapping between the fragment geometry and corresponding asymptotic parameters. Moreover, we develop a PyTorch implementation of the previously JAX-based differentiable molecular force field (DMFF),[56] enabling the differentiable calculation of the resulting long-range force field. This program offers the basic infrastructure to deploy the hybrid ML/physical model in real MD simulations. Finally, we will demonstrate the consistency between our long-range PES and the symmetry-adapted perturbation theory (SAPT)[57,58] results and demonstrate the transferability of the long-range parameters obtained using our model. Through such work, we obtain an accurate long-range model, thus laying out the foundation for the development of a generic range-separation force field for proteins.

2. THEORY AND COMPUTATIONAL METHODS

2.1. Long-Range Force Field. In this study, the long-range components of the force field consist of electrostatic and dispersion contributions. The electrostatic component encom-

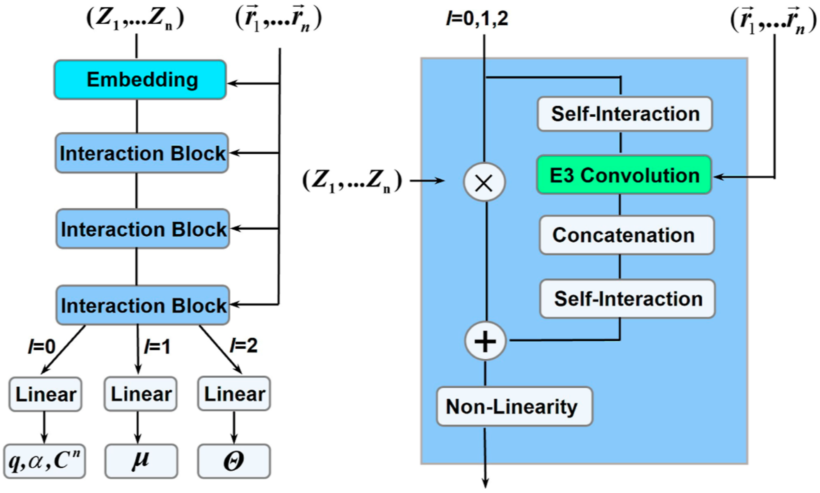


Figure 2. Schematic architecture of the neural network for predicting atomic scalar and tensor parameters.

passes both the permanent electrostatic energy and the electronic polarization energy. Consequently, the long-range interactions may be formulated as follows

$$E_{\mathrm{lr}} = E_{\mathrm{es}} + E_{\mathrm{pol}} + E_{\mathrm{disp}} \tag{1}$$

As illustrated in Figure 1, each term of long-range interaction can be computed utilizing differentiable multipolar polarizable force field calculator and asymptotic parameters, which serve as a direct counterpart to the corresponding term in DFT-SAPT.[57,58] All the asymptotic parameters, including atomic multipole moments, polarizability, and dispersion coefficients, can be computed through quantum mechanical (QM) calculations with small protein fragments and further predicted utilizing the E3NN model. As most NN models are based on the PyTorch, we develop the Torch-based DMFF as the force field calculator, which works in conjunction with the E3NN-based asymptotic parameters to automatically calculate long-range energy and forces. In the following section, we provide a detailed explanation of each long-range term along with the E3NN architecture.

2.1.1. Permanent Electrostatic Energy. The permanent electrostatic energy is described using multipole expansion as follows

$$E_{\mathrm{es}} = \sum_{i<j} \sum_{\mathrm{tu}} Q_t^i T_{\mathrm{tu}}^{ij} Q_{\mathrm{u}}^j \tag{2}$$

Here, Q_t^i refers to the atomic multipole moments (in spherical harmonics) at atom i with rank t, which include the harmonic monopole (charge), dipole, and quadrupole moments. All the permanent atomic multipole moments are calculated using the BS-ISA[59] in this work. We also calculated the atomic multipole moments through Hirshfeld partition approach,[60] a cheaper scheme for comparison. As all of the atomic multipole moments are fitted using the equivariant E3NN model, the rotation symmetry of these tensors is already satisfied by construction. Therefore, their definition does not rely on predefined local frames that rotates with the molecule as done in previous work.[61] The T_{tu}^{ij} refers to the multipole interaction tensor between atom i with multipole moments at rank t and j with multipole moments at rank u, the expression of which is detailed in Appendix F of ref 24.

2.1.2. Electronic Polarization Energy. To account for the polarization energy, we introduce an induced point dipole moment for each atom i. The final polarization energy is defined as follows

$$E_{\mathrm{pol}} = \sum_{i\neq j} \mu_{\mathrm{t}}^{i,\mathrm{ind}} T_{\mathrm{tu}}^{ij} Q_{\mathrm{u}}^{j} f_{\mathrm{t,u}}^{i,j} + \sum_{i<j} \mu_t^{i,\mathrm{ind}} T_{\mathrm{tu}}^{ij} \mu_{\mathrm{u}}^{j,\mathrm{ind}} f_{t,\mathrm{u}}^{i,j} \tag{3}$$

The first term refers to the induced-permanent interactions, while the second term refers to the induced−induced dipole interactions. To obtain the induced dipole moments $\mu_{\mathrm{t}}^{i,\mathrm{ind}}$, the induced dipole is solved full self-consistently until the induced dipoles at each atom reach convergence (the maximum value of $\frac{\partial(E_{\mathrm{es}}+E_{\mathrm{pol}})}{\partial \mu_t^{\mathrm{ind}}} < 0.01$ eV/(e·Å)), and the following condition is satisfied.

$$\mu_t^{i,\mathrm{ind}} = \alpha_{i,\mathrm{iso}}\Big(\sum_j T_{\mathrm{tu}}^{ij} Q_{\mathrm{u}}^{j} f_{t,\mathrm{u}}^{i,j} + \sum_j T_{\mathrm{tu}}^{ij} \mu_{\mathrm{u}}^{j,\mathrm{ind}} f_{t,\mathrm{u}}^{i,j} \Big) \tag{4}$$

To balance accuracy and computational cost, we assume that the atomic polarizability is isotropic and truncated at the dipole−dipole level. Atomic dipole−dipole isotropic polarizability $\alpha_{i,\mathrm{iso}}$ was obtained by distributing the TD-DFT charge density susceptibility matrix to atoms using the ISA-Pol method.[23] The localization scheme is chosen as the Lillestolen−Wheatley scheme,[62] and the weight coefficient is set as 0.01 during the ISA-Pol calculations.

To circumvent the polarization catastrophe at short distances, Thole's damping scheme is employed, which is identical to the approach utilized in the multipole and induced dipole (MPID) model.[63] For example, when considering the induced dipole-permanent charge interaction, the damping function can be defined as follows

$$f_{\mathrm{ind-dip,charge}}^{i,j} = 1 - (1 + a_{ij}u_{ij} + 0.5a_{ij}^2 u_{ij}^2)\mathrm{e}^{(-a_{ij}u_{ij})} \tag{5}$$

where a_{ij} for a pair ij is defined as $a_i + a_j$, while the term u_{ij} for the same pair ij is defined as $r_{ij}/(\alpha_i\alpha_j)^{-1/6}$. The parameter a for all atoms is set as 0.333 Å^{-1}, which is the default value of MPID. Other damping functions between multipole moments can also be found in Table 1 of ref 63.

2.1.3. Dispersion Energy. The dispersion energy is calculated using the following expression

$$E_{\mathrm{disp}} = \sum_{i<j} -\frac{C_{ij}^{6}}{r_{ij}^{6}} - \frac{C_{ij}^{8}}{r_{ij}^{8}} - \frac{C_{ij}^{10}}{r_{ij}^{10}} \tag{6}$$

As demonstrated in the original ISA-Pol paper, the isotropic C_6 and C_8 cross terms can be well approximated by using the geometric mean combination rule. The error of the combination rule on isotropic C_{10} is also small enough when aug-cc-PVTZ or larger basis set is employed.[23] Thus, in this equation, we define the C_{ij}^{n} as $C_{ij}^{n} = \sqrt{C_i^n C_j^n}$, where C_i^n represents the isotropic atomic dispersion coefficients calculated from imaginary-frequency-dependent isotropic polarizability using the Casimir−Polder relationship. Given that the damping function's effect can be ignored at relatively large distances and our focus in this work is on long-range interactions, we did not include the damping function in the dispersion energy calculations in this work. The optimization of the damping function would be done together with the optimization of other short-range interactions, which is left to future work.

2.1.4. E3NN for Asymptotic Parameters. In order to establish the relationship between fragment geometry and tensorial asymptotic parameters, we adopt the architecture of a SOTA E3NN called neural equivariant interatomic potentials (NequIP).[39] NequIP is the first equivariant tensor graph convolutional neural network. During the message passing process, previous neural networks such as SchNet were limited to aggregating scalar information, thus restricting their application to the prediction of scalar properties, such as molecule energy. In contrast, NequIP aggregates both scalar and tensor information during the message passing process, thereby providing a more information-rich description of atomic environments. Furthermore, the E(3)-equivariant features employed in NequIP are inherently suitable for predicting tensorial properties such as atomic dipole moments and atomic quadrupole moments, which are central to the tasks in this study. The architecture of the neural network for predicting atomic scalar and tensor parameters is depicted in Figure 2. Following the original NequIP architecture, we constructed the representation model based on the E3 convolutional kernel and the "message passing" mechanism to describe the local environment of atoms within a molecule. The NequIP framework utilizes the $l = 0$ tensor output from the representation model to construct the atomic neural network and employs a summation of atomic contributions to represent the molecular energy. In this study, we introduce independent linear regression models to map the $l = 0$, 1, and 2 tensors from the representation model to the corresponding atomic scalar properties, such as charge, and atomic tensor properties, such as atomic quadrupole moments. The E3NN model is then interfaced with the differentiable multipolar polarizable force field calculator (i.e., DMFF) to compute the energy and force.

As mentioned in Figure 2 in the original NequIP paper,[39] the atomic number Z_i of the center atom i is first encoded using a trainable embedding network, represented as $\mathbf{x}_i^0$. Subsequently, the environment is encoded using the relative position from atom i to it is neighboring atom j

$$S_m^l(\vec{r}_{ij}) = R(r_{ij}) Y_m^{(l)}(\hat{r}_{ij}) \tag{7}$$

The radial function $R(r_{ij})$ is a multilayer perceptron

$$R(r_{ij}) = W_n \sigma(\ldots \sigma(W_2 \sigma(W_1 B(r_{ij})))) \tag{8}$$

where $B(r_{ij})$ is a radial basis embedding of the interatomic distance, W_i is the weight matrices, and $\sigma(x)$ is the activation function.

The interaction blocks are designed to introduce interactions between different neighboring atoms with a ResNet-style update

$$\mathbf{x}_i^{k+1} = f(\mathbf{x}_i^k) + \text{self-interaction}(\mathbf{x}_\mathbf{i}^\mathbf{k}) \tag{9}$$

The function f represents a series of operations, including self-interaction, convolutional message passing, MLPs, and nonlinearity.[39] In the convolution operation, the equivariant network considers the case where $l_{\max}$ is greater than 0, while the invariant network handles only the $l = 0$ part. The convolutional layer $\mathcal{L}$, which is used in the equivariant network to facilitate interaction with filter f acting on input u and producing output o, is denoted as o: $l_\mathrm{u} \otimes l_\mathrm{f} \to l_\mathrm{o}$ and is defined using the following equation[39]

$$\mathcal{L}_{acm_\mathrm{o}}^{l_\mathrm{o},p_\mathrm{o},l_\mathrm{f},p_\mathrm{f},l_\mathrm{u},p_\mathrm{u}}(\vec{r}_i, V_{acm_\mathrm{u}}^{l_\mathrm{u},p_\mathrm{u}}) = \sum_{m_\mathrm{f},m_\mathrm{u}} C_{l_\mathrm{u},m_\mathrm{u},l_\mathrm{f},m_\mathrm{f}}^{l_\mathrm{o},m_\mathrm{o}} \sum_{j\in S} R(r_{ij})_{c,l_\mathrm{o},p_\mathrm{o},l_\mathrm{f},p_\mathrm{f},l_\mathrm{u},p_\mathrm{u}} Y_{m_\mathrm{f}}^{l_\mathrm{f}}(\hat{r}_{ij}) V_{bcm_\mathrm{u}}^{l_\mathrm{u},p_\mathrm{u}} \tag{10}$$

The indices i and $j \in S$ denote the center atom of the convolution and the neighboring atoms of i within a cutoff distance, respectively. The $V_{bcm_\mathrm{u}}^{l_\mathrm{u},p_\mathrm{u}}$ denotes the feature vectors of the input u that comprise a direct sum of irreducible representation of the O(3) symmetry group. The "rotation order" $l_\mathrm{u} = 0, 1, 2, \cdots$is a non-negative integer and parity is one of $p_\mathrm{u} \in (-1, 1)$, which together label the irreducible representations of O(3). The indices b, c, m_u correspond to the atoms, the channels (elements of the feature vector), and the representation index which takes values $m \in [-l, l]$, respectively. Furthermore, C indicates the Clebsch−Gordan coefficients. It should be noted that the placement of indices into subscript and superscript does not carry a specific meaning. Moreover, only output tensors with $l_\mathrm{o} \leq l_{\max}$ are taken into account.

Finally, for the output block which comprises a set of two atomwise self-interaction layers, the $l = 0$ features of the final convolution are transmitted and utilized to predict the scalar properties including the charge, atomic polarizability, and dispersion coefficients (C6, C8, and C10). Concurrently, the $l = 1$ and $l = 2$ features of final convolution are transmitted and utilized to predict the atomic dipole moment and quadrupole moment, respectively.

2.1.5. Construction of Asymptotic Parameter Database. In this study, we utilized the BS-ISA[59] and ISA-Pol approaches,[23] as implemented in Camcasp7,[27,64] to compute asymptotic atomic parameters. However, the method ISA-Pol is computationally intensive, generally limiting its application to molecules with no more than dozens of heavy atoms. Furthermore, the chemical and conformational complexity of proteins grows exponentially with increasing atom numbers, making it impossible to sample an entire protein. In this work, we rely on the locality of asymptotic parameters, assuming that atom parameters can be fully determined by their local environment, so they can be computed in small fragments and safely transferred to larger proteins. To systematically fragment the protein systems, we adopted the AQML[65] approach, as

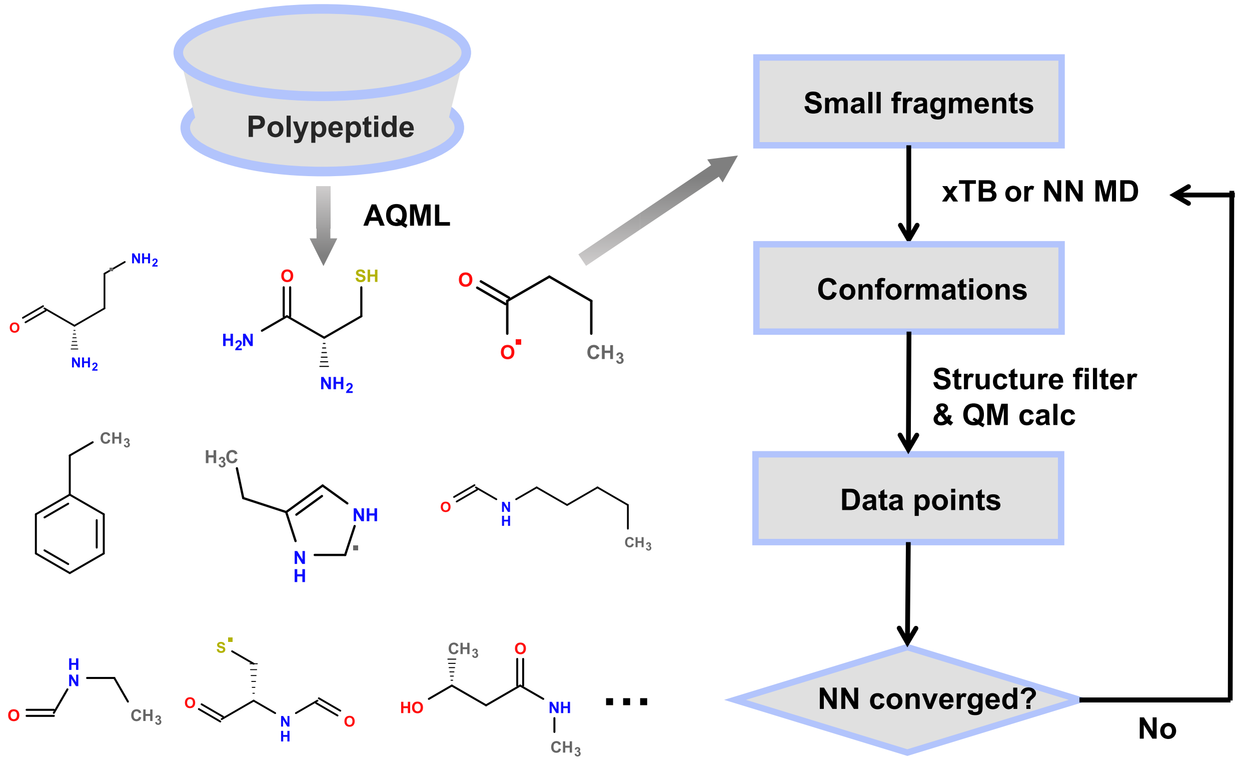


Figure 3. Schematic data construction workflow for asymptotic parameters.

depicted in Figure 3. With this methodology, we decomposed the high-dimensional chemical space of proteins into fragments containing up to 8 heavy atoms (362 of them in total), which is much easier to cover systematically. We subsequently implemented an active learning strategy to sample representative conformations from xTB MD trajectories and further sampled from NequIP-based MD trajectories at temperatures ranging from 100 to 800 K. As a result, the conformations in our database adequately cover all relevant regions of the potential energy surface for nonreactive MD simulations of protein fragments within a reasonable temperature range at the DFT level.

During the active learning process, we trained four NequIP potential models at the DFT level of theory and selected new data points based on the maximum force variance criterion. The reference QM calculations were performed using ωb97xD/6-31G*, as implemented in Gaussian 16.[66] To avoid unphysical structures that could result in non-converged ISA-Pol calculations, we incorporated a structural filter within the active learning workflow. This filter eliminates structures with bond lengths less than 0.8 or more than 1.4 times their equilibrium values.

The E3NN architecture utilized in this work demonstrated high data efficiency in previous work;[39] it requires only 133 structures for fitting water and ice potential surfaces, while other previous neural network (NN) architectures need 140,000 structures to achieve comparable accuracy.[36] Benefiting from both the active learning technique and the SOTA NN architecture, we achieved an accuracy of 17 meV for energy and 26 meV/Å for forces with the NequIP models, and our training set was enriched with approximately 78,000 data points sampled within the temperature range of 100−800 K. The accuracy of our NN has achieved a chemical accuracy (typically within 43 meV). To further demonstrate the comprehensiveness of our data set in covering the chemical and conformational space of protein fragments for nonreactive MD simulations, NequIP-based MD simulations were performed on five proteins fragments for 50 ps at 800 K. Figure S1 illustrates the changes in root-mean-square deviation (RMSD) and conformations on these fragments. Significant conformation changes (such as dihedral angle flips) were observed during the MD simulations. Yet, no unstable phenomena such as chemical bond breaking occurred during the MD simulation, showing the adequacy of our samplings on the training of E(3)-equivariant models. Thus, our data set ensures a comprehensive coverage of the related local chemical and conformational space in nonreactive protein simulations. In addition, the NequIP model training procedure is also elaborated in the Supporting Information.

With the data set available, we computed all the asymptotic parameters using TDDFT implemented in Psi4,[67] at the PBE0/aug-cc-pVTZ level and in conjunction with the BS-ISA and ISA-Pol algorithms. Although a structural filter was employed to ensure the validity of the conformations within the data set, a small subset of structures still resulted in nonconvergent BS-ISA calculations, culminating in a final data set of approximately 74,000 data points. Additionally, for comparison, atomic multipole moments were also computed using DFT calculations at the PBE0/aug-cc-pVTZ level coupled with the Hirshfeld method,[60] as implemented in Multiwfn,[68] and ALDA xc kernel was used. Owing to the lower computational demand of the Hirshfeld approach compared to BS-ISA and ISA-pol, the Hirshfeld method was used to derive a test set of atomic multipole moments for 20 types of amino acid dipeptides. This test set serves to assess the transferability of the E3NN model, which has been trained on smaller fragments with up to 8 heavy atoms. Further details on

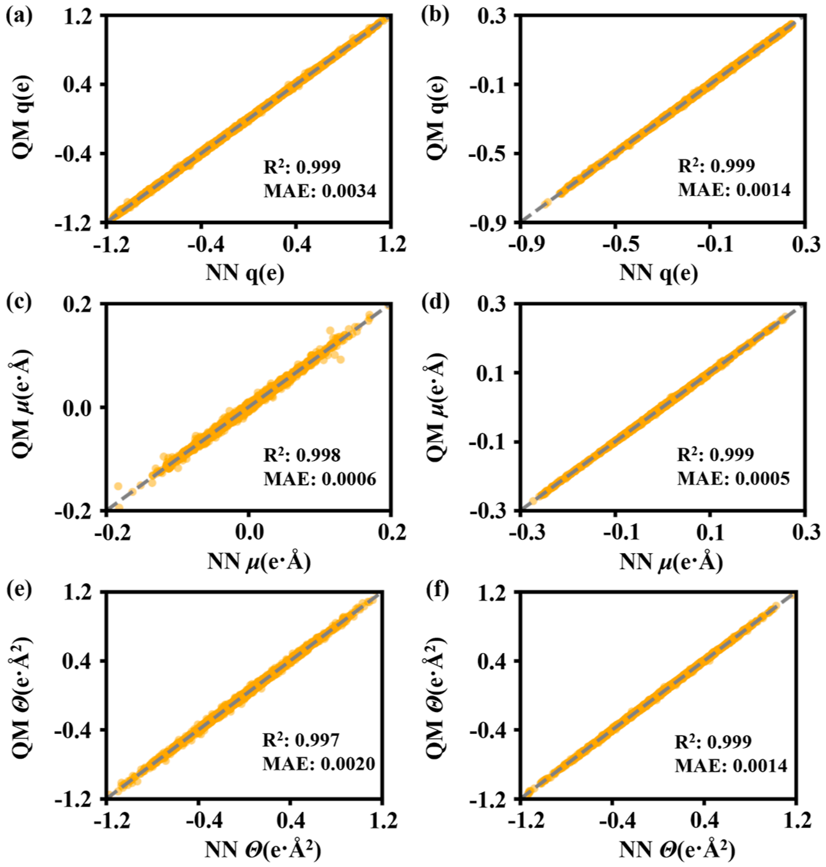


Figure 4. Correlation between the E3NN-predicted multipole moments [(a)(b) charge (q), (c)(d) dipole (μ), and (e)(f) quadrupole (Θ in spherical harmonics format)] and the QM counterparts for the test sets. The QM calculations are performed at the PBE0/aug-cc-pVTZ level. The QM multipole moments in (a), (c), and (e) are obtained using the BS-ISA approach, while the results in (b), (d), and (f) are obtained using the Hirshfeld partition.

CamCASP and Multiwfn calculations are provided in the Supporting Information.

3. RESULTS AND DISCUSSION

3.1. E3NN Accuracy on Atomic Multipole Moments.

For multipole moments, we partitioned the data set into a training set and a test set using a random 9:1 split and trained this data with E3NN. Figure 4 illustrates the correlations of charge, atomic dipole moment, and traceless quadrupole moments between the NN-predicted values and the QM counterparts for the test sets. The QM data presented in Figure 4a,c,e were obtained using the BS-ISA approach. In contrast, the data in Figure 4b,d,f were generated using the Hirshfeld method. Notably, the Hirshfeld charges are generally lower than the BS-ISA-derived charges, which is a known feature of the Hirshfeld method, as discussed in the literature.[69,70] On the other hand, this is compensated for by larger higher-ranking moments; therefore, atomic dipole moments derived from the Hirshfeld method tend to be higher than those obtained via the BS-ISA approach. In terms of E3NN accuracy, for either BS-ISA or Hirshfeld multipole moments, the E3NN prediction exhibits an excellent R^2 value of approximately 0.999. Moreover, the mean absolute error (MAE) for the predicted monopole moment (charges) is on the order of 10^{-3}e. For dipole moments, the MAE is in the order of 10^{-4}e·Å, and for quadrupole moments, the MAE is about 1.5 10^{-3}e·Å^2. The E3NN models achieve a slightly superior performance in predicting the Hirshfeld multipoles compared to those obtained via BS-ISA. This is probably due to the fact that the BS-ISA multipoles display more geometric fluctuations compared to the Hirshfeld results and thus are more difficult to fit (see below for more details). Nevertheless, the accuracies of both fittings are adequate to give excellent descriptions of the long-range electrostatic interactions, as we will show below.

To further analyze the performance of E3NN, we present the MAE and the distribution ranges of multipole moments for different elements using both the BS-ISA and Hirshfeld methods, as detailed in Tables S1 and S2. It is evident that the MAE values exhibit noticeable variations across different elements for all multipole moments. The magnitude of the error generally shows a consistent trend with the width of the distribution range of the corresponding quantities. For instance, the range for the quadrupole moment of sulfur (S) using both the BS-ISA and Hirshfeld methods is approximately 2.5 e/Å^2, which is notably higher than that for other elements; the range of the quadrupole moment on other elements does not exceed 1.5 e/Å^2. This indicates that accurately predicting the quadrupole moments for sulfur poses a greater challenge, resulting in an MAE of roughly 5 me/Å^2. This simply indicates that the *S* parameters experience more structure-dependent variations and thus are more difficult to model. Nevertheless, as we stated above, the overall accuracy is still quite satisfactory.

3.2. E3NN Accuracy on Polarizability and Dispersion Coefficients.

Similar to the multipole moments, the atomic dipole−dipole isotropic polarizability $\boldsymbol{\alpha}_{\mathbf{iso}}$ and isotropic atomic dispersion coefficients (C_6, C_8, and C_{10}) data are also partitioned into a training set and a test set using a random 9:1 split. These scalar properties are also fitted using E3NN, and Figure 5 illustrates the correlation between the prediction and the corresponding ISA-Pol results on the test set. In contrast to permanent atomic multipole moments, which are determined exclusively by the static electron density, the polarizability and dispersion coefficients are derived from the density response at first order in the perturbation operator using the frequency-dependent charge-density susceptibilities.[23] Therefore, the computation and the localization of the response matrix and the fitting process all pose greater challenges. These response-related parameters are usually not included in the existing databases but are critical in force field development as they contribute a large portion of long-range nonbonding interactions.

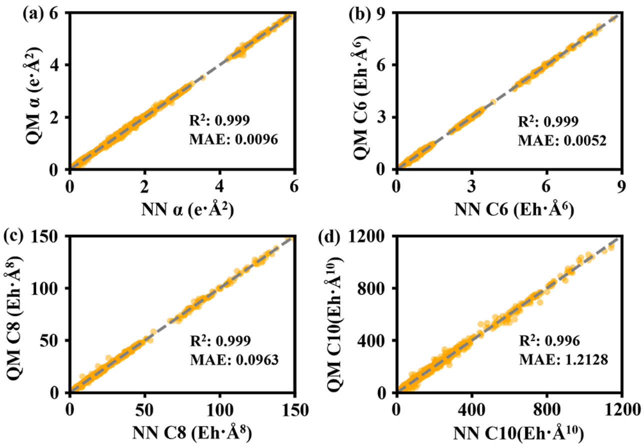


Figure 5. Correlation between the E3NN-predicted properties [(a) isotropic polarizability (α_{iso}) and (b−d) dispersion coefficients (C_6, C_8, C_{10})] and their QM counterparts on the test sets. The polarizability and dispersion coefficients are obtained at the PBE0/aug-cc-pVTZ level with the ISA-Pol approach, and GRAC shift is considered during the DFT calculation.

Notwithstanding these challenges, our work demonstrates that the R^2 values for the E3NN-predicted atomic isotropic polarizability and dispersion coefficients surpass 0.995. The mean absolute error (MAE) for these parameters is below 0.2% of the respective distribution ranges. Detailed in Table S3 are the MAE and the distribution ranges of atomic polarizability and dispersion coefficients on different elements. Similar to the electrostatic case, we also observe that S (sulfur) atoms exhibit the broadest range on both polarizability and dispersion coefficients, resulting in a relatively higher MAE. The worst scenario happens for the C_{10} of S, which shows a MAE of 9.1 $E_h{\cdot}\text{Å}^{10}$. This error is still merely 1% of the distribution range, showing that E3NN can capture the geometry fluctuations of atomic parameters with remarkable accuracy.

3.3. Transferability of E3NN.

In our training and testing data set, only small fragments are included due to the limitation of computational cost. However, as these parameters are meant to be used in the simulation of large proteins, the transferability of these parameters in molecules of different sizes is a critical issue.

To assess the transferability of our E3NN model, we generate a test data set composed of 20 amino acid dipeptides, each capped with acetyl (ACE) and *N*-methyl (NME) groups. For each dipeptide, we uniformly sample 100 unique conformations from a 50 ps xTB MD trajectory at 300 K. This dipeptide test set possesses a much larger chemical diversity compared to our training set, which is limited to protein fragments with no more than eight heavy atoms.

Due to the prohibitive computational cost of ISA-pol calculations on dipeptides, we only assess the correlation of the Hirshfeld multipole moments between the E3NN model and the QM results. As shown in the Figure 6, the R^2 for charge, atomic dipole moment, and quadrupole moments all exceeded 0.999. The MAEs for charge and atomic quadrupole moment at the 10^{-3}e level and 10^{-3}e·Å^2, respectively, and the atomic dipole moment at the 10^{-4}e·Å level. Such good agreement indicates that the asymptotic parameters are highly localized in closed-shell organic molecules and thus can be determined using the local environment that is within no more than four to five bonds. Therefore, the long-range model can be trained using tiny fragments and then safely transferred to larger biomolecules such as proteins. This result is encouraging, as it shows that it is possible to drastically reduce the seemingly vast chemical space of organic molecules by fragmentation, as one would expect by intuition.

3.4. Accuracy of Long-Range Energy.

While the accuracy of parameters is important by itself, an even more important gauge is the accuracy of the interaction energy, which eventually determines the behavior of the MD simulation. To evaluate the accuracy of our long-range force field, we constructed a data set using SAPT, which encompasses all combinations of dimers formed by fragments with 1−8 heavy atoms. Dimer scan is performed to sample the conformations over a range of distances from 3.5 to 6.5 Å, which we find in the asymptotic region. In this region, the interatomic distances are larger than the typical van der Waals contact range, so the charge penetration and Pauli repulsion effects can be safely neglected. Approximately 15000 DFT-

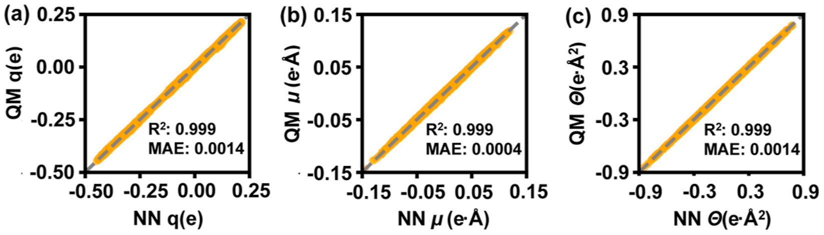


Figure 6. Correlation between the E3NN-predicted multipole moments [(a) charge (q), (b) dipole (μ), and (c) quadrupole (Θ)] and their QM counterparts for the dipeptide test set. The multipole moments were obtained at the PBE0/aug-cc-pVTZ level with the Hirshfeld partition approach.

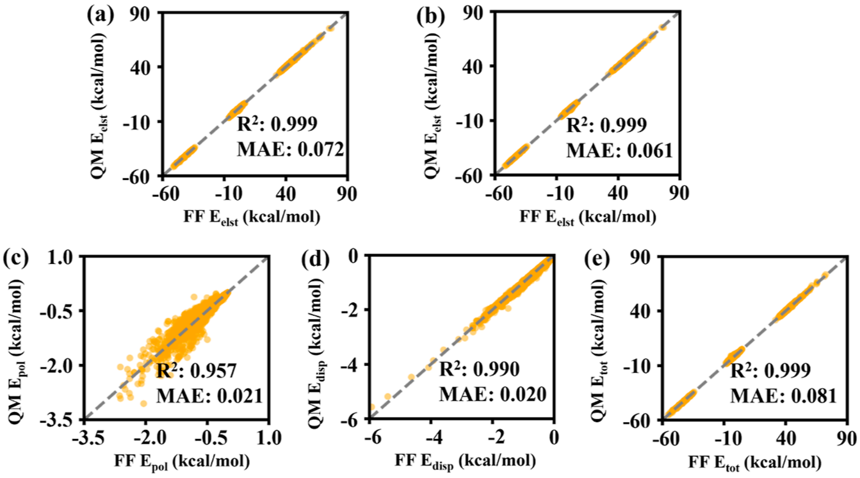


Figure 7. Correlation between long-range energies [(a,b) permanent electrostatic energy, (c) electronic polarization energy, (d) dispersion energy component, and (e) total energy calculated by our force field and their DFT-SAPT counterparts]. The E3NN models of (a), (c), (d), and (e) were trained using the data derived from the BS-ISA and ISA-Pol methods, while the E3NN model of (b) was trained based on data obtained via the Hirshfeld approach. All the energy components for DFT-SAPT were calculated at the PBE0/aug-cc-pVTZ level, and GRAC shift is considered during the DFT calculation.

SAPT calculations were carried out to verify the electrostatic/polarization and dispersion energy components. Details of the DFT-SAPT calculations are provided in the Supporting Information. The long-range DFT-SAPT components including electrostatic(E(1)pol), polarization(E(2)ind + E(2)ind-exch), and dispersion energies(E(2)disp + E(2)disp-exch) are then compared with the classical model with E3NN-predicted parameters. Through this test, we systematically examine how these asymptotic parameters perform in a classical force field.

Figure 7a,b shows the comparison of the electrostatic energies between the force field and the DFT-SAPT results. The atomic multipole moments employed for the electrostatic energy in Figure 7a were determined using the E3NN model trained on the BS-ISA data set, while those in Figure 7b were derived using the E3NN model trained on the Hirshfeld data set. The electrostatic energy based on both BS-ISA and Hirshfeld multipole moments exhibited equally high accuracy, with R^2 values of 0.999 and MAEs of 0.072 and 0.061 kcal/mol, respectively. The error from the Hirshfeld-E3NN model is slightly lower, in consistent with the fact that the Hirshfeld multipoles can be fitted slightly better. Figure 7c,d demonstrates the accuracies of the polarization energy and dispersion energy, respectively. Only BS-ISA/ISA-Pol results are shown since at present, the Hirshfeld method cannot be used as an alternative to the BS-ISA in the ISA-Pol algorithm to provide atomic response parameters. The prediction accuracy of our force field for polarization and dispersion energy is lower than that for electrostatic energy. Some data points deviate from the diagonal line in Figure 7c, and the R^2 for polarization is 0.957. This may due to the approximation of the atomic polarizability as isotropic and the truncation at the dipole level. Considering that the development of a PyTorch-based polarization energy calculator that incorporates anisotropic atomic dipole polarizability is challenging, we intend to tackle this task in the next phase of our research. Meanwhile, the deviation of data points in Figure 7d is much smaller, and the R^2 for dispersion energy is as high as 0.99. Given that both the polarization and dispersion energies vary within a relatively narrower range and the MAE is only 0.02 kcal/mol for both components, the main source of error is still the electrostatic energy. We also compare the correlation of total intermolecular interactions in Figure 7e. Again, excellent agreement is observed for total intermolecular interactions, with an R^2 value of 0.999 and a MAE of 0.081 kcal/mol.

It is noted that all long-range parameters are computed using single molecules in the gas phase, which is an essential approximation. One concern is that strong intermolecular interactions (such as hydrogen bonds) can induce charge transfer between molecules, which changes the long-range interaction. To test how important this effect is, we also demonstrate the accuracy of our long-range force fields on fragments that have hydrogen bonds. Figure 8a−c illustrates the total, electrostatic, polarization, and dispersion energies between monomer A and B, as calculated by SAPT2 + (3)δMP2 and NN-based long-range force field at various distances. Figure S2 also presents these four energy components for four other dimers at varying distances. In all cases, monomer A consists of two protein fragments with hydrogen bonds, while monomer B comprises a single protein fragment. For all the dimers, excellent agreement between SAPT2 + (3)δMP2 and the long-range force field for all different energy components is observed when the distances between two monomers are larger than 3.5 Å, with the maximum error on total energy only about 1 kcal/mol. All these results show that while charge transfer does exist in hydrogen bonded protein systems, its effects are mainly short-ranged. It can be largely neglected and in long-range (beyond 3.5 Å) at the level of chemical accuracy.

Moreover, it can also be argued that the linear response computed in vacuum can be different from the response in solvated environments. For example, due to the Pauli repulsion of surrounding solvents, the response of solute in a strong field can be suppressed, leading to nonlinear behaviors. To examine the importance of such effects, we also computed the SAPT2 + (3)δMP2 and our force field based total, electrostatic, polarization, and dispersion energy between solvated fragments and a sodium ion in Figure 8b−d, as well as in Figure S3. For all five test cases, excellent agreement between SAPT2 + (3)δMP2 and long-range force field on all energy components is also observed when the distance between two monomers is

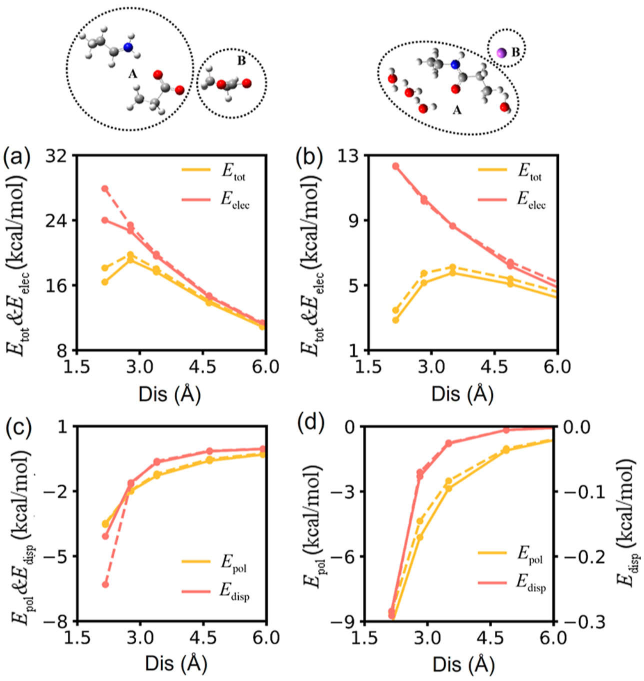


Figure 8. Comparison of the total, electrostatic, polarization, and dispersion energies for various dimers between the SAPT (solid line) and the NN-based long-range force fields (dotted line). For a dimer in (a) and (c), monomer A is composed of two protein fragments that have hydrogen bonds, while monomer B consists of a single protein fragment. For the dimer in (b) and (d), monomer A is composed of a protein fragment surrounded by four water molecules, whereas monomer B is a sodium ion. All the energy components for SAPT were calculated at the SAPT2 + (3)δMP2/aug-cc-pVTZ level.

>3.5 Å, with the maximum error on total energy being only about 0.9 kcal/mol. Once again, these excellent agreements show that in this preliminary study, the change of effective polarizability in solvated environment can be neglected. It is acknowledged that a more systematic investigation is needed in bulk simulation to draw a more definite conclusion, which is left to future work.

In principle, the impact of both charge transfer and solvation effects to asymptotic parameters can be accounted for by extending the training data set from single molecule fragments to many-body clusters. However, the increased size of the training cluster renders the long-range parameter calculations excessively expensive. The chemical and conformational space of many-body clusters is also considerably larger than that of single fragments, making it much more challenging to construct the database. Considering that the remaining error shown in this work is reasonably low, we are inclined to keep the simplicity of the training data set, which we expect to be accurate enough for nonreactive MD simulations.

4. CONCLUSIONS

In summary, we have established a comprehensive database of asymptotic parameters, including atomic multipole moments, atomic polarizabilities, and dispersion coefficients, to accurately describe the long-range interactions in proteins. By harnessing the power of active learning, our database, with merely 78,000 data points in total, fully captures the essential local chemical and conformational variations of proteins, demonstrating excellent data efficiency. The database of asymptotic parameters is further trained with E3NN, resulting in a model with remarkable accuracy and transferability. The R^2 values reach 0.999 across both protein fragments and amino acid dipeptide test sets for all components. Additionally, a PyTorch-based multipolar polarizable force field calculator enables us to estimate the long-range interaction energy with E3NN-predicted asymptotic parameters differentiably. We further compared the computed energies with those derived using DFT-SAPT, showing excellent agreement between them. Such a long-range force field gives a reliable nonbonding potential for peptide molecules beyond 3.5 Å, laying a solid foundation for the future development of a complete range-separation protein force field.

In the future, we will focus on the short-range part of the force field on the basis of the model derived in this work. Since the long-range potential is available now, we will be able to train the short-range interactions using small cluster quantum data at the CW level of theory. The active learning strategy and the fragment conformation samplings achieved in this work also pave the road to the establishment of an efficient sampling algorithm for small fragment clusters. Furthermore, the PyTorch-based force field calculator implemented in this work warrants an easy integration of force field models with state-of-the-art ML infrastructures. Combining all these techniques, we aim to build a chemically accurate force field for proteins and eventually for general organic molecules.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.4c00337.

Details of NN training, Multiwfn, CamCASP, and DFT-SAPT calculation and MAE of NN for all the asymptotic parameters with different elements or heavy atoms (PDF)

■ AUTHOR INFORMATION

Corresponding Author

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

Authors

Zheng Cheng − *School of Mathematical Sciences, Peking University, Beijing 100871, China; AI for Science Institute, Beijing 100084, P. R. China;* orcid.org/0000-0003-2737-606X

Hangrui Bi − *DP Technology, Beijing 100080, P. R. China; School of Mathematical Sciences, Peking University, Beijing 100871, China*

Siyuan Liu − *DP Technology, Beijing 100080, P. R. China*

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

Alston J. Misquitta − *School of Physics and Astronomy, Queen Mary, University of London, London E1 4NS, U.K.*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.4c00337

Notes

The authors declare no competing financial interest.

The E3NN code for training asymptotic parameters is accessible at https://github.com/zhengcheng233/e3_layer.git. A Pytorch-based implementation of JAX-dependent DMFF can be found at https://github.com/zhengcheng233/dmff_torch.git. Additionally, the SAPT database along with the E3NN model is available at https://github.com/zhengcheng233/lr_database.git. The database containing the asymptotic parameters will be available at https://aissquare.com/datasets/detail?pageType=datasets&name=Protein_longrange_parameters_dataset&id=245. Any additional codes and database not listed here are available from the authors upon request.

■ ACKNOWLEDGMENTS

The authors gratefully acknowledge the funding support from the National Natural Science of Foundation of China (grant number: 22103048) and Shenzhen Bay Laboratory (grant number: SZBL2021080601013). The authors thank Jingxiang Zou from Innovation Laboratory for Sciences and Technologies of Energy Materials of Fujian Province (IKKEM) and Bing Huang in University of Basel for helpful discussion. The authors also acknowledge EPSRC grant EP/X036863/1 which funds part of the PHYMOL: Physics, Accuracy and Machine Learning: Towards the next-generation of Molecular Potentials MSCA Doctoral Network (Horizon Europe Project 101073474).

■ REFERENCES

(1) Dill, K. A.; MacCallum, J. L. The protein-folding problem, 50 years on. *Science* **2012**, *338*, 1042−1046.

(2) Kuhlman, B.; Bradley, P. Advances in protein structure prediction and design. *Nat. Rev. Mol. Cell Biol.* **2019**, *20*, 681−697.

(3) Torbeev, V. Y.; Raghuraman, H.; Hamelberg, D.; Tonelli, M.; Westler, W. M.; Perozo, E.; Kent, S. B. Protein conformational dynamics in the mechanism of HIV-1 protease catalysis. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 20982−20987.

(4) Elsässer, B.; Goettig, P. Mechanisms of proteolytic enzymes and their inhibition in QM/MM studies. *Int. J. Mol. Sci.* **2021**, *22*, 3232.

(5) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B. L.; Grubmüller, H.; MacKerell, A. D. CHARMM36: An improved force field for folded and intrinsically disordered proteins. *Biophys. J.* **2017**, *112*, 175a−176a.

(6) Xu, Y.; Huang, J. Validating the CHARMM36m protein force field with LJ-PME reveals altered hydrogen bonding dynamics under elevated pressures. *Commun. Chem.* **2021**, *4*, 99.

(7) Lu, C.; Wu, C.; Ghoreishi, D.; Chen, W.; Wang, L.; Damm, W.; Ross, G. A.; Dahlgren, M. K.; Russell, E.; Von Bargen, C. D.; et al. OPLS4: Improving Force Field Accuracy on Challenging Regimes of Chemical Space. *J. Chem. Theory Comput.* **2021**, *17*, 4291−4300.

(8) Winkler, L.; Galindo-Murillo, R.; Cheatham, T. E., III Structures and Dynamics of DNA Mini-Dumbbells Are Force Field Dependent. *J. Chem. Theory Comput.* **2023**, *19*, 2198−2212.

(9) Sanavia, T.; Birolo, G.; Montanucci, L.; Turina, P.; Capriotti, E.; Fariselli, P. Limitations and challenges in protein stability prediction upon genome variations: towards future applications in precision medicine. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 1968−1979.

(10) Vennelakanti, V.; Nazemi, A.; Mehmood, R.; Steeves, A. H.; Kulik, H. J. Harder, better, faster, stronger: Large-scale QM and QM/MM for predictive modeling in enzymes and proteins. *Curr. Opin. Struct. Biol.* **2022**, *72*, 9−17.

(11) Magalhães, R. P.; Fernandes, H. S.; Sousa, S. F. Modelling enzymatic mechanisms with QM/MM approaches: current status and future challenges. *Isr. J. Chem.* **2020**, *60*, 655−666.

(12) Robertson, M. J.; Tirado-Rives, J.; Jorgensen, W. L. Improved peptide and protein torsional energetics with the OPLS-AA force field. *J. Chem. Theory Comput.* **2015**, *11*, 3499−3509.

(13) Ponder, J. W.; Case, D. A. Force fields for protein simulations. *Adv. Protein Chem.* **2003**, *66*, 27−85.

(14) Vanommeslaeghe, K.; MacKerell, A., Jr. CHARMM additive and polarizable force fields for biophysics and computer-aided drug design. *Biochim. Biophys. Acta, Gen. Subj.* **2015**, *1850*, 861−871.

(15) Warshel, A.; Levitt, M. Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J. Mol. Biol.* **1976**, *103*, 227−249.

(16) Lyne, P. D.; Hodoscek, M.; Karplus, M. A hybrid QM- MM potential employing Hartree- Fock or density functional methods in the quantum region. *J. Phys. Chem. A* **1999**, *103*, 3462−3471.

(17) Shurki, A.; Warshel, A. Structure Function Correlations of Proteins using MM, QM MM, and Related Approaches: Methods, Concepts, Pitfalls, and Current Progress. *Adv. Protein Chem.* **2003**, *66*, 249−313.

(18) Filippi, C.; Buda, F.; Guidoni, L.; Sinicropi, A. Bathochromic shift in green fluorescent protein: a puzzle for QM/MM approaches. *J. Chem. Theory Comput.* **2012**, *8*, 112−124.

(19) Hu, L.; Soderhjelm, P.; Ryde, U. Accurate reaction energies in proteins obtained by combining QM/MM and large QM calculations. *J. Chem. Theory Comput.* **2013**, *9*, 640−649.

(20) Beierlein, F. R.; Michel, J.; Essex, J. W. A simple QM/MM approach for capturing polarization effects in protein- ligand binding free energy calculations. *J. Phys. Chem. B* **2011**, *115*, 4911−4926.

(21) Varnai, C.; Bernstein, N.; Mones, L.; Csányi, G. Tests of an adaptive QM/MM calculation on free energy profiles of chemical reactions in solution. *J. Phys. Chem. B* **2013**, *117*, 12202−12211.

(22) Gao, J. Energy components of aqueous solution: Insight from hybrid QM/MM simulations using a polarizable solvent model. *J. Comput. Chem.* **1997**, *18*, 1061−1071.

(23) 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*, 153−220.

(24) Stone, A. *The Theory of Intermolecular Forces*; oUP oxford, 2013.

(25) Schmidt, J.; Yu, K.; McDaniel, J. G. Transferable next-generation force fields from simple liquids to complex materials. *Acc. Chem. Res.* **2015**, *48*, 548−556.

(26) McDaniel, J. G.; Schmidt, J. Next-generation force fields from symmetry-adapted perturbation theory. *Annu. Rev. Phys. Chem.* **2016**, *67*, 467−488.

(27) 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.

(28) 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.

(29) 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.

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

(31) 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.

(32) Shapeev, A. V. Moment tensor potentials: A class of systematically improvable interatomic potentials. *Multiscale Model. Simul.* **2016**, *14*, 1153−1173.

(33) Thompson, A. P.; Swiler, L. P.; Trott, C. R.; Foiles, S. M.; Tucker, G. J. Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials. *J. Comput. Phys.* **2015**, *285*, 316−330.

(34) Schütt, K.; Kindermans, P.-J.; Sauceda Felix, H. E.; Chmiela, S.; Tkatchenko, A.; Müller, K.-R. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in Neural Information Processing Systems*; Curran Associates, Inc., 2017; pp 991−1001.

(35) 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.

(36) 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.

(37) 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.

(38) 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.

(39) Batzner, S.; Musaelian, A.; Sun, L.; Geiger, M.; Mailoa, J. P.; Kornbluth, M.; Molinari, N.; Smidt, T. E.; Kozinsky, B. E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat. Commun.* **2022**, *13*, 2453.

(40) Musaelian, A.; Batzner, S.; Johansson, A.; Sun, L.; Owen, C. J.; Kornbluth, M.; Kozinsky, B. Learning local equivariant representations for large-scale atomistic dynamics. *Nat. Commun.* **2023**, *14*, 579.

(41) Schütt, K.; Unke, O.; Gastegger, M. Equivariant message passing for the prediction of tensorial properties and molecular spectra. *International Conference on Machine Learning*, 2021; pp 9377−9388.

(42) Christensen, A. S.; Bratholm, L. A.; Faber, F. A.; Anatole von Lilienfeld, O. FCHL revisited: Faster and more accurate quantum machine learning. *J. Chem. Phys.* **2020**, *152*, 044107.

(43) 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.

(44) Cheng, Z.; Du, J.; Zhang, L.; Ma, J.; Li, W.; Li, S. Building quantum mechanics quality force fields of proteins with the generalized energy-based fragmentation approach and machine learning. *Phys. Chem. Chem. Phys.* **2022**, *24*, 1326−1337.

(45) Wang, H.; Yang, W. Toward building protein force fields by residue-based systematic molecular fragmentation and neural network. *J. Chem. Theory Comput.* **2019**, *15*, 1409−1417.

(46) Morawietz, T.; Behler, J. A density-functional theory-based neural network potential for water clusters including van der Waals corrections. *J. Phys. Chem. A* **2013**, *117*, 7356−7366.

(47) 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.

(48) 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.

(49) Chen, J.; Yu, K. PhyNEO: A Neural-Network-Enhanced Physics-Driven Force Field Development Workflow for Bulk Organic Molecule and Polymer Simulations. *J. Chem. Theory Comput.* **2023**, *20*, 253−265.

(50) Eastman, P.; Behara, P. K.; Dotson, D. L.; Galvelis, R.; Herr, J. E.; Horton, J. T.; Mao, Y.; Chodera, J. D.; Pritchard, B. P.; Wang, Y.; et al. SPICE, A Dataset of Drug-like Molecules and Peptides for Training Machine Learning Potentials. *Sci. Data* **2023**, *10*, 11.

(51) Smith, J. S.; Zubatyuk, R.; Nebgen, B.; Lubbers, N.; Barros, K.; Roitberg, A. E.; Isayev, O.; Tretiak, S. The ANI-1ccx and ANI-1x data sets, coupled-cluster and density functional theory properties for molecules. *Sci. Data* **2020**, *7*, 134.

(52) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; Von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **2014**, *1*, 140022−140027.

(53) Hoja, J.; Medrano Sandonas, L.; Ernst, B. G.; Vazquez-Mayagoitia, A.; DiStasio, R. A.; Tkatchenko, A. QM7-X, a comprehensive dataset of quantum-mechanical properties spanning the chemical space of small organic molecules. *Sci. Data* **2021**, *8*, 43.

(54) Donchev, A. G.; Taube, A. G.; Decolvenaere, E.; Hargus, C.; McGibbon, R. T.; Law, K.-H.; Gregersen, B. A.; Li, J.-L.; Palmo, K.; Siva, K.; et al. Quantum chemical benchmark databases of gold-standard dimer interaction energies. *Sci. Data* **2021**, *8*, 55.

(55) Zhong, Y.; Yu, H.; Gong, X.; Xiang, H. A General Tensor Prediction Framework Based on Graph Neural Networks. *J. Phys. Chem. Lett.* **2023**, *14*, 6339−6348.

(56) 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.

(57) Heßelmann, A. DFT-SAPT intermolecular interaction energies employing exact-exchange Kohn−Sham response methods. *J. Chem. Theory Comput.* **2018**, *14*, 1943−1959.

(58) Misquitta, A. J.; Stone, A. J.; Price, S. L. Accurate induction energies for small organic molecules. 2. Development and testing of distributed polarizability models against SAPT (DFT) energies. *J. Chem. Theory Comput.* **2008**, *4*, 19−32.

(59) 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.

(60) Hirshfeld, F. L. Bonded-atom fragments for describing molecular charge densities. *Theor. Chim. Acta* **1977**, *44*, 129−138.

(61) Shi, Y.; Xia, Z.; Zhang, J.; Best, R.; Wu, C.; Ponder, J. W.; Ren, P. Polarizable atomic multipole-based AMOEBA force field for proteins. *J. Chem. Theory Comput.* **2013**, *9*, 4046−4063.

(62) Lillestolen, T.; Wheatley, R. First-principles calculation of local atomic polarizabilities. *J. Phys. Chem. A* **2007**, *111*, 11141−11146.

(63) 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*, 161702.

(64) Misquitta, A. J. CamCasp-bin. 2024, https://github.com/ajmisquitta/camcasp-bin.

(65) Huang, B.; von Lilienfeld, O. A. Quantum machine learning using atom-in-molecule-based fragments selected on the fly. *Nat. Chem.* **2020**, *12*, 945−951.

(66) Frisch, M. J. et al. *Gaussian 16* Revision C.01; Gaussian Inc: Wallingford CT, 2016.

(67) Smith, D. G.; Burns, L. A.; Simmonett, A. C.; Parrish, R. M.; Schieber, M. C.; Galvelis, R.; Kraus, P.; Kruse, H.; Di Remigio, R.; Alenaizan, A.; et al. PSI4 1.4: Open-source software for high-throughput quantum chemistry. *J. Chem. Phys.* **2020**, *152*, 184108.

(68) Lu, T.; Chen, F. Multiwfn: A multifunctional wavefunction analyzer. *J. Comput. Chem.* **2012**, *33*, 580−592.

(69) Verstraelen, T.; Vandenbrande, S.; Heidar-Zadeh, F.; Vanduyfhuys, L.; Van Speybroeck, V.; Waroquier, M.; Ayers, P. W. Minimal basis iterative stockholder: atoms in molecules for force-field development. *J. Chem. Theory Comput.* **2016**, *12*, 3894−3912.

(70) Lu, T.; Chen, F. Atomic dipole moment corrected Hirshfeld population method. *J. Theor. Comput. Chem.* **2012**, *11*, 163−183.