The Journal of Chemical Physics ARTICLE pubs.aip.org/aip/jcp

Toward a general neural network force field for protein simulations: Refining the intramolecular interaction in protein

Cite as: J. Chem. Phys. 159, 024118 (2023); doi: 10.1063/5.0142280
Submitted: 12 January 2023 • Accepted: 22 June 2023 •
Published Online: 11 July 2023

Pan Zhang and Weitao Yang[a)]

AFFILIATIONS

Department of Chemistry, Duke University, Durham, North Carolina 27708, USA

Note: This paper is part of the JCP Special Topic on Machine Learning Hits Molecular Simulations.
[a)]**Author to whom correspondence should be addressed:** weitao.yang@duke.edu

ABSTRACT

Molecular dynamics (MD) is an extremely powerful, highly effective, and widely used approach to understanding the nature of chemical processes in atomic details for proteins. The accuracy of results from MD simulations is highly dependent on force fields. Currently, molecular mechanical (MM) force fields are mainly utilized in MD simulations because of their low computational cost. Quantum mechanical (QM) calculation has high accuracy, but it is exceedingly time consuming for protein simulations. Machine learning (ML) provides the capability for generating accurate potential at the QM level without increasing much computational effort for specific systems that can be studied at the QM level. However, the construction of general machine learned force fields, needed for broad applications and large and complex systems, is still challenging. Here, general and transferable neural network (NN) force fields based on CHARMM force fields, named CHARMM-NN, are constructed for proteins by training NN models on 27 fragments partitioned from the residue-based systematic molecular fragmentation (rSMF) method. The NN for each fragment is based on atom types and uses new input features that are similar to MM inputs, including bonds, angles, dihedrals, and non-bonded terms, which enhance the compatibility of CHARMM-NN to MM MD and enable the implementation of CHARMM-NN force fields in different MD programs. While the main part of the energy of the protein is based on rSMF and NN, the nonbonded interactions between the fragments and with water are taken from the CHARMM force field through mechanical embedding. The validations of the method for dipeptides on geometric data, relative potential energies, and structural reorganization energies demonstrate that the CHARMM-NN local minima on the potential energy surface are very accurate approximations to QM, showing the success of CHARMM-NN for bonded interactions. However, the MD simulations on peptides and proteins indicate that more accurate methods to represent protein–water interactions in fragments and non-bonded interactions between fragments should be considered in the future improvement of CHARMM-NN, which can increase the accuracy of approximation beyond the current mechanical embedding QM/MM level.

I. INTRODUCTION

Molecular dynamics (MD) simulation is an effective and widely used technique to investigate proteins, such as protein folding and unfolding, protein–ligand binding, and protein conformational and structural analysis.[1–5] One of the most important factors related to the accuracy of results from MD simulations is the force field.[6–8] The traditional molecular mechanical (MM) force fields outperform the quantum mechanical (QM) or QM/MM force fields because they have much simpler energy representations, which are much more efficient on large protein systems. The potential energy described by MM force fields usually consists of bond, angle, dihedral, improper dihedral, and non-bonded energies such as electrostatic energy and van der Waals energy.[9] The parameters of all energy terms are generally fitted to experimental data, such as macroscopic properties, spectroscopic and crystallographic data, and quantum mechanical (QM) calculations.[10] Four major families of general force fields, AMBER,[11–14] CHARMM,[15–18] GROMOS,[19–22] and OPLS,[23–26] are widely used in MD simulations, and many improvements have been achieved to increase the applicability of these force fields over the years. In addition to the traditional fixed-charge MM force fields, polarizable force fields were developed to account

for the polarization effect[27] based on fluctuating charges,[28] Drude oscillators,[29] inducible dipoles,[30] and methods including multipole electrostatics.[31]

Even though the current protein force fields are applicable to many problems, they are still not accurate enough and could fail in some situations.[32,33] For example, protein folding pathways and thermodynamics cannot be identified very well,[34,35] and the strength of protein–water interactions is not balanced well compared to protein–protein interactions.[36,37] In addition, most force fields have deficiencies in simulating intrinsically disordered proteins (IDPs), which lack well-defined three-dimensional structures but have full functionalities.[38,39] Since an IDP can only be characterized with an ensemble of flexible interconverting conformations, the MD simulation approach would be an ideal approach.[40–42] This suggests the urgency of developing more accurate force fields. The current force fields generally generate a too compact ensemble of IDP,[43–46] and two types of strategies were generally applied to improve force fields for IDP.[47,48] The first one is adjusting backbone dihedral or energy correction map (CMAP) parameters to avoid the high propensity of secondary structures for α-helix or β-sheet, such as ff14IDPSFF,[49] CHARMM36IDPSFF,[50,51] OPLSIDPSFF,[52] RSFF2C,[53] and ESFF.[54] The second strategy is refining protein–water interactions by adjusting non-bonded parameters to achieve a better description of protein–water interactions, such as a99SB-UCB,[55] ff03ws,[37] and KBFF.[56–58] Moreover, specific coarse-grained force fields were also developed for IDP,[59,60] such as AWSEM-IDP[61] and MOFF-IDP.[62] With the continuous development of force fields, the simulation results for IDP were improved, but there are still unsolved issues, such as the limitation of studying the post-translation modifications in IDP and the difficulty of obtaining balanced local and global structural features.[41,48] This is caused by the simplicity and limitations of force field forms in MM representations. According to an investigation on the structural reorganization energy defined by the energy difference between the MM optimal structure and the QM optimal structure, the minimization of QM optimal structures with the major AMBER, CHARMM, GROMOS, and OPLS force fields all resulted in large reorganization energies, which suggests that the MM force field forms need to be improved.[63]

Recently, machine learning techniques were widely and successfully integrated into force field development because of their capability to describe potential energy surfaces (PESs) at a high level of theory at a low computational cost.[64–68] Although machine learning methods can be utilized to adjust MM force field parameters,[69–72] our focus is primarily on the construction of general force fields with machine learning models because of the potential improvement of functional form. The common machine learning models used to build force fields are divided into two categories: kernel-based methods and artificial neural networks (NNs). One example of kernel-based methods is the Gaussian approximation potential, which is approximated by the Gaussian process regression on different descriptors and kernels,[73] such as local atomic density and the SOAP kernel.[74] Another kernel-based method is gradient domain machine learning (GDML), which is trained to predict the force directly rather than the energy.[75–78] For artificial NNs, high-dimensional neural network potential (HDNNP) was the first descriptor-based neural network potential (NNP) that uses atomic-centered symmetry functions as descriptors,[79,80] and subsequent developments included the consideration of long-range interaction and non-local phenomena.[81–83] Some different NNPs were also developed based on different descriptors such as ANI,[84] TensorMol,[85] deep potential,[86] QM/MM-NN,[87,88] water NN,[89] and embedded atom NN.[90,91] Another type of NN is end-to-end NN, which directly uses the Cartesian coordinates and nuclear charges as inputs instead of descriptors.[92] End-to-end NN is usually based on graph NN,[93] like message-passing NN,[94] and some prominent examples are deep tensor NN,[95,96] SchNet,[97,98] and PhysNet.[99,100] Machine learning provides a powerful and promising way to develop force fields with QM accuracy and excellent efficiency.

Despite the success of machine learning force fields in a plethora of chemical problems, including electronic effects, thermodynamics, reactions, and spectroscopies,[101] the development of general, transferable, and scalable force fields for proteins is still very challenging for the following reasons. First, to obtain plenty of data for the training of a machine learning model, the calculation of QM reference energy is infeasible for proteins since they could have thousands of atoms. Second, the size of proteins for MD simulations could range from short peptides to large proteins, which increases the difficulty of training a general machine learning force field that is applicable to different systems. To address these issues, fragmentation methods can be utilized to systematically express the total energy of large systems with the individual energy of small fragments.[102] Some common fragmentation methods include divide and conquer,[103] many-body expansion,[104–107] systematic molecular fragmentation (SMF),[108–110] the effective fragment potential approach,[111] electrostatically embedded generalized molecular fractionation with conjugate caps (EE-GMFCC),[112–115] the X-pol method,[116–119] and the energy-based fragmentation method.[120,121] In 2019, Hao and Yang developed residue-based systematic molecular fragmentation (rSMF) to enable the construction of general fragments for any proteins, and the neural network force field (NNFF) was trained on glycine and alanine dipeptides to demonstrate the transferability to mixed alanine and glycine polypeptides.[122] More machine learned force fields based on similar fragmentation methods were developed,[123–125] but a comprehensive and transferable force field that can be applied to general protein simulations is still not available.

In this work, we apply the rSMF method to obtain fragments that are trained with NN and develop an NNFF based on the CHARMM force field, named CHARMM-NN. The CHARMM-NN is based on atom types, and the input features are similar to the variables in CHARMM energy terms, such as bonds, angles, dihedrals, and non-bonded features, which is highly compatible with the original MM simulations. The CHARMM-NN is first validated by the geometric data and structural reorganization energies, and more simulations of peptides and proteins are performed to test the quality of this force field.

II. METHODS

A. Residue-based systematic molecular fragmentation

We approach the construction of the protein force field by partitioning the protein into embedded fragments and developing the NN description of fragments with input data from QM calculations.[122] Based on the original SMF method,[108,126] residue-based systematic molecular fragmentation (rSMF) was developed by Wang

and Yang in 2019,[122] which generates fragments that can be applied to any protein. The fragments are based on amino acid residues, and different levels of rSMF can be applied. The rSMF at level 1 generates dipeptides, defined by one side chain group and two peptide bonds, as the basic fragments. A higher level of rSMF can be applied to increase the accuracy of fragmentation, but the sampling effort as well as the computation of QM energies and forces will increase significantly. Therefore, the rSMF used in this work is at level 1. Using the rSMF method, a protein with N residues capping with an acetyl (ACE) group on the *N* terminus and an N-methyl amide (NME) group on the C terminus can be divided into *N* dipeptides and $N-1$ ACE-NME fragments. As an example, a tripeptide P^3 is fragmented, as shown in Fig. 1, where A_1 and A_2 are the corresponding dipeptides for each amino acid in the tripeptide, and ACE-NME is the peptide bond between A_1 and A_2.

The test of the rSMF method on homogeneous and heterogeneous alanine and glycine polypeptides suggests that this method can provide stability of errors with increasing size, which is highly desirable and necessary in protein force field development since the protein size could be arbitrarily large. Based on this behavior, we construct NN models for the ACE-NME fragment and 24 dipeptide fragments, including three dipeptides for different protonation states of His (Hse, Hsd, and Hsp), two protonated states of negative charged amino acid dipeptides (AspH and GluH), and 19 other natural amino acid dipeptides (Gly, Ala, Val, Cys, Pro, Leu, Ile, Met, Trp, Phe, Ser, Thr, Tyr, Asn, Gln, Lys, Arg, Asp, and Glu). Even though these fragments are enough for some proteins, many proteins contain disulfide bonds formed by two cysteines, which cannot be described by the fragments. Therefore, to enable the applicability on more proteins, we complement the rSMF method with two more fragments: one is the dipeptide fragment of cysteine with an SMe group connected to the S atom, named Cys-SMe, and another is the dimethyl disulfide (DMDS) fragment that accounts for the overlap between two cysteines that form a disulfide bond. As shown in Fig. 2, the molecule of two cysteines connected by a disulfide bond can be fragmented into two Cys-SMe dipeptides and a DMDS fragment. The CHARMM-NN includes 27 fragments in total, and each fragment has its own NN model.

The hydrogen atoms are capped by the broken bonds according to

$$x(H) = x(i) + \frac{r(i)+r(H)}{r(i)+r(j)}[x(j)-x(i)], \tag{1}$$

where $x(H)$ is the coordinates of capping hydrogen atoms, $x(i)$ and $x(j)$ are the coordinates of atom i and atom j that form the bond to be broken in forming the fragment, respectively, and $r(i)$, $r(j)$, and $r(H)$ are the standard covalent radius for elements i, j, and H, respectively. Since the NN is trained on the difference between QM energy and MM energy, the total energy of a protein with n amino acid residues, denoted as P^n, can be represented as

$$\begin{aligned}
E[P^n] &= E_b[P^n] + E_{nb}[P^n] \\
&= \sum_{i=1}^{n} E_{\mathrm{QM}}[A_i] - \sum_{i=1}^{n-1} E_{\mathrm{QM}}[(\mathrm{ACE}-\mathrm{NME})_i] \\
&\quad - \sum_{i=1}^{m} E_{\mathrm{QM}}[\mathrm{DMDS}_i] + E_{\mathrm{nb,MM}}[P^n] \\
&= \sum_{i=1}^{n} (E_{\mathrm{MM}}[A_i] + E_{\mathrm{NN}}[A_i]) \\
&\quad - \sum_{i=1}^{n-1} (E_{\mathrm{MM}}[(\mathrm{ACE}-\mathrm{NME})_i] + E_{\mathrm{NN}}[(\mathrm{ACE}-\mathrm{NME})_i]) \\
&\quad - \sum_{i=1}^{m} (E_{\mathrm{MM}}[\mathrm{DMDS}_i] + E_{\mathrm{NN}}[\mathrm{DMDS}_i]) + E_{\mathrm{nb,MM}}[P^n] \\
&\approx \sum_{i=1}^{n} E_{\mathrm{NN}}[A_i] - \sum_{i=1}^{n-1} E_{\mathrm{NN}}[(\mathrm{ACE}-\mathrm{NME})_i] \\
&\quad - \sum_{i=1}^{m} E_{\mathrm{NN}}[\mathrm{DMDS}_i] + E_{\mathrm{MM}}[P^n].
\end{aligned} \tag{2}$$

The NN correction energy is the sum of the energies of all dipeptide fragments minus the energies from ACE-NME and DMDS, and the total energy is the sum of the MM energy and the NN correction energy. This is an approximation because the nonbonded interactions between the capped hydrogen atoms and other atoms in the dipeptide fragment and ACE-NME fragment are not completely the same, even though the additional bond, angle, and dihedral energy terms that are related to the capped hydrogen atoms can be exactly canceled each other. Based on this approximation, we can run normal MD simulations with C36m force fields and apply the additional correction from NNFF, which is very convenient for integration into existing MD programs without the requirement to modify core codes.

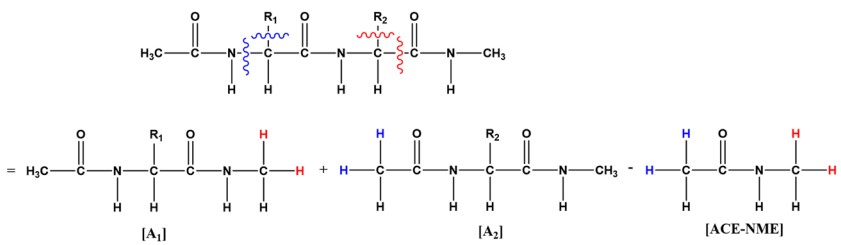

FIG. 1. The fragmentation of a tripeptide with rSMF at level 1.

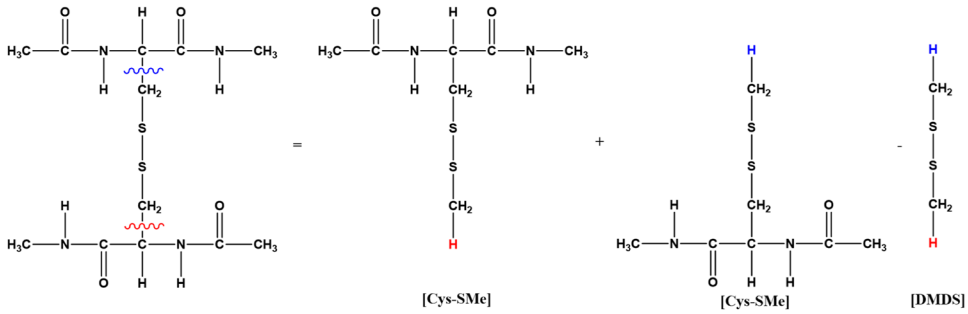

FIG. 2. The fragmentation of two cysteines connected by a disulfide bond.

B. Atom type based NN and MM based input features

For each fragment, the total correction energy from NN is the sum of atomic NN energy, expressed as

$$E_{\mathrm{NN}}^{\mathrm{tot}}[A] = \sum_{i=1}^{n} \sum_{j=1}^{m_i} E_{\mathrm{NN}}^{i}(G_j^i), \tag{3}$$

where n is the total number of atom types in fragment A, m_i is the number of atoms for atom type i, E_{NN}^{i} is the NN energy for atom type i, and G_j^i is the input vector for atom j in atom type i. The atom types are exactly the same as the atom types in the CHARMM force fields for each fragment. For example, since glycine dipeptide has atom types NH1, H, CT2, HB, CT3, HA, C, and O, the atomic NN with these eight atom types is trained. The input features G_j^i are also obtained based on atom types, and they are similar to the variables in MM potential energy terms, represented as

$$G_j^i = \left\{ \mathbf{d}_j, \boldsymbol{\theta}_j, \cos\boldsymbol{\phi}_j + 2, \sin\boldsymbol{\phi}_j + 2, X_j, Y_j \right\}^i, \tag{4}$$

where $\mathbf{d}_j$ is the vector of bond lengths containing atom j, $\boldsymbol{\theta}_j$ is the vector of angles containing atom j, and $\boldsymbol{\phi}_j$ is the vector of dihedral angles containing atom j. The +2 in the dihedral angle variables makes the variables in range of [1, 3], so they have non-zero values for the dihedrals included in an atom, which differentiate the zero value from the dihedrals not included in the atom. Two more input variables X_j and Y_j are the inputs accounting for electrostatic and van der Waals interactions, expressed as

$$X_j = \sum_k \frac{1}{r_{jk}}, \tag{5}$$

and

$$Y_j = \sum_k \left(\frac{2}{r_{jk}} \right)^6, \tag{6}$$

where k are the atoms that do not form a bond, angle, and dihedral angle with atom j in the fragment, and r_{jk} is the distance between atom j and atom k. For the example of CT3 atom type in glycine dipeptide, the $\mathbf{d}_j$ inputs contain the CT3-NH1, CT3-HA, and CT3-C bonds; the $\boldsymbol{\theta}_j$ inputs contain the NH1-CT3-HA, HA-CT3-HA, HA-CT3-C, CT3-NH1-H, CT3-NH1-C, CT3-C-O, and CT3-C-NH1 angles; and the $\boldsymbol{\phi}_j$ inputs contain the HA-CT3-NH1-H, HA-CT3-NH1-C, HA-CT3-C-NH1, HA-CT3-C-O, CT3-NH1-C-O, CT3-NH1-C-CT2, CT3-C-NH1-H, and CT3-C-NH1-CT2 dihedral angles. The total number of input variables for the glycine CT3 atom type is thus 28, including three bonds, seven angles, 16 variables for dihedral angles, and two non-bonded input variables. The input features for other atom types and other fragments are obtained similarly. Using the MM based input features can reduce the computational cost not only because of their simplicity but also because they can be directly obtained or derived from every step of MD simulations when those MM variables are calculated. Therefore, the atom type based NN and MM based input features are well compatible with the MD simulations with MM force fields, which is the baseline before the NN correction is applied.

C. Computational details

In this work, the C36m force fields[18] were used for all MM calculations, and the B3LYP-GD3BJ/6-31 + G(d) force fields[127–131] were used for all QM calculations. Even though some force field parameters may be fitted with a higher level of QM theory, we calculated QM energy with B3LYP-GD3BJ/6-31 + G(d) for the consideration of computational efficiency on large amounts of data. Future development can certainly use higher level QM theory with the same ML method. To construct the dataset for all fragments, two strategies were combined. First, normal mode sampling (NMS)[84,132,133] was performed on the dataset of stationary points for all fragments. The initial data were constructed using two approaches. In the first approach, 24 dipeptide fragments were extracted from the NOMAD dipeptide dataset that is optimized with the PBE + vdW method,[134] and the Cys-SMe initial structures were obtained by replacing the hydrogen atoms with SMe groups for all Cys dipeptides. The initial structures of all dipeptide fragments were optimized with B3LYP-GD3BJ/6-31 + G(d), and similar structures were removed based on root mean square deviation (RMSD) and energy. For the two connecting fragments, optimization resulted in 2 and 1 local minima for the ACE-NME and DMDS fragments, respectively. This approach provides the initial conformers at local minima or stationary points. In the second approach, the backbone dihedrals ϕ and ψ were each constrained into 20 windows, leading to 400 windows for dipeptide fragments, and 20 windows were obtained by constraining the C–C–N–C and C–S–S–C dihedrals for the ACE-NME and DMDS fragments, respectively. The QM optimizations were performed with

constraints to obtain the 400 and 20 optimal structures at each window for dipeptide fragments and connecting fragments, respectively. This approach provides conformers that span the whole conformation space. From the QM optimization for the initial structures in the two approaches, the normal mode coordinates and the corresponding force constants can be obtained, and the NMS was performed according to ANI-1 work.[84] The displacement R_i for normal mode coordinate i was calculated as

$$R_i = \pm\sqrt{\frac{3c_i N_a k_b T}{K_i}}, \tag{7}$$

where c_i is the uniformly distributed random number for normal mode coordinate i with the constraint that the sum of all c_i is between 0 and 1, N_a is the number of atoms in the fragment, k_b is Boltzmann's constant, T is the temperature, and K_i is the force constant of normal mode coordinate i. The sign is determined based on a Bernoulli distribution with a probability equal to 0.5. The new structures were obtained by applying the displacements to the optimized structures, and the maximum perturbation distance allowed along each normal mode is 0.35 Å. During the generation of non-equilibrium structures with NMS, the structures that are similar to all previously accepted structures were not accepted if the RMSD $< 0.0075\ N$ Å,[135] where N is the number of heavy atoms used to calculate the RMSD. Second, adaptive samplings based on the disagreement of the NN ensemble were performed to sample more comprehensive data points. For each fragment, two NNs were trained and used to predict energies in MD simulations. The prediction discrepancy of two NNs is calculated by

$$\sigma_E = \sqrt{(E_1 - \bar{E})^2 + (E_2 - \bar{E})^2}, \tag{8}$$

where E_1 and E_2 are the predicted energies from the first NN and the second NN, respectively, and $\bar{E}$ is the average predicted energy of two NNs. If σ_E is larger than a pre-defined tolerance, the structure will be considered a new data point and added to the dataset. Several cycles of adaptive sampling were performed to ensure the NN is not encountering many new configurations in MD simulations. For the MD simulations in adaptive sampling, a temperature of 500 K was used to ensure NN could be safely used in the MD simulations at 300 K.[101] The parameters of MD simulations are described as follows. The Verlet cutoff scheme[136] was used for neighbor searching. The cutoff for short range Coulomb interaction and van der Waals interaction was set at 1.2 nm, while the force switch was performed for van der Waals interaction with the switch distance starting at 1.0 nm. The smooth particle mesh Ewald (PME) method[137,138] was applied to calculate long-range electrostatic interaction with 0.16 nm Fourier grid spacing and cubic interpolation. Velocity rescaling[139] was used for temperature coupling, and the time constant for coupling was set at 0.1 ps. To control pressure at 1 bar, the Parrinello–Rahman barostat[140] was used with isotropic pressure coupling. The time constant for pressure coupling was set as 2 ps, and the compressibility was set as 4.5×10^{-5} bar^{-1}. The periodic boundary condition (PBC) was applied in all directions. The leap-frog integrator was used with the 0.5 fs integration step because of the high temperature, and no constraint was added to the bonds involving hydrogen. The simulation time was set at 15 ns, and the two new structures obtained from NN prediction uncertainty were set to be at least 5 ps away from each other to avoid high correlation.[141]

TABLE I. Number of data points, RMSEs of energies and forces for training set and testing set, and R^2 for testing set from CHARMM-NN (E: kcal/mol, F: kcal/mol/Å).

	Data	E train	F train	E test	F test	R^2
ACE-NME	2 499	0.44	1.71	0.49	1.84	0.99
DMDS	1 180	0.16	1.05	0.13	0.93	1.00
Gly	6 381	0.57	2.08	0.78	2.16	0.99
Ala	7 836	0.67	2.09	0.85	2.16	0.98
Val	12 620	0.86	2.29	1.09	2.37	0.98
Cys	10 629	1.04	2.53	1.29	2.64	0.98
Pro	5 829	0.74	2.18	1.02	2.29	0.97
Leu	12 493	0.84	2.16	1.12	2.27	0.97
Ile	13 705	0.84	2.20	1.06	2.28	0.98
Met	11 128	0.90	2.23	1.18	2.27	0.97
Trp	14 275	0.80	2.30	1.15	2.32	0.98
Phe	11 713	0.84	2.70	1.12	2.77	0.98
Hse	10 882	0.91	2.28	1.23	2.35	0.97
Hsd	10 428	0.83	2.25	1.13	2.28	0.98
Cys-SMe	11 510	0.96	2.30	1.30	2.47	0.97
Ser	9 167	0.81	2.29	1.00	2.35	0.98
Thr	14 717	0.89	2.47	1.14	2.64	0.97
Tyr	10 023	0.88	2.83	1.22	2.92	0.97
Asn	11 005	0.90	2.53	1.19	2.64	0.98
Gln	13 009	0.91	2.34	1.17	2.37	0.97
Lys	20 008	1.34	3.00	2.08	3.79	0.96
Arg	23 587	1.35	2.81	1.86	2.97	0.97
Hsp	15 491	0.73	2.16	1.07	2.20	0.99
Asp	11 278	1.04	2.92	1.80	3.35	0.98
Glu	12 387	1.07	2.91	1.68	3.50	0.98
AspH	11 926	0.89	2.42	1.13	2.47	0.98
GluH	12 381	0.90	2.35	1.15	2.34	0.96
Average		0.86	2.35	1.16	2.48	0.98

After new data were sampled from each cycle of NMS and adaptive sampling, the MM and QM single point energies and forces were calculated, and the NN models were trained on the updated dataset. The dataset was divided into a training set and a testing set, with around 90% data for training and 10% data for testing before adaptive samplings and after new data are sampled in adaptive samplings. For the atomic NN architecture, two hidden layers were used for each atom type in all fragments. The nodes of the hidden layers were determined differently for different sizes of fragments, which are 16,

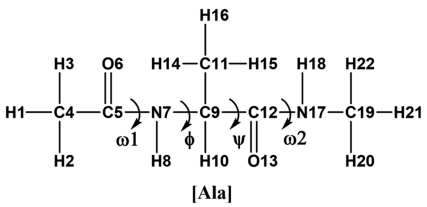

FIG. 3. Structure of Ala dipeptide fragment with labels on atoms.

TABLE II. Geometric data from C36m, CHARMM-NN, and B3LYP on three typical alanine dipeptides, $C7_{eq}$, $C7_{ax}$, and C5.

	$C7_{eq}$			$C7_{ax}$			C5		
	C36m	CHARMM-NN	B3LYP	C36m	CHARMM-NN	B3LYP	C36m	CHARMM-NN	B3LYP
					Dihedral angles (deg)				
ϕ	−77.1	−78.9	−82.3	78.0	75.9	73.4	−158.1	−159.5	−158.1
ψ	71.3	73.6	74.5	−56.8	−55.8	−56.1	161.5	161.7	161.8
$\omega 1$	178.2	179.4	−178.7	174.0	174.5	175.4	179.3	178.8	178.2
$\omega 2$	−179.7	−175.4	−174.7	−179.3	−177.8	−178.6	179.2	179.0	177.4
					Bonds (Å)				
C4–C5	1.481	1.517	1.515	1.480	1.519	1.516	1.482	1.519	1.518
C5–N7	1.341	1.361	1.360	1.343	1.360	1.358	1.337	1.360	1.361
C5–O6	1.224	1.236	1.236	1.225	1.237	1.237	1.223	1.231	1.232
N7–C9	1.449	1.467	1.467	1.457	1.473	1.475	1.441	1.450	1.449
C9–C11	1.543	1.525	1.523	1.547	1.537	1.535	1.545	1.542	1.540
C9–C12	1.530	1.551	1.548	1.524	1.543	1.544	1.516	1.531	1.533
C12–O13	1.229	1.231	1.231	1.228	1.232	1.233	1.230	1.231	1.232
C12–N17	1.346	1.356	1.356	1.346	1.352	1.351	1.348	1.356	1.355
N17–C19	1.444	1.456	1.454	1.444	1.452	1.453	1.445	1.456	1.457
N7–H8	0.992	1.012	1.011	0.993	1.010	1.010	0.996	1.014	1.015
N17–H18	1.002	1.018	1.017	1.004	1.017	1.019	0.994	1.011	1.010
					Angles (deg)				
C4–C5–N7	117.0	116.5	116.4	116.4	116.3	116.1	116.8	115.8	115.9
C5–N7–C9	123.6	123.2	123.1	126.1	125.9	126.8	122.6	122.0	122.2
N7–C9–C12	112.9	110.7	110.0	114.0	113.5	114.2	108.3	107.0	107.1
C9–C12–N17	116.9	114.4	114.0	117.5	116.2	116.6	117.5	115.4	115.7
C12–N17–C19	122.8	121.7	121.4	122.9	121.8	121.3	121.8	121.7	121.7

24, or 32 for heavy atoms and 8, 12, or 16 for hydrogen atoms. The loss function contains the energy and force contributions, defined as

$$L = \sum_i \left(|E_i^{\mathrm{ref}} - E_i^{\mathrm{NN}}|^2 + \frac{a}{3N_i} \sum_j |F_{ij}^{\mathrm{ref}} - F_{ij}^{\mathrm{NN}}|^2 \right) + b \sum_k \omega_k^2, \quad (9)$$

where a is the relative weight of force in training with respect to energy, N_i is the number of atoms in molecule i, j is the specific atom in molecule i, b is the hyperparameter for regularization term that is aimed to prevent overfitting, and ω_k is the NN parameters. In this work, the a was set to 1, and the hyperparameter b was searched with different values in training to obtain the best results. All NN training was performed in TensorFlow.[142] The input features were scaled based on the maximum values such that the range is [0, 1], and the reference energies were translated by subtracting the average values. The Adam algorithm[143] was used to optimize the NN models with the default parameters, for which the learning rate is 0.001 and the exponential decay rates for the first and second moment estimates are 0.9 and 0.999, respectively. The forces of NN were cal-

TABLE III. Potential energies (kcal/mol) with respect to $C7_{eq}$ from C36m, CHARMM-NN, and B3LYP, and structural reorganization energies (kJ/mol) with respect to B3LYP reference from C36m and CHARMM-NN for six alanine dipeptide conformers.

	Relative potential energies			Reorganization energies	
Conformers (ϕ, ψ)	C36m	CHARMM-NN	B3LYP	C36m	CHARMM-NN
$C7_{eq}$ (−82.3, 74.5)	0.00	0.00	0.00	16.99	1.71
$C7_{ax}$ (73.4, −56.1)	1.27	2.13	2.10	14.03	2.10
C5 (−158.1, 161.8)	−0.67	1.75	1.62	12.63	1.37
α_L (69.0, 22.3)	4.18	5.10	5.30	22.49	1.57
β_2 (−105.3, 9.8)	1.21	3.12	3.13	20.35	2.39
α' (−166.3, −43.1)	5.41	6.80	6.80	31.55	0.93

culated with automatic differentiation. The training was based on mini-batches, in which the batch size is a hyperparameter that could be 64, 256, or 1024. The maximum number of epochs is 50 000, and the early stopping was employed based on the testing error to alleviate overfitting.

All MD simulations to test CHARMM-NN on dipeptides, peptides, and proteins were performed in the modified GROMACS 2021.6,[144] in which the NN corrections were calculated with the TensorFlow C API, and all QM energies and forces were calculated by Gaussian 16.[145] The basic MD parameters, such as cutoff scheme, temperature coupling, and pressure coupling, are the same as they were in the adaptive samplings. In the simulations to validate CHARMM-NN, the bonds involving hydrogens were constrained by the LINCS algorithm[146] with a fourth order and one iteration. The integration time step was, therefore, set at 1 or 2 fs. All systems were first solvated by creating a cubic water box with the length of the longest distance in the protein plus 2 nm buffer and neutralizing it with enough sodium or chloride counter-ions. Next, the solvated systems were minimized with the steepest descent algorithm and equilibrium with a 500 ps NVT simulation using a v-rescale thermostat followed by a 1 ns NPT simulation using a Parrinello–Rahman barostat. Starting from the QM-optimized structures, the conformers of all dipeptides were minimized with the steepest descent algorithm for up to 50 000 steps for C36m and CHARMM-NN. The structural reorganization energies were calculated by subtracting the energy at the last step of minimization, which is the MM-optimized structure, from the energy at the first step of minimization, which is the QM-optimized structure. All production MD simulations were performed in solution with the TIP3P model, and a v-rescale thermostat and Parrinello–Rahman barostat were used. For the simulations of Gly_3, Ala_3, Val_3, and Ala_4, the total simulation time was set as 1 μs, in which the first 100 ns were discarded and the coordinates were saved every 10 ps. The J-coupling constants were calculated with Karplus equations, and the parameters were extracted following the previous work.[147] Slightly different results would be obtained if different sets of Karplus parameters are applied.

All data related to this work can be found at https://figshare.com/projects/CHARMM-NN/163756. The energy, force, and coordinates of the sampled structures, the CHARMM-NN force field parameters, the codes to run MD simulations with CHARMM-NN, training data and scripts, MD simulation files, and analysis scripts are provided.

III. RESULTS AND DISCUSSION

The CHARMM-NN training and testing accuracy are measured with the root mean squared error (RMSE), defined as

$$E^{\mathrm{RMSE}} = \sqrt{\frac{1}{N}\sum_i \left(E_i^{\mathrm{ref}} - E_i^{\mathrm{NN}}\right)^2}, \tag{10}$$

and

$$F^{\mathrm{RMSE}} = \sqrt{\frac{1}{N}\sum \frac{1}{3N_i}\left|\mathbf{F}_i^{\mathrm{ref}} - \mathbf{F}_i^{\mathrm{NN}}\right|^2}, \tag{11}$$

where N is the number of samples and R^2 is defined as

$$R^2 = 1 - \frac{\sum_i \left(E_i^{\mathrm{ref}} - E_i^{\mathrm{NN}}\right)^2}{\sum_i \left(E_i^{\mathrm{ref}} - \bar{E}_i^{\mathrm{ref}}\right)^2}, \tag{12}$$

where $\bar{E}_i^{\mathrm{ref}}$ is the average of the reference energies for sample i. The number of data points for each fragment and the training quality of CHARMM-NN are shown in Table I. For the smallest fragment DMDS, the RMSEs of energies and forces for training and testing sets are below 0.2 kcal/mol and around 1.0 kcal/mol/Å, respectively, and the R^2 for the testing set is 1, which shows the best training accuracy across all fragments. Another connecting fragment, ACE-NME, also has relatively low errors in energies and forces, which are below 0.5 kcal/mol and 2.0 kcal/mol/Å, respectively. For dipeptide fragments, the errors in energies and forces are typically larger. The errors are lower for neutral dipeptides compared to charged dipeptides. For the simplest dipeptides Gly and Ala, the RMSEs of energies and forces for the testing set are below 1.0 kcal/mol and equal to 2.16 kcal/mol/Å, respectively, which are smaller than other neutral dipeptides. The RMSEs of energies for other neutral dipeptides are between 1.0 and 1.3 kcal/mol, and the RMSEs of forces are in the range of [2.2, 3.0] kcal/mol/Å. For the negatively charged dipeptides Asp and Glu, the errors are slightly higher, with energy errors of around 1.8 kcal/mol and force errors of around 3.5 kcal/mol/Å. The positively charged dipeptides Lys and Arg have similar behaviors, and the Lys dipeptide has the largest error among all fragments. For Lys dipeptide, even though the RMSEs of energies and forces for the testing set are 2.08 kcal/mol and 3.79 kcal/mol/Å, respectively, the R^2 is equal to 0.96, which still demonstrates the quality of the fitting. The average RMSEs of energies and forces are 0.86 kcal/mol and 2.35 kcal/mol/Å for the training set, respectively, and 1.16 kcal/mol and 2.48 kcal/mol/Å for the testing set, respectively, with the R^2 equal to 0.98. Overall, the training of CHARMM-NN is accurate for all fragments.

TABLE IV. Number of conformers at energy minima and average structural reorganization energies (kJ/mol) for 14 dipeptide fragments with respect to MP2/cc-PVTZ reference structures from the YMPJ conformer database.

	No. minima	C36m	CHARMM-NN
Ala	10	32.72	7.07
Asn	12	36.76	7.40
Cys	23	36.83	7.36
Gln	20	41.55	10.25
Gly	8	24.33	6.87
Ile	24	39.63	7.70
Leu	26	37.83	10.02
Met	56	41.37	12.07
Phe	26	39.34	10.15
Pro	5	33.10	7.07
Ser	23	35.94	7.10
Thr	17	42.96	8.13
Tyr	16	42.04	6.72
Val	14	39.88	7.38
Average		38.82	9.01

Next, we examined the detailed performance of CHARMM-NN on dipeptides. One typical system is alanine dipeptide, which has many conformations and has received systematic studies.[148] The graphical structure of Ala dipeptide is shown in Fig. 3. We first compare the geometric information optimized from C36m, CHARMM-NN, and B3LYP. As displayed in Table II, the CHARMM-NN outperforms C36m on all three Ala conformers, $C7_{eq}$, $C7_{ax}$, and C5. Most angles and dihedral angles from CHARMM-NN are closer to the B3LYP geometries compared to C36m, and almost all bonds optimized from CHARMM-NN are better than C36m. The large corrections occur on C4–C5 and C5–N7 bonds, for which the bond lengths from C36m are generally much smaller than B3LYP structures, and on the C9–C11 bond, which typically has a large bond length from C36m.

Besides the geometric evaluation, we also compared the relative potential energies and reorganization energies between the C36m and CHARMM-NN force fields. As shown in Table III, six Ala conformers, $C7_{eq}$, $C7_{ax}$, C5, α_L, β_2, and α', were calculated because the optimization with B3LYP can be obtained for these conformers so that direct comparison can be executed. The optimized backbone dihedrals from QM calculations are shown in Table III for all conformers. According to B3LYP calculations, the $C7_{eq}$ conformer has the lowest energy, which is selected as the baseline for comparison, and the relative energies are listed in Table III, in which the C5 and α' conformers have the lowest and highest relative energies, respectively. For the conformers minimized with C36m, the errors of relative energies are between 0.8 and 1.4 kcal/mol for $C7_{ax}$, α_L, and α' conformers compared to B3LYP. For C5 and β_2 conformers, the errors are up to 2.29 and 1.92 kcal/mol, respectively, and the C36m optimized C5 tends to be slightly more stable than $C7_{eq}$. Using CHARMM-NN, the relative energies are very consistent with B3LYP results for all conformers, and the largest error is only 0.2 kcal/mol. The accurate relative energies from CHARMM-NN demonstrate that the local minima on the PES of Ala dipeptide are close to QM PES, which can be very important for the simulation samplings being similar to QM results. In addition to the relative potential energies, we also compared the reorganization energies between C36m and CHARMM-NN for these six conformers of Ala dipeptide. In Table III, we can observe that the reorganization energies with respect to B3LYP range from 12.63 to 31.55 kJ/mol for C36m but only between 0.93 and 2.39 kJ/mol for CHARMM-NN. The minimized structures from CHARMM-NN are much closer to B3LYP optimal structures than C36m.

Besides the few tests of reorganization energies for Ala dipeptides with respect to B3LYP, which is the reference QM method that we train the CHARMM-NN on, more thorough validations of the reorganization energies were performed on the YMPJ conformer database,[149,150] which includes the conformers of dipeptides optimized with MP2/cc-PVTZ. As shown in Table IV, the 14 neutral dipeptides have different numbers of conformers at energy minima. For C36m, the reorganization energies for almost all dipeptides are larger than 30 kJ/mol, except for the Gly dipeptide (24.33 kJ/mol), which is the smallest dipeptide. The reorganization energies can be higher than 40 kJ/mol for Gln, Met, Thr, and Tyr dipeptides. For CHARMM-NN, the reorganization energies for most dipeptides are below 10 kJ/mol, except for Gln, Leu, Met, and Phe dipeptides. The Met dipeptide, which includes the most conformers, has the largest reorganization energy (12.07 kJ/mol), but it is still much better than the C36m results. The average reorganization energy for C36m is 38.82 kJ/mol, which is close to the work from König and Riniker.[63] Even though the CHARMM-NN is trained with B3LYP, it outperforms C36m on the dataset of MP2/cc-PVTZ as well, and the average

TABLE V. *J*-coupling constants (Hz) for Gly_3, Ala_3, Val_3, and Ala_4 (\: No experimental values).

			$^3J(H_N,H_\alpha)$	$^3J(H_N,C')$	$^3J(H_\alpha,C')$	$^3J(C',C')$	$^3J(H_N,C_\beta)$	$^1J(N,C_\alpha)$	$^2J(N,C_\alpha)$	$^3J(H_N,C_\alpha)$
Gly_3	Residue 2	CHARMM-NN	6.04	0.48	4.28	0.76	2.26	11.02	7.79	0.61
		Exp.	5.89	1.10	4.01	0.26	\	12.17	10.45	0.78
	Residue 3	CHARMM-NN	6.04	0.48	4.25	0.77	2.28	11.18	7.67	0.59
		Exp.	5.87	0.99	3.90	\	\	12.77	9.05	0.61
Ala_3	Residue 2	CHARMM-NN	6.69	1.63	2.66	0.91	1.56	10.44	7.14	0.48
		Exp.	5.68	1.13	1.84	0.25	2.39	11.34	9.14	0.70
	Residue 3	CHARMM-NN	6.88	1.66	2.60	0.96	1.53	10.48	7.05	0.49
		Exp.	6.52	1.29	2.14	\	2.02	11.47	8.45	0.65
Val_3	Residue 2	CHARMM-NN	8.34	3.04	2.45	1.86	0.56	9.62	6.88	0.58
		Exp.	7.94	0.58	2.42	0.34	1.38	10.80	8.35	0.77
	Residue 3	CHARMM-NN	3.59	3.39	3.55	1.01	0.54	11.06	7.24	0.16
		Exp.	7.91	1.01	2.45	\	1.40	11.02	7.80	0.75
Ala_4	Residue 2	CHARMM-NN	6.45	1.57	2.73	0.87	1.63	10.47	7.11	0.46
		Exp.	5.62	1.15	1.87	0.19	2.36	11.39	9.17	0.68
	Residue 3	CHARMM-NN	6.57	1.53	2.69	0.86	1.64	10.40	7.14	0.48
		Exp.	5.89	1.11	1.95	\	2.24	11.33	8.56	0.60
	Residue 4	CHARMM-NN	6.90	1.62	2.72	0.95	1.52	10.44	6.97	0.48
		Exp.	6.56	1.26	2.24	\	1.99	11.53	8.37	0.60

reorganization energy is only 9.01 kJ/mol. The low reorganization energy with respect to QM is critical to obtaining accurate thermodynamic properties like free energy that are close to QM since the expected variance is exponentially related to the reorganization energies for the free-energy estimation.[63]

To validate the performance of CHARMM-NN in MD simulations, the J-coupling constants were calculated from the MD simulations for Gly_3, Ala_3, Val_3, and Ala_4, and the results are shown in Table V. For Gly_3, the J-coupling constants are close to the experimental values for residues 2 and 3. The largest error is on ${}^2J(N, C_\alpha)$, which has a deviation of 2.66 and 1.38 Hz from experiments for residue 2 and residue 3, respectively. However, the experimental values of 10.45 and 9.05 Hz can never be obtained because the maximum value is less than 9 Hz with the parameters for the Karplus equation. For Ala_3, the deviations for ${}^3J(H_N, H_\alpha)$, ${}^3J(H_\alpha, C')$, and ${}^3J(H_N, C_\beta)$ are around 0.82 to 1.01 Hz for residue 2, but they are much better for residue 3. Other J-coupling constants have similar errors for residue 2 and residue 3, and the largest deviation is 2.00 Hz on ${}^2J(N, C_\alpha)$ for residue 2. For Val_3, the ${}^3J(H_N, H_\alpha)$ is pretty good for residue 2 with a difference of 0.40 Hz, but it is terrible for residue 3, which has an error of 4.32 Hz. In contrast, the ${}^1J(N, C_\alpha)$ and ${}^2J(N, C_\alpha)$ have a better agreement with experiments on residue 3 compared to residue 2. For Ala_4, the results are similar to Ala_3 in that the ${}^3J(H_N, H_\alpha)$, ${}^3J(H_\alpha, C')$, and ${}^3J(H_N, C_\beta)$ have less errors from residue 2 to residue 4. Overall, Gly_3 has good results for all J-coupling constants except ${}^2J(N, C_\alpha)$. The Ala_3 and Ala_4 have similar results, suggesting that the simulations on systems with different sizes provide stable results for CHARMM-NN. The results of Val_3 are good for some J-coupling constants, but they can also have a large error on others. In addition, MM force fields still have better agreement with experimental J-coupling values than the current CHARMM-NN, mainly because it only has corrections for intramolecular interactions of dipeptides, while the protein–protein and protein–solvent non-covalent interactions have not been addressed yet.

We also ran several simulations on several folded proteins with CHARMM-NN. The folded proteins we tested are ubiquitin (PDB: 1UBQ[151]), crambin (PDB: 1EJG[152]), GB3 (PDB: 1P7E[153]), and lysozyme (PDB: 135L[154]). As displayed in Fig. 4, the potential of the mean force plots for the folded proteins is similar to the normal Ramachandran plots. The dominating regions include α region ($-160° < \phi < -20°$ and $-120° < \psi < 50°$) in which the right-handed

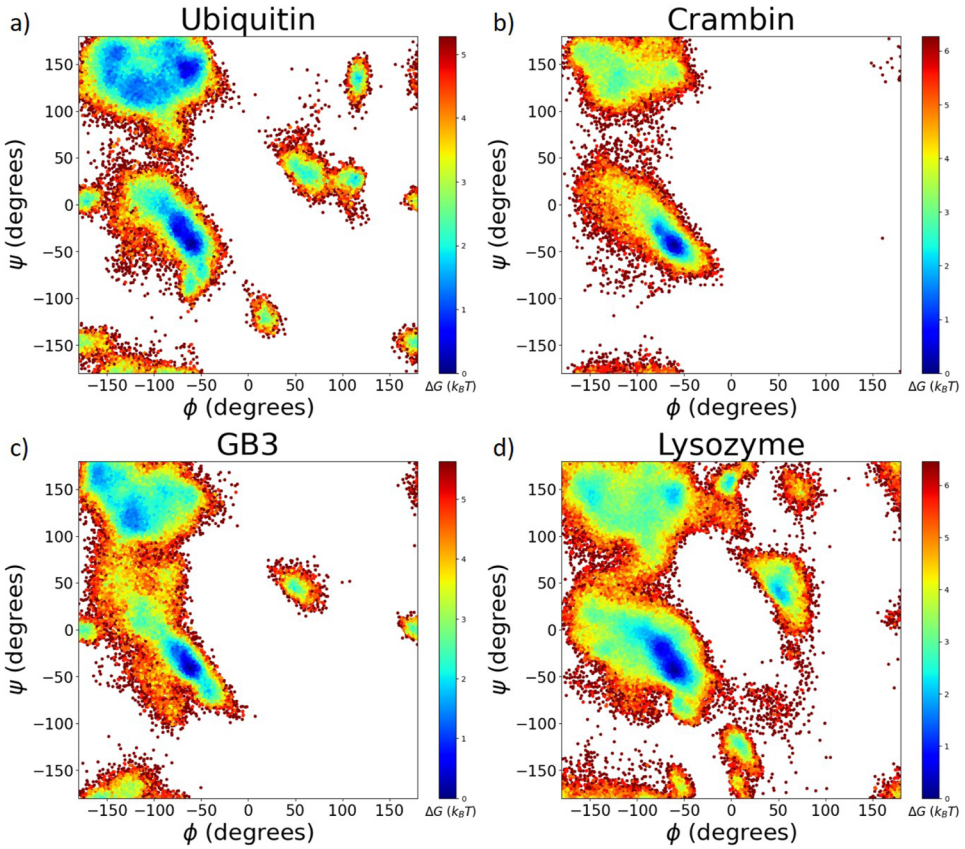

FIG. 4. Ramachandran plots for four folded proteins from MD simulations using CHARMM-NN force fields: (a) ubiquitin: some conformations with $100° < \phi < 120°$ are sampled; (b) crambin: all conformations are within α and β regions; (c) GB3: all conformations are within α and β regions; (d) lysozyme: some conformations with $0° < \phi < 30°$ are sampled.

α helix α_R ($-100^\circ < \phi < -30^\circ$ and $-67^\circ < \psi < -7^\circ$) are sampled most, β region ($-180^\circ < \phi < -90^\circ$ and $50^\circ < \psi < 180^\circ$ plus $-180^\circ < \phi < -90^\circ$ and $-180^\circ < \psi < -120^\circ$ plus $160^\circ < \phi < 180^\circ$ and $110^\circ < \psi < 180^\circ$), ppII region ($-90^\circ < \phi < -20^\circ$ and $50^\circ < \psi < 180^\circ$ plus $-90^\circ < \phi < -20^\circ$ and $-180^\circ < \psi < -120^\circ$), and α_L region ($30^\circ < \phi < 100^\circ$ and $7^\circ < \psi < 67^\circ$).[18] The crambin and GB3 have almost all their points located in these regions, showing the reliability of CHARMM-NN in sampling conformational space. However, results for ubiquitin and lysozyme show several unfavored regions, such as the regions with $100^\circ < \phi < 120^\circ$ for ubiquitin and the regions with $0^\circ < \phi < 30^\circ$ for lysozyme. The non-ideal results from the Ramachandran plots of some folded proteins and the J-coupling constants of some peptides indicate that the CHARMM-NN is still not ready to be used for general proteins. The difficulty of CHARMM-NN is because of the following reasons. First, it is only trained with all data in the vacuum state, and no background charges are included in QM calculations and NN inputs; therefore, the method can only approach the limit of accuracy at the mechanical embedding QM/MM level without considering the MM charge contribution to the QM region. We do include solvent in all MD simulations of peptides and proteins, while the protein–solvent interactions are based on the MM force fields and not improved with machine learning models. Protein–water interactions are crucial to MD simulations of proteins. However, the protein–water interactions cannot be easily included in the training of NN because the solvent environment of small fragments is largely different from the solvent environment of proteins, and there is yet no general sampling method that can be applied to sample enough small fragments surrounding with solvent to resemble the actual solvent environment in all proteins. Second, the non-bonded interactions between fragments are still calculated at the MM level, which may cause the CHARMM-NN corrections to be imbalanced on proteins.

IV. CONCLUSIONS

In summary, we constructed machine learned force fields based on C36m force fields, named CHARMM-NN, by using the rSMF method to generate the elementary fragments that can form any kind of protein. The dataset was generated and enlarged by combining NMS and adaptive sampling based on prediction uncertainty. The CHARMM-NN force fields use the atom-type based NN to calculate energies and forces, and the input features can be obtained or simply derived from the traditional MM variables. The high compatibility between CHARMM-NN and MM force fields enables the convenient implementation of CHARMM-NN in all MD programs without the need to modify core codes. The training error for CHARMM-NN is low for all 27 fragments, and the validations on dipeptides demonstrate that CHARMM-NN can result in good geometric data similar to QM calculations and much lower reorganization energies than traditional MM force fields. For the MD simulations with CHARMM-NN on several peptides and proteins, some results are acceptable, but some results are not sufficiently accurate because the current CHARMM-NN can only achieve corrections at the mechanical embedding QM/MM level, suggesting that the CHARMM-NN needs to be improved further. Future directions include considering the comprehensive solvent effect in data sampling and constructing machine learned force fields to describe non-bonded interactions between fragments beyond the mechanical embedding level.

ACKNOWLEDGMENTS

This work has been supported by the National Institutes of Health (Grant No. R01-GM061870).

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Pan Zhang: Data curation (lead); Formal analysis (lead); Investigation (lead); Software (lead); Validation (lead); Visualization (lead); Writing – original draft (lead). **Weitao Yang**: Conceptualization (lead); Funding acquisition (lead); Methodology (lead); Project administration (lead); Resources (lead); Supervision (lead); Writing – review & editing (lead).

DATA AVAILABILITY

The data that support the findings of this study are available within the article.

REFERENCES

[1] M. Karplus and J. A. McCammon, "Molecular dynamics simulations of biomolecules," Nat. Struct. Biol. **9**, 646–652 (2002).
[2] S. A. Adcock and J. A. McCammon, "Molecular dynamics: Survey of methods for simulating the activity of proteins," Chem. Rev. **106**, 1589–1615 (2006).
[3] D. M. Zuckerman, "Equilibrium sampling in biomolecular simulations," Annu. Rev. Biophys. **40**, 41–62 (2011).
[4] W. F. van Gunsteren, X. Daura, N. Hansen, A. E. Mark, C. Oostenbrink, S. Riniker, and L. J. Smith, "Validation of molecular simulation: An overview of issues," Angew. Chem., Int. Ed. **57**, 884–902 (2018).
[5] S. A. Hollingsworth and R. O. Dror, "Molecular dynamics simulation for all," Neuron **99**, 1129–1143 (2018).
[6] A. D. Mackerell, "Empirical force fields for biological macromolecules: Overview and issues," J. Comput. Chem. **25**, 1584–1604 (2004).
[7] W. L. Jorgensen and J. Tirado-Rives, "Potential energy functions for atomic-level simulations of water and organic and biomolecular systems," Proc. Natl. Acad. Sci. U. S. A. **102**, 6665–6670 (2005).
[8] S. Riniker, "Fixed-charge atomistic force fields for molecular dynamics simulations in the condensed phase: An overview," J. Chem. Inf. Model. **58**, 565–578 (2018).
[9] X. Zhu, P. E. M. Lopes, and A. D. MacKerell, "Recent developments and applications of the CHARMM force fields," Wiley Interdiscip. Rev.: Comput. Mol. Sci. **2**, 167–185 (2012).
[10] T. Fröhlking, M. Bernetti, N. Calonaci, and G. Bussi, "Toward empirical force fields that match experimental observables," J. Chem. Phys. **152**, 230902 (2020).
[11] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, "A second generation force field for the simulation of proteins, nucleic acids, and organic molecules," J. Am. Chem. Soc. **117**, 5179–5197 (1995).
[12] J. Wang, P. Cieplak, and P. A. Kollman, "How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules?," J. Comput. Chem. **21**, 1049–1074 (2000).

[13]J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling, "ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB," J. Chem. Theory Comput. **11**, 3696–3713 (2015).
[14]C. Tian, K. Kasavajhala, K. A. A. Belfon, L. Raguette, H. Huang, A. N. Migues, J. Bickel, Y. Wang, J. Pincay, and Q. Wu, "ff19SB: Amino-acid-specific protein backbone parameters trained against quantum mechanics energy surfaces in solution," J. Chem. Theory Comput. **16**, 528–552 (2019).
[15]A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, and M. Karplus, "All-atom empirical potential for molecular modeling and dynamics studies of proteins," J. Phys. Chem. B **102**, 3586–3616 (1998).
[16]A. D. Mackerell, M. Feig, and C. L. Brooks, "Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations," J. Comput. Chem. **25**, 1400–1415 (2004).
[17]R. B. Best, X. Zhu, J. Shim, P. E. M. Lopes, J. Mittal, M. Feig, and A. D. MacKerell, "Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain χ_1 and χ_2 dihedral angles," J. Chem. Theory Comput. **8**, 3257–3273 (2012).
[18]J. Huang, S. Rauscher, G. Nawrocki, T. Ran, M. Feig, B. L. de Groot, H. Grubmüller, and A. D. MacKerell, "CHARMM36m: An improved force field for folded and intrinsically disordered proteins," Nat. Methods **14**, 71–73 (2017).
[19]L. D. Schuler, X. Daura, and W. F. Van Gunsteren, "An improved GROMOS96 force field for aliphatic hydrocarbons in the condensed phase," J. Comput. Chem. **22**, 1205–1218 (2001).
[20]C. Oostenbrink, A. Villa, A. E. Mark, and W. F. Van Gunsteren, "A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6," J. Comput. Chem. **25**, 1656–1676 (2004).
[21]T. A. Soares, P. H. Hünenberger, M. A. Kastenholz, V. Kräutler, T. Lenz, R. D. Lins, C. Oostenbrink, and W. F. van Gunsteren, "An improved nucleic acid parameter set for the GROMOS force field," J. Comput. Chem. **26**, 725–737 (2005).
[22]N. Schmid, A. P. Eichenberger, A. Choutko, S. Riniker, M. Winger, A. E. Mark, and W. F. van Gunsteren, "Definition and testing of the GROMOS force-field versions 54A7 and 54B7," Eur. Biophys. J. **40**, 843–856 (2011).
[23]W. L. Jorgensen and J. Tirado-Rives, "The OPLS force field for proteins. Energy minimizations for crystals of cyclic peptides and crambin," J. Am. Chem. Soc. **110**, 1657–1666 (1988).
[24]W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives, "Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids," J. Am. Chem. Soc. **118**, 11225–11236 (1996).
[25]G. A. Kaminski, R. A. Friesner, J. Tirado-Rives, and W. L. Jorgensen, "Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides," J. Phys. Chem. B **105**, 6474–6487 (2001).
[26]M. J. Robertson, J. Tirado-Rives, and W. L. Jorgensen, "Improved peptide and protein torsional energetics with the OPLS-AA force field," J. Chem. Theory Comput. **11**, 3499–3509 (2015).
[27]C. M. Baker, "Polarizable force fields for molecular dynamics simulations of biomolecules," Wiley Interdiscip. Rev.: Comput. Mol. Sci. **5**, 241–254 (2015).
[28]S. Patel and C. L. Brooks III, "Fluctuating charge force fields: Recent developments and applications from small molecules to macromolecular biological systems," Mol. Simul. **32**, 231–249 (2006).
[29]J. A. Lemkul, J. Huang, B. Roux, and A. D. MacKerell, "An empirical polarizable force field based on the classical Drude oscillator model: Development history and recent applications," Chem. Rev. **116**, 4983–5013 (2016).
[30]J. Wang, P. Cieplak, J. Li, J. Wang, Q. Cai, M. Hsieh, H. Lei, R. Luo, and Y. Duan, "Development of polarizable models for molecular mechanical calculations II: Induced dipole models significantly improve accuracy of intermolecular interaction energies," J. Phys. Chem. B **115**, 3100–3111 (2011).
[31]Y. Shi, Z. Xia, J. Zhang, R. Best, C. Wu, J. W. Ponder, and P. Ren, "Polarizable atomic multipole-based amoeba force field for proteins," J. Chem. Theory Comput. **9**, 4046–4063 (2013).
[32]K. Lindorff-Larsen, P. Maragakis, S. Piana, M. P. Eastwood, R. O. Dror, and D. E. Shaw, "Systematic validation of protein force fields against experimental data," PLoS One **7**, e32131 (2012).
[33]P. S. Nerenberg and T. Head-Gordon, "New developments in force fields for biomolecular simulations," Curr. Opin. Struct. Biol. **49**, 129–138 (2018).
[34]S. Piana, K. Lindorff-Larsen, and D. E. Shaw, "How robust are protein folding simulations with respect to force field parameterization?," Biophys. J. **100**, L47–L49 (2011).
[35]K. A. McKiernan, B. E. Husic, and V. S. Pande, "Modeling the mechanism of CLN025 beta-hairpin formation," J. Chem. Phys. **147**, 104107 (2017).
[36]P. S. Nerenberg, B. Jo, C. So, A. Tripathy, and T. Head-Gordon, "Optimizing solute-water van der Waals interactions to reproduce solvation free energies," J. Phys. Chem. B **116**, 4524–4534 (2012).
[37]R. B. Best, W. Zheng, and J. Mittal, "Balanced protein–water interactions improve properties of disordered proteins and non-specific protein association," J. Chem. Theory Comput. **10**, 5113–5124 (2014).
[38]C. J. Oldfield and A. K. Dunker, "Intrinsically disordered proteins and intrinsically disordered protein regions," Annu. Rev. Biochem. **83**, 553–584 (2014).
[39]V. N. Uversky, "Introduction to intrinsically disordered proteins (IDPs)," Chem. Rev. **114**, 6557–6560 (2014).
[40]S.-H. Chong, P. Chatterjee, and S. Ham, "Computer simulations of intrinsically disordered proteins," Annu. Rev. Phys. Chem. **68**, 117–134 (2017).
[41]S. Bhattacharya and X. Lin, "Recent advances in computational protocols addressing intrinsically disordered proteins," Biomolecules **9**, 146 (2019).
[42]W. Wang, "Recent advances in atomic molecular dynamics simulation of intrinsically disordered proteins," Phys. Chem. Chem. Phys. **23**, 777–784 (2021).
[43]R. B. Best and J. Mittal, "Protein simulations with an optimized water model: Cooperative helix formation and temperature-induced unfolded state collapse," J. Phys. Chem. B **114**, 14916–14923 (2010).
[44]R. B. Best and J. Mittal, "Free-energy landscape of the GB1 hairpin in all-atom explicit solvent simulations with different force fields: Similarities and differences," Proteins: Struct., Funct., Bioinf. **79**, 1318–1328 (2011).
[45]J. Henriques, C. Cragnell, and M. Skepö, "Molecular dynamics simulations of intrinsically disordered proteins: Force field evaluation and comparison with experiment," J. Chem. Theory Comput. **11**, 3420–3431 (2015).
[46]Z. A. Levine and J.-E. Shea, "Simulations of disordered proteins and systems with conformational heterogeneity," Curr. Opin. Struct. Biol. **43**, 95–103 (2017).
[47]J. Huang and A. D. MacKerell, "Force field development and simulations of intrinsically disordered proteins," Curr. Opin. Struct. Biol. **48**, 40–48 (2018).
[48]J. Mu, H. Liu, J. Zhang, R. Luo, and H.-F. Chen, "Recent force field strategies for intrinsically disordered proteins," J. Chem. Inf. Model. **61**, 1037–1047 (2021).
[49]D. Song, W. Wang, W. Ye, D. Ji, R. Luo, and H.-F. Chen, "ff14IDPs force field improving the conformation sampling of intrinsically disordered proteins," Chem. Biol. Drug Des. **89**, 5–15 (2017).
[50]H. Liu, D. Song, H. Lu, R. Luo, and H.-F. Chen, "Intrinsically disordered protein-specific force field CHARMM36IDPSFF," Chem. Biol. Drug Des. **92**, 1722–1735 (2018).
[51]H. Liu, D. Song, Y. Zhang, S. Yang, R. Luo, and H.-F. Chen, "Extensive tests and evaluation of the CHARMM36IDPSFF force field for intrinsically disordered proteins and folded proteins," Phys. Chem. Chem. Phys. **21**, 21918–21931 (2019).
[52]S. Yang, H. Liu, Y. Zhang, H. Lu, and H. Chen, "Residue-specific force field improving the sample of intrinsically disordered proteins and folded proteins," J. Chem. Inf. Model. **59**, 4793–4805 (2019).
[53]W. Kang, F. Jiang, and Y.-D. Wu, "Universal implementation of a residue-specific force field based on CMAP potentials and free energy decomposition," J. Chem. Theory Comput. **14**, 4474–4486 (2018).
[54]D. Song, H. Liu, R. Luo, and H.-F. Chen, "Environment-specific force field for intrinsically disordered and ordered proteins," J. Chem. Inf. Model. **60**, 2257–2267 (2020).

[55]P. S. Nerenberg and T. Head-Gordon, "Optimizing protein–solvent force fields to reproduce intrinsic conformational preferences of model peptides," J. Chem. Theory Comput. **7**, 1220–1230 (2011).
[56]D. Mercadante, S. Milles, G. Fuertes, D. I. Svergun, E. A. Lemke, and F. Gräter, "Kirkwood–Buff approach rescues overcollapse of a disordered protein in canonical protein force fields," J. Phys. Chem. B **119**, 7975–7984 (2015).
[57]E. A. Ploetz, S. Karunaweera, N. Bentenitis, F. Chen, S. Dai, M. B. Gee, Y. Jiao, M. Kang, N. L. Kariyawasam, N. Naleem, S. Weerasinghe, and P. E. Smith, "Kirkwood–Buff-derived force field for peptides and proteins: Philosophy and development of KBFF20," J. Chem. Theory Comput. **17**, 2964–2990 (2021).
[58]E. A. Ploetz, S. Karunaweera, and P. E. Smith, "Kirkwood–Buff-derived force field for peptides and proteins: Applications of KBFF20," J. Chem. Theory Comput. **17**, 2991–3009 (2021).
[59]S. Kmiecik, D. Gront, M. Kolinski, L. Wieteska, A. E. Dawid, and A. Kolinski, "Coarse-grained protein models and their applications," Chem. Rev. **116**, 7898–7936 (2016).
[60]A. P. Latham and B. Zhang, "Unifying coarse-grained force fields for folded and disordered proteins," Curr. Opin. Struct. Biol. **72**, 63–70 (2022).
[61]H. Wu, P. G. Wolynes, and G. A. Papoian, "AWSEM-IDP: A coarse-grained force field for intrinsically disordered proteins," J. Phys. Chem. B **122**, 11115–11125 (2018).
[62]A. P. Latham and B. Zhang, "Maximum entropy optimized force field for intrinsically disordered proteins," J. Chem. Theory Comput. **16**, 773–781 (2020).
[63]G. König and S. Riniker, "On the faithfulness of molecular mechanics representations of proteins towards quantum-mechanical energy surfaces," Interface Focus **10**, 20190121 (2020).
[64]V. Botu, R. Batra, J. Chapman, and R. Ramprasad, "Machine learning force fields: Construction, validation, and outlook," J. Phys. Chem. C **121**, 511–522 (2017).
[65]T. D. Huan, R. Batra, J. Chapman, S. Krishnan, L. Chen, and R. Ramprasad, "A universal strategy for the creation of machine learning-based atomistic force fields," npj Comput. Mater. **3**, 37 (2017).
[66]T. Mueller, A. Hernandez, and C. Wang, "Machine learning for interatomic potential models," J. Chem. Phys. **152**, 050902 (2020).
[67]S. Manzhos and T. Carrington, "Neural network potential energy surfaces for small molecules and reactions," Chem. Rev. **121**, 10187–10217 (2021).
[68]I. Poltavsky and A. Tkatchenko, "Machine learning force fields: Recent advances and remaining challenges," J. Phys. Chem. Lett. **12**, 6551–6564 (2021).
[69]Y. Li, H. Li, F. C. Pickard, B. Narayanan, F. G. Sen, M. K. Y. Chan, S. K. R. S. Sankaranarayanan, B. R. Brooks, and B. Roux, "Machine learning force field parameters from ab initio data," J. Chem. Theory Comput. **13**, 4492–4503 (2017).
[70]E. Heid, M. Fleck, P. Chatterjee, C. Schröder, and A. D. MacKerell, "Toward prediction of electrostatic parameters for force fields that explicitly treat electronic polarization," J. Chem. Theory Comput. **15**, 2460–2469 (2019).
[71]O. Demerdash, U. R. Shrestha, L. Petridis, J. C. Smith, J. C. Mitchell, and A. Ramanathan, "Using small-angle scattering data and parametric machine learning to optimize force field parameters for intrinsically disordered proteins," Front. Mol. Biosci. **6**, 64 (2019).
[72]R. Galvelis, S. Doerr, J. M. Damas, M. J. Harvey, and G. De Fabritiis, "A scalable molecular force field parameterization method based on density functional theory and quantum-level machine learning," J. Chem. Inf. Model. **59**, 3485–3493 (2019).
[73]V. L. Deringer, A. P. Bartók, N. Bernstein, D. M. Wilkins, M. Ceriotti, and G. Csányi, "Gaussian process regression for materials and molecules," Chem. Rev. **121**, 10073–10141 (2021).
[74]A. P. Bartók, R. Kondor, and G. Csányi, "On representing chemical environments," Phys. Rev. B **87**, 184115 (2013).
[75]S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, and K.-R. Müller, "Machine learning of accurate energy-conserving molecular force fields," Sci. Adv. **3**, e1603015 (2017).
[76]S. Chmiela, H. E. Sauceda, K.-R. Müller, and A. Tkatchenko, "Towards exact molecular dynamics simulations with machine-learned force fields," Nat. Commun. **9**, 3887 (2018).
[77]H. E. Sauceda, S. Chmiela, I. Poltavsky, K.-R. Müller, and A. Tkatchenko, "Molecular force fields with gradient-domain machine learning: Construction and application to dynamics of small molecules with coupled cluster forces," J. Chem. Phys. **150**, 114102 (2019).
[78]H. E. Sauceda, M. Gastegger, S. Chmiela, K.-R. Müller, and A. Tkatchenko, "Molecular force fields with gradient-domain machine learning (GDML): Comparison and synergies with classical force fields," J. Chem. Phys. **153**, 124109 (2020).
[79]J. Behler and M. Parrinello, "Generalized neural-network representation of high-dimensional potential-energy surfaces," Phys. Rev. Lett. **98**, 146401 (2007).
[80]J. Behler, "Atom-centered symmetry functions for constructing high-dimensional neural network potentials," J. Chem. Phys. **134**, 074106 (2011).
[81]S. A. Ghasemi, A. Hofstetter, S. Saha, and S. Goedecker, "Interatomic potentials for ionic systems with density functional accuracy based on charge densities obtained by a neural network," Phys. Rev. B **92**, 045131 (2015).
[82]T. W. Ko, J. A. Finkler, S. Goedecker, and J. Behler, "General-purpose machine learning potentials capturing nonlocal charge transfer," Acc. Chem. Res. **54**, 808–817 (2021).
[83]T. W. Ko, J. A. Finkler, S. Goedecker, and J. Behler, "A fourth-generation high-dimensional neural network potential with accurate electrostatics including non-local charge transfer," Nat. Commun. **12**, 398 (2021).
[84]J. S. Smith, O. Isayev, and A. E. Roitberg, "ANI-1: An extensible neural network potential with DFT accuracy at force field computational cost," Chem. Sci. **8**, 3192–3203 (2017).
[85]K. Yao, J. E. Herr, D. W. Toth, R. McKintyre, and J. Parkhill, "The TensorMol-0.1 model chemistry: A neural network augmented with long-range physics," Chem. Sci. **9**, 2261–2269 (2018).
[86]L. Zhang, J. Han, H. Wang, R. Car, and E. Weinan, "Deep potential molecular dynamics: A scalable model with the accuracy of quantum mechanics," Phys. Rev. Lett. **120**, 143001 (2018).
[87]L. Shen, J. Wu, and W. Yang, "Multiscale quantum mechanics/molecular mechanics simulations with neural networks," J. Chem. Theory Comput. **12**, 4934–4946 (2016).
[88]L. Shen and W. Yang, "Molecular dynamics simulations with quantum mechanics/molecular mechanics and adaptive neural networks," J. Chem. Theory Comput. **14**, 1442–1455 (2018).
[89]H. Wang and W. Yang, "Force field for water based on neural network," J. Phys. Chem. Lett. **9**, 3232–3240 (2018).
[90]Y. Zhang, C. Hu, and B. Jiang, "Embedded atom neural network potentials: Efficient and accurate machine learning with a physically inspired representation," J. Phys. Chem. Lett. **10**, 4962–4967 (2019).
[91]Y. Zhang, C. Hu, and B. Jiang, "Accelerating atomistic simulations with piecewise machine-learned *ab initio* potentials at a classical force field-like cost," Phys. Chem. Chem. Phys. **23**, 1815–1821 (2021).
[92]J. Zhang, Y.-K. Lei, Z. Zhang, J. Chang, M. Li, X. Han, L. Yang, Y. I. Yang, and Y. Q. Gao, "A perspective on deep learning for molecular modeling and simulations," J. Phys. Chem. A **124**, 6745–6763 (2020).
[93]M. K. Matlock, M. Hoffman, N. L. Dang, D. L. Folmsbee, L. A. Langkamp, G. R. Hutchison, N. Kumar, K. Sarullo, and S. J. Swamidass, "Deep learning coordinate-free quantum chemistry," J. Phys. Chem. A **125**, 8978–8986 (2021).
[94]C. A. Grambow, L. Pattanaik, and W. H. Green, "Deep learning of activation energies," J. Phys. Chem. Lett. **11**, 2992–2997 (2020).
[95]K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, and A. Tkatchenko, "Quantum-chemical insights from deep tensor neural networks," Nat. Commun. **8**, 13890 (2017).
[96]J. Lu, C. Wang, and Y. Zhang, "Predicting molecular energy using force-field optimized geometries and atomic vector representations learned from an improved deep tensor neural network," J. Chem. Theory Comput. **15**, 4113–4121 (2019).
[97]K. T. Schütt, H. E. Sauceda, P. J. Kindermans, A. Tkatchenko, and K. R. Müller, "SchNet—A deep learning architecture for molecules and materials," J. Chem. Phys. **148**, 241722 (2018).
[98]M. Gastegger, K. T. Schütt, and K.-R. Máller, "Machine learning of solvent effects on molecular spectra and reactions," Chem. Sci. **12**, 11473–11483 (2021).
[99]O. T. Unke and M. Meuwly, "PhysNet: A neural network for predicting energies, forces, dipole moments, and partial charges," J. Chem. Theory Comput. **15**, 3678–3693 (2019).
[100]J. Lu, S. Xia, J. Lu, and Y. Zhang, "Dataset construction to explore chemical space with 3D geometry and deep learning," J. Chem. Inf. Model. **61**, 1095–1104 (2021).

[101]O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schütt, A. Tkatchenko, and K.-R. Müller, "Machine learning force fields," Chem. Rev. **121**, 10142–10186 (2021).
[102]M. A. Collins and R. P. A. Bettens, "Energy-based molecular fragmentation methods," Chem. Rev. **115**, 5607–5642 (2015).
[103]W. Yang, "Direct calculation of electron density in density-functional theory," Phys. Rev. Lett. **66**, 1438–1441 (1991).
[104]E. E. Dahlke and D. G. Truhlar, "Electrostatically embedded many-body expansion for large systems, with applications to water clusters," J. Chem. Theory Comput. **3**, 46–53 (2007).
[105]N. J. Mayhall and K. Raghavachari, "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).
[106]R. M. Richard and J. M. Herbert, "A generalized many-body expansion and a unified view of fragment-based methods in electronic structure theory," J. Chem. Phys. **137**, 064113 (2012).
[107]K. Raghavachari and A. Saha, "Accurate composite and fragment-based quantum chemical models for large molecules," Chem. Rev. **115**, 5643–5677 (2015).
[108]V. Deev and M. A. Collins, "Approximate *ab initio* energies by systematic molecular fragmentation," J. Chem. Phys. **122**, 154102 (2005).
[109]M. A. Addicoat and M. A. Collins, "Accurate treatment of nonbonded interactions within systematic molecular fragmentation," J. Chem. Phys. **131**, 104103 (2009).
[110]M. A. Collins, "Molecular forces, geometries, and frequencies by systematic molecular fragmentation including embedded charges," J. Chem. Phys. **141**, 094108 (2014).
[111]M. S. Gordon, D. G. Fedorov, S. R. Pruitt, and L. V. Slipchenko, "Fragmentation methods: A route to accurate calculations on large systems," Chem. Rev. **112**, 632–672 (2012).
[112]D. W. Zhang and J. Z. H. Zhang, "Molecular fractionation with conjugate caps for full quantum mechanical calculation of protein–molecule interaction energy," J. Chem. Phys. **119**, 3599–3605 (2003).
[113]A. M. Gao, D. W. Zhang, J. Z. H. Zhang, and Y. Zhang, "An efficient linear scaling method for ab initio calculation of electron density of proteins," Chem. Phys. Lett. **394**, 293–297 (2004).
[114]X. He and J. Z. H. Zhang, "A new method for direct calculation of total energy of protein," J. Chem. Phys. **122**, 031103 (2004).
[115]J. Liu, T. Zhu, X. Wang, X. He, and J. Z. H. Zhang, "Quantum fragment based *ab initio* molecular dynamics for proteins," J. Chem. Theory Comput. **11**, 5897–5905 (2015).
[116]J. Gao, "Toward a molecular orbital derived empirical potential for liquid simulations," J. Phys. Chem. B **101**, 657–663 (1997).
[117]J. Gao, "A molecular-orbital derived polarization potential for liquid water," J. Chem. Phys. **109**, 2346–2354 (1998).
[118]W. Xie and J. Gao, "Design of a next generation force field: The X-POL potential," J. Chem. Theory Comput. **3**, 1890–1900 (2007).
[119]W. Xie, M. Orozco, D. G. Truhlar, and J. Gao, "X-POL potential: An electronic structure-based force field for molecular dynamics simulation of a solvated protein in water," J. Chem. Theory Comput. **5**, 459–467 (2009).
[120]W. Li, S. Li, and Y. Jiang, "Generalized energy-based fragmentation approach for computing the ground-state energies and properties of large molecules," J. Phys. Chem. A **111**, 2193–2199 (2007).
[121]S. Hua, W. Li, and S. Li, "The generalized energy-based fragmentation approach with an improved fragmentation scheme: Benchmark results and illustrative applications," ChemPhysChem **14**, 108–115 (2013).
[122]H. Wang and W. Yang, "Toward building protein force fields by residue-based systematic molecular fragmentation and neural network," J. Chem. Theory Comput. **15**, 1409–1417 (2019).
[123]Z. Wang, Y. Han, J. Li, and X. He, "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).
[124]Z. Cheng, J. Du, L. Zhang, J. Ma, W. Li, and S. Li, "Building quantum mechanics quality force fields of proteins with the generalized energy-based fragmentation approach and machine learning," Phys. Chem. Chem. Phys. **24**, 1326–1337 (2022).
[125]K. Liao, S. Dong, Z. Cheng, W. Li, and S. Li, "Combined fragment-based machine learning force field with classical force field and its application in the NMR calculations of macromolecules in solutions," Phys. Chem. Chem. Phys. **24**, 18559–18567 (2022).
[126]M. A. Collins and V. A. Deev, "Accuracy and efficiency of electronic energies from systematic molecular fragmentation," J. Chem. Phys. **125**, 104104 (2006).
[127]R. Krishnan, J. S. Binkley, R. Seeger, and J. A. Pople, "Self-consistent molecular orbital methods. XX. A basis set for correlated wave functions," J. Chem. Phys. **72**, 650–654 (1980).
[128]C. Lee, W. Yang, and R. G. Parr, "Development of the Colle–Salvetti correlation-energy formula into a functional of the electron density," Phys. Rev. B **37**, 785–789 (1988).
[129]A. D. Becke, "Density-functional thermochemistry. I. The effect of the exchange-only gradient correction," J. Chem. Phys. **96**, 2155–2160 (1992).
[130]A. D. Becke and E. R. Johnson, "A density-functional model of the dispersion interaction," J. Chem. Phys. **123**, 154101 (2005).
[131]S. Grimme, S. Ehrlich, and L. Goerigk, "Effect of the damping function in dispersion corrected density functional theory," J. Comput. Chem. **32**, 1456–1465 (2011).
[132]J. S. Smith, B. Nebgen, N. Lubbers, O. Isayev, and A. E. Roitberg, "Less is more: Sampling chemical space with active learning," J. Chem. Phys. **148**, 241733 (2018).
[133]J. S. Smith, B. T. Nebgen, R. Zubatyuk, N. Lubbers, C. Devereux, K. Barros, S. Tretiak, O. Isayev, and A. E. Roitberg, "Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning," Nat. Commun. **10**, 2903 (2019).
[134]X. Hu, M.-O. Lenz-Himmer, and C. Baldauf, "Better force fields start with better data: A data set of cation dipeptide interactions," Sci. Data **9**, 327 (2022).
[135]J. Hoja, L. Medrano Sandonas, B. G. Ernst, A. Vazquez-Mayagoitia, R. A. DiStasio, Jr., and A. Tkatchenko, "QM7-X, a comprehensive dataset of quantum-mechanical properties spanning the chemical space of small organic molecules," Sci. Data **8**, 43 (2021).
[136]H. Grubmüller, H. Heller, A. Windemuth, and K. Schulten, "Generalized Verlet algorithm for efficient molecular dynamics simulations with long-range interactions," Mol. Simul. **6**, 121–142 (1991).
[137]T. Darden, D. York, and L. Pedersen, "Particle mesh Ewald: An N log(N) method for Ewald sums in large systems," J. Chem. Phys. **98**, 10089–10092 (1993).
[138]U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen, "A smooth particle mesh Ewald method," J. Chem. Phys. **103**, 8577–8593 (1995).
[139]G. Bussi, D. Donadio, and M. Parrinello, "Canonical sampling through velocity rescaling," J. Chem. Phys. **126**, 014101 (2007).
[140]M. Parrinello and A. Rahman, "Polymorphic transitions in single crystals: A new molecular dynamics method," J. Appl. Phys. **52**, 7182–7190 (1981).
[141]P. Zhang, L. Shen, and W. Yang, "Solvation free energy calculations with quantum mechanics/molecular mechanics and machine learning models," J. Phys. Chem. B **123**, 901–908 (2019).
[142]M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Man'e, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Vi'egas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, TensorFlow large-scale machine learning on heterogeneous systems, software, tensorflow.org, 2015.
[143]D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv:1412.6980 (2014).
[144]M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl, "GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers," SoftwareX **1-2**, 19–25 (2015).
[145]M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D.

Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox, Gaussian 16 Revision A.03, Gaussian Inc., Wallingford CT, 2016.

[146]B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije, "LINCS: A linear constraint solver for molecular simulations," J. Comput. Chem. **18**, 1463–1472 (1997).

[147]J. Graf, P. H. Nguyen, G. Stock, and H. Schwalbe, "Structure and dynamics of the homologous series of alanine peptides: A joint molecular dynamics/NMR study," J. Am. Chem. Soc. **129**, 1179–1189 (2007).

[148]V. Mironov, Y. Alexeev, V. K. Mulligan, and D. G. Fedorov, "A systematic study of minima in alanine dipeptide," J. Comput. Chem. **40**, 297–309 (2019).

[149]Y. Yuan, M. J. L. Mills, P. L. A. Popelier, and F. Jensen, "Comprehensive analysis of energy minima of the 20 natural amino acids," J. Phys. Chem. A **118**, 7876–7891 (2014).

[150]M. K. Kesharwani, A. Karton, and J. M. L. Martin, "Benchmark *ab initio* conformational energies for the proteinogenic amino acids through explicitly correlated methods. assessment of density functional methods," J. Chem. Theory Comput. **12**, 444–454 (2016).

[151]S. Vijay-Kumar, C. E. Bugg, and W. J. Cook, "Structure of ubiquitin refined at 1.8 Å resolution," J. Mol. Biol. **194**, 531–544 (1987).

[152]C. Jelsch, M. M. Teeter, V. Lamzin, V. Pichon-Pesme, R. H. Blessing, and C. Lecomte, "Accurate protein crystallography at ultra-high resolution: Valence electron distribution in crambin," Proc. Natl. Acad. Sci. U. S. A. **97**, 3171–3176 (2000).

[153]T. S. Ulmer, B. E. Ramirez, F. Delaglio, and A. Bax, "Evaluation of backbone proton positions and dynamics in a small protein by liquid crystal NMR spectroscopy," J. Am. Chem. Soc. **125**, 9179–9191 (2003).

[154]K. Harata, "X-ray structure of monoclinic Turkey egg lysozyme at 1.3 Å resolution," Acta Crystallogr., Sect. D: Biol. Crystallogr. **49**, 497–504 (1993).