Building Machine Learning Force Fields of Proteins with Fragment-based Approach and Data Transfer

Zheng Cheng, Jiahui Du, Lei Zhang, Jing Ma,* Wei Li,* and Shuhua Li*

ABSTRACT: We combined our generalized energy-based fragmentation (GEBF) approach and transfer learning technique to construct machine learning force fields (MLFFs) for proteins only from quantum mechanics (QM) calculations of small subsystems. Using a kernel-based model called Gaussian Approximation Potential (GAP), our protocol can automatically generate training sets with high efficiency. To facilitate the construction of training sets for various proteins, a protein's data library is created to store all data of subsystems generated from trained proteins. With this data library, for a new protein only its subsystems with new topological types are required for the construction of the corresponding training set. With two polypeptides, 4ZNN and 1XQ8 segment, as examples, the energies and forces predicted by GEBF-GAP are in good agreement with those from QM calculations, and dihedral angle distributions from GEBF-GAP molecular dynamics (MD) simulations can also well reproduce those from *ab initio* MD simulations. In addition, with the training set generated from GEBF-GAP, we also demonstrate that GEBF-MLFFs can also be constructed by neural network (NN) methods with full QM quality. Therefore, the present work provides an efficient and systematic way to build force fields for biological systems like proteins with QM accuracy.

1. INTRODUCTION

Molecular dynamics (MD) simulation has emerged as an important tool to understand how the structure of a protein molecule determines its function in a cell. Currently, MD simulations with the classical force fields[1-6] have been widely applied for large biomolecules including proteins.[7,8] However, the accuracy of classical force fields is still insufficient for reliable descriptions of some proteins. For example, the α-helical propensity is underestimated by the AMBER99SB force field compared to the corresponding experimental values.[9] The classical force fields cannot accurately describe temperature-dependent folding.[10]Nowadays, the machine learning (ML) method has been increasingly applied to develop more accurate atomistic potentials with very general functional forms than the conventional force fields with physically inspired functional forms.[11-20] The resulting machine learning potentials, also called ML force fields (MLFFs), have been demonstrated to be quite successful for a variety of different systems.[21-34] By "learning" from reference data sets obtained from QM calculations for a given system or a type of systems, MLFFs may reach similar accuracy as QM methods at a cost which is orders of magnitude less than that required for QM calculations of the same system.

Due to the chemical complexities of proteins and the high computational costs of QM methods for large systems, building MLFFs for proteins remains a great challenge. Energy-based fragmentation (EBF) approaches[35-45] provide a practical and attractive solution to overcome these two difficulties. With this approach, the ground-state MLFF of a large system can be obtained as the linear combination of MLFF trained from small subsystems, which are representations of different local regions of a large system. In previous studies, a residue-based neural work (NN) approach[46,47] was proposed to construct NN potentials for 20 types of amino acid capped with an acetyl group (ACE) and *N*-methyl amid group (NME) and 1 type of ACE-NME, as shown in Figure 1. Then, the MLFFs of a protein are expressed as the linear combination of these NN potentials. The resulting ML potentials represent the first step towards *ab initio* quality protein force fields. However, the local regions on these subsystems are not same as the target system. Thus, these potentials are not yet accurate enough, with the root-mean-square errors (RMSEs) for the energy and forces of $(Ala)_9$ being 0.15 kcal/(mol·atom) and 4.75 kcal/(mol·Å), respectively, with respect to reference density functional theory (DFT) data.[47] Based on the generalized energy-based fragmentation (GEBF) approach developed by our group[35], we also constructed MLFFs for alkanes with the linear combination of MLFFs of small subsystems trained individually in our previous work.[48] Our previous scheme may be suitable for simple biomolecules like cellulose. However, proteins have twenty types of amino acid residues and too many different types of subsystems will be generated. It is difficult to construct MLFFs of all kinds of subsystems individually according to the previous fragment-based ML scheme.[46-48]

In this work, we propose a new protocol to construct MLFFs for proteins with full QM accuracy only from QM calculations on small subsystems. To circumvent the difficulty of MLFFs construction for enormous types of subsystems in previous fragment-based ML schemes,[46-48] a new strategy is adopted here by fitting the energy (or forces) of a given protein as the summation of atomic contributions from QM calculations of various subsystems. To facilitate the construction of MLFF for various proteins, a protein's data library is created to store all data of subsystems generated from trained proteins. For a new protein, a subset of subsystems with the same topological types that are already in the protein's data library can be directly taken as a part of the training set, together with some newly generated subsystems. To automatically collect the training set, an online active learning[48] is adopted here to generate these new subsystems for studied protein. Then, full-QM quality GEBF-MLFF can be constructed using either kernel model like GAP[12] or NN model like Deep Potential[17] with the training set generated by GEBF-

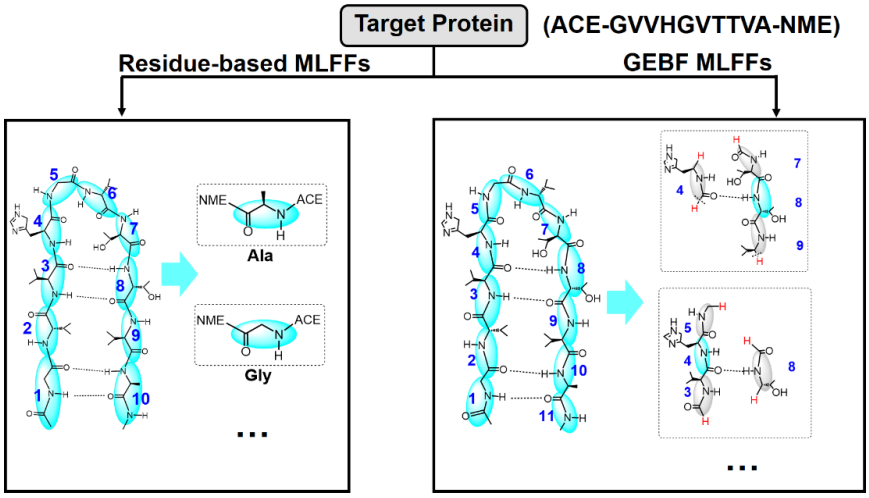


Figure 1. Fragmentation scheme utilized in the construction of MLFFs. In our GEBF method, fragments are capped with their environmental fragments or hydrogen atoms if necessary. In the previous residue-based method, fragments are capped with an acetyl group (ACE) and *N*-methylamine group (NME).

ML protocol. Our protocol is applied on two polypeptides (4ZNN and 1XQ8 segment) to construct the corresponding GEBF-MLFFs and their accuracy and efficiency are validated with reference QM calculations. The results indicate that the GEBF-MLFFs can reproduce QM results very well at speeds several orders of magnitude faster than *ab initio* calculations. We expect that this protocol will greatly promote the development of fast and accurate MLFFs for various biological systems.

The remainder of this paper is organized as follows. In Section 2, we describe the theoretical foundations of GEBF-MLFFs and the data transfer approach. In Section 3, the accuracy and computation costs are demonstrated by applying the GEBF-ML method to two polypeptides. In Section 4, a brief summary is presented.

2. METHODOLOGY

2.1. The GEBF-ML Force Fields. To automatically construct the subsystems on the training set, the GEBF approach developed by our group is adopted. The generation of subsystems for a polypeptide 4ZNN is also illustrated in Figure 1, we will generate various subsystems, each of which contains a fragment and its neighboring fragments and capping hydrogen atoms if necessary (in grey oval). Clearly, subsystems constructed in this way are better representations of the local chemical environment of different regions in a protein than those in the residue-based NN approach.

As the differences between QM and PM6 methods are generally much smaller than absolute QM values, the errors of ML models can be reduced if the PM6 method is used as the baseline.[49,50] In this work, this strategy is adopted, and an atomic ML model called GAP[12] based on kernel ridge regression with the SOAP kernels[51] (see details in the Sec.1 of the supporting information) is chosen to learn the energy difference of all subsystems for the studied proteins. To illustrate the advantage and disadvantage of this approach, GEBF-MLFFs of the 1XQ8 segment is also constructed by learning the QM energy data of corresponding subsystems directly. The approach to construct GEBF-MLFFs from QM energies is similar to the method for total energy difference prediction explained below, details can be seen in the Sec.1 of the supporting information. In addition, the Deep Potential method[19] is also adopted to construct GEBF-MLFFs with NN methods. In the GAP or Deep Potential method, the energy difference ΔE_m^{ML} of the *m*th subsystem with S_m atoms are described as the summation of atomic energy e_i^m ,

$$\Delta E_m^{\mathrm{ML}} = E_m^{\mathrm{DFT}} - E_m^{\mathrm{PM6}} = \sum_{i \in S_m} e_i^m \qquad (1)$$

For GAP, the atomic energy e_i can be express as

$$e_i = \sum_{i_B}^{N_B} w_{i_B} K(\mathbf{X}_i, \mathbf{X}_{i_B}) \qquad (2)$$

Here, N_B is the number of representative local atomic environments, w_{i_B} is the weight factor. SOAP is used to describe the local atomic environment $\mathbf{X}_i$ and the kernel *K*.

For Deep Potential, the atomic energy e_i can be represented as a neural network. Using a two hidden layer feedforward NN as an example, the atomic energy can be expressed as

$$e_i = \sum_{k=1}^{N_2} w_k^{23} \tanh\left[\sum_{j=1}^{N_1} w_{jk}^{12} \tanh\left(\sum_{u=1}^{N_0} w_{uj}^{01} G_u + b_j^1\right) + b_k^2\right] + b_0 \qquad (3)$$

Where G is the output of a local embedding network to describe the atomic environment.[19] $w_{uj}^{01}(w_{jk}^{12})$ is the weight factor that connects node *u* (*j*) in the previous layer and node *j* (*k*) in the current layer, w_k^{23} is the weight factor that connects node *k*

in the second hidden layer and output layer, b_j^1, b_k^2 and b^0 is the bias weight factor, N_0, N_1 and N_2 is the number of nodes for input layer and two hidden layers, respectively.

During the training, the energy differences for a set of reference subsystems are fitted by GAP or Deep Potential to determine the weight factor and bias factor. After training, the weight and bias factors are fixed. Although only energy differences of subsystems are trained, the energy contribution of each atom with different local environments in subsystems can be predicted by the atomic ML models. Based on the similarity of atomic environments between subsystems and the target protein, the total energy difference of the target system is obtained with the summation of atomic contribution e_i directly.

$$\Delta E^{\mathrm{ML}} = \sum_{i=1}^{N} e_i \tag{4}$$

Here, N is the number of atoms for the target protein and e_i is the atomic contribution of atom *i* with the local environment in the target protein

The total energy of the target system is the combination of the energy difference ΔE^{ML} and the PM6 energy E^{PM6} (taken as the baseline)

$$E = \Delta E^{\mathrm{ML}} + E^{\mathrm{PM6}} \tag{5}$$

The PM6 energy of the target system with *M* subsystems are evaluated with the GEBF method by the linear combination of subsystem energy E_m (C_m is the coefficient of each subsystem)

$$E^{\mathrm{PM6}} = \sum_m C_m \left(E_m^{\mathrm{PM6}} - \sum_{A \in S_m} \sum_{B > A \in S_m} \frac{Q_A Q_B}{\mathbf{R}_{AB}} \right) + \sum_A \sum_{B>A} \frac{Q_A Q_B}{\mathbf{R}_{AB}} \tag{6}$$

Details of subsystem construction and determination of coefficients are explained in Sec.3 of supporting information. The long-range nonbonded interactions between each subsystem and background charges on distant atoms are treated as the Coulomb interaction. The point charges are obtained from the natural population analysis (NPA) of subsystems, which are generated from the initial structure (extended structure generated from peptide sequence using Amber 16 program[52]) used in the online training process. After training, the point charges are assumed to be constant like in traditional force fields[1-4]. $\mathbf{r}_A$ and Q_A denote the coordinate of atom A and the point charge locating on atom A, respectively.

2.2. Data Transfer Approach. Because a subset of subsystems generates from a protein may have the same topological structure in chemical space as those from another protein, we may introduce instance-based transfer learning[53] to avoid redundant QM calculations on these subsystems. The flowchart of the scheme is shown in Figure 2. In our approach, we create a protein's data library, which contains all data of subsystems generated from trained proteins. Starting from a given conformer of a new protein, MD simulation with NVT ensemble is performed based on the GEBF-MLFFs. As GAP can give internal uncertainty of Gaussian process regression model, which both consider the sampling density in the conformation space and the actual shape of the potential energy landscape, the ML model is chosen as GAP when the training sets are constructed. During the simulation, subsystems are automatically generated using our GEBF approach. If the subsystem types are already in the data library (the details of subsystem discrimination can be found in the Sec.4 of the supporting information), the corresponding sub-datasets are loaded to the training set. Otherwise, online active learning[48] (see details in Sec.5 of supporting information) is employed to select the representative subsystem conformers. When the training set is updated, the GEBF-ML force fields are also renewed to fit the energies and forces of conformers explored by online training.

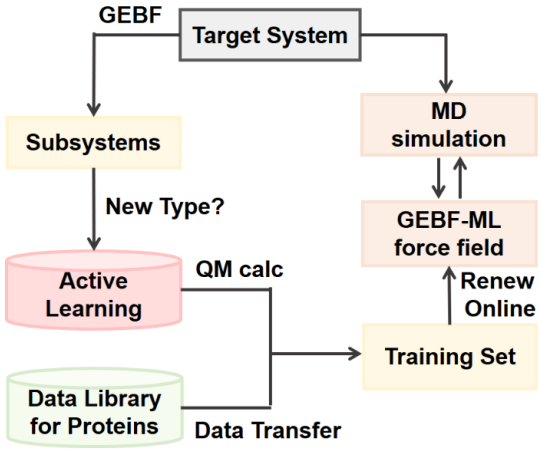


Figure 2. Scheme diagram of the GEBF-ML method. Training sets are constructed from relevant sub-datasets from the protein's data library and some subsystems from online active learning.

3. RESULTS AND DISCUSSION

3.1 Details of Training Sets and Testing Sets Construction. As a proof of concept, MLFFs of two polypeptides, 4ZNN (ACE-GVVHGVTTVA-NME) and 1XQ8 segment (ACE-GVVHGVATVA-NME) are constructed by our GEBF-ML scheme. First, online GEBF-GAP based MD simulations are performed on 4ZNN to generate the training set of subsystems, the GEBF-PM6 method is used as the baseline. During the 1-ns MD simulation at 500 K, QM calculations are carried out for only 0.15% of generated subsystems. The number of subsystem configurations and subsystem types (the definition of subsystem type can be seen in Sec.4 of the supporting information) in the training set are 8320 and 74, respectively. After the training set of 4ZNN has been constructed, all subsystems in the training set are divided into sub-datasets according to their topological types and stored in the data library. When we construct the training set for the 1XQ8 segment, we load the corresponding sub-datasets in the data library to it. As 4ZNN and 1XQ8 segments differ from each other by only one amino acid residue, about 4000 configurations (54 types of subsystems) are loaded from the data library. Then, online active learning is performed for the 1XQ8 segment to sample new subsystems during the 1ns GEBF-GAP based MD simulation at 500 K. For the 1XQ8 segment, only 0.01% of newly generated subsystems are needed for QM calculations during

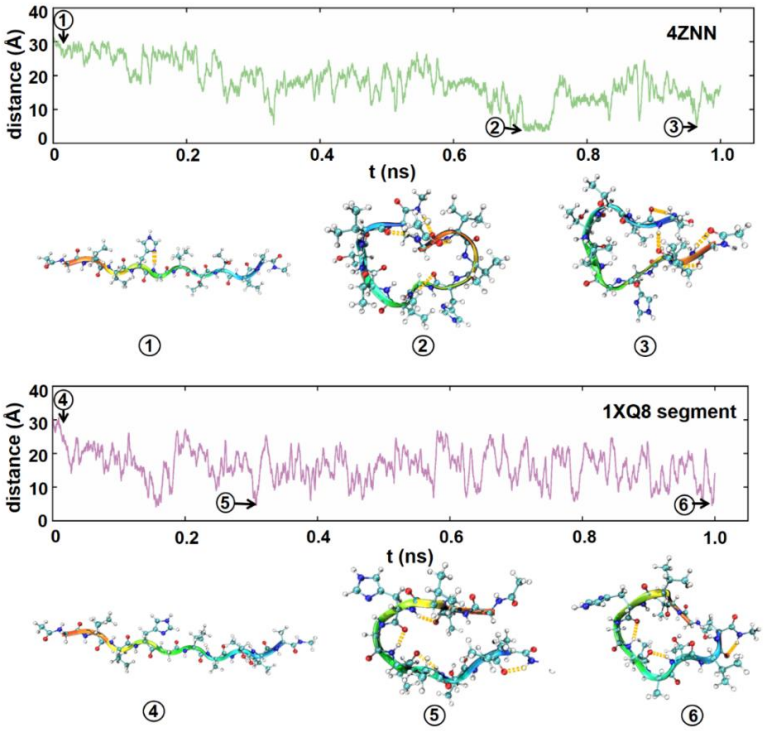


Figure 3. End-to-end distance of 4ZNN and 1XQ8 segment during the GEBF-GAP based MD simulations using GEBF-PM6 as the baseline.

the online active learning. The total number of subsystem configurations and subsystem types in the training set are 5020 and 65, respectively. The fraction of QM calculations for the 1XQ8 segment is much smaller than that for the 4ZNN, since a large number of subsystems generated from 4ZNN can be reused. Thus, our GEBF-ML scheme shows high efficiency for building the training set. It is also worth mentioning that with a data library for all trained subsystems, QM calculations on many subsystems are avoided when MLFFs on new proteins are constructed. The subsystem data is simply transferred according to their topology structures and the advantage of this approach is robust and intuitive. More advanced data transfer techniques may also be adopted if the aim is to construct MLFFs for a new protein with as few configurations as possible.

After the training sets have been constructed, the GEBF-GAP force fields are constructed from the QM energy differences. As GEBF-GAP based MD simulations show small energy drift (less than 0.001 kcal/(mol·atom·ps)) at the microcanonical (NVE) ensemble for both two polypeptides (see details in Sec.6 of the supporting information.), 1-ns GEBF-GAP based MD simulation using a Langevin thermostat[54] are performed at 300 K with a timestep of 1 fs in the canonical (NVT) ensemble. During the MD simulation, no QM calculations are performed and the MLFFs are not renewed online anymore. Figure 3 shows the changes of end-to-end distances between C_α atoms of the first and the last amino acid residues during the MD simulation. Three representative structures at different times are also plotted in Figure 4. As the trajectories show large conformation changes of the polypeptides from the chain-like extended structure to the folded one, 1000 structures for both two target systems are randomly sampled from the trajectories as testing set to evaluate the performance of our MLFFs. The electronic structure calculations on testing sets are carried out at the ωB97XD/6-31G* level with the Gaussian 16 package,[55] and the GEBF-PM6 calculations were performed with MOPAC package[56] and our LSQC program.[57]

3.2. Relative Energy Prediction and Structure Optimization. After MLFFs have been constructed, we first show the applicability of the MLFFs on relative energy prediction. The energies of are conformers in testing sets are calculated with GEBF-GAP, PM6, ff14SB and ωB97XD/6-31G*. Here, GEBF-PM6 is used as the baseline for GEBF-GAP and the energy of the first conformer in the test set was taken as zero. For six conformers (the structures of conformers are plotted in Sec.7 of supporting information.) randomly chosen from the testing sets, the absolute deviations of relative energies (relative to the ωB97XD/6-31G* results) are shown in Figure 4a. One can note that the largest deviations are less than 6 kcal/mol for GEBF-GAP results, but are much larger (more than 18 kcal/mol) for PM6 and ff14SB results. Clearly, PM6 and ff14SB methods cannot correctly predict the relative stability of different conformers if these conformers are close in energies. The results indicate that our MLFFs method could be used to search for the low-energy conformers of systems under study.

Further, we also test whether our MLFFs are suitable for structure optimization. The conformers with the lowest energy predicted by GEBF-GAP (using GEBF-PM6 method as baseline) in test sets are optimized with the BFGS algorithm[58] (implemented in ASE package[59]). Figure 4b shows optimized structures obtained with GEBF-GAP and ωB97XD/6-31G* for 4ZNN and 1XQ8 segments. The root-mean-square deviation (RMSD) between DFT and MLFF results is 0.31 Å and 0.36 Å

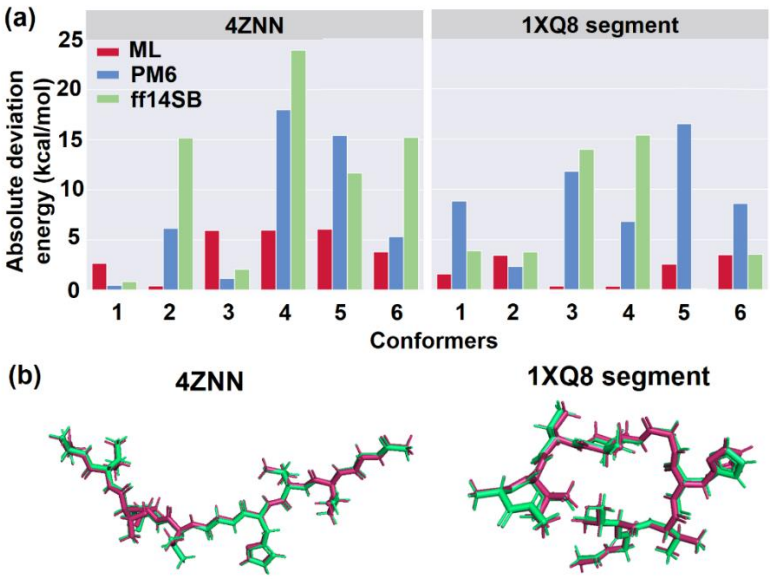


Figure 4. (a) The comparison of the absolute deviations of the GEBF-GAP, PM6, and ff14SB relative energies (relative to the ωB97XD/6-31G* values) among 6 conformers. For both systems, the energy of the first conformer is taken as zero for each method. (b) Optimized structures of 4ZNN and 1XQ8 segment. The superposition between the structure obtained with GEBF-GAP (red) and the DFT-optimized structure (green) is shown for both systems.

on 4ZNN and 1XQ8 segment, respectively. The geometrical parameters obtained with our MLFFs are very close to the corresponding values from the ωB97XD method. In addition, the geometries optimized with PM6 and ff14SB are also calculated for comparison. At respectively optimized structures, the absolute energy deviations predicted by GEBF-GAP, PM6, ff14SB (relative to the ωB97XD/6-31G* results) are 4.14, 13.96, 21.33 kcal/mol, respectively, for 4ZNN, and 0.85, 20.40, 24.60 kcal/mol, respectively, for 1XQ8 segment. Among these three methods, only the relative energies of MLFFs at their optimized structures are in good agreement with those from ωB97XD.

Although our MLFFs are not trained from NMR structures of proteins, we also applied GEBF-GAP and ωB97XD/6-31G* method to obtain the optimized structure with the NMR structure of 1XQ8 segment as the initial geometry. The optimized structures are shown in Figure S1. The RMSD between the MLFFs optimized structure and the reference DFT optimized structure is only 0.27 Å, which suggests that the present MLFFs are also reliable for geometry optimizations outside the MD trajectories.

3.3. Verification of Accuracy for GEBF-MLFFs During the MD Simulation. To investigate the applicability of our MLFFs on MD simulations. We first performed 20-ps MD simulations with GEBF-GAP, ff14SB and PM6 methods, respectively, the GEBF-PM6 method is used as the baseline for GEBF-GAP. MD simulations with ωB97XD/6-31G* are also carried out for comparison. Figure 5 display the dihedral angle distribution calculated with the GEBF-GAP and ωB97XD/6-31G* method. For each backbone dihedral φ, ψ and ω, histograms are accumulated for all amino acid residues except Gly. The results suggest that the distributions obtained from the GEBF-GAP and ωB97XD/6-31G* methods are very close to each other. The distributions predicted by the ff14SB and PM6 methods are plotted in Figure S2 and S3, respectively. The dihedral distributions from these two methods are quite different from the ωB97XD/6-31G* methods. For dihedrals φ and ψ, the shapes of distribution show a great difference when compared with the results from ωB97X-D/6-31G*. For dihedral angle ω, the peak intensity predicted by ff14SB is 20 % larger than the ωB97X-D/6-31G* result, and the deviation of the location of peak predicted by PM6 method from the ωB97X-D/6-31G* one reaches 10°. One can conclude that the dihedral angle distributions from GEBF-GAP are much more accurate than those from the ff14SB and PM6 methods.

Table 1. The Root Mean Squared Errors (RMSEs) of the MLFFs energies [in kcal/(mol·atom)], and forces [in kcal/(mol·Å)] (with respect to the conventional ωB97X-D/6-31G* results) for the testing set, all MLFFs are constructed from QM energy difference.

System	4ZNN	1XQ8 segment
RMSE E[a]	0.025	0.022
RMSE F[a]	1.5	1.5
RMSE E[b]	0.021	0.018
RMSE F[b]	1.3	1.3

[a]GEBF-MLFFs using GAP as ML model, [b]GEBF-MLFFs using NN as ML model.

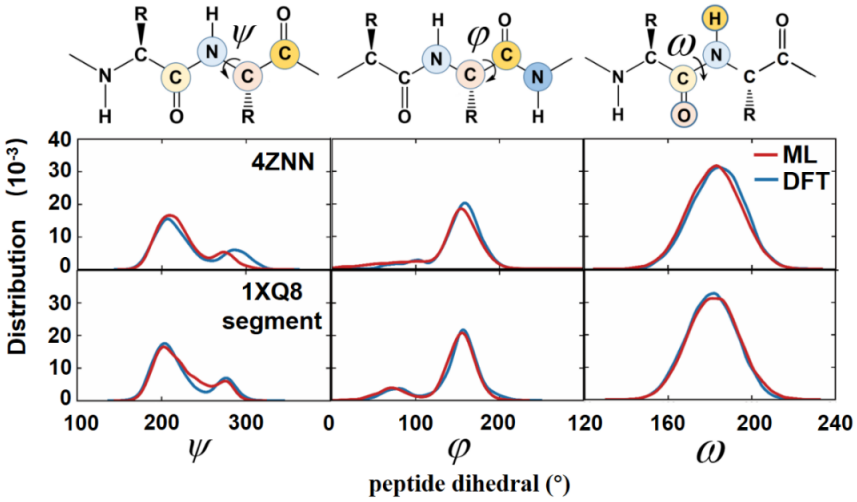


Figure 5. Backbone peptide dihedral distributions of 4ZNN (top) and 1XQ8 segment (bottom) obtained from 20 ps trajectories with reference DFT (blue solid line) and ML (red solid line). Distributions of dihedral angles, φ, ψ and ω are shown from left to right, respectively.

Then, we evaluate the accuracy of our MLFFs on the testing sets, which are randomly sampled from 1-ns GEBF-GAP based trajectories at 300K. As the GEBF-PM6 method is employed as the baseline of the MLFFs, the accuracy of GEBF-PM6 with respect to conventional PM6 is first evaluated. The mean absolute errors (MAEs) of energies between GEBF-PM6 and PM6 on the testing set are only 0.003 kcal/(mol·atom). 10 conformers of the 4ZNN and 1XQ8 segment are randomly chosen from the testing set. The deviations of GEBF-PM6 energies relative to conventional PM6 ones are listed in Table S2. The maximum deviation is only about 0.008 kcal/(mol·atom). Thus, the errors of GEBF-PM6 results with respect to the conventional PM6 ones are negligible for the two polypeptides. In Table 1, the root mean squared errors (RMSEs) of energy and forces obtained with the GEBF-GAP, relative to the conventional ωB97XD/6-31G* are shown. For both two systems, the RMSEs of energy and forces for GEBF-GAP results are about 0.024 kcal/(mol·atom) and 1.5 kcal/(mol·Å), respectively. For comparison, the RMSEs of PM6 and ff14SB force field results in energies and forces, relative to the conventional ωB97XD/6-31G* results, are also shown in Table S3. For two polypeptides, the RMSEs with ff14SB are 0.13 kcal/(mol·atom) and 12 kcal/(mol·Å), respectively. The RMSEs with PM6 are 0.06 kcal/(mol·atom) and 14 kcal/(mol·Å), respectively. These results indicate that our MLFFs are much more accurate than the PM6 or ff14SB method.

Table 2. The Root Mean Squared Errors (RMSEs) of the MLFFs energies [in kcal/(mol·atom)], and forces [in kcal/(mol·Å)] (with respect to the conventional ωB97X-D/6-31G* results) for the testing set on 1XQ8 segment, all MLFFs are constructed from QM energies.

Method	*N*st	RMSE E	RMSE F
GEBF-GAP	5020	0.045	2.5
GEBF-NN	5020	0.028	2.1
GEBF-GAP	6405	0.040	2.3
GEBF-NN	6405	0.026	1.8

We also use NN to parametrize subsystems, with the training set generated using the GEBF-ML protocol. DeePMD-kit[19] is adopted to construct the GEBF-NN force field from QM energy differences. Table 1 also shows the accuracy of GEBF-NN on the testing sets. For both two polypeptides, the RMSEs of the energy and forces for GEBF-NN results are about 0.020 kcal/(mol·atom) and 1.3 kcal/(mol·Å). Both GEBF-GAP and GEBF-NN could predict the energies and forces with full QM quality. The high accuracy of GEBF-NN also indicates that with the training set generated from GEBF-GAP, our MLFFs can also be constructed by NNs.

Although only PM6 calculations on small subsystems are needed for the GEBF-PM6 method and the computation cost of PM6 calculations is smaller than some MLFFs on small molecules,[60] we also construct the GEBF-MLFFs from QM energies directly. Using 1XQ8 segment as an example, the root mean squared errors (RMSEs) of energy and forces obtained with the GEBF-MLFFs, relative to the conventional ωB97XD/6-31G* are shown in Table 2. Using the same training sets (5020 subsystem configurations), the RMSEs of energy and forces for GEBF-GAP on the testing set are about 0.045 kcal/(mol·atom) and 2.5 kcal/(mol·Å), respectively. The RMSEs for GEBF-NN are about 0.028 kcal/(mol·atom) and 2.1 kcal/(mol·Å), respectively. Both GEBF-GAP and GEBF-NN constructed from QM energies show slightly larger errors than those constructed with energy differences. Fortunately, the accuracy of GEBF-MLFFs constructed from QM energies can further increase by adding more subsystem configurations in training sets. Using 5020 subsystem configurations as preliminary training sets, GEBF-GAP based MD simulation was performed to sampling more subsystem conformers. With 6405 subsystem configurations as the training set, the RMSEs for GEBF-GAP are 0.04 kcal/(mol·atom) and 2.3 kcal/(mol·Å), respectively. While the RMSEs for GEBF-NN are 0.026 kcal/(mol·atom) and 1.8 kcal/(mol·Å), respectively. The accuracies of both GEBF-MLFFs are increased by adding more data points in the training set. Moreover, the GEBF-NN constructed from QM energies shows similar accuracy as GEBF-MLFFs constructed from en-

ergy differences. Thus, accurate GEBF-MLFFs can be constructed either from energy differences between QM and PM6 methods with relatively small training sets or from QM energies of subsystems with more subsystem configurations in the training set.

3.4. MD Simulation with GEBF-MLFFs Constructed from QM Energies. With the GEBF-NN force field constructed from QM energies, 1-ns MD simulations using a Langevin thermostat have been performed for 1XQ8 segment at 300 K with a timestep of 1 fs. To quantitatively describe the conformational changes, the RMSDs with respect to the initial structure of the 1XQ8 segment during the simulation are shown in Figure 6a. The RMSD increases rapidly and reaches the maximum value of 10 Å during the MD simulation. Thus, the trajectory also shows large conformation changes during the MD simulation. To evaluate the accuracy of the GEBF-NN force field during the simulation, 1000 configurations are evenly sampled from the 1ns trajectory and the mean absolute error (MAE) with respect to traditional ωB97X-D/6-31G* are calculated on each configuration. Figure 6b shows the time evolution of MAE for forces during the 1-ns MD simulations. One can see that the MAEs on almost all configurations are less than 2 kcal/(mol·Å). Thus, GEBF-NN force fields can also be constructed from QM energies with full QM quality if enough subsystem configurations are added in training sets.

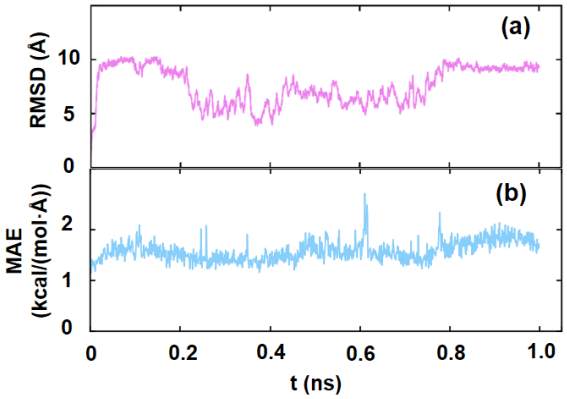


Figure 6. (a) Time evolution of the RMSD with respect to the initial structure during the GEBF-NN MD simulation of the 1XQ8 segment. (b) Time evolution of the MAE for forces during the GEBF-NN MD simulation of the 1XQ8 segment. GEBF-NN is constructed from QM energies directly.

Finally, it is also necessary to demonstrate the computational cost and scalability of our GEBF and GEBF-MLFF approach by comparing them with the conventional QM method and classical force field. Here, we take polypeptides ACE-$(Ala)_n$-NME (n = 50, 100 and 150) as examples. All calculations are carried out on 48-core Intel Xeon Platinum 8163 2.5 GHz CPU. In Table 3, we present the total CPU time required for ACE-$(Ala)_n$-NME (n = 50, 100 and 150) at different theoretical levels (ωB97XD and GEBF-ωB97XD with 6-31G* basis set, ff14SB, GEBF-GAP and GEBF-NN with or without GEBF-PM6 as baseline).

First, we compared the scalability of different methods. It can be seen from Table 3 that all the tested methods (except the conventional DFT method) show linear scaling behavior. For instance, the total CPU time required by GEBF-GAP constructed from energy difference is about 1.9 and 2.9 times for ACE-$(Ala)_{100}$-NME and ACE-$(Ala)_{150}$-NME than that for ACE-$(Ala)_{50}$-NME (39.52 seconds).

Then, we compare the computational cost of our GEBF method and GEBF-MLFFs with the QM method and traditional force field. As summarized in Table 3, for all three polypeptides, the computational cost of GEBF-DFT is smaller than the conventional DFT method and the acceleration ratio on ACE-$(Ala)_{150}$-NME is about 20. Thus, the computational costs of QM calculations are highly reduced during the training set construction. For all GEBF-MLFFs, the acceleration ratios are at least three magnitude orders than the full QM calculations. Thus, our GEBF-MLFFs show the low computational cost, with respect to QM methods. However, MLFFs are slower than the ff14SB force field, because much more parameters are needed to describe MLFFs.

Table 3. Total CPU time (in seconds) required for ACE-$(Ala)_n$-NME (n= 50, 100, and 150) energy and forces calculations with the DFT, GEBF-DFT, GEBF-GAP, GEBF-NN and ff14SB methods. DFT calculations are carried out at the ωB97XD/6-31G* level.

Method	n=50	n=100	n=150
DFT	69732	364847	1013869
GEBF-DFT	17058	34322	51532
GEBF-GAP[a]	33.12	64.42	95.42
GEBF-GAP[b]	39.52	77.31	114.82
GEBF-NN[a]	4.52	7.12	10.50
GEBF-NN[b]	10.92	20.02	29.89
ff14SB	0.01	0.02	0.04

[a]MLFFs constructed from QM energies, [b]MLFFs constructed from energy differences

Further, the computational costs of GEBF-MLFFs constructed from QM energies or energy differences are also compared in Table 3. One can see that the computational costs of two GEBF-GAP force fields are similar. For all tested polypeptides, the total CPU time required by GEBF-GAP constructed from energy difference is about 1.2 times than that from QM energies. However, for cost-effective ML potential like Deep Potential, the computational cost of GEBF-PM6 cannot be neglected, the total CPU time required by GEBF-NN constructed from energy differences is about 3 times than that from QM energies. GEBF-GAP construct from energy differences may be appropriate for the online training process and nanosecond-scale MD simulations. If microseconds-scale MD simulations are needed, it is more convenient to construct GEBF-NNs from QM energies directly.

4. CONCLUSIONS

In summary, we developed a general GEBF-ML protocol to automatically construct MLFFs for proteins with QM accuracy. Using GAP as the ML model, our protocol can automatically generate training sets with high efficiency. Moreover, for a given protein, only QM calculations on small subsystems containing a few residues are required in the construction of training

sets. To facilitate the construction of training sets for various proteins, we create a protein's data library, which contains all data of subsystems generated from trained proteins. With this protein's data library, for a new protein only its subsystems with new topological structures are required for the construction of the corresponding training sets. Using two polypeptides 4ZNN and 1XQ8 segment, as examples. The accuracy of the constructed GEBF-GAP for both systems is validated by comparing the conformational energies, optimized structure, and MD simulation results with those from conventional DFT results. Our results show that GEBF-GAP can lead to quite accurate energies and forces similar to those from full QM calculations, and dihedral angle distributions from GEBF-GAP MD simulations are in good agreement with those from *ab initio* MD simulations. In addition, we also demonstrated that full QM quality GEBF-NN force fields can also be constructed using the training sets generated by GEBF-GAP. Thus, this work provides an efficient and systematic way to build MLFF for proteins, we also expected GEBF-ML protocol could be used for polymer materials and complex biological systems in aqueous solutions in the future.

ASSOCIATED CONTENT

Supporting Information. Computational efficiency, additional ML results, additional MD results, fragmentation scheme, the construction of GEBF subsystems. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Authors

Jing Ma – *Key Laboratory of Mesoscopic Chemistry of Ministry of Education, Institute of Theoretical and Computational Chemistry, School of Theoretical and Computational Chemistry, School of Chemistry and Chemical Engineering, Nanjing University, Nanjing, 210093, P. R. China;* orcid.org/0000-0001-5848-9775; Email: majing@nju.edu.cn.

Wei Li – *Key Laboratory of Mesoscopic Chemistry of Ministry of Education, Institute of Theoretical and Computational Chemistry, School of Theoretical and Computational Chemistry, School of Chemistry and Chemical Engineering, Nanjing University, Nanjing, 210093, P. R. China;* orcid.org/0000-0001-7801-3463; Email: wli@nju.edu.cn.

Shuhua Li – *Key Laboratory of Mesoscopic Chemistry of Ministry of Education, Institute of Theoretical and Computational Chemistry, School of Theoretical and Computational Chemistry, School of Chemistry and Chemical Engineering, Nanjing University, Nanjing, 210093, P. R. China;* orcid.org/0000-0001-6756-057X; Email: shuhua@nju.edu.cn.

Authors

Zheng Cheng – *Key Laboratory of Mesoscopic Chemistry of Ministry of Education, Institute of Theoretical and Computational Chemistry, School of Theoretical and Computational Chemistry, School of Chemistry and Chemical Engineering, Nanjing University, Nanjing, 210093, P. R. China;*

Jiahui Du – *Key Laboratory of Mesoscopic Chemistry of Ministry of Education, Institute of Theoretical and Computational Chemistry, School of Theoretical and Computational Chemistry, School of Chemistry and Chemical Engineering, Nanjing University, Nanjing, 210093, P. R. China;*

Lei Zhang – *Key Laboratory of Mesoscopic Chemistry of Ministry of Education, Institute of Theoretical and Computational Chemistry, School of Theoretical and Computational Chemistry, School of Chemistry and Chemical Engineering, Nanjing University, Nanjing, 210093, P. R. China;*

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (Grants Nos. 21833002, 22033004, 21873046, and 22073043). Part of the calculations were performed using computational resources on an IBM Blade cluster system from the High Performance Computing Center (HPCC) of Nanjing University. Prof. Gárbor Csányi is greatly acknowledged for fruitful discussion.

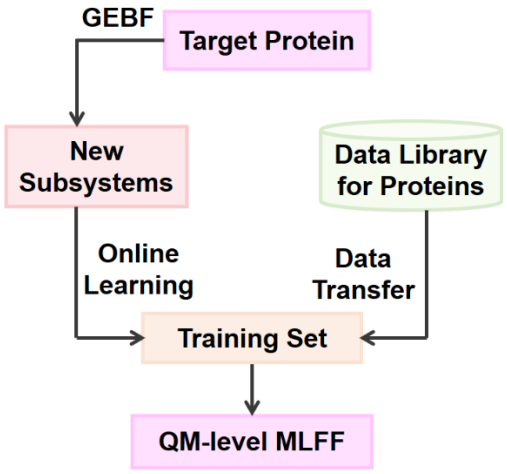
REFERENCES

(1) Bjelkmar, P.; Larsson, P.; Cuendet, M. A.; Hess, B.; Lindahl, E. Implementation of the CHARMM Force Field in GROMACS: Analysis of Protein Stability Effects from Correction Maps, Virtual Interaction Sites, and Water Models. *J. Chem. Theory Comput*. **2010**, *6*, 459-466.
(2) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber biomolecular simulation programs. *J. Comput. Chem*. **2005**, *26*, 1668-1688.
(3) Eichenberger, A. P.; Allison, J. R.; Dolenc, J.; Geerke, D. P.; Horta, B. A. C.; Meier, K.; Oostenbrin, C.; Schmid, N.; Steiner, D.; Wang, D.; van Gunsteren, W. F. The GROMOS++ Software for the Analysis of Biomolecular Simulation Trajectories. *J. Chem. Theory Comput.* **2011**, *7*, 3379-3390.
(4) Jorgensen, W. L.; Tirado. R. J. The OPLS Force Field for Proteins. Energy Minimizations for Crystals of Cyclic Peptides and Crambin. *J. Am. Chem. Soc.* **1998**, *110*, 1657-1666.
(5) 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.
(6) Lamoureux, G.; Roux, B. Modeling induced polarization with classical Drude oscillators: Theory and molecular dynamics simulation algorithm. *J. Chem. Phys*. **2003**, *119*, 3025-3039.
(7) Abel, R.; Wang, L.; Harder, E. D.; Berne, B. J.; Friesner, R. A. Advancing Drug Discovery through Enhanced Free Energy Calculations. *Acc. Chem. Res*. **2017**, *50*, 1625-1632.
(8) Nerenberg, P. S.; Gordon-Heard, T. New developments in force fields for biomolecular simulations. *Curr, Opin. Struct. Biol.* **2018**, *49*, 129-138.
(9) Best, R. B.; Buchete, N. V.; Hummer, G. Are current Molecular Dynamics Force Fields too Helical? *Biophys. J.* **2008**, *108*, 132696.
(10) Lindorff-Larsen, K.; Maragakis, P. and Piana, S. and Eastwood, M. P.; Dror, R. O.; Shaw, D. E. Systematic Validation of Protein Force Fields against Experimental Data. *PLoS One* **2012**, *7*, e32131.
(11) Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98*, No. 156401.
(12) 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*, No. 136403.
(13) Thompson, A.; Swiler, L.; Trott, C.; Foiles, S.; Tucker, G. Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials. *J. Comput. Phys*. **2015**, *285*, 316.
(14) Shapeev, A. V. Moment Tensor Potentials: A Class of Systematically Improvable Interatomic Potentials. *Multiscale Model. Simul.* **2016**, *14*, 1153-1173.
(15) Li, Z.; Kermode, J. R.; De Vita, A. Molecular Dynamics with On-the-Fly Machine Learning of Quantum-Mechanical Forces. *Phys. Rev. Lett.* **2015**, *114*, No. 096405.

(16) Schütt, K. T.; Kessel, P.; Gastegger, M.; Nicoli, K. A.; Tkatchenko, A.; Müller, K.R. SchNetPack: A Deep Learning Toolbox for Atomistic Systems. *J. Chem. Theory Comput*, **2019**, *15*, 448-455.
(17) Zhang, L.; Han, J.; Wang, H.; Car, R.; E, W. Deep Potential Molecular Dynamics: A Scalable Model with the Accuracy of Quantum Mechanics. *Phys. Rev. Lett*. **2018**, *120*, No. 143001.
(18) 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.
(19) Zhang, L.; Han, J.; Wang, H.; Saidi, W. A.; Car, R. End-to-End Symmetry Preserving Inter-Atomic Potential Energy Model for Finite and Extended *Systems. Proceedings of the 32nd International Conference on Neural Information Processing Systems* **2018**, 4441-4451.
(20) Drautz, R. Atomic cluster expansion of scalar, vectorial, and tensorial properties including magnetism and charge transfer. *Phys. Rev. B*. **2020**, *102*, 024104.
(21) Gastegger, M.; Marquetand, P. High-dimensional Neural Network Potentials for Organic Reactions and an Improved Training Algorithm. *J. Chem. Theory Comput*. **2015**, *11*, 2187-2198.
(22) Gastegger, M.; Behler, J.; Marquetand, P. Machine learning molecular dynamics for the simulation of infrared spectra. *Chem. Sci.* **2017**, *8*, 6924-6935.
(23) Westermayr, J.; Gastegger, M.; Marquetand, P. Combining SchNet and SHARC: The SchNarc Machine Learning Approach for Excited-State Dynamics. *J. Phys. Chem. Lett*. **2020**, *11*, 3828-3834.
(24) Deringer, V. L.; Bernstein, N.; Csányi, G; Mahmoud, C. B.; Ceriotti, M.; Wilson, M.; Drabold, D. A.; Elliott, S. R. Origins of structural and electronic transitions in disordered silicon. *Nature*. **2021**, *589*, 59-64.
(25) Huang, B.; von Lilienfeld, O. A. Quantum machine learning using atom-in-molecule-based fragments elected on the fly. *Nat Chem* **2020**, *12*, 945-951.
(26) Cheng, B.; Engel, E. A.; Behler, J.; Dellago, C.; Ceriotti, M. Ab Initio Thermodynamics of Liquid and Solid Water. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 1110-1115.
(27) 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.
(28) Jinnouchi, R.; Lahnsteiner, J.; Karsai, F.; Kresse, G.; Bokdam, M. Phase Transitions of Hybrid Perovskites Simulated by Machine Learning Force Fields Trained on the Fly with Bayesian Inference. *Phys. Rev. Lett.* **2019**, *122*, 225701.
(29) Liu, Z.; Lin, L.; Jia, Q.; Cheng, Z.; Jiang, Y.; Guo, Y.; Ma, J. Transferable Multilevel Attention Neural Network for Accurate Prediction of Quantum Chemistry Properties via Multitask Learning. *J. Chem. Inf. Model.* **2021**, *61*, 1066-1082.
(30) Yang, M.; Bonati, L.; Polino, D.; Parrinello, M. Using metadynamics to build neural network potentials for reactive events: the case of urea decomposition in water. *Catalysis Today*, **2021**, 18.
(31) Kang, P.; Shang, C.; Liu, Z. Large-Scale Atomic Simulation via Machine Learning Potentials Constructed by Global Potential Energy Surface Exploration. *Acc. Chem. Res*. **2020**, *53*, 2119-2129.
(32) Niu, H.; Bonati, L.; Piaggi, P. M.; Parrinello, M. Ab initio phase diagram and nucleation of gallium. *Nat, Commun*. **2020**, *11*, 2654.
(33) Liu, C.; Shi, J.; Gao, H.; Wang, J.; Han, Y.; Lu, X.; Wang, H.; Xing, D.; Sun, J. Mixed Coordination Silica at Megabar Pressure. *Phys. Rev. Lett.* **2021**, *126*, No.035701.
(34) Zhang, L.; Wang, H.; Car, R.; E, W. Phase Diagram of a Deep Potential Water Model. *Phys. Rev. Lett.* **2021**, *126*, No.236001.
(35) Li, W.; Dong, H.; Ma, J.; Li, S. Structures and Spectroscopic Properties of Large Molecules and Condensed-Phase Systems Predicted by Generalized Energy-Based Fragmentation Approach. *Acc. Chem. Res*. **2021**, *54*, 169-181.
(36) Zhao, D.; Shen, X.; Cheng, Z.; Li, W.; Dong, H.; Li, S. Accurate and Efficient Prediction of NMR Parameters of Condensed-Phase Systems with the Generalized Energy-Based Fragmentation Method. *J. Chem. Theory Comput*. **2020**, *16*, 2995-3005.
(37) Collins, M. A.; Cvitkovic, M. W.; Bettens, R. P. A. The Combined Fragmentation and Systematic Molecular Fragmentation Methods. *Acc. Chem. Res.* **2014**, *47*, 2776-2785.
(38) Ganesh, V.; Dongare, R. K.; Balanarayan, P.; Gadre, S. R. Molecular Tailoring Approach for Geometry Optimization of Large Molecules: Energy Evaluation and Parallelization Strategies. *J. Chem. Phys.* **2006**, *125*, 104109.
(39) Dahlke, D. E.; Truhlar, D. G. Electrostatically Embedded Many-Body Expansion for Large Systems, with Applications to Water Clusters. *J. Chem. Theory Comput.* **2006**, *3*, 46-53.
(40) Gordon, M. S.; Fedorov, D. G.; Pruitt, S. R.; Slipchenko, L. V. Fragmentation Methods: A Route to Accurate Calculations on Large Systems. *Chem. Rev*. **2011**, *112*, 632-672.
(41) He, X.; Zhang, J. Z. H. The generalized molecular fractionation with conjugate caps/molecular mechanics method for direct calculation of protein energy. *J. Chem. Phys*. **2006**, *124*, No. 184703.
(42) Bettens, R. P. A.; Lee, A. M. A New Algorithm for Molecular Fragmentation in Quantum Chemical Calculations. *J. Phys. Chem. A* **2006**, *110*, 8777-8785.
(43) Huang, L.; Massa, L.; Karle, J. Kernel energy method illustrated with peptides. *Int. J. Quantum Chem*. **2005**, *103*, 808-817.
(44) Richard, R. M.; Herbert, J. M. A generalized many-body expansion and a unified view of fragment-based methods in electronic structure theory. *J. Chem. Phys.* **2012**, *137*, No. 064113.
(45) Mayhall, N. J.; Raghavachari, K. Many-Overlapping-Body (MOB) Expansion: A Generalized Many Body Expansion for Non-disjoint Monomers in Molecular Fragmentation Calculations of Covalent Molecules. *J. Chem. Theory Comput.* **2012**, *8*, 2669-2675.
(46) Wang, H.; Yang, W. Toward Building Protein Force Fields by Residue-Based Systematic Molecular Fragmentation and Neural Network. *J. Chem. Theory Comput.* **2018**, *15*, 1409-1417.
(47) Wang, Z.; Han, Y.; Li, J. He, X. Combining the Fragmentation Approach and Neural Network Potential Energy Surfaces of Fragments for Accurate Calculation of Protein Energy. *J. Phys. Chem. B* **2020**, *124*, 3027-3035.
(48) Cheng, Z.; Zhao, D.; Ma, J.; Li, W. and Li, S. An On-the-Fly approach to Construct Generalized Energy-Based Fragmentation Machine Learning Force Fields of Complex Systems. *J. Phys. Chem. A* **2020**, *124*, 5007-5014.
(49) Moussa, J. E. Comment on "Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *109*, 059801.
(50) Dral, P. Quantum Chemistry in the Age of Machine Learning. *J. Phys. Chem. Lett.* **2020**, *11*, 2336-2347.
(51) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Phys. Rev. B* **2013**, *87*, No. 184115.
(52) Nguyen, H.; Maier, J.; Huang, H.; Perrone, V.; Simmerling, C. Folding Simulations for Proteins with a Physics-Based Force and Implicit Solvent. *J. Am. Chem. Soc.* **2014**, *136*, 13959-13962.
(53) Pan S. J. and Yang, Q. A survey on transfer learning. IEEE Transactions on knowledge and data engineering. **2010**, *22*, 1345-1359.
(54) Langevin, P. Sur la theories du mouvement brownien. *C. R. Acad. Sci. Pairs*. **1908**, *146*, 530-533.
(55) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma,

K.; Farkas, O.; Foresman, J. B.; Fox, D. J. Gaussian 16; Gaussian, Inc.: Wallingford CT, 2016.
(56) Stewart, J.J.P. Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements. *J Mol Model*. **2007**, *13*, 1173-1213.
(57) Li, W.; Chen, C.; Zhao, D.; Li, S. LSQC: Low scaling quantum chemistry program. *Int. J. Quantum Chem.* **2015**, *115*, 641-646.
(58) Fletcher, R. Practical Methods of Optimization; *John Wiley & Sons:* **2013**.
(59) Larsen, A. H.; Mortensen, J. J.; Blomqvist, J.; Castelli, I. E.; Christensen, R.; Dular, M.; Friis, J.; Groves, M. N.; Hammer, B. and Hargus, C. The atomic simulation environment – a Python library for working with atoms. *J. Phys.: Condens. Matter*. **2017**, *29*, 273002.
(60) Folmsbee, D.; Hutchison, G. Assessing Conformer Energies using Electronic Structure and Machine Learning Methods. *Int. J. Quantum Chem.* **2021**, *121*, No. e26381.

Table of Contents



Molecular dynamic simulation based on quantum mechanics (QM) can give highly accurate results but at high computational costs. Herein, we propose a protocol for the first time to construct machine learning force fields with QM quality at the cost of some QM calculations on subsystems not stored in a data library. This work takes an important step into the practical computational study of biological systems with QM accuracy.