Building machine learning force fields of proteins with fragment-based approach and transfer learning

Zheng Cheng[1], Jiahui Du[1], Lei Zhang[1], Jing Ma[1],★ Wei Li[1],★ & Shuhua Li[1],★

[1]Key Laboratory of Mesoscopic Chemistry of Ministry of Education, Institute of Theoretical and Computational Chemistry, School of Chemistry and Chemical Engineering, Nanjing University, Nanjing, 210023, China. ★e-mail: shuhua@nju.edu.cn; majing@nju.edu.cn; wli@nju.edu.cn

Molecular dynamic (MD) simulation plays an essential role in understanding protein functions at atomic level. At present, MD simulations on proteins are mainly based on classical force fields. However, the accuracy of classical force fields for proteins is still insufficient for accurate descriptions of their structures and dynamical properties. Here we present a novel protocol to construct machine learning force field (MLFF) for a given protein with full quantum mechanics (QM) accuracy. In this protocol, the energy of the target system is obtained by fitting energies of its various subsystems constructed with the generalized energy-based fragmentation (GEBF) approach. 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. With this protein's data library, for a new protein only its subsystems with new topological types are required for the construction of the corresponding MLFF. This protocol is illustrated with two polypeptides, 4ZNN and 1XQ8 segment, as examples. The energies and forces predicted from this MLFF are in good agreement with those from density functional theory calculations, and dihedral angle distributions from GEBF-MLFF MD simulations can also well reproduce those from *ab initio* MD simulations. Therefore, this GEBF-ML protocol is expected to be an efficient and systematic way to build force fields for proteins and other biological systems with QM accuracy.

Molecular dynamics (MD) simulation is an import tool to understand how the structure of a protein molecule determines its function in a cell. It is generally believed that *ab initio* MD simulations based on quantum mechanics (QM) methods can be used to obtain accurate and reliable simulation results. However, due to the high computational cost of QM calculations for large biomolecules, *ab initio* MD simulations are not available for proteins. 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.[11-19] The resulting machine learning potentials, also called as ML force fields (MLFFs), have been demonstrated to be quite successful for a variety of different systems.[20-27] In MLFFs, the total energy of a target system is generally expressed as a sum of local environment-dependent atomic energies. 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 to do QM calculations for the same systems.

Due to the chemical complexities of proteins and high computational costs of QM methods for large systems, building MLFFs for proteins remains a great challenge. Energy-based fragmentation (EBF) approaches[28-38] provide a practical and attractive solution to achieve QM calculations of large molecules including proteins. With this approach, the ground-state energy of a large system can be evaluated as the linear combination of ground-state energies of small subsystems, which are representation of different local regions of a large system. The combination of ML technique with fragment-based approaches enables the construction of the MLFFs for a large system with only QM energies (or forces) of small subsystems. In previous studies, a residue-based neural network (NN) approach has been developed to

construct preliminary MLFFs for proteins.[39,40] In their approach, a protein is first fragmented into various dipeptides, in which an amino acid is capped with two terminal groups (acetyl group and *N*-methyl amide group), as shown in Figure 1. Then, the total energy of a protein is expressed as the linear combination of the energies of all these dipeptides, each of which can be represented as NN potentials. The resulting ML potentials represent the first step towards *ab initio* quality protein force fields. However, 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.[39] Obviously, the accuracy of MLFFs for proteins should be significantly improved so that accurate simulations of their structures and dynamical properties are available.

In this work, we develop a general protocol for constructing MLFFs for proteins with full QM quality. This is achieved by fitting atomic energies from QM calculations on subsystems with realistic local chemical environments and taking the long-range interactions outside various subsystems into account. To simplify the parametrization in the construction of MLFFs, the nonparametric Gaussian approximation potentials (GAP)[12] proposed by Csányi is chosen to learn the ground-state energies of various subsystems and the energy of the target protein is predicted by GAP directly as the summation of atomic contributions. The generation of subsystems for a protein segment 4ZNN is illustrated in Figure 1. With the generalized energy-based fragmentation (GEBF) approach,[28] 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 representation of the local chemical environment of different regions in a protein than those in residue-based neural network (NN) approach (also shown in Figure 1). However, a general protein may contain as many as 20 different types of amino-acid residues, thus the number of different topological types of subsystems (with three or more residues) that could be constructed for different proteins is enormous. Thus, a cost-effective practical strategy for building the MLFF of a given protein is to fit the energy (or forces) of this protein as the summation of atomic contributions from QM calculations of various subsystems for the studied protein. Because a subset of

subsystems generate from a protein may have the same topological structure in chemical space as those from another protein, we may introduce transfer learning[41] to avoid redundant QM calculations on these subsystems. In our approach, we create a protein's data library, which contains all data of subsystems generated from trained proteins. For a new protein, a subset of subsystems with 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. An online active learning[42] is adopted here to generate these new subsystems for the studied protein. This protocol is applied on two polypeptides (4ZNN and 1XQ8 segment) to construct the corresponding MLFFs and their accuracy and efficiency are validated with reference QM calculations. Our results indicate that this GEBF-ML force field 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.

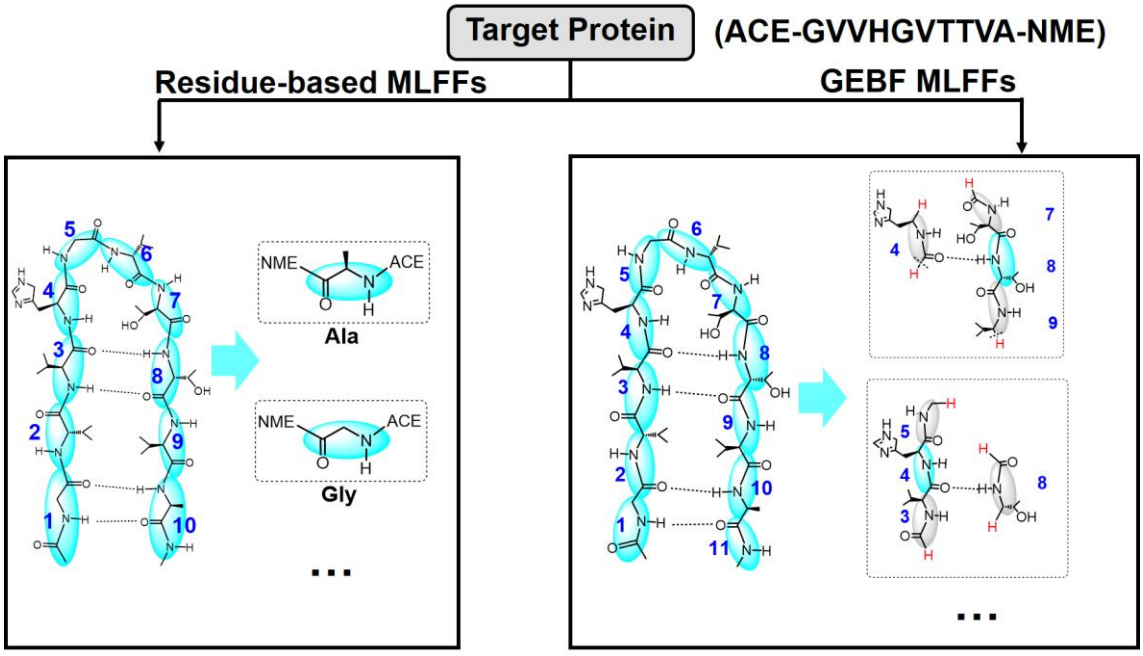

Figure 1. Fragmentation scheme utilized in the construction of MLFFs. In our GEBF method, fragments are capped with its environmental fragments. In previous residue-based method, fragments are capped with an acetyl group (ACE) and *N*-methylamide group (NME).

Results.

Accuracy and efficiency of machine learning force field. As a proof of concept, MLFFs of two polypeptides, 4ZNN segment (ACE-GVVHGVTTVA-NME) and 1XQ8 segment (ACE-GVVHGVATVA-NME), are constructed by our GEBF-ML scheme. Both protein segments are capped with ACE and NME. First, online machine learning MD simulations are performed on 4ZNN to generate the training set of subsystems. During the 1-ns MD simulation at 500 K, QM calculations are carried out for only 0.15% of generated subsystems and the number of subsystems with different topological types or different configurations in the training set is only 8147, as shown in Table 1. After the MLFFs of 4ZNN have been constructed, all subsystems of 4ZNN in the training set are divided into sub-datasets according to their topological types and stored in the data library. When we construct MLFFs for 1XQ8 segment, we load the corresponding sub-datasets in the data library to the training set. As the 4ZNN and 1XQ8 segments differ from each other by only one amino acid residue, about 4000 subsystems are loaded from the data library. Then, online active learning is performed for the 1XQ8 segment to sample new subsystems, only 0.009 % of newly generated subsystems are needed for QM calculations, and the total number of subsystems in the training set is only 4810 (Table 1). The fraction of QM calculations for 1XQ8 segment is much smaller than that for 4ZNN segment, since a large number of subsystems generated from 4ZNN can be reused. The online training process shows high sampling efficiency for building the training set. It is worth mentioning that the data library will be continuously expanded when more proteins are trained, and much less QM calculations on subsystems may be required for the construction of MLFFs for any new protein.

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 *ω*B97X-D/6-31G* results) for the test set, fraction x_1(%) of the QM calculations during the online active learning and numbers of subsystems N_{st} for the training set.

System	4ZNN	1XQ8 segment
RMSE *E*	0.025	0.022
RMSE *F*	1.475	1.482
x_1	0.145	0.009
N_{st}	8147	4810

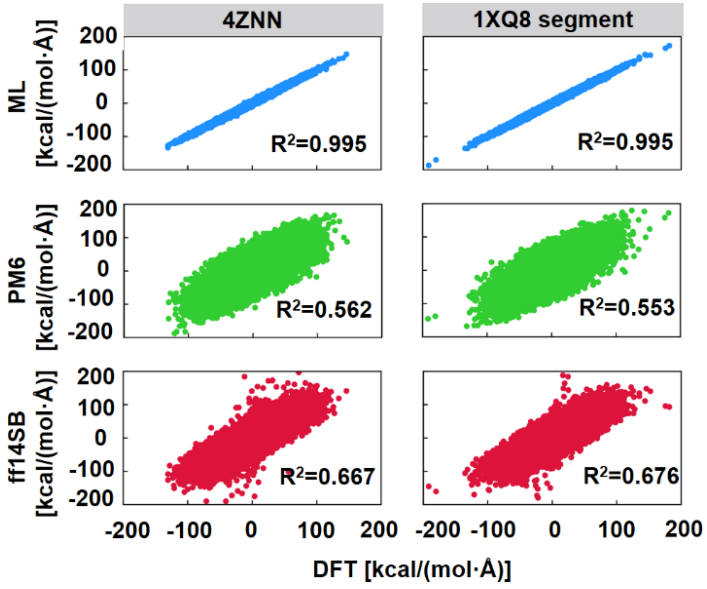

Figure 2. The comparisons of correlations between the forces from MLFFs, PM6, and ff14SB, and the *ω*B97XD/6-31G* ones.

The accuracy of the generated MLFFs is then evaluated. As the GEBF-PM6 method is employed as the baseline of the MLFFs, 10 randomly chosen conformers of 4ZNN and 1XQ8 segments are calculated with the GEBF-PM6 method to verify whether the GEBF-PM6 method can reproduce the conventional PM6 results. The deviations of GEBF-PM6 energies relative to conventional PM6 ones are listed in Table S1. Our calculations show that the mean absolute errors (MAEs) of energies for both two systems are only 0.003 kcal/(mol·atom). Thus, the errors of GEBF-PM6 results with respect to the conventional PM6 ones are negligible for the two polypeptides. Then, 1000 structures are randomly chosen from the trajectories at 300 K for both systems as test sets, and the accuracy of our MLFFs is evaluated on them. The RMSEs between the MLFF results and the conventional ωB97XD/6-31G* calculations are summarized in Table 1. The RMSEs are less than 0.025 kcal/(mol·atom) and 1.5 kcal/(mol·Å), respectively, indicating that the MLFFs could accurately predict the energies and forces for both systems. For comparison, the RMSEs of PM6 and ff14SB force field results in energies and forces, relative to the conventional ωB97X-D/6-31G* results, are also shown in Table S2. 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, and the previous MLFF from the residue-based neural network approach. To further show the accuracy of MLFFs, Figure 2 plots the correlations between the forces from MLFFs (top), PM6 (center), and ff14SB (bottom) and the ωB97XD/6-31G* ones for all configurations in test sets. The coefficient of determination (R^2) between these results and ωB97XD/6-31G* results is 0.995 (MLFFs), much higher than 0.56 for PM6 or 0.67 for ff14SB. The whole range of force amplitudes predicted by MLFFs is almost the same with that from reference ωB97XD/6-31G* calculations. Therefore, our GEBF-ML

protocol can automatically build MLFFs of these two polypeptides with full QM quality.

Relative energy prediction and structure optimization. To show the applicability of the MLFFs on relative energy prediction, we compare the relative energies for all conformers in the test set from MLFFs, PM6, ff14SB with the ωB97XD/6-31G* data. Here, the energy of the first conformer calculated with each method was taken as zero. The MAEs of relative energies predicted by MLFFs are 3.20 and 2.93 kcal/mol for 4ZNN and 1XQ8 segments, respectively, relative to the ωB97XD/6-31G* results. Relative to the ωB97XD/6-31G* results, the MAEs predicted by PM6 for both systems are 7.22 and 7.34 kcal/mol, respectively, and the MAEs predicted by ff14SB are 22.60 and 14.38 kcal/mol, respectively. Thus, both PM6 and ff14SB results are much less accurate than the present MLFF ones. For six structures randomly chosen from the test sets, the absolute deviations of relative energies (relative to the ωB97XD/6-31G* results) are shown in Figure 3a. One can note that the largest deviations are less than 6 kcal/mol for MLFF 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 the GEBF-MLFF method could be used to search for the low-energy conformers of systems under study.

Further, to test if our MLFF could also be suitable for structure optimization, the conformers with the lowest energy predicted by MLFFs in test sets are considered as initial geometries. Figure 3b shows optimized structures obtained with MLFFs 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 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 MLFFs, 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. Therefore, the MLFFs can be directly used to obtain *ab initio* quality optimized structures for proteins.

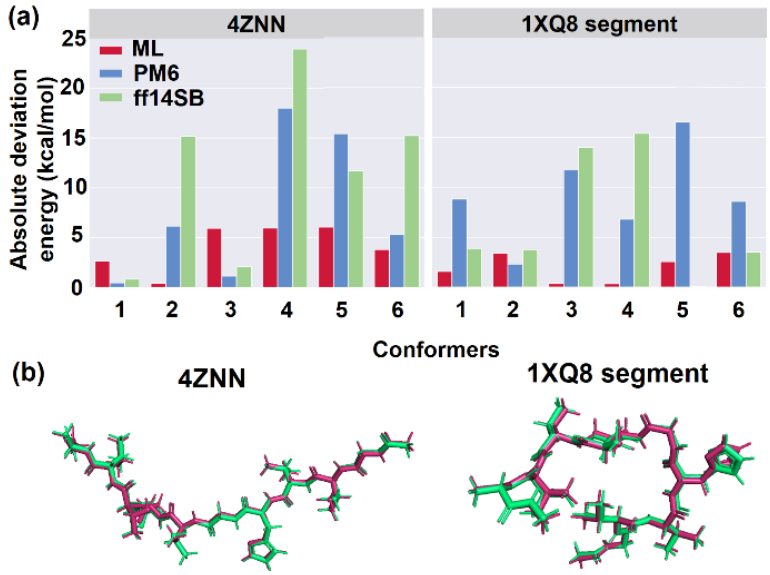

Figure 3. (a) The comparisons of the absolute deviations of the MLFF, PM6, and ff14SB relative energies (relative to the ωB97XD/6-31G* values) among 6 conformers. For both systems, the energy of the first conformer is taken as zero for each method. (b) Optimized structures of 4ZNN and 1XQ8 segment. The superposition between the structure obtained with our MLFFs (red) and the DFT-optimized structure (green) is shown for both systems.

Molecular Dynamics Simulation. To investigate the applicability of our MLFFs on MD simulation. we first perform MD simulations for two polypeptides in the microcanonical (NVE) ensemble. Figure S1 shows total energy fluctuations whose initial velocities are consistent with T = 300 K. As shown in

Figure S1b and S1d, the energy drifts are negligible during the GEBF-PM6 simulation for both systems. For 4ZNN, Figure S1a shows that the energy drift is about 0.001 kcal/(mol·atom·ps) during the MLFF-based MD simulation. For 1XQ8 segment the energy drift is even smaller during the MLFF-based MD simulation, as shown in Figure S1c. The energy drift of our MLFFs is much less than those in the AIMD simulations (for example, 0.023 kcal/(mol·atom·ps)[43,44] for sodium-ion batteries) and in eReaxFF reactive force field MD simulations [0.01kcal/(mol·atom·ps)].[45] Thus, our GEBF-MLFF could be employed for long-time MD simulations to investigate the conformational changes of two systems under study.

Then, MLFF-based MD simulations using a Langevin thermostat[46] are performed at 300 K with a timestep of 1 fs in the canonical (NVT) ensemble. To verify the accuracy of our MLFFs, we have performed 20-ps MD simulations with MLFFs, ff14SB and PM6 methods, respectively. MD simulations with ωB97X-D/6-31G* are also carried out for comparison. Figure 4 displays the dihedral angle distributions calculated with the MLFFs and ωB97X-D/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 MLFFs and ωB97X-D/6-31G* methods are very close to each other. The distributions predicted by the ff14SB and PM6 methods are plotted on Figure S2 and S3, respectively. The dihedral distributions from these two methods are quite different from the ωB97X-D/6-31G* results. For dihedrals φ and ψ, the shapes of distribution show 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 MLFFs are much more accurate than those from the ff14SB and PM6 methods.

After the accuracy of the MLFFs for MD simulation is validated, we perform 1-ns MD simulations for both systems using a Langevin thermostat at 300 K, starting at their chain-like structures. The end-to-end distances between the C_α atoms of the first and the last amino acid residues during 1-ns MD simulations are plotted in Figure 5. One can see that the end-to-end distances decrease rapidly in the first 0.2 ns and reach the minimum values about 4 Å during the rest of the simulation time. Three representative structures at different times are plotted in Figure 5. The results show that the conformation of the polypeptides gradually changes from the chain-like extended structure to the folded one, indicating a large conformational change during the MD simulations. Although the conformational changes are quite large and complex during the simulation for these two systems, MD simulations based on GEBF-MLFFs can be used to explore different regions of the potential energy surface with high accuracy. It can be expected that this GEBF-MLFF is applicable for accurately investigating the folding process of similar biological systems.

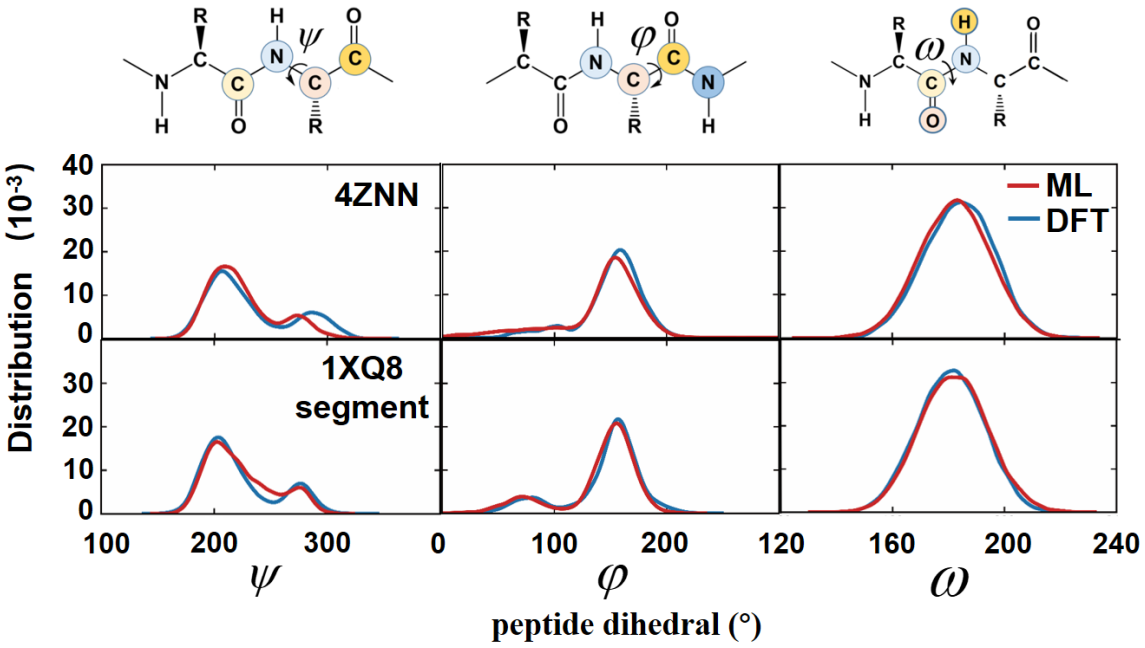

Figure 4. Backbone peptide dihedral distributions of 4ZNN (top) and 1XQ8 segment (bottom) obtained from 20-ps trajectories with DFT MD simulations (blue solid line) and MLFFs (red solid line), respectively. Distributions of dihedral angles, φ, ψ, and ω are shown from left to right, respectively.

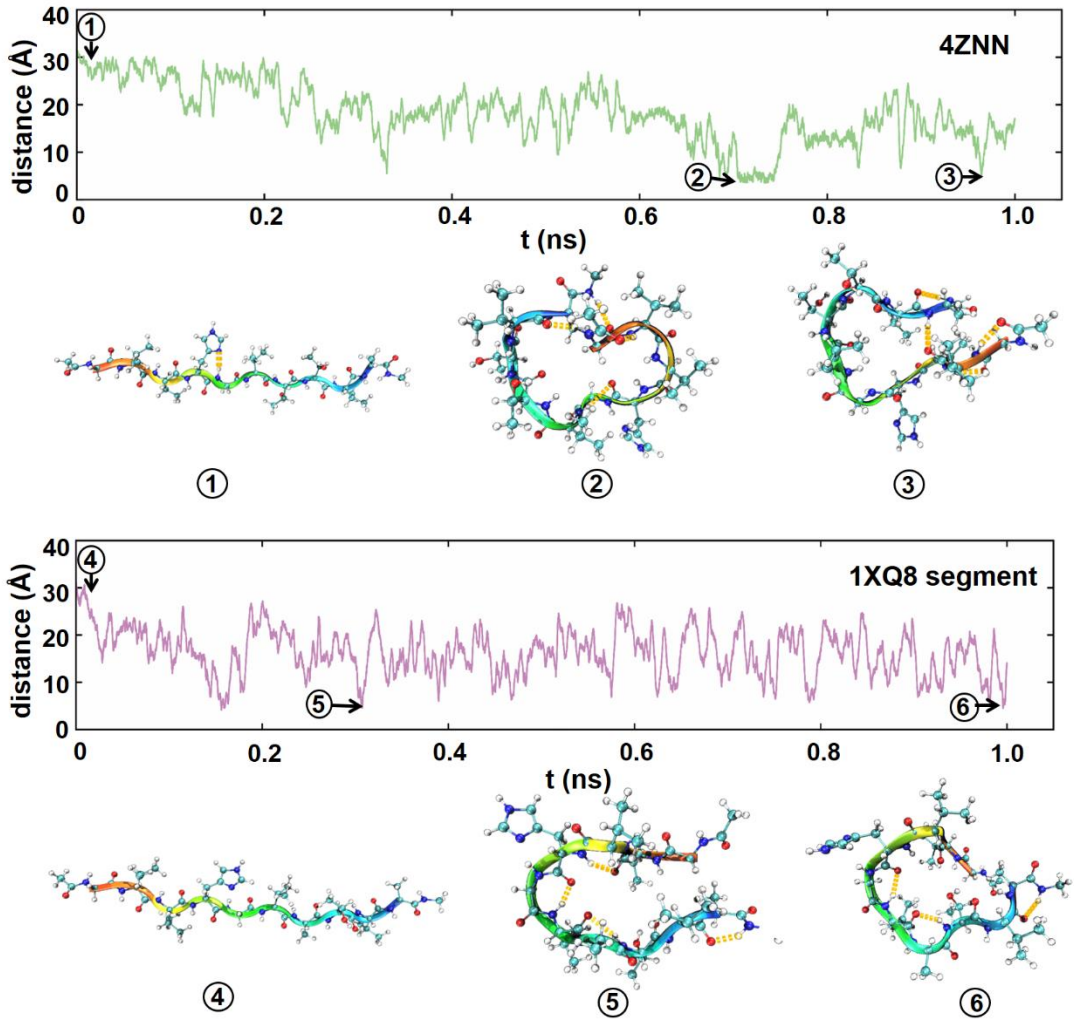

Figure 5. End-to-end distance of 4ZNN and 1XQ8 segment during MLFF-based MD simulations.

Discussion.

In summary, we have developed a general GEBF-ML protocol to automatically construct MLFFs for proteins with quantum mechanics accuracy. For a given protein, only QM calculations on small subsystems containing a few residues are required in the construction of MLFFs. To facilitate the construction of MLFFs 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 MLFF. This protocol was tested on two polypeptides 4ZNN and 1XQ8 segment. The accuracy of the constructed GEBF-MLFFs 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-MLFFs can lead to quite accurate energies and forces similar to those from full QM calculations, and dihedral angle distributions from GEBF-MLFF MD simulations are in good agreement with those from *ab initio* MD simulations. With this strategy, we expect that the GEBF-MLFFs with full QM accuracy can also easily be developed for other biological systems. Future work will aim to construct the GEBF-MLFFs for large proteins and other complex biological systems in vacuum and in aqueous solution. Eventually, GEBF-MLFF-based simulations are expected to be available for various biological systems in the physiological environments.

Methods

Computational details. The electronic structure calculations of these structures were carried out at the ωB97X-D/6-31G* level with the Gaussian 16 package[47], and the PM6 calculations were performed with MOPAC package.[48] The distance threshold and the maximum number of fragments in a subsystem are chosen as 3.0 Å and 4, respectively. Parameters of smooth overlap of atomic positions (SOAP) for 4ZNN and 1XQ8 segments are both listed in Table S3. The geometries were optimized with the BFGS algorithm[49] (implemented in ASE package[50]). The MD simulations were performed using the ASE package at NVT ensemble and the integration timestep is set as 1 fs.

GEBF-ML force field. In our GEBF-ML force field, we employ an atomic ML model (GAP) to learn the energy difference of each subsystem between ωB97X-D/6-31G* and PM6 method, which is expressed as follows,

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

where S_m is the *m*th subsystem, e_i^m is the atomic energy of the *i*th atom in the *m*th subsystem. After training, we can easily get the energy contribution of each atom with different local environments in subsystems. 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 all atomic contributions as shown below,

$$\Delta E^{\mathrm{ML}} = \sum_{i=1}^{N} e_i \tag{2}$$

Here, N is the number of atoms in the target system, e_i is the energy of the *i*th atom in the target system. The total energy of the target system is the combination of the energy difference and the PM6 energy (taken as the baseline)

$$E = \Delta E^{\mathrm{ML}} + E^{\mathrm{PM6}} \tag{3}$$

The PM6 energy of the target system is calculated using the GEBF-PM6 method as shown below,

$$E^{\mathrm{PM6}} = \sum_m C_m \left(E_m^{\mathrm{PM6}} - \sum_{A \in S_m} \sum_{B>A \in S_m} \frac{Q_A Q_B}{\mathbf{R}_{AB}} \right) + \sum_A \sum_{B>A} \frac{Q_A Q_B}{\mathbf{R}_{AB}} \tag{4}$$

In the GEBF method, the ground-state energy of a target system is obtained as the linear combination of ground-state energies of a series of small subsystems (including primary and derivative subsystems). The details of subsystem construction and discrimination can be found in the Sec.4 of the supporting information. E_m and C_m are the energy and coefficient of the *m*th subsystem, respectively, and *M* is the number of subsystems. 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 primary subsystems, which are generated from the first configuration during the MD simulation (and assumed to be constant for all of other configurations). $\mathbf{r_A}$ and Q_A denote the coordinate of atom A and the point charge locating on atom A, respectively. All ML models are based on kernel ridge regression with the SOAP kernels. Details of ML models are provided in the Sec. 5 of supporting information.

Outline of the MLFF construction To automatically construct machine learning force fields with high accuracy and efficiency, the GEBF-ML scheme was developed with active learning and transfer learning. The flowchart of the scheme is shown in Figure 6, in which the energy of the target system is predicted by "learning" from a subset of subsystems in the data library and some newly generated subsystems from online active learning. The details of each module in the flowchart are given below.

Starting from a given conformer, MD simulation with NVT ensemble at 500 K is performed based on the GEBF-ML force fields. During the simulation, subsystems are generated using our GEBF approach. If the subsystem types are already in the data library, the corresponding sub-datasets are loaded to the training set. Otherwise, online active learning (see details in Sec 6 of supporting information) is employed to select the representative subsystems. 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.

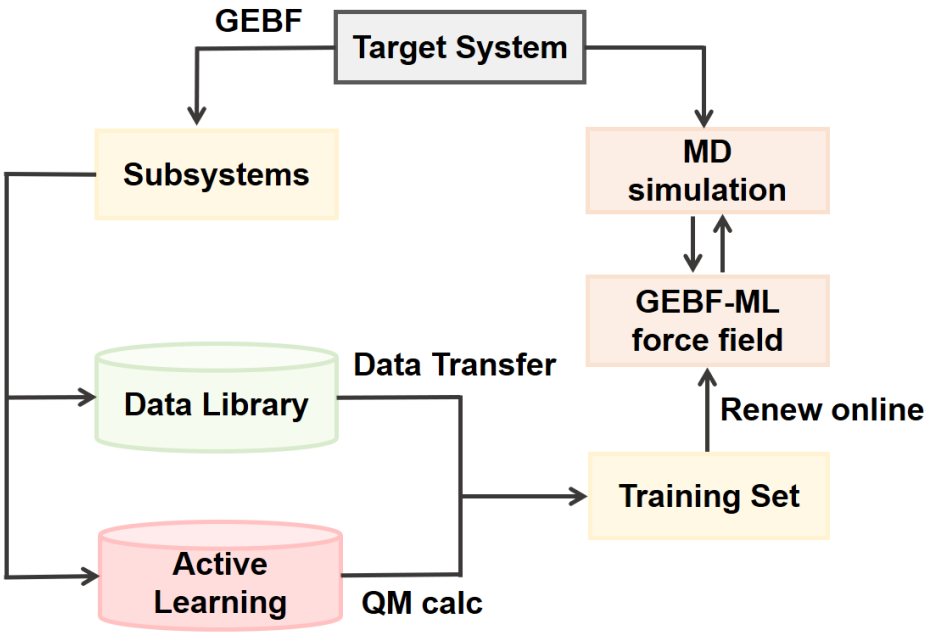

Figure 6. Scheme diagram of the GEBF-ML method. Training sets are constructed from relevant sub-datasets from the protein's data library and some subsystems from online active learning.

Data availability. The data that support the findings of this study are available from the corresponding author upon reasonable request.

Code availability. Codes are available from the corresponding upon reasonable request.

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*. **6**, 459-466 (2010).

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*. **26**, 1668-1688 (2005).

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. GROMOS++ Software for the Analysis of Biomolecular Simulation Trajectories. *J. Chem. Theory Comput.* **7**, 3379-3390 (2011).

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.* **110**, 1657-1666 (1988).

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*. **9**, 4046-4063 (2013).

6. Lamoureux, G. & Roux, B. Modeling induced polarization with classical Drude oscillators: Theory and molecular dynamics simulation algorithm. *J. Chem. Phys*. **119**, 3025-3039 (2003).

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*. **50**, 1625-1632 (2017).

8. Nerenberg, P. S. & Gordon-Heard, T. New developments in force fields for biomolecular simulations. *Curr. Opin. Struct. Biol*. **49**, 129-138 (2018).

9. Best, R. B., Buchete, N. V. & Hummer, G. Are current Molecular Dynamics Force Fields too Helical? *Biophys. J*. **108**, 132696 (2008).

10. Lindorff-Larsen, K., Maragakis, P., Piana, S., Eastwood, M. P., Dror, R. O. & Shaw, D. E. Systematic Validation of Protein Force Fields against Experimental Data. *PLoS One* **7**, e32131 (2012).

11. Behler, J. & Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **98**, No. 156401 (2007).

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*. **104**, No. 136403 (2010).

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*. **285**, 316 (2015).

14. Shapeev, A. V. Moment Tensor Potentials: A Class of Systematically Improvable Interatomic Potentials. *Multiscale Model. Simul.* **14**, 1153-1173 (2016).

15. Li, Z., Kermode, J. R. & De Vita, A. Molecular Dynamics with On-the-Fly Machine Learning of Quantum-Mechanical Forces. *Phys. Rev. Lett.* **114**, No. 096405 (2015).

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*, **15**, 448-455 (2019).

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*. **120**, No. 143001 (2018).

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*. **8**, 3192-3203 (2017).

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* 4441-4451 (2018).

20. Gastegger, M. & Marquetand, P. High-dimensional Neural Network Potentials for Organic Reactions and an Improved Training Algorithm. *J. Chem. Theory Comput*. **11**, 2187-2198 (2015).

21. Gastegger, M., Behler, J. & Marquetand, P. Machine learning molecular dynamics for the simulation of infrared spectra. *Chem. Sci.* **8**, 6924-6935 (2017).

22. Westermayr, J., Gastegger, M. & Marquetand, P. Combining SchNet and SHARC: The SchNarc Machine Learning Approach for Excited-State Dynamics. *J. Phys. Chem. Lett*. **11**, 3828-3834 (2020).

23. 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*. **589**, 59-64 (2021).

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

25. 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.* **116**, 1110-1115 (2019).

26. 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*. **10**, 4962-4967 (2019).

27. 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.* **122**, No. 225701 (2019).

28. 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*. **54**, 169-181 (2021).

29. 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*. **16**, 2995-3005 (2020).

30. Collins, M. A., Cvitkovic, M. W. & Bettens, R. P. A. The Combined Fragmentation and Systematic Molecular Fragmentation Methods. *Acc. Chem. Res.* **47**, 2776-2785 (2014).

31. 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.* **125**, No. 104109 (2006).

32. Dahlke, D. E. & Truhlar, D. G. Electrostatically Embedded Many-Body Expansion for Large Systems, with Applications to Water Clusters. *J. Chem. Theory Comput.* **3**, 46-53 (2006).

33. Gordon, M. S., Fedorov, D. G., Pruitt, S. R. & Slipchenko, L. V. Fragmentation Methods: A Route to Accurate Calculations on Large Systems. *Chem. Rev*. **112**, 632-672 (2011).

34. 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*. **124**, No. 184703 (2006).

35. Bettens, R. P. A. & Lee, A. M. A New Algorithm for Molecular Fragmentation in Quantum Chemical Calculations. *J. Phys. Chem. A* **110**, 8777-8785 (2006).

36. Huang, L., Massa, L. & Karle, J. Kernel energy method illustrated with peptides. *Int. J. Quantum Chem*. **103**, 808-817 (2005).

37. 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.* **137**, No. 064113 (2012).

38. Mayhall, N. J. & 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.* **8**, 2669-2675 (2012).

39. Wang, H. & Yang, W. Toward Building Protein Force Fields by Residue-Based Systematic Molecular Fragmentation and Neural Network. *J. Chem. Theory Comput.* **15**, 1409-1417 (2018).

40. 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*

124, 3027-3035 (2020).

41. Pan S. J. & Yang, Q. A survey on transfer learning. IEEE Transactions on knowledge and data engineering. **22**, 1345-1359 (2010).

42. Cheng, Z., Zhao, D., Ma, J., Li, W. & Li, S. An On-the-Fly approach to Construct Generalized Energy-Based Fragmentation Machine Learning Force Fields of Complex Systems. *J. Phys. Chem. A* **124**, 5007-5014 (2020).

43. Liu, J., Zhang, C., Xu, L. & Ju, S. Borophene as a promising anode material for sodium-ion batteries with high capacity and high rate capability using DFT. *RSC Adv*. **8**, 17773-17785 (2018).

44. Lv, X., Xu, Z., Li, J., Chen, J. & Liu, Q. First-principles molecular dynamics investigation on Na_3AlF_6 molten salt. *J. Fluorine Chem*. **185**, 42-47 (2016).

45. Islam, M. M., Kolesov, G., Verstraelen, T., Kaxiras, E. & van Duin, A. C. T. eReaxFF: A Pseudoclassical Treatment of Explicit Electrons with Reactive Force Field Simulations. *J. Chem. Theory Comput*. **12**, 3463-3472 (2016).

46. Langevin, P. Sur la theories du mouvement brownien. *C. R. Acad. Sci. Pairs*. **146**, 530-533 (1908).

47. 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.

48. Stewart, J.J.P. Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements. *J Mol Model*. **13**, 1173-1213 (2007).

49. Fletcher, R. Practical Methods of Optimization; *John Wiley & Sons*: **2013**.

50. Larsen, A. H., Mortensen, J. J., Blomqvist, J., Castelli, I. E., Christensen, R., Dular, M., Friis, J., Groves, M. N., Hammer, B. & Hargus, C. The atomic simulation environment – a Python library for working with atoms. *J. Phys.: Condens. Matter*. **29**, 273002 (2017).

Additional information

Acknowledgements 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 his fruitful discussions.

Supplementary information is available for this paper at ….

Competing interests: The authors declare no competing financial interest.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to S.H.L

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.