Published in final edited form as:

J Chem Theory Comput. 2019 February 12; 15(2): 1409–1417. doi:10.1021/acs.jctc.8b00895.

Toward Building Protein Force Fields by Residue-Based Systematic Molecular Fragmentation and Neural Network

Hao Wang and

Department of Chemistry, Duke University, Durham, North Carolina 27708, United States

Weitao Yang*

Department of Chemistry and Department of Physics, Duke University, Durham, North Carolina 27708, United States Key Laboratory of Theoretical Chemistry of Environment, Ministry of Education, School of Chemistry and Environment, South China Normal University, Guangzhou 510006, China

Abstract

Accurate force fields are crucial for molecular dynamics investigation of complex biological systems. Building accurate protein force fields from quantum mechanical (QM) calculations is challenging due to the complexity of proteins and high computational costs of QM methods. In order to overcome these two difficulties, here we developed the residue-based systematic molecular fragmentation method to partition general proteins into only 20 types of amino acid dipeptides and one type of peptide bond at level 1. The total energy of proteins is the combination of the energies of these fragments. Each type of the fragments is then parametrized using neural network (NN) representation of the QM reference. Adopting NN representation can circumvent the limitation of the analytic form of classical molecular mechanics (MM) force fields. Using MM force fields as the baseline, our method adds NN representation of QM corrections at the length scale of amino acid dipeptides. We tested our force fields for both homogeneous and heterogeneous polypeptides. Energy and forces predicted by our force fields compare favorably with full QM calculations from tripeptides to decapeptides. Our development provides an efficient and accurate method of building protein force fields fully from ab initio QM calculations.

GRAPHICAL ABSTRACT

*Corresponding Author weitao.yang@duke.edu.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the [ACS Publications website](https://pubs.acs.org) at DOI: [10.1021/acs.jctc.8b00895](https://doi.org/10.1021/acs.jctc.8b00895). Analysis of the approximation in eq 6, details of NN structure and rSMF at level 2 (PDF)

Notes

The authors declare no competing financial interest.

I. INTRODUCTION

Molecular dynamics (MD) simulations strongly depend on the accuracy of underlying force fields. Accurate protein force fields play an important role in the MD investigations of enzyme catalysis,^{1,2} protein folding,^{3,4} and protein–ligand binding.⁵ The development of protein force fields has seen tremendous progress in the past few decades. Commonly used all-atom protein force fields include but are not limited to CHARMM,⁶ AMBER,⁷ GROMOS,⁸ and OPLS.⁹ Protein force fields can be built either from fitting to experimental data or from quantum mechanical (QM) calculations. Classical protein molecular mechanics (MM) force fields divide the total energy into bond energy, angle energy, dihedral energy, electrostatic interaction energy, and van der Waals energy.² In conventional protein force field parametrization, some of these energy terms are fitted to QM calculations and some are fitted to experimental data. For example, in the AMBER ff94 force field the van der Waals energy is fitted to reproduce densities and enthalpies of vaporization of organic liquids while the dihedral energy term is fitted to relative energies of alanine and glycine dipeptide at the level of second-order Møller–Plesset perturbation theory.¹⁰ An important advance in the development of classical protein force fields in the past few years is the characterization of electronic polarization using classical models.^{11–17} It is now possible to directly obtain the parameters of polarizable force fields from linear response quantum mechanical calculations.¹⁸ Popular polarizable protein force fields, including but not limited to the Drude polarizable force field¹⁹ and the AMOEBA force field,^{20,21} have been successfully applied to both peptides and full proteins of different secondary structures. For example, the

J-coupling value from the AMOEBA force field for (Ala)₅ peptides was in good agreement with experimental values.²¹ However, there are still significant challenges in protein force fields. For example, the AMBER ff03 and CHARMM27 force fields favor helical structure; they can fold villin correctly, but are not able to fold the WW domain containing β -sheet.^{22,23} The current protein force fields also have limited ability to model the temperature dependency of the conformational propensities.²²

The main problem of conventional MM force fields is due to the limited MM force field form. Thus, we are particularly interested in going beyond the form of the conventional MM force fields to achieve higher accuracy. Force fields built on accurate QM calculations have been shown to be very successful for describing bulk water.^{24–28} Our aim is to build an accurate protein force field completely from QM calculations, which makes the force field consistent. Building protein force fields fully from QM calculations is challenging due to the complexity of proteins and high computational costs of QM methods. Fragmentation-based methods provide a possible route to circumvent these difficulties. Fragmentation-based methods partition large systems into small fragments, and the total energy of large systems is the combination of those of small fragments. QM calculations of small fragments can be easily handled. Thus, these fragmentation-based methods can reduce the computational costs with good accuracy. Many fragmentation-based methods have been developed in the literature such as the divide-and-conquer method,²⁹ many-body expansion,^{30–34} effective fragment potential approach,³⁵ systematic molecular fragmentation,^{36–38} electrostatically embedded generalized molecular fractionation with conjugate caps (EE-GMFCC),^{39–42} X-pol method,^{43–46} and energy-based fragmentation method.^{47,48} In order to construct force fields, it is necessary to parametrize the energy of these small fragments. However, proteins are diverse and direct applications of available fragmentation-based methods will produce too many different types of fragments to be easily parametrized. This motivates us to develop the residue-based systematic molecular fragmentation (rSMF) method in this work. At level 1 of the theory, the rSMF method generates only 21 types of small fragments for any general proteins, which makes it easier to parametrize and makes it transferable among different proteins.

To go beyond the limitation of traditional force fields, we use neural network (NN). Recent development in NN has made significant progress in diverse chemical applications, including determining NMR parameters in solid-state materials,⁴⁹ building the force field for water,²⁸ simulations of infrared spectra,⁵⁰ crystal structure prediction,⁵¹ quantitative structure–activity relationship,⁵² and enhancing the efficiency of ab initio QM/MM simulations.^{53–55} NN-based potentials are drawing more attention in MD simulations due to its high accuracy and flexibility. High-dimensional NN potentials (HDNNP) were developed by Behler and Parrinello.⁵⁶ In their work, high-dimensional potentials are expressed as the sum of atomic energy with feedforward NN representation. This strategy simplifies the structure of NNs and has been successfully applied to many problems. Gastegger and Marquetand applied HDNNP for the prediction of organic reactions.⁵⁷ Parrinello et al. combined metadynamics and HDNNP for high-pressure phases of silicon simulations.⁵⁸ Behler et al. applied HDNNP for simulations of solid–liquid interfaces.⁵⁹ Input features representing chemical environment for NN are of vital importance and should satisfy translational, rotational, and permutational invariance. HDNNP adopts radial and angular

symmetry functions to represent chemical environment.^{56,60} Many other features are developed in the literature including Coulomb matrix,^{61–63} bispectrum,^{64,65} permutation invariant polynomial,^{66,67} bag of bonds,⁶⁸ and Fourier series of atomic radial distribution functions.⁶⁹

Section II first presents the rSMF method for a tripeptide and then for a general protein at level 1. Section III describes the details of the adaptive construction of the training set of NN. Section IV shows the results of applying the rSMF method to both homogeneous polypeptides and heterogeneous polypeptides and discusses them. Section V summarizes the conclusions.

II. METHOD

In this work, we developed the rSMF method to partition a protein into small fragments and used NN representation of fragment energies for protein force field construction. Our rSMF method shares the same spirit of the original SMF method. Here we only summarize the main idea of the SMF method. Readers are referred to original papers for details.^{36–38} In the SMF method, atoms are typically grouped based on the functional group, though not always. Each atom group (i.e., fragment) contains more functional groups as the fragmentation level increases. The accuracy of predicted energy of the whole molecule, which is the combination of the energies of fragments, also increases with fragmentation level increasing. The lowest fragmentation level is called level 1, which normally contains two functional groups in one fragment. There are normally three functional groups in one fragment at level 2. We now describe how to decompose a protein into fragments using the rSMF method. For simplicity, we take a tripeptide as an example to show our decomposition method at level 1. A given tripeptide P^3 is represented as

$$P^3 = ACE - A_1 - A_2 - NME \quad (1)$$

where we capped protein termini with an acetyl group (ACE) and an *N*-methyl amide group (NME). This capping choice is mainly for symmetry consideration so that the boundary fragments are of the same form of inner fragments.

A graphical representation of rSMF decomposition of a tripeptide P^3 at level 1 is shown in Figure 1. This decomposition is called level 1 because there are at most two units in one fragment, which is similar to level 1 in the original SMF method. Note that each amino acid residue and peptide bond is separately treated as one unit in our method. Hydrogen atoms are used to terminate broken bonds in the following way:

$$X(H) = X(i) + \frac{r(i) + r(H)}{r(i) + r(j)}[X(j) - X(i)] \quad (2)$$

where $X(H)$ is the coordinates of capping hydrogen atoms, $X(i)$ is the coordinates of atom i , $X(j)$ is the coordinates of atom j , r denotes the standard covalent radius for each element, i is in the current atom, and $i \cdots j$ is the bond to be broken in forming the fragment. A tripeptide

P^3 is decomposed into three fragments A_1 , A_2 , and ACE–NME at level 1, where fragments A_1 and A_2 are amino acid dipeptides. Compared with the original form of the SMF method, there are two major differences. The first is how to group atoms. In our rSMF method, atoms are grouped based on residues rather than functional groups. The second is that we used MM force fields for nonbonded interactions instead of low level fragmentation. But calculating nonbonded interactions with low level fragmentation using QM methods may be important to further improve the accuracy, especially for closely separated fragments. This may be added in future work. We compared our rSMF method at level 2 (details in the Supporting Information) with the EE-GMFCC method, which also decomposes proteins into fragments based on residues. The main difference is the choice of capping atoms to broken bonds and places of broken bonds. The EE-GMFCC method cuts peptide bonds and uses conjugate caps to replace broken bonds. Thus, each fragment generated in the EE-GMFCC method has two side chains; i.e., each fragment is an amino acid tripeptide. However, only about half of the fragments generated in the rSMF method at level 2 have two side chains (tripeptides) and the remaining fragments only have one side chain (dipeptides). Fragments with two side chains will have up to 400 types for protein systems. Thus, it will be more challenging to construct the NN representation library for fragments with two side chains.

The total energy of P^3 ($E[P^3]$) consists of bonded ($E_b[P^3]$) and nonbonded ($E_{nb}[P^3]$) energies, expressed as

$$E[P^3] = E_b[P^3] + E_{nb}[P^3] \quad (3)$$

We evaluate bonded energies at QM level as

$$E_b[P^3] = E_{QM}[A_1] + E_{QM}[A_2] - E_{QM}[\text{ACE} - \text{NME}] \quad (4)$$

where each fragment is evaluated at QM level. Nonbonded energies containing all interactions not included in any single fragment are evaluated at MM level using protein MM force fields. This scheme shows QM corrections in our work focus on the length scale of amino acid dipeptides. As will be shown later, though focusing on amino acid dipeptides, our method compares well with corresponding full QM calculations of polypeptides. NN representation of bonded energies is used for the force field purpose. Each fragment energy calculated at QM level has two contributions:

$$E_{QM}[A_i] = E_{MM}[A_i] + E_{NN}[A_i] \quad (5)$$

where $E_{QM}[A_i]$ is the fragment energy evaluated at QM level, $E_{MM}[A_i]$ is the fragment energy evaluated with MM force fields, and $E_{NN}[A_i]$ is the NN energy corrections. For a general protein consisting of n amino acid residues, $P^n = \text{ACE}-A_1-\cdots-A_n-\text{NME}$, the total energy is calculated as

$$\begin{aligned}
 E[\mathbf{P}^n] &= E_b[\mathbf{P}^n] + E_{nb}[\mathbf{P}^n] \\
 &= \sum_{i=1}^n E_{QM}[A_i] - \sum_{i=1}^{n-1} E_{QM}[(ACE - NME)_{i,i+1}] + E_{nb,MM}[\mathbf{P}^n] \\
 &= \sum_{i=1}^n (E_{MM}[A_i] + E_{NN}[A_i]) - \sum_{i=1}^{n-1} (E_{MM}[(ACE - NME)_{i,i+1}] \\
 &\quad + E_{NN}[(ACE - NME)_{i,i+1}]) + E_{nb,MM}[\mathbf{P}^n] \\
 &\approx \sum_{i=1}^n E_{NN}[A_i] - \sum_{i=1}^{n-1} E_{NN}[(ACE - NME)_{i,i+1}] + E_{MM}[\mathbf{P}^n]
 \end{aligned}
 \tag{6}$$

where $(ACE-NME)_{i,i+1}$ is the ACE–NME fragment between residue i and residue $i+1$. At the last step in eq 6 we made a key approximation that the bonded fragment energy evaluated at MM level ($E_{MM}[A_i]$ and $E_{MM}[(ACE-NME)_{i,i+1}]$) and the nonbonded energy evaluated at MM level ($E_{nb,MM}[\mathbf{P}^n]$) are combined as the total energy of the protein at MM level ($E_{MM}[\mathbf{P}^n]$). This approximation is justified since careful investigations show that additional bond, angle, and dihedral energy terms related with capping hydrogen atoms in $E_{MM}[A_i]$, $E_{MM}[A_{i+1}]$, and $E_{MM}[(ACE-NME)_{i,i+1}]$ cancel each other exactly (this approximation is analyzed in detail in the Supporting Information). From another perspective, this approximation can also be regarded as unifying the MM-level fragmentation at level 1 as the total energy at MM level.

The idea that the bonded energy is evaluated at QM level while the nonbonded energy is evaluated at MM level has already been developed in the original SMF paper.⁷⁰ Here we further developed this idea by decomposing the QM-level bonded energy into MM contributions and NN corrections (eq 5). Then unify the MM-level bonded energy and nonbonded energy as the total energy at MM level. Thus, our NN corrections can be conveniently added on top of current popular MM force fields. The most appealing feature of our method is that we only need to construct NNs for 21 different types of small fragments, including 20 amino acid dipeptides and 1 ACE–NME fragment, for application to any protein systems.

The NN energy correction to each fragment ($E_{NN}[A_i]$ or $E_{NN}[(ACE-NME)_{i,i+1}]$) is expressed as the sum of atomic energy e_a :

$$E_{NN}[A_i/(ACE - NME)_{i,i+1}] = \sum_a e_a \tag{7}$$

The schematic representation of NN energy corrections of fragments is shown in Figure 2. The atomic energy e_a represented by two hidden layer feedforward NN is given as

$$e_a = \sum_{k=1}^{N_2} w_k^{23} \tanh \left[\sum_{j=1}^{N_1} w_{jk}^{12} \tanh \left(\sum_{i=1}^{N_0} w_{ij}^{01} G_i + b_j^1 \right) + b_k^2 \right] + b_0 \quad (8)$$

where G_i is the input variable for node i , w_{ij}^{01} (w_{jk}^{12}) is the weight parameter that connects node i (j) in the previous layer and node j (k) in the current layer, w_k^{23} is the weight parameter that connects node k in the second hidden layer and output layer, b_j^1 and b_k^2 are bias weight parameters of two hidden layers, b_0 is the bias weight parameter of output layer, N_0 is the number of nodes in input layer, and N_1 and N_2 are the number of nodes in two hidden layers.

To properly represent the atomic environment, many different features have been developed in the literature. In our previous study, we found a power spectrum is efficient in representing the atomic environment for water.²⁸ We apply it to protein systems in this work. A power spectrum is a subset of bispectrum, which was developed by Csányi and coworkers.^{64,65} To construct a power spectrum, we project the nuclear charge density $\rho(\mathbf{r}) = \sum_i \delta(\mathbf{r} - \mathbf{r}_i) Z_i$ onto screened evenly distributed Gaussian radial functions, and spherical harmonic functions:

$$\rho(r) = \sum_n \sum_{l=0} \sum_{m=-l}^l c_{nlm} f_c(r) g_n(r) Y_{lm}(\theta, \phi) \quad (9)$$

where $f_c(r)$ is the radial screening function, $g_n(r)$ is radial basis functions (evenly distributed Gaussian radial functions in this work), and $Y_{lm}(\theta, \phi)$ is spherical harmonic functions. Then the power spectrum is calculated as

$$\mathbf{c}_{nl}^\dagger \mathbf{c}_{nl} = \sum_{m=-l}^l c_{nlm}^* c_{nlm} \quad (10)$$

Details of NN structures and our choice of radial screening function and basis functions can be found in the Supporting Information.

III. COMPUTATIONAL DETAILS

The NN optimization object function, in weighted mean squared deviation, is defined as

$$L = \sum_i \left(|E_{\text{ref},i} - E_i|^2 + \frac{\alpha}{3N_i} |\mathbf{F}_{\text{ref},i} - \mathbf{F}_i|^2 \right) + \frac{1}{N_{\text{para}}} \sum_i w_i^2 \quad (11)$$

where α is the relative weight of forces with respect to energy and is set to 1 in this work. The last term is a regularization penalty added to prevent overfitting, where $\{w_i\}$ are NN

parameters. N_i is the number of atoms for each configuration, and N_{para} is the number of NN parameters. In this work, we trained NNs for the glycine dipeptide, the alanine dipeptide, and the ACE–NME fragment, separately. In order to cover all representative configurations of glycine and alanine dipeptides in the training set of NNs, we used two-dimensional umbrella sampling with backbone dihedral angle ϕ and ψ as the two collective variables. For the ACE–NME fragment, we used one-dimensional umbrella sampling with the only dihedral angle in the backbone as the collective variable. In both cases, each sampling window is spaced 18° apart with a force constant of $80 \text{ kcal/mol}\cdot\text{rad}^2$. Thus, total 400 sampling windows are used for each amino acid dipeptide and 20 sampling windows are used for the ACE–NME fragment. We used the NN ensemble method of Marquetand and coworkers⁵⁰ to run MD simulations. Two sets of NNs are trained for each amino acid dipeptide and the ACE–NME fragment. In the NN ensemble method, the energy and forces are computed as the average of the two sets of NN:

$$\bar{E} = \frac{1}{2}(E_1 + E_2) \quad (12)$$

$$\bar{\mathbf{F}} = \frac{1}{2}(\mathbf{F}_1 + \mathbf{F}_2) \quad (13)$$

MD simulations are carried out using the average forces. The prediction uncertainty of the NN ensemble is measured by

$$\sigma_E = \sqrt{(E_1 - \bar{E})^2 + (E_2 - \bar{E})^2} \quad (14)$$

The NN ensemble method has been proven to effectively reduce the error of energy and force prediction.⁵⁰ The final NN is constructed in an iterative way. First, configuration snapshots are taken from MM simulations using umbrella sampling to construct initial NNs. Second, MD simulations are carried out using umbrella sampling to enlarge the training set adaptively. In this process, we set an energy threshold E_t . When E_{MM} is less than E_t , MD simulations are carried out with NN ensemble corrections. Otherwise, MD simulations are carried out with pure MM force fields. The motivation of setting an energy threshold in the sampling process is that NN trained from the first several rounds may be very rough due to the limited number of snapshots in the training set. Direct application of such NN corrections at each MD step may lead to sampling abnormal configurations. These abnormal configurations may never be encountered in normal MD simulations. Setting up an energy threshold and gradually increasing it in the process of training set construction may alleviate this problem. In this process of training set construction, when σ_E is larger than 3 kcal/mol, the current snapshot is taken out and added into the training set. We iterate this step until σ_E is always less than 3 kcal/mol. Then we increase E_t and repeat. The final stopping E_t depends on our requirement. The flowchart is shown in Figure 3. In the production run, only when E_{MM} is less than E_t NN corrections are applied to the corresponding dipeptide.

Otherwise, pure MM force fields are applied to that dipeptide. The NN correction to (ACE–NME)_{*i,i+1*} is added only when both its nearby residues *i* and *i+1* are corrected by NN. The final E_t value for both glycine and alanine dipeptides is increased to be 45 kcal/mol, so few configurations encountered in MD production run are without NN corrections.

IV. RESULTS AND DISCUSSION

The accuracy of NN for the glycine dipeptide, the alanine dipeptide, and the ACE–NME fragment using the CHARMM22/CMAP force field^{71,72} as the MM baseline is shown in Table 1. Errors of energy and forces are measured using the root-mean-squared deviation (RMSD).

$$E_{\text{RMSD}} = \sqrt{\frac{1}{N} \sum_i (E_i - E_{\text{ref},i})^2} \quad (15)$$

$$F_{\text{RMSD}} = \sqrt{\frac{1}{N_f} \sum_i |\mathbf{F}_i - \mathbf{F}_{\text{ref},i}|^2} \quad (16)$$

where $E_{\text{ref},i}$ is the QM reference of energy for configuration *i*, $F_{\text{ref},i}$ is the QM reference of force for configuration *i*, *N* is the number of configurations, and N_f is the number of force components in all configurations. In order to show whether the predicted energy of NN is unbiased, we also calculated the mean signed error (MSE) in energy:

$$\bar{E} = \frac{1}{N} \sum_i (E_i - E_{\text{ref},i}) \quad (17)$$

If the MSE is close to zero, this indicates the predicted energy of NN is unbiased. We trained two sets of NNs for each type of fragments. For the glycine dipeptide, we used 889 configurations for training and 87 configurations for testing. For configurations in the training set, the RMSD of the energy error between QM and MM calculations before training is 7.97 kcal/mol and the RMSD of the force error is 25.73 kcal/mol/Å. For configurations in the testing set, the RMSD of the energy error before training is 7.26 kcal/mol and the RMSD of the force error is 24.05 kcal/mol/Å. In the calculation of the RMSD of the energy error between QM and MM results, the average of the energy error has been deducted since only the relative energy matters. The NN training accuracy of energy is about 0.7 kcal/mol, and the training accuracy of force is about 2.6 kcal/mol/Å.

For the alanine dipeptide, we used 1661 configurations for training and 188 configurations for testing. For configurations in the training set, the RMSD of the energy error between QM and MM calculations before training is 12.27 kcal/mol and the RMSD of the force error is 31.77 kcal/mol/Å. For configurations in the testing set, the RMSD of the energy error before training is 10.09 kcal/mol and the RMSD of the force error is 27.78 kcal/mol/Å. The NN

training accuracy of energy is about 0.9 kcal/mol, and the training accuracy of force is about 2.9 kcal/mol/Å.

For the ACE–NME fragment, we used 596 configurations for training and 81 configurations for testing. For configurations in the training set, the RMSD of the energy error between QM and MM calculations before training is 7.28 kcal/mol and the RMSD of the force error is 28.82 kcal/mol/Å. For configurations in the testing set, the RMSD of the energy error before training is 7.38 kcal/mol and the RMSD of the force error is 29.41 kcal/mol/Å. The NN training accuracy of energy is about 0.4 kcal/mol, and the training accuracy of force is about 1.6 kcal/mol/Å. All the NN testing set results for the three fragments show similar accuracies compared with the corresponding training set results. The MSEs in energy for the three fragments are all close to zero, which indicates our NN potential energy surface is unbiased.

We then applied our method of NN-corrected MM forces fields to alanine polypeptides (ACE–(Ala)_{*n*}–NME with *n* from 2 to 9) and mixed alanine/glycine polypeptides (Figure 4). Alanine polypeptides represent homogeneous systems. Mixed alanine and glycine polypeptides represent heterogeneous systems. We compared our results with two references. One is full QM calculations for the polypeptides, and the other is QM calculations using rSMF with nonbonded interactions evaluated using MM force fields. The latter reference tests the accuracy limit of our NN corrections. All our QM calculations are done with B3LYP-GD3BJ/6–31+g* using Gaussian 09,^{73–76} in which D3 dispersion with Becke–Johnson damping corrections are added to the B3LYP functional.^{77,78} For each polypeptide, we run *NVTMD* simulations for 10 ps at 300 K with the integration time step of 1 fs using CHARMM 22/CMAP force fields^{71,72} corrected by NNs. The Berendsen temperature coupling algorithm with a relaxation time of 0.1 ps is used to control the temperature.⁷⁹ Although the Berendsen thermostat cannot sample the correct canonical ensemble, it is widely used for system equilibrium due to its efficiency in relaxing systems to the target temperature. Our main purpose is to extract snapshots from MD simulations for NN training. Using the Berendsen temperature coupling algorithm does not affect our conclusions. When we use our method to calculate physical properties of large protein systems in the next step, we will adopt the correct thermostat, such as the Nosé–Hoover thermostat.^{80,81} MD simulations are performed with our in-house program QM4D.⁸² Configuration snapshots are saved every 10 steps for analysis.

Data for alanine polypeptides are shown in Table 2 and Figure 5. Data for mixed alanine and glycine polypeptides are shown in Table 3 and Figure 5. As we can see from Figure 5, the force error between fragmentation QM and full QM calculations increases linearly for small polypeptides and becomes stable for large polypeptides. This is mainly because the error caused by nonbonded interactions becomes stable when the system size is beyond a certain value. But the force error between NN and fragmentation QM or full QM calculations does not show an obvious linear increase for small alanine polypeptides. This is because of the error cancellation in the approximation introduced in eq 6. We observed that for large mixed alanine and glycine polypeptides our NN force field shows better accuracy than fragmentation QM calculations. In principle, the latter is our accuracy limit. We believe the better accuracy in comparison with full QM calculations is also due to the error cancellation

in the approximation we introduced in eq 6. For alanine polypeptides, the energy error between our NN force field and full QM calculations is about 1.2 kcal/mol/fragment, and the energy error between fragmentation QM calculations and full QM calculations is about 0.6 kcal/mol/fragment. The force error between our NN force field and full QM calculations is about 4 kcal/mol/Å, and the force error between fragmentation QM calculations and full QM calculations is about 2.5 kcal/mol/Å. For mixed alanine and glycine polypeptides, the energy error between our NN force fields and full QM calculations is about 0.6 kcal/mol/fragment, and the energy error between fragmentation QM calculations and full QM calculations is about 0.6 kcal/mol/fragment. The force error between our NN force field and full QM calculations is about 5 kcal/mol/Å, and the force error between fragmentation QM calculations and full QM calculations is about 3 kcal/mol/Å. Our testing results for both homogeneous polypeptides and heterogeneous polypeptides show high precision, considering the chemical complexity of polypeptides. The stability of errors with increasing system sizes is desirable and necessary for the development of force fields. In the current implementation, nonbonded interactions are evaluated at the MM level. But QM corrections to the nonbonded interactions can be added in a similar way as corrections added to bonded energy. In the original form of the SMF method, short-range nonbonded interactions are evaluated with a low level fragmentation method.³⁸ This strategy can further improve the accuracy.

In the current work, we focus on the fragmentation at rSMF level 1. Though increasing the fragmentation level can further improve the accuracy, it is much harder to build an NN library to parametrize each fragment at a higher fragmentation level. There are up to 400 different types of fragments at rSMF level 2. For this reason, we first plan to continue with rSMF level 1 for further development.

V. CONCLUSION

In summary, we have developed the rSMF-based NN correction method to MM force fields to build an accurate protein force field. Only QM calculations of small amino acid dipeptides and peptide bonds are needed to construct the protein force field, leading to the low computational cost in our approach and the feasibility for using high accuracy QM methods as input calculations. Our QM corrections focus on the length scale of amino acid dipeptides, and nonbonded interactions beyond dipeptides are evaluated with MM force fields. Our method scales linearly with system size. We tested our protein force field with both homogeneous polypeptides and heterogeneous polypeptides. The accuracy of our method is comparable with full QM calculations, and it shows stable energy and force errors with increasing system sizes. The most appealing feature of our method at level 1 is that the parametrization of only 21 different types of fragments including 20 amino acid dipeptides and 1 ACE–NME fragment is needed for any protein systems. Our method provides a new way to build protein force fields, which circumvents the limited form of conventional MM force fields and holds the promise of much higher accuracy.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding

Financial support from the National Institutes of Health (Grant R01 GM061870–13) is gratefully acknowledged.

REFERENCES

- (1). Karplus M; McCammon JA Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol* 2002, 9, 646. [PubMed: 12198485]
- (2). Adcock SA; McCammon JA Molecular Dynamics: Survey of Methods for Simulating the Activity of Proteins. *Chem. Rev* 2006, 106, 1589. [PubMed: 16683746]
- (3). Sugita Y; Okamoto Y Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett* 1999, 314, 141.
- (4). Swope WC; Pitera JW; Suits F Describing Protein Folding Kinetics by Molecular Dynamics Simulations. 1. Theory. *J. Phys. Chem. B* 2004, 108, 6571.
- (5). Woo H-J; Roux B Calculation of absolute protein–ligand binding free energy from computer simulations. *Proc. Natl. Acad. Sci. U. S. A* 2005, 102, 6825. [PubMed: 15867154]
- (6). Brooks BR; et al. CHARMM: The biomolecular simulation program. *J. Comput. Chem* 2009, 30, 1545. [PubMed: 19444816]
- (7). Case D; et al. AMBER 2018; University of California: San Francisco, 2018.
- (8). Scott WRP; Hünenberger PH; Tironi IG; Mark AE; Billeter SR; Fennen J; Torda AE; Huber T; Krüger P; van Gunsteren WF The GROMOS Biomolecular Simulation Program Package. *J. Phys. Chem. A* 1999, 103, 3596.
- (9). Jorgensen WL; Maxwell DS; 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.
- (10). Cornell WD; Cieplak P; Bayly CI; Gould IR; Merz KM; Ferguson DM; Spellmeyer DC; Fox T; Caldwell JW; Kollman PA A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc* 1995, 117, 5179.
- (11). Rappe AK; Goddard WA Charge equilibration for molecular dynamics simulations. *J. Phys. Chem* 1991, 95, 3358.
- (12). Rick SW; Stuart SJ; Berne BJ Dynamical fluctuating charge force fields: Application to liquid water. *J. Chem. Phys* 1994, 101, 6141.
- (13). Drude P *The Theory of Optics*; Longmans, Green, and Co: New York, 1902.
- (14). Huang K; Born M *Dynamic Theory of Crystal Lattices*; Oxford University Press: Oxford, U.K, 1954.
- (15). Straatsma TP; McCammon JA Molecular Dynamics Simulations with Interaction Potentials Including Polarization Development of a Noniterative Method and Application to Water. *Mol. Simul* 1990, 5, 181.
- (16). Vesely FJ N-particle dynamics of polarizable Stockmayer-type molecules. *J. Comput. Phys* 1977, 24, 361.
- (17). Van Belle D; Froeyen M; Lippens G; Wodak SJ Molecular dynamics simulation of polarizable water by an extended Lagrangian method. *Mol. Phys* 1992, 77, 239.
- (18). Wang H; Yang W Determining polarizable force fields with electrostatic potentials from quantum mechanical linear response theory. *J. Chem. Phys* 2016, 144, 224107. [PubMed: 27305996]
- (19). Lopes PEM; Huang J; Shim J; Luo Y; Li H; Roux B; MacKerell AD Polarizable Force Field for Peptides and Proteins Based on the Classical Drude Oscillator. *J. Chem. Theory Comput* 2013, 9, 5430. [PubMed: 24459460]
- (20). Ren P; Ponder JW Consistent treatment of inter- and intramolecular polarization in molecular mechanics calculations. *J. Comput. Chem* 2002, 23, 1497. [PubMed: 12395419]

- (21). Shi Y; Xia Z; Zhang J; Best R; Wu C; Ponder JW; Ren P Polarizable Atomic Multipole-Based AMOEBA Force Field for Proteins. *J. Chem. Theory Comput* 2013, 9, 4046. [PubMed: 24163642]
- (22). Lindorff-Larsen K; Maragakis P; Piana S; Eastwood MP; Dror RO; Shaw DE Systematic Validation of Protein Force Fields against Experimental Data. *PLoS One* 2012, 7, e32131. [PubMed: 22384157]
- (23). Piana S; Lindorff-Larsen K; Shaw D How Robust Are Protein Folding Simulations with Respect to Force Field Parameterization? *Biophys. J* 2011, 100, L47. [PubMed: 21539772]
- (24). 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. [PubMed: 26592277]
- (25). Babin V; Medders GR; Paesani F Development of a “First Principles” Water Potential with Flexible Monomers. II: Trimer Potential Energy Surface, Third Virial Coefficient, and Small Clusters. *J. Chem. Theory Comput* 2014, 10, 1599. [PubMed: 26580372]
- (26). Reddy SK; Straight SC; Bajaj P; Huy Pham C; Riera M; Moberg DR; Morales MA; Knight C; Götz AW; 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. [PubMed: 27875875]
- (27). Cisneros GA; Wikfeldt KT; Ojamäe L; Lu J; Xu Y; Torabifard H; Bartok AP; Csányi G; Molinero V; Paesani F Modeling Molecular Interactions in Water: From Pairwise to Many-Body Potential Energy Functions. *Chem. Rev* 2016, 116, 7501. [PubMed: 27186804]
- (28). Wang H; Yang W Force Field for Water Based on Neural Network. *J. Phys. Chem. Lett* 2018, 9, 3232. [PubMed: 29775313]
- (29). Yang W Direct calculation of electron density in density-functional theory. *Phys. Rev. Lett* 1991, 66, 1438. [PubMed: 10043209]
- (30). Hankins D; Moskowitz JW; Stillinger FH Water Molecule Interactions. *J. Chem. Phys* 1970, 53, 4544.
- (31). Mayhall NJ; Raghavachari K Many-Overlapping-Body (MOB) Expansion: A Generalized Many Body Expansion for Nondisjoint Monomers in Molecular Fragmentation Calculations of Covalent Molecules. *J. Chem. Theory Comput* 2012, 8, 2669. [PubMed: 26592112]
- (32). Raghavachari K; Saha A Accurate Composite and Fragment-Based Quantum Chemical Models for Large Molecules. *Chem. Rev* 2015, 115, 5643. [PubMed: 25849163]
- (33). Richard RM; Herbert JM A generalized many-body expansion and a unified view of fragment-based methods in electronic structure theory. *J. Chem. Phys* 2012, 137, 064113. [PubMed: 22897261]
- (34). Dahlke EE; Truhlar DG Electrostatically Embedded Many-Body Expansion for Large Systems, with Applications to Water Clusters. *J. Chem. Theory Comput* 2007, 3, 46. [PubMed: 26627150]
- (35). Gordon MS; Fedorov DG; Pruitt SR; Slipchenko LV Fragmentation Methods: A Route to Accurate Calculations on Large Systems. *Chem. Rev* 2012, 112, 632. [PubMed: 21866983]
- (36). Deev V; Collins MA Approximate ab initio energies by systematic molecular fragmentation. *J. Chem. Phys* 2005, 122, 154102. [PubMed: 15945620]
- (37). Addicoat MA; Collins MA Accurate treatment of nonbonded interactions within systematic molecular fragmentation. *J. Chem. Phys* 2009, 131, 104103.
- (38). Collins MA Molecular forces, geometries, and frequencies by systematic molecular fragmentation including embedded charges. *J. Chem. Phys* 2014, 141, 094108. [PubMed: 25194365]
- (39). Zhang DW; Zhang JZH Molecular fractionation with conjugate caps for full quantum mechanical calculation of protein-molecule interaction energy. *J. Chem. Phys* 2003, 119, 3599.
- (40). Gao AM; Zhang DW; Zhang JZ; Zhang Y An efficient linear scaling method for ab initio calculation of electron density of proteins. *Chem. Phys. Lett* 2004, 394, 293.
- (41). He X; Zhang JZH A new method for direct calculation of total energy of protein. *J. Chem. Phys* 2005, 122, 031103.
- (42). Liu J; Zhu T; Wang X; He X; Zhang JZH Quantum Fragment Based ab Initio Molecular Dynamics for Proteins. *J. Chem. Theory Comput* 2015, 11, 5897. [PubMed: 26642993]

- (43). Xie W; Gao J Design of a Next Generation Force Field: The X-POL Potential. *J. Chem. Theory Comput* 2007, 3, 1890. [PubMed: 18985172]
- (44). Xie W; Orozco M; Truhlar DG; Gao J X-Pol Potential: An Electronic Structure-Based Force Field for Molecular Dynamics Simulation of a Solvated Protein in Water. *J. Chem. Theory Comput* 2009, 5, 459. [PubMed: 20490369]
- (45). Gao J Toward a Molecular Orbital Derived Empirical Potential for Liquid Simulations. *J. Phys. Chem. B* 1997, 101, 657.
- (46). Gao J A molecular-orbital derived polarization potential for liquid water. *J. Chem. Phys* 1998, 109, 2346.
- (47). Li W; Li S; Jiang Y Generalized Energy-Based Fragmentation Approach for Computing the Ground-State Energies and Properties of Large Molecules. *J. Phys. Chem. A* 2007, 111, 2193. [PubMed: 17388268]
- (48). Hua S; Li W; Li S The Generalized Energy-Based Fragmentation Approach with an Improved Fragmentation Scheme: Benchmark Results and Illustrative Applications. *ChemPhysChem* 2013, 14, 108. [PubMed: 23239545]
- (49). Cuny J; Xie Y; Pickard CJ; Hassnani AA Ab Initio Quality NMR Parameters in Solid-State Materials Using a High-Dimensional Neural-Network Representation. *J. Chem. Theory Comput* 2016, 12, 765. [PubMed: 26730889]
- (50). Gastegger M; Behler J; Marquetand P Machine learning molecular dynamics for the simulation of infrared spectra. *Chem. Sci* 2017, 8, 6924. [PubMed: 29147518]
- (51). Fayos J; Cano FH Crystal-Packing Prediction by Neural Networks. *Cryst. Growth Des* 2002, 2, 591.
- (52). So S-S; Karplus M Evolutionary Optimization in Quantitative Structure-Activity Relationship: An Application of Genetic Neural Networks. *J. Med. Chem* 1996, 39, 1521. [PubMed: 8691483]
- (53). Shen L; Wu J; Yang W Multiscale Quantum Mechanics/Molecular Mechanics Simulations with Neural Networks. *J. Chem. Theory Comput* 2016, 12, 4934. [PubMed: 27552235]
- (54). Wu J; Shen L; Yang W Internal force corrections with machine learning for quantum mechanics/molecular mechanics simulations. *J. Chem. Phys* 2017, 147, 161732. [PubMed: 29096448]
- (55). Shen L; Yang W Molecular Dynamics Simulations with Quantum Mechanics/Molecular Mechanics and Adaptive Neural Networks. *J. Chem. Theory Comput* 2018, 14, 1442. [PubMed: 29438614]
- (56). Behler J; Parrinello M Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett* 2007, 98, 146401. [PubMed: 17501293]
- (57). Gastegger M; Marquetand P High-Dimensional Neural Network Potentials for Organic Reactions and an Improved Training Algorithm. *J. Chem. Theory Comput* 2015, 11, 2187. [PubMed: 26574419]
- (58). Behler J; Marto ák R; Donadio D; Parrinello M Metadynamics Simulations of the High-Pressure Phases of Silicon Employing a High-Dimensional Neural Network Potential. *Phys. Rev. Lett* 2008, 100, 185501. [PubMed: 18518388]
- (59). Natarajan SK; Behler J Neural network molecular dynamics simulations of solid-liquid interfaces: water at low-index copper surfaces. *Phys. Chem. Chem. Phys* 2016, 18, 28704. [PubMed: 27722603]
- (60). Behler J Representing potential energy surfaces by high-dimensional neural network potentials. *J. Phys.: Condens. Matter* 2014, 26, 183001. [PubMed: 24758952]
- (61). Rupp M; Tkatchenko A; Müller K-R; von Lilienfeld OA Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett* 2012, 108, 058301. [PubMed: 22400967]
- (62). Hansen K; Montavon G; Biegler F; Fazli S; Rupp M; Scheffler M; von Lilienfeld OA; Tkatchenko A; Müller K-R Assessment and Validation of Machine Learning Methods for Predicting Molecular Atomization Energies. *J. Chem. Theory Comput* 2013, 9, 3404. [PubMed: 26584096]
- (63). Faber F; Lindmaa A; von Lilienfeld OA; Armiento R Crystal structure representations for machine learning models of formation energies. *Int. J. Quantum Chem* 2015, 115, 1094.

- (64). Bartók AP; Payne MC; Kondor R; Csányi G Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons. *Phys. Rev. Lett* 2010, 104, 136403. [PubMed: 20481899]
- (65). Bartók AP; Kondor R; Csányi G On representing chemical environments. *Phys. Rev. B: Condens. Matter Mater. Phys* 2013, 87, 184115.
- (66). Jiang B; Guo H Permutation invariant polynomial neural network approach to fitting potential energy surfaces. *J. Chem. Phys* 2013, 139, 054112. [PubMed: 23927248]
- (67). Jiang B; Guo H Permutation invariant polynomial neural network approach to fitting potential energy surfaces. III. Molecule-surface interactions. *J. Chem. Phys* 2014, 141, 034109. [PubMed: 25053303]
- (68). Hansen K; Biegler F; Ramakrishnan R; Pronobis W; von Lilienfeld OA; Müller K-R; Tkatchenko A Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space. *J. Phys. Chem. Lett* 2015, 6, 2326. [PubMed: 26113956]
- (69). von Lilienfeld OA; Ramakrishnan R; Rupp M; Knoll A Fourier series of atomic radial distribution functions: A molecular fingerprint for machine learning models of quantum chemical properties. *Int. J. Quantum Chem* 2015, 115, 1084.
- (70). Collins MA; Deev VA Accuracy and efficiency of electronic energies from systematic molecular fragmentation. *J. Chem. Phys* 2006, 125, 104104. [PubMed: 16999512]
- (71). MacKerell AD; et al. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* 1998, 102, 3586. [PubMed: 24889800]
- (72). Mackerell AD; Feig M; Brooks CL Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J. Comput. Chem* 2004, 25, 1400. [PubMed: 15185334]
- (73). Frisch MJ; et al. Gaussian 09, revision D.01; Gaussian, Inc: Wallingford, CT, 2016.
- (74). Becke AD Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys* 1993, 98, 5648.
- (75). Lee C; Yang W; Parr RG Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B: Condens. Matter Mater. Phys* 1988, 37, 785.
- (76). Stephens PJ; Devlin FJ; Chabalowski CF; Frisch MJ Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *J. Phys. Chem* 1994, 98, 11623.
- (77). Johnson ER; Becke AD A post-Hartree-Fock model of intermolecular interactions: Inclusion of higher-order corrections. *J. Chem. Phys* 2006, 124, 174104. [PubMed: 16689564]
- (78). Grimme S; Ehrlich S; Goerigk L Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem* 2011, 32, 1456. [PubMed: 21370243]
- (79). Berendsen HJC; Postma JPM; van Gunsteren WF; DiNola A; Haak JR Molecular dynamics with coupling to an external bath. *J. Chem. Phys* 1984, 81, 3684.
- (80). Nosé S A unified formulation of the constant temperature molecular dynamics methods. *J. Chem. Phys* 1984, 81, 511–519.
- (81). Hoover WG Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A: At., Mol., Opt. Phys* 1985, 31, 1695–1697.
- (82). Hu H; Hu X; Yang W An integrated and versatile quantum mechanical/molecular mechanical simulation package. 2016 <http://www.qm4d.info/>.

Figure 1.

Graphical representation of the residue-based systematic molecular fragmentation of a tripeptide at level 1. Bonds with arrow marks are broken in fragmentation.

Figure 2.

Schematic representation of an N atom fragment energy as the sum of atomic energies e_i . $\{G_i\}$ is the power spectrum for each atom.

Figure 3. Flowchart of neural network (NN) update process for dipeptides. E_t is the energy threshold. US MD means molecular dynamics simulations with umbrella sampling. Only when the molecular energy is less than E_t NN corrections are added. Otherwise MD is run with molecular mechanics force fields. σ_E measures uncertainty of energy prediction. NN is updated unless σ_E is always less than 3 kcal/mol in the MD simulations.

Figure 4.
Structures of mixed alanine/glycine polypeptides, from tripeptide to decapeptide.

Figure 5.
Energy and force errors of alanine polypeptides and mixed alanine/glycine polypeptides.

Table 1.

Energy and Force Accuracies of Neural Networks for Glycine Dipeptide, Alanine Dipeptide, and ACE2013NME Fragment^a

	glycine dipeptide				alanine dipeptide				ACE-NME fragment			
	NN1	training	testing	NN2	NN1	training	testing	NN2	NN1	training	testing	NN2
energy (kcal/mol)	0.73	0.95	0.71	0.85	0.92	1.07	0.92	1.02	0.39	0.51	0.42	0.53
force (kcal/mol/Å)	2.59	3.06	2.58	3.10	2.88	3.12	2.88	3.14	1.66	1.81	1.66	1.84
MSE (kcal/mol)	0.08	0.09	0.00	-0.02	0.10	0.06	-0.03	-0.14	-0.01	-0.09	0.03	-0.03

^aThe energy and force accuracies are measured with the root-mean-squared deviation. MSE represents the mean signed error in energy. For each type of fragment there are two neural networks, which will be averaged in molecular dynamics simulations.

Table 2.Alanine Polypeptide Data^a

alanine polypeptides	no. of fragments	NN-QM		NN-frag QM		frag QM-QM	
		E_{RMSD}	F_{RMSD}	E_{RMSD}	F_{RMSD}	E_{RMSD}	F_{RMSD}
3	3	1.69	4.27	1.27	4.20	0.50	0.96
4	5	1.14	3.80	0.87	3.76	0.45	1.62
5	7	0.83	4.35	0.43	3.88	0.63	2.15
6	9	1.36	4.51	0.72	3.89	0.70	2.73
7	11	0.96	4.61	0.56	3.89	0.46	2.56
8	13	1.34	4.74	0.95	4.29	0.42	2.36
9	15	1.04	4.75	0.56	4.10	0.51	2.97
10	17	1.27	4.41	0.68	3.98	0.63	2.54

^aResults of fragmentation with neural network are compared with full QM calculations and QM calculations with fragmentation (non-bonded interactions evaluated using molecular mechanics force fields). E_{RMSD} is the root-mean-squared deviation (RMSD) of energy in units of kcal/mol/fragment. F_{RMSD} values are RMSDs of forces in units of kcal/mol/Å.

Table 3.Mixed Alanine/Glycine Polypeptide Data^a

mixed alanine/glycine polypeptides	no. of fragments	NN-QM		NN-frag QM		frag QM-QM	
		E_{RMSD}	F_{RMSD}	E_{RMSD}	F_{RMSD}	E_{RMSD}	F_{RMSD}
3	3	0.93	3.38	0.64	3.32	0.38	1.19
4	5	0.74	3.82	0.35	3.65	0.68	1.55
5	7	0.86	4.04	0.41	3.86	0.61	1.63
6	9	0.70	4.92	0.25	4.02	0.67	3.10
7	11	0.54	5.00	0.29	4.11	0.60	3.33
8	13	0.51	4.91	0.38	4.06	0.75	3.17
9	15	0.61	4.92	0.39	4.13	0.77	3.41
10	17	0.59	5.13	0.34	4.28	0.71	2.88

^aResults of fragmentation with neural network are compared with full QM calculations and QM calculations with fragmentation (non-bonded interactions evaluated using molecular mechanics force fields). E_{RMSD} is the root-mean-squared deviation (RMSD) of energy in units of kcal/mol/fragment. F_{RMSD} values are the RMSDs of forces in units of kcal/mol/Å.