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

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 field (MLFF) for proteins only with quantum mechanics (QM) calculations of small 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. With this data library, for a new protein only its subsystems with new topological types are required for the construction of the corresponding MLFF. With two polypeptides, 4ZNN and 1XQ8 segment, as examples, the energies and forces predicted by MLFF are in good agreement with those from QM calculations, and dihedral angle distributions from GEBF-MLFF molecular dynamics (MD) simulations can also well reproduce those from *ab initio* MD simulations. Therefore, the present work provides an efficient and systematic way to build force fields for biological systems like proteins with QM accuracy.

Molecular dynamics (MD) simulation has emerged as an import 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.[11-19] 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] 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 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 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 representation of different local regions of a large system. In previous studies, a residue-based neural work (NN) approach[39,40] was proposed to construct NN potentials for 20 types of amino acid capped with 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 is 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 with 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.[39]

In this work, we propose a 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 scheme,[39,40] 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. 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. Thus, structure optimization and MD simulations on complex proteins can be performed with high QM accuracy and low computational costs.

To automatically construct the subsystems on training set, a fragmentation method called generalized energy-based fragmentation (GEBF) approach developed by our group [28] 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 representation of the local chemical environment of different regions in a protein than those in residue-based NN approach. Using PM6 method as baseline, an atomic ML model called GAP[12] based on kernel ridge regression with the SOAP kernels[41] (see details in the Sec.2 of the supporting information) is chosen to learn the energy difference of all primary subsystems for the studied protein. In GAP, the energy difference ΔE_m^{ML} of the mth subsystem with S_m atoms are described as the summation

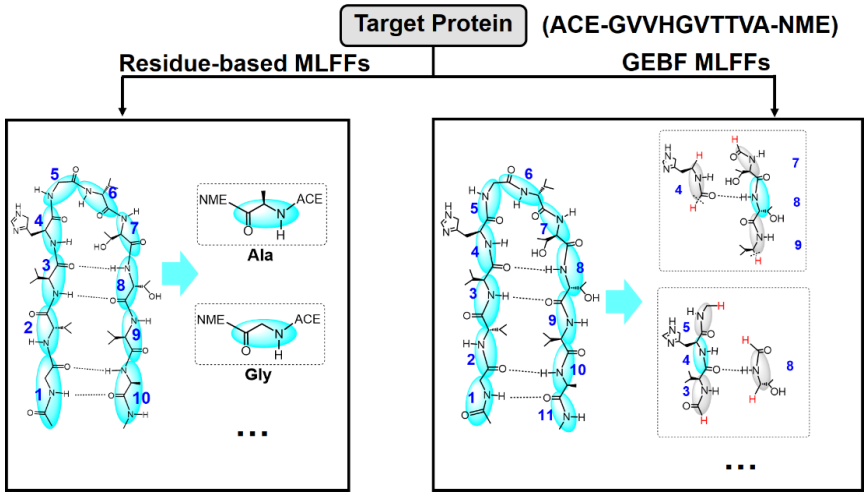


Figure 1. Fragmentation scheme utilized in the construction of MLFFs. In our GEBF method, fragments are capped with its environmental fragments or hydrogen atoms if necessary. In previous residue-based method, fragments are capped with an acetyl group (ACE) and *N*-methylamide group (NME).

of atomic energy e_i^m ,

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

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 with N atoms are obtained with the summation of atomic contribution e_i .

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

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)

$$E = \Delta E^{\mathrm{ML}} + E^{\mathrm{PM6}} \qquad (3)$$

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

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

Details of subsystem construction and determination of coefficients are explained in the 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 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.

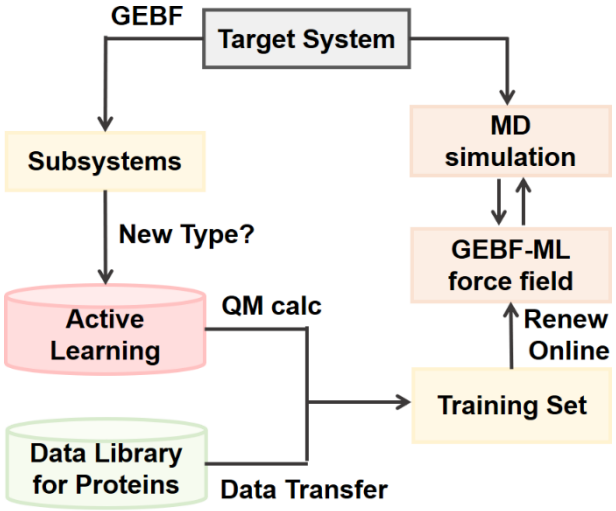


Figure 2. Scheme diagram of the GEBF-ML method. Training sets are constructed from relevant sub-datasets from the protein's data library and some subsystems from online active learning.

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[42] 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-ML force fields. During the simulation, subsystems are 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[43] (see details in Sec5 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.

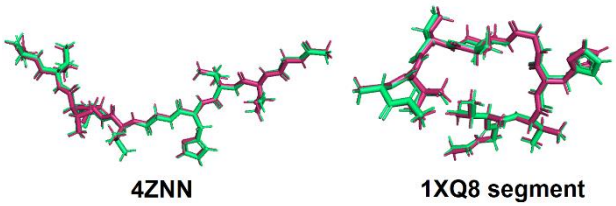


Figure 3. Optimized structures of 4ZNN and 1XQ8 segment. The superposition between the structure obtained with our MLFFs (red) and the DFT-optimized structure (green) is shown for both systems.

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. The training set of subsystems are generated from a 1ns canonical (NVT) MD trajectory at 500K based on the protein library and online active learning. For 4ZNN, QM calculations are carried out for only 0.15% of generated subsystems. As 4ZNN and 1XQ8 segments differ from each other by only one amino acid residue, a large number of subsystems generated from 4ZNN can be reused. For 1XQ8 segment, only 0.01% of newly generated subsystems are needed for QM calculations during the online active learning. Thus, our GEBF-ML scheme shows high efficiency for MLFFs construction. The testing set of two target systems are randomly sampled from a 1ns NVT GEBF-ML MD trajectories at 300 K. The mean absolute errors (MAEs) of energies between GEBF-PM6 and PM6 on testing set are only 0.003 kcal/(mol·atom). The RMSEs between the energy and forces of MLFF and ωB97XD/6-31G* results on testing set are less than 0.024 kcal/(mol·atom) and 1.5 kcal/(mol·Å), respectively. Thus, the MLFFs could predict the energies and forces with near-QM quality. (see details in the Sec.6 of the supporting information.)

To test whether MLFFs are suitable for structure optimization. The conformers with the lowest energy predicted by MLFFs in test sets are optimized with the BFGS algorithm[44] (implemented in ASE package[45]). Figure 3 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.

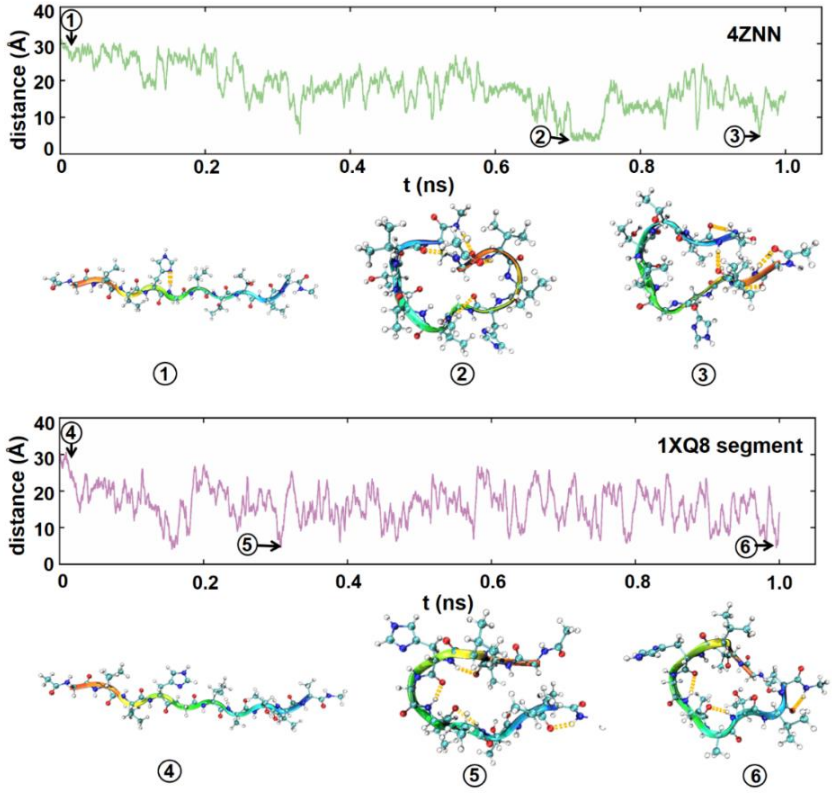


Figure 4. End-to-end distance of 4ZNN and 1XQ8 segment during ML-based MD simulations.

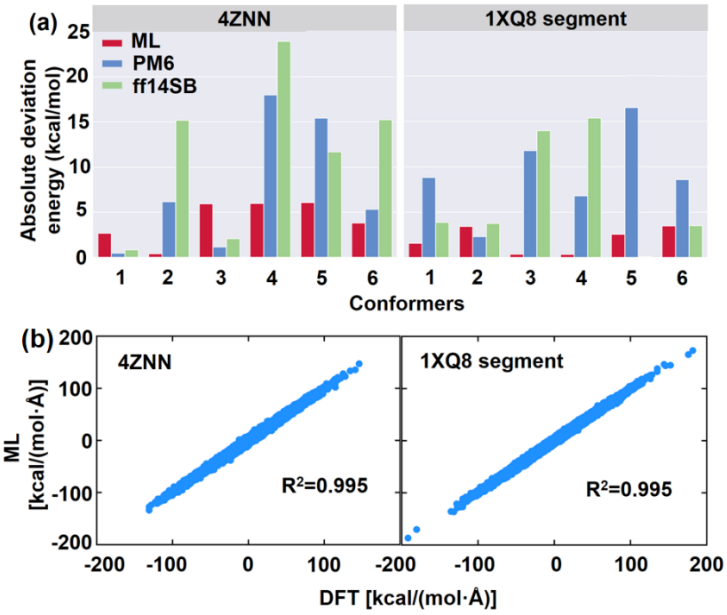


Figure 5. (a) The comparison of the absolute deviations of the MLFF, PM6, and ff14SB relative energies (relative to the ωB97XD/6-31G* values) among 6 conformers. (b) The comparisons of correlations between the forces from MLFFs and the ωB97XD/6-31G* ones.

Then, we investigate the applicability of our MLFFs on MD simulation. As MLFF-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 the Sec.7 of the supporting information.), long-time 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. Staring from the chain-like structures for both two systems, 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 4. 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 4. 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.

To verify the performance of our MLFFs in all conformation space during the MD simulations, first, we compare the relative energies for six conformers randomly chosen from the MLFF-based trajectories. Here, the energies of the six conformers are calculated with MLFFs, PM6, ff14SB and ωB97XD/6-31G*. The energy of the first conformer calculated with each method was taken as zero. The absolute deviations of relative energies (relative to the ωB97XD/6-31G* results) are shown in Figure 5a. 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. Then, the correlations between the forces from MLFFs and the ωB97XD/6-31G* one for 100 conformers randomly chosen form the trajectories are plotted in Figure 5b. The coefficient of determination (R^2) between these results and ωB97XD/6-31G* results is 0.995 (MLFFs), indicating that the forces predicted by MLFFs are almost the same with that from reference ωB97XD/6-31G* calculations. The correlation between the forces from PM6 and ff14SB and the ωB97XD/6-31G* ones are also plotted in Figure S1, the R^2 is 0.56 for PM6 and 0.67 for ff14SB, both are much small than the MLFFs.

Finally, we also 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 6 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. Thus, MD simulations based on GEBF-MLFFs can be used to explore different regions of the potential energy surface with high accuracy.

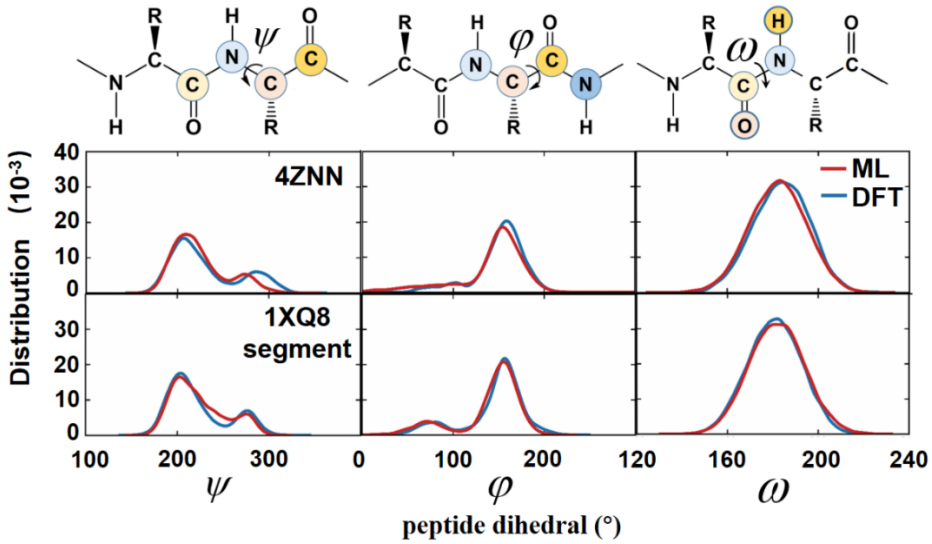


Figure 6. Backbone peptide dihedral distributions of 4ZNN (top) and 1XQ8 segment (bottom) obtained from 20 ps trajectories with reference DFT (blue solid line) and ML (red solid line). Distributions of dihedral angles, φ, ψ and *ω* are shown from left to right, respectively

In summary, we developed a general GEBF-ML protocol to automatically construct MLFFs for proteins with QM 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. 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 solution 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) 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.

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

(22) 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.

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

(24) 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.

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

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

(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.* **2019**, *122*, 225701.

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

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

(30) 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.

(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.* **2006**, *125*, 104109.

(32) 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.

(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*. **2011**, *112*, 632-672.

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

(35) 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.

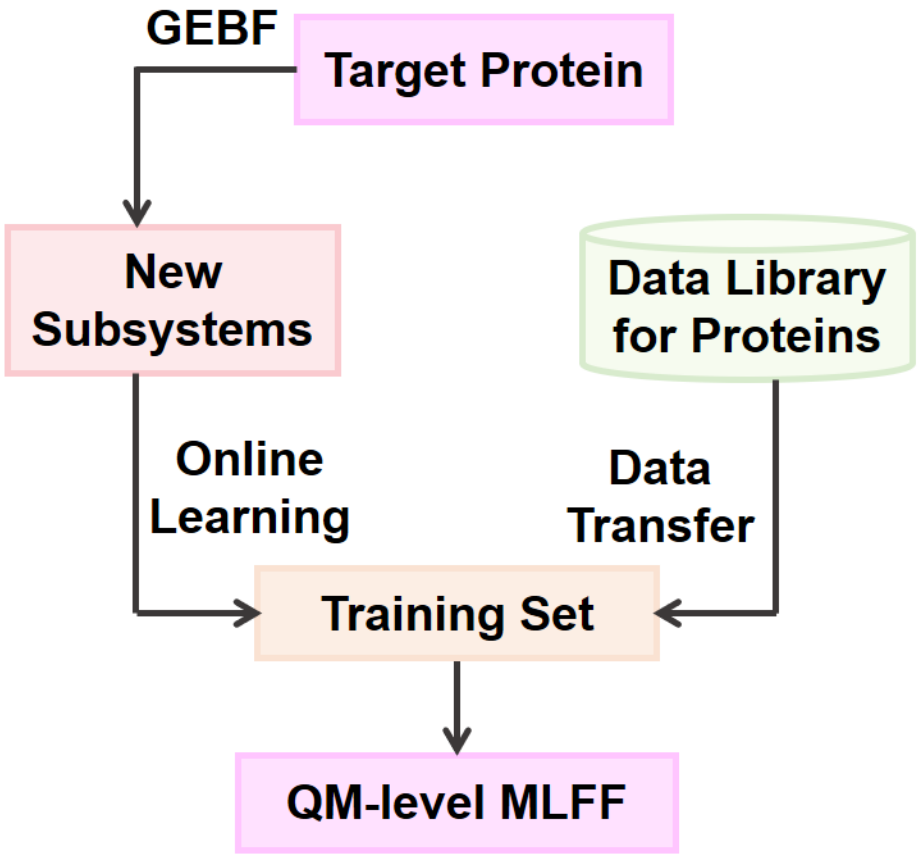
(36) Huang, L.; Massa, L.; Karle, J. Kernel energy method illustrated with peptides. *Int. J. Quantum Chem*. **2005**, *103*, 808-817.

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

(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.* **2012**, *8*, 2669-2675.
(39) 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.
(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* **2020**, *124*, 3027-3035.
(41) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Phys. Rev. B* **2013**, *87*, No. 184115.
(42) Pan S. J. and Yang, Q. A survey on transfer learning. IEEE Transactions on knowledge and data engineering. **2010**, *22*, 1345-1359.
(43) 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.
(44) Fletcher, R. Practical Methods of Optimization; *John Wiley & Sons:* **2013**.
(45) 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.
(46) Langevin, P. Sur la theories du mouvement brownien. *C. R. Acad. Sci. Pairs*. **1908**, *146*, 530-533.

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 field with QM quality at the cost of some QM calculations on subsystems not stored in data library. This work takes an important step into the practical computational study of biological systems with QM accuracy.