Electronic Supplementary Information for:

Building quantum mechanics quality force fields of proteins with the generalized energy-based fragmentation approach and machine learning

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

Key Laboratory of Mesoscopic Chemistry of Ministry of Education, Institute of Theoretical and Computational Chemistry, School of Chemistry and Chemical Engineering, Nanjing University, Nanjing, 210023, P. R. China. E-mail: shuhua@nju.edu.cn; majing@nju.edu.cn; wli@nju.edu.cn

Contents

S1. The Gaussian approximation potential (GAP) and GEBF-ML methodology without PM6.

For the description of the relationship between potential energy and structure, GAP pioneered by Bartók et al[1,2] is adopted in this work. The energy of a system E with N atoms is decomposed into atomic energies E_i,

$$E = \sum_{i=1}^{N} e_i = \sum_{i=1}^{N} \sum_{i_B=1}^{N_B} w_{i_B} K(\mathbf{X}_i, \mathbf{X}_{i_B}) \tag{1}$$

In eq 1, each e_i is expressed as a linear combination of the kernel function $K(\mathbf{X}_i, \mathbf{X}_{i_B})$ and weight factors w_{i_B}. The kernel is used to measure the similarity between the local configuration of atom *i* and the reference local configuration i_B. The smooth overlap of atomic positions (SOAP)[1] is used to describe the local atomic environment $\mathbf{X}_i$ and the kernel *K*. In SOAP, a local density of atom *i* from its neighbors within a radius R_{cut} is expressed as

$$\rho_i(\mathbf{r}) = \Sigma_{j=1}^{N} \frac{1}{(\sqrt{2\sigma_{atom}\pi})} \exp(-\frac{|\mathbf{r}-\mathbf{r}_{ij}|^2}{2\sigma_{atom}^2}) f_{cut}(r_{ij}) \tag{2}$$

$$f_{\text{cut}} = \begin{cases} 0.5 \cdot [1 + \cos(\frac{\pi R_{AB}}{R_{\text{cut}}})] & \mathrm{R}_{AB} \le R_{\text{cut}} \\ 0 & \mathrm{R}_{AB} > R_{\text{cut}} \end{cases} \tag{3}$$

Here, f_{cut} is a cutoff function in which the cutoff radius R_{cut} reflects the spatial scale of the interactions, $\mathbf{r}$ is the position vector of atom *i*, σ_{atom} is the hyperparameter and $\mathbf{r}_{ij}$ is the interatomic distance. R_{AB} is the distance between atoms A and B. The atomic neighbor density is expanded in terms of radial basis functions and spherical harmonic functions as

$$\rho_i(\mathbf{r}) = \sum_{l=1}^{L_{\max}} \sum_{m=-l}^{l} \sum_{n=1}^{N_R^l} c_{nlm}^i x_{nl}(r) Y_{lm}(\hat{\mathbf{r}}) \tag{4}$$

To keep the rotational invariance and to avoid rotational decoupling, the element of the descriptor is expressed as

$$p_{n_1 n_2 l}^i = \sum\nolimits_{m=-l}^{l} c_{n_1 lm}^{i^*} c_{n_2 lm}^i \tag{5}$$

Each vector $\mathbf{X}_i$ collects all coefficients $p^i_{n_1 n_2 l}$ (see eq 4) for the atomic neighbor density $\rho_i(\mathbf{r})$. In addition, the dot-product kernel is defined as

$$K(\mathbf{X}_i, \mathbf{X}_{i_B}) = \left(\sum_j \mathbf{X}_{i,j} \mathbf{X}_{i_B,j} \right)^{\varsigma} \tag{6}$$

It approaches 1 or 0 if two configurations are almost identical or totally different, respectively. Here, ς is a parameter to control the sharpness of the function *K*.

According to Eq 1, we can describe the energy and forces for a given structure as $\boldsymbol{\phi}\mathbf{w}$. Here, $\mathbf{w} = \{w_{i_B}\}$ and $\boldsymbol{\phi}$ is a matrix containing $K(\mathbf{X}_i, \mathbf{X}_{i_B})$ and its derivatives with respect to the coordinates. All training structures can be summarized as $\boldsymbol{\Phi}\mathbf{w}$, where $\boldsymbol{\Phi}$ collects $\boldsymbol{\phi}$ for all training structures. The parameters $\mathbf{w}$ and the uncertainty $\boldsymbol{\sigma}$ are simultaneously determined as

$$\mathbf{w} = \frac{1}{\sigma_v^2} \left(\frac{1}{\sigma_w^2} \mathbf{I} + \frac{1}{\sigma_v^2} \boldsymbol{\Phi}^T \boldsymbol{\Phi} \right)^{-1} \boldsymbol{\Phi}^T \mathbf{Y} \tag{7}$$

$$\boldsymbol{\sigma} = \boldsymbol{\phi} \left(\frac{1}{\sigma_w^2} \mathbf{I} + \frac{1}{\sigma_v^2} \boldsymbol{\Phi}^T \boldsymbol{\Phi} \right)^{-1} \boldsymbol{\phi}^T \tag{8}$$

Here, $\mathbf{Y}$ collects energies and forces in all training structures and $\mathbf{I}$ is a unit matrix. The symbols σ_v and σ_w are optimized iteratively by the evidence approximation[3] to balance the accuracy and robustness of the machine learning (ML) force field. The uncertainty $\boldsymbol{\sigma}$ is used to decide whether the quantum mechanics (QM) calculations are needed or not during the online active learning (see Sec.5).

The SOAP parameters for the two proteins are listed in Table S1.

Table S1 The SOAP parameters for two protein segments

system	4ZNN	1XQ8 segment
R_c	3.0	3.0
σ_{atom}	0.4	0.4
ζ	2.5	2.5
N_R^l	6	6
$L_{\max}$	3	3

If the PM6 method is not used as the baseline, the energy of the *m*th subsystem with S_m atoms are first extracted from the energy E_m^{DFT} by removing the Coulomb and Van der Waals interactions, then the remaining term is described as the summation of atomic energy e_i^m,

$$E_m^{\text{ML}} = E_m^{\text{DFT}} - (\sum_{A\in S_m}\sum_{B>A\in S_m}(\frac{Q_A Q_B}{R_{AB}} + \frac{C_{A,B}^{(12)}}{R_{AB}^{12}} - \frac{C_{A,B}^{(6)}}{R_{AB}^{6}}) f_{\text{cut}}^{'}) = \sum_{i\in S_m} e_i^m \tag{9}$$

$$f_{\text{cut}}^{'} = \begin{cases} 0.5\cdot[1-\cos(\frac{\pi R_{AB}}{R_c})] & \text{R}_{AB} \le R_{\text{cut}} \\ 1 & \text{R}_{AB} > R_{\text{cut}} \end{cases} \tag{10}$$

Here, the point charges are obtained from the natural population analysis (NPA) of subsystems, which are generated from the initial structure (extended structure generated from peptide sequence using Amber 16 program) used in the online training process. After training, the point charges are assumed to be constant like in traditional force fields. $\mathbf{r}_A$ and Q_A denote the coordinate of atom A and the point charge locating at atom A, respectively. R_c denotes the cutoff distances used in the ML model. $C_{A,B}^{(12)}$ and $C_{A,B}^{(6)}$ denote the pairwise dispersion coefficients in ff14SB force fields.[4] For van der Waals interactions, only interactions from nonbonded atom pairs defined in ff14SB force field are calculated.

After the training, the total energy of the target system is obtained as the summation of atomic contribution e_i and long-range interactions (including Coulomb and van der Waals interactions).

$$E^{\mathrm{ML}} = \sum_{i=1}^{N} e_i + \sum_{A} \sum_{B>A} \left(\frac{Q_A Q_B}{R_{AB}} + \frac{C_{A,B}^{(12)}}{R_{AB}^{12}} - \frac{C_{A,B}^{(6)}}{R_{AB}^{6}} \right) f_{\mathrm{cut}}^{'} \tag{11}$$

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

S2. Force correlations between GEBF-NN and QM methods

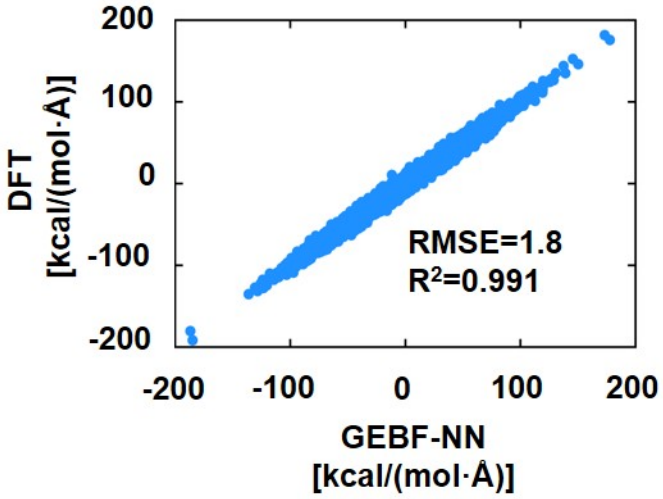

Fig S1. The comparison of correlations between the forces from generalized energy-based fragmentation based neural network (GEBF-NN) and the ωB97X-D/6-31G(d) ones. GEBF-NN force fields are trained directly from QM energies.

S3. Subsystem construction and coefficients determination

The main procedure of constructing subsystems in the generalized energy-based fragmentation (GEBF) method in this work are summarized as follows. (1) Divide the target system into various fragments. (2) For each fragment, construct a primitive subsystem by adding its neigboring environmental fragments within a distance threshold ζ. In the GEBF scheme, the distance between two fragments I and J is defined as the nearest distance between atoms in fragment I and atoms in fragment J. To control the size of primitive subsystems, we limit the maximum number of fragments in a subsystem as η. Hydrogen atoms are added to subsystems for valence saturation to avoid dangling bonds. In this work, the parameters ζ

and η are chosen as 3.0 Å and 4, respectively. (3) Once all primitive subsystems are obtained, derivative subsystems with their coefficients are constructed with the inclusion-exclusion principle, to cancel the overloading of primitive subsystems. The GEBF calculation is denoted as GEBF(ζ, η). Here, ACE-$(Ala)_3$-NME was used as an example to illustrate our subsystem construction.

As shown in Fig. S2 (a), the ACE-$(Ala)_3$-NME was first divided into five fragments, the box model of the molecular is shown in Fig. S2(b).

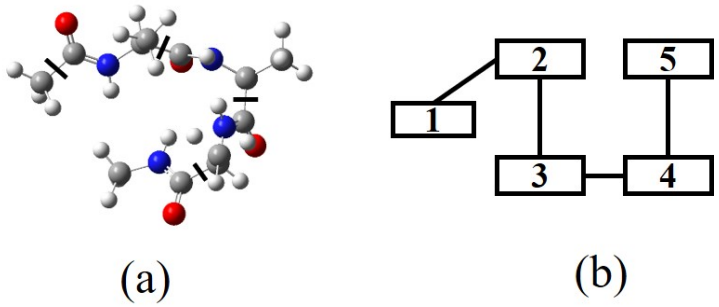

Fig. S2 Fragmentation scheme of ACE-$(Ala)_3$-NME: (a) Molecular structure of ACE-$(Ala)_3$-NME and four C-C bonds (denoted in solid line) are cut to generate five fragments; (b) box model of five fragments. The solid lines represent covalent single bonds.

(1) For each fragment (denoted as a central fragment), several neighboring (environmental) fragments were added to construct its primitive subsystem with ζ and η being 3.0 Å and 4, respectively. Hydrogen atoms are added for valence saturation. All the primitive subsystems are listed in Fig. S3.

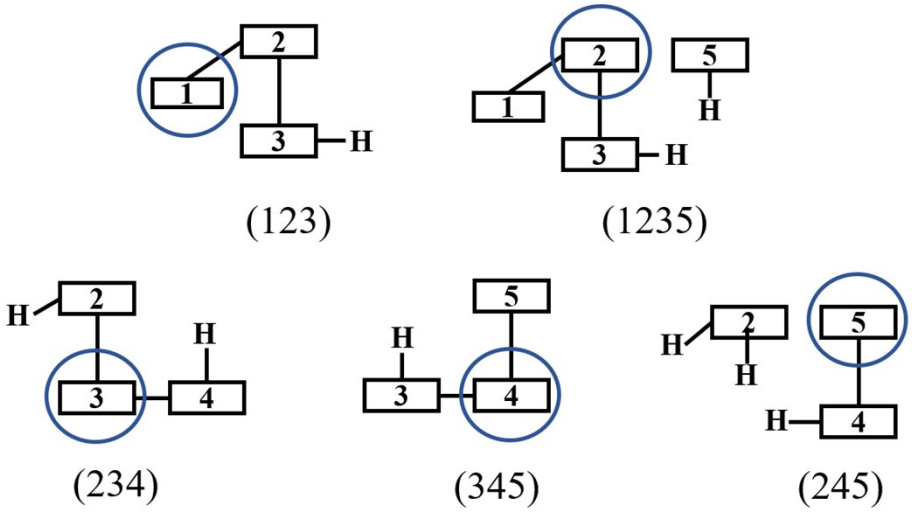

Fig. S3 Primitive subsystems of the ACE-$(Ala)_3$-NME, each of which contains a central fragment (inside

the circle) and its environmental fragments. The fragment indices in each subsystem are listed in parentheses.

(2) Delete the redundant small primitive subsystems, which are included in larger ones. For the ACE-$(Ala)_3$-NME, subsystem (123) is deleted because that it is included in the subsystem (1235). The retained primitive subsystems and their coefficients are shown in Fig. S4.

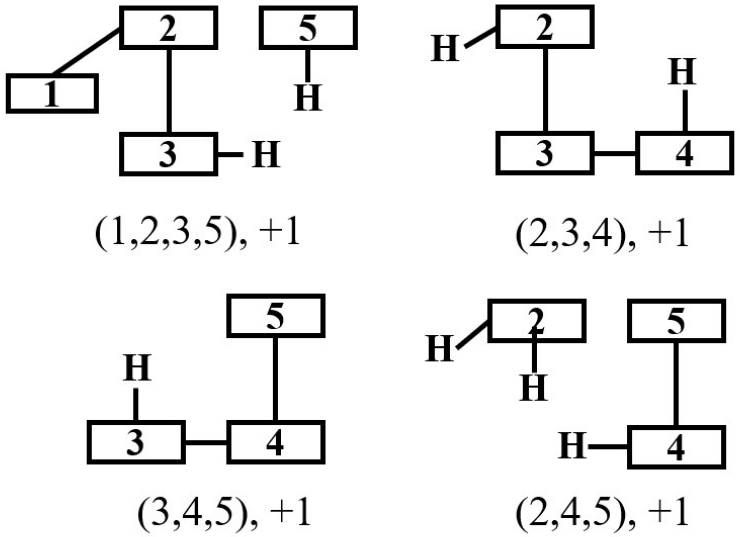

Fig. S4 The retained primitive subsystems. Fragment indices in each subsystem are listed in parentheses, and the coefficients are denoted after the parentheses.

(3) Build a series of derivative subsystems with the inclusion-exclusion principle to cancel the overlapping of primitive subsystems. All derivative subsystems and their coefficients are shown in Fig. S5

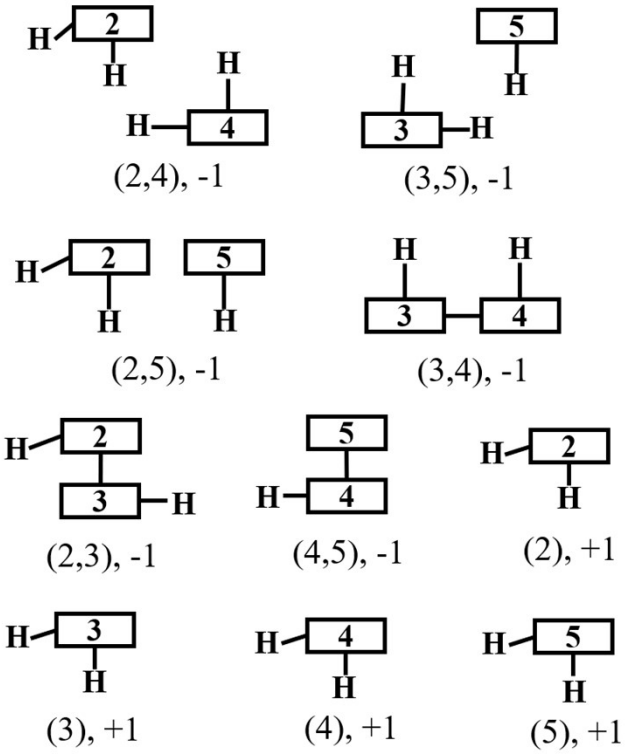

Fig. S5 The derivative subsystems are generated with the inclusion-exclusion principle. Fragment indices in each subsystem are listed in parentheses, and the coefficients are denoted after the parentheses.

S4. Subsystems Discrimination

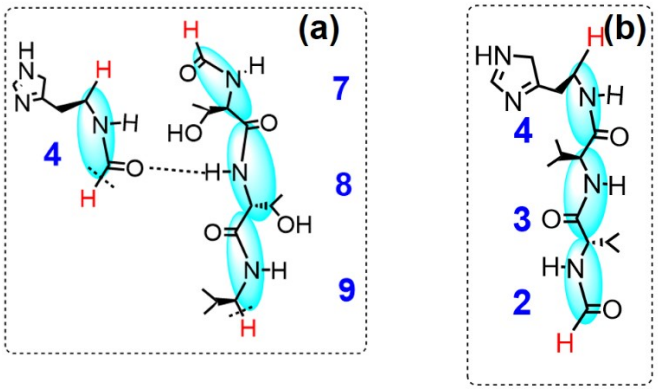

Fig. S6 Different subsystem type (a) His-Thr-Thr-Val-0-1-2-1 and (b) Val-Val-His-1-2-1

When constructing the data library, subsystems are discriminated against according to their bond types and amino acid type. Using the subsystems in Fig. S6 as examples, the subsystem in Fig. S6(a) can be

denoted as His-Thr-Thr-Val-0-1-2-1. The primary subsystem in Fig. S6(b) can be denoted as Val-Val-His-1-2-1. Here, the fragments are named by their residue name and sorted by the order from the ACE terminal. 0, 1, or 2 denotes the number of connected fragments for each fragment.

S5. Details of the online active learning

When collected the training set for a new protein, if the subsystems types are already in a data library, the corresponding sub-datasets were added as part of the training set for the new protein. During the new proteins training process, the data library aimed to store subsystems for all possible proteins is also expanded.

The details of the decision on whether to perform QM calculation or not are shown in Fig. S7. The Gaussian Process model gives the energy, forces, and Bayesian errors of forces on target systems. First, if the maximum Bayesian error $\sigma_{\max}$ ($\sigma_{\max} = \max\{\boldsymbol{\sigma}\}$)on the target system is larger than twice the threshold $\varepsilon_{\text{Bayes}}$, we will check the maximum Bayesian error σ_i on each subsystem. For subsystem *i,* if the maximum Bayesian error on its center fragment is larger than twice the $\varepsilon_{\text{Bayes}}$ or the subsystem type is not in the data library, QM calculations will be performed. It avoids instabilities caused by less-accuracy forces during the molecular dynamics (MD) simulation. Next, our program examines the previous QM calculation step. If the current step is within *m*th MD steps from the previous QM calculations, The QM calculation step will always be skipped. Here, *m* is defined as $m = \max(10, 0.18/\kappa)$ and κ (in Hartree/Å) is maximum forces error between ML force fields (MLFFs) results and QM results on the target system at the last QM calculation. This operation avoids too dense sampling during the MD simulation. Otherwise, if the maximum Bayesian error on the target system is larger than $\varepsilon_{\text{Bayes}}$, QM calculations will be performed on systems whose maximum Bayesian errors on central fragments are larger than ε_{Bayes} or their subsystem types are not in the data library. Finally, if QM calculations are performed on subsystems from more than 5 newly target conformations or the maximum forces error κ is larger than 0.036 Hartree/Å, MLFFs are updated using the difference between the density functional theory (DFT) result and PM6 result. The local configurations are first chosen by the values of forces error and then filtered using CUR matrix approximation.[5]

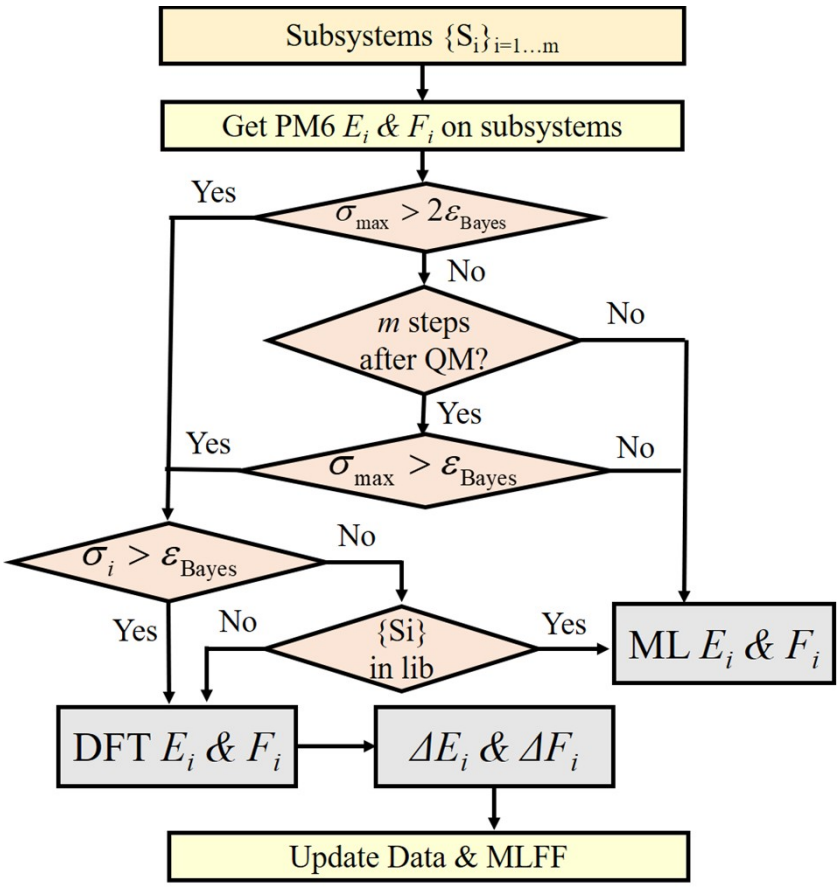

Fig S7. Details of the online training process. In our scheme, subsystem QM calculations are performed and added to the training sets when the maximum Bayesian error on the target system is larger than the threshold.

Fig. S8 shows how to set the Bayesian threshold $\varepsilon_{\text{Bayes}}$ automatically. Here, the threshold is set to zero at the beginning, if the number of data points in the training set is zero. To measure the lowest currently attainable Bayesian error, at the *I*th MD step just after the retraining of the force field, the maximum value of the Bayesian errors for the forces on the target system is stored as $\sigma_{\text{max},I}$. The threshold is updated to be the average of the last 10 $\sigma_{\text{max},I}$ if their relative standard deviation is less than 0.2.

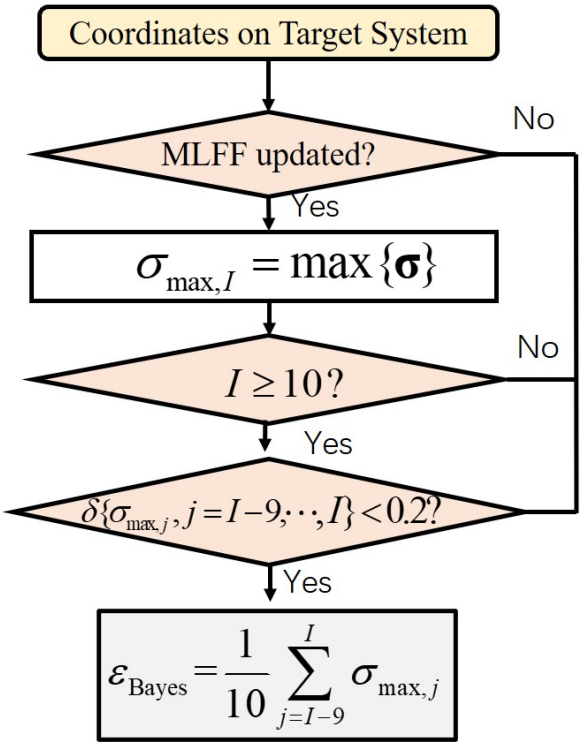

Fig. S8 Flowchart of the criterion setting step for Bayesian error threshold.

S6. Time evolutions of the total energies of proteins

We first perform MD simulations for two polypeptides in the microcanonical (NVE) ensemble. Although the net forces of capping hydrogen atoms calculated with the GEBF-PM6 method are very small, their forces are added to the corresponding carbon atoms which are replaced by these capping hydrogen atoms.[6] Fig. S9 shows total energy fluctuations whose initial velocities are consistent with T = 300 K. As shown in Fig. S9b and S9d, the energy drifts are negligible during the GEBF-PM6 simulation for both systems. For 4ZNN, Fig. S9a shows that the energy drift is about 0.001 kcal/(mol·atom·ps) during the MLFF-based MD simulation. For the 1XQ8 segment, the energy drift is even smaller during the MLFF-based MD simulation, as shown in Fig. S9c. The energy drift of our MLFFs is much less than those in the *ab initio* MD (AIMD) simulations (for example, 0.023 kcal/(mol·atom·ps)[7,8] for sodium-ion batteries) and in eReaxFF reactive force field MD simulations [0.01kcal/(mol·atom·ps)].[9] Thus, our GEBF-MLFF could be employed for long-time MD simulations to investigate the conformational changes of the two systems under study.

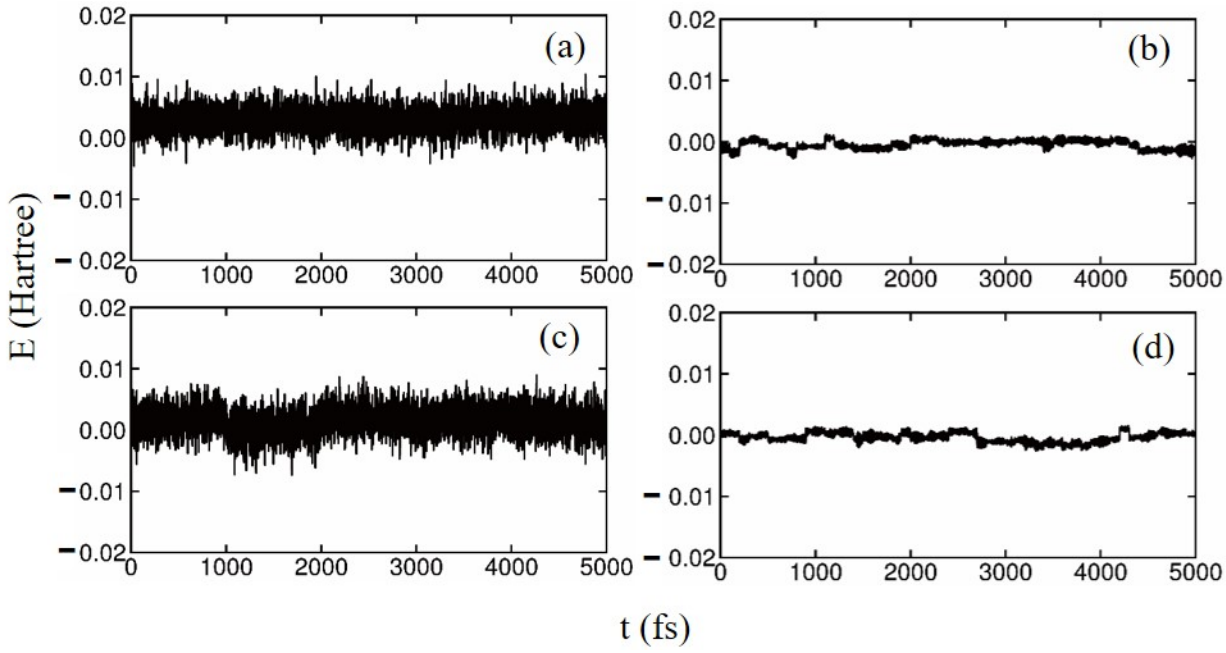

Fig. S9 The total energy fluctuations as functions of time in generalized energy-based fragmentation machine learning (GEBF-ML) (a, c) and GEBF-PM6 (b, d) molecular dynamics (MD) simulations of 4ZNN (a, b) and 1XQ8 segment (c, d) at the NVE ensemble with a time step of 1 fs. The initial velocities are consistent with T = 300 K. The zero energy is chosen to the total energy of the initial structure. The GEBF-ML MD simulations were performed using the force field without any retraining.

S7. Accuracy of PM6 methods

Table S2. The deviations (in kcal/mol) of the GEBF(3,4)-PM6 energies for ten 4ZNN and 1XQ8 segments, with respect to the conventional PM6 methods.

conformers	4ZNN	1XQ8 segment
1	0.073	-0.297
2	-0.394	0.795
3	0.271	0.340
4	-0.547	-0.237
5	0.246	0.574
6	0.052	-0.501
7	-0.393	0.350
8	-0.935	-0.787
9	-0.066	0.401
10	1.156	-0.038

References

1 A. P. Bartók, R. Kondor, and G. Csányi, *Phys. Rev. B,* 2013, **87**, 184115.

2 A. P. Bartók, and G. Csányi, *Int. J. Quantum Chem*., 2015, **115**, 1051-1057.

3 D. J. MacKay, *Neural Comput*., 1992, **4**, 415-447.

4 J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser and C. Simmerling, *J. Chem. Theory Comput*., 2015, **11**, 3696-3713.

5 M. W. Mahoney and P. Drineas, *Proc. Natl. Acad. Sci. USA*, 2009, **106**, 697.

6 M. Xu, T. Zhu and J. Z. H. Zhang, *Front. Chem*., 2018, **6**, 189.

7 M. M. Islam, G. Kolesov, T. Verstraelen, E. Kaxiras and A. C. T. van Duin, *J. Chem. Theory Comput*., 2016, **12**, 3463-3472.

8 J. Liu, C. Zhang, L. Xu, and S. Ju, *RSC Adv*., 2018, **8**, 17773-17785.

9 X. Lv, Z. Xu, J. Li, J. Chen and Q. Liu, *J. Fluorine Chem*., 2016, **185**, 42-47.