An On-the-fly Approach to Construct Generalized Energy–Based Fragmentation Machine Learning Force Fields of Complex Systems

Zheng Cheng[†], Dongbo Zhao[†,‡], Jing Ma[†], Wei Li[†,*], and Shuhua Li[†,*]

[†]Institute of Theoretical and Computational Chemistry, School of Chemistry and Chemical Engineering, Nanjing University, Nanjing 210023, People's Republic of China

[‡]Kuang Yaming Honors School, Nanjing University, 210023, People's Republic of China

ABSTRACT: An on-the-fly fragment-based machine learning (ML) approach was developed to construct the machine learning force field for large complex systems. In this approach, the energy, forces, and molecular properties of the target system are obtained by combining machine learning force fields of various subsystems with the generalized energy-based fragmentation (GEBF) approach. Using nonparametric Gaussian process (GP) model, all the force fields of subsystems are automatically generated online without data selection and parameter optimization. With the GEBF-ML force field constructed for a normal alkane, $C_{60}H_{122}$, long-time molecular dynamics (MD) simulations are performed on different sizes of alkanes, and the predicted energy, forces, and molecular properties (dipole moment) are favorably comparable with full quantum mechanics (QM) calculations. The predicted IR spectra also show excellent agreement with the direct *ab initio* MD results. Our results demonstrate that the GEBF-ML method provides an automatic and efficient way to build force fields for a broad range of complex systems such as biomolecules and supramolecular systems.

1. Introduction

Molecular dynamics (MD) simulations have achieved great success at the classical force field level. However, more accurate *ab initio* MD (AIMD) simulations can be only applied to very small systems on a time scale of picosecond mainly because that the computational cost of full quantum mechanics (QM) calculations increases rapidly with the system size. In the last decade, energy-based fragmentation approaches have been developed to greatly accelerate the QM calculation.[1–10] However, it is still very expensive to perform long-time AIMD simulations of large systems even using those approaches.

Recently, the applications of machine learning (ML) to chemical problems have attracted great interests because of their substantially reduced cost.[11–14] The high-dimensional neural network potentials (HDNNPs) was proposed by Parrinello and Behler.[15] An alternative ML model, Gaussian approximation potential (GAP), was introduced by Csányi and Bartók.[16] Both the HDNNPs and GAP are several orders of magnitude faster than conventional density functional theory (DFT) calculations, and have been applied to small molecules,[17] molecular clusters,[18] metal,[19,20] bulk materials,[21–24] surfaces,[25] liquid,[26–28] aqueous electrolyte solutions,[29] and solid-liquid interfaces.[30] Although the HDNNPs and GAP have shown high accuracy and acceleration ratio, it is still expensive to build ML-based force fields for large systems, since full QM calculations of large molecules are required.

In the recent years, the HDNNPs were combined with energy-based fragmentation approaches.[31–35] The force fields of methanol clusters[31] and water clusters[32] have been built by combining neural networks (NNs) with the many-body expansion method. The energies and forces of polypeptides have been predicted by combining the NNs with systematic molecular fragmentation (SMF)[33] and generalized molecular fractionation with conjugate caps (GMFCC)[34] methods. The excited states of large systems were calculated by combining the multilayer fragmentation method and NNs.[35] Using the fragmentation approaches in the NNs, fast QM calculations for some large systems are feasible. However, the NN parameters need to be carefully optimized to avoid overfitting, and the reference dataset for each type of subsystems is iteratively constructed with a large number of trial-and-error steps.[36]

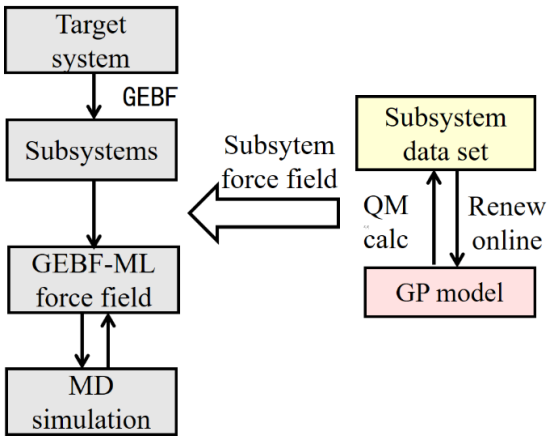

Figure 1. Schematic diagram of the GEBF-ML method. MLFF was automatically generated during the MD simulations at the cost of QM calculations of small subsystems. Force field on the target system are obtained from the subsystems using GEBF. In this work, as a proof-of concept, the electrostatic embedding is not considered.

In this work we combined the generalized energy-based fragmentation (GEBF) approach[37] and machine learning (ML) method to construct the GEBF-ML force fields for large molecules. As schematically shown in Figure 1, our GEBF-ML scheme can substantially reduce the costs of QM calculations on reference dataset, and accelerate the GEBF-AIMD simula-

tions by several orders of magnitude. It is achieved by combining subsystem force fields to build the ML force field for large systems, and thus full QM calculations of large molecules are avoided. Different from previous fragment-based NNPs, in this work, the nonparametric Gaussian process model is employed to construct the ML force fields for various types of subsystems without parameter optimization. Moreover, to generate reference dataset automatically, a robust on-the-fly algorithm,[36,38] is employed and modified to fit our GEBF-ML scheme. The GEBF-ML approach is applied to normal alkanes to verify its accuracy, efficiency, and robustness. The approach is expected to be applicable to more complex systems, such as molecular aggregates, polypeptides, proteins, and supramolecular systems.

2. Methodology

2.1. Gaussian approximation potential.

In this work, the smooth overlap of atomic positions (SOAP),[39] proposed by Bartók et al. and implemented in the QUIP package,[40] is used to describe the local atomic environment in molecular systems. SOAP tries to form a local density of atom i from its neighbors within a radius R_{cut} as,

$$\rho_i(\mathbf{r}) = \sum_{j=1}^{N_\alpha} \delta(\mathbf{r} - \mathbf{r}_{ij})\, f_{\text{cut}}(r_{ij}) \tag{1}$$

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, and $\mathbf{r}_{ij}$ is inter-atomic distance. To avoid discontinuity, the δ function is replaced by a normalized Gaussian function, so that eq 1 is rewritten as follows,

$$\rho_{\text{i}}(\mathbf{r}) = \sum_{j=1}^{N_\alpha} \frac{1}{\sqrt{2\sigma_{atom}\pi}} \exp\left(-\frac{|\mathbf{r} - \mathbf{r}_{\text{ij}}|^2}{2\sigma_{atom}^2}\right) f_{cut}(r_{ij}) \tag{2}$$

The atomic neighbor density is then expanded in terms of radial basis functions and spherical harmonic functions as

$$\rho_{\text{i}}(\mathbf{r}) = \sum_{l=1}^{L_{\max}} \sum_{m=-l}^{l} \sum_{n=1}^{N_R^l} c_{nlm}^i x_{nl}(r) Y_{lm}(\hat{\mathbf{r}}) \tag{3}$$

To keep the rotational invariance and avoid the rotational decoupling,[33] the element of the descriptor is expressed as:

$$p_{n_1 n_2 l}^i = \sum_{m=-l}^{l} c_{n_1 lm}^{i^*} c_{n_2 lm}^i \tag{4}$$

To interpolate the atomic energy in the SOAP space, the nonparametric Gaussian process regression[11] is adopted, where a set of N_{B} local reference structures $\{\rho_{i_B} \mid i_B = 1, \cdots, N_B\}$ are chosen and the local energy U_i of atom i is approximately obtained by fitting a set of coefficients $\{\mathrm{w}_{i_{\text{B}}} \mid i_B = 1, \cdots, N_B\}$:

$$U_i = \sum_{i_B=1}^{N_B} \mathrm{w}_{i_B} K(\mathbf{X}_i, \mathbf{X}_{i_{\text{B}}}) \tag{5}$$

Each vector $\mathbf{X}_i$ collects all coefficients $p_{n_1 n_2 l}^i$ (see eq 4) for the atomic neighbor density ρ_i. The kernel function K is used to measure the similarity between a local configuration of interest $\rho_i(\mathbf{r})$ and a reference configuration $\rho_{i_B}(\mathbf{r})$. It approaches 1 or 0 if two configurations are almost identical or totally different, respectively. In addition, the dot-product kernel is defined as

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

where ζ is a parameter to control the sharpness of the function K.

The energies and forces for a set of reference datasets labeled by a superscript $\alpha = 1, \cdots, N_{st}$ are fitted to determine the coefficients w_{i_B} and the covariance Σ. The total energy is written as a sum of atomic energies and fulfilled in a least square sense as,

$$U^\alpha \equiv \sum_{i=1}^{N^\alpha} U_i^\alpha = \sum_{i_B=1}^{N_B} \mathrm{w}_{i_B} \sum_{i=1}^{N^\alpha} K(\mathbf{X}_i^\alpha, \mathbf{X}_{i_B}) \tag{7}$$

In eq 7, U^α is the actual QM energy, U_i^α is the local energy of atom i in the structure α, and N^α is the number of atoms in a structure α. The partial derivative of the total energy leads to the forces,

$$\mathrm{f}_{j,k}^\alpha = -\frac{\partial U^\alpha}{\partial r_{jk}} = -\sum_{i_B=1}^{N_B} \mathrm{w}_{i_B} \sum_{i=1}^{N^\alpha} \frac{\partial K(\mathbf{X}_i^\alpha, \mathbf{X}_{i_B})}{\partial r_{jk}} \tag{8}$$

Here, $r_{j,k}^\alpha$ and $\mathrm{f}_{j,k}^\alpha$ are the kth (k = 1, 2, 3) component of the Cartesian coordinates and forces of atom j in the structure α, respectively.

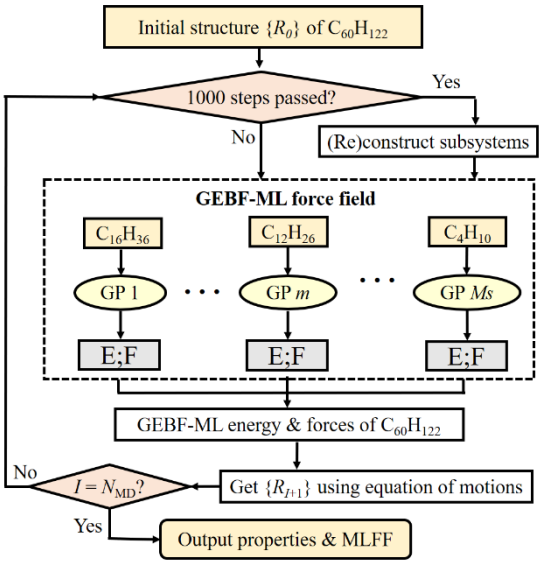

Figure 2. Flowchart of on-the-fly GEBF-ML force field generation scheme. In our scheme, the energy and force of the target molecule ($C_{60}H_{122}$) is obtained from the combination of energies and forces predicted by the respective individual ML model of different types of subsystems ($C_{16}H_{36}$, ($C_{12}H_{26}$+C_4H_{10}), $C_{16}H_{34}$, eta). All different types of ML force fields are updated individually on-the-fly.

In practice, the total energy and forces in eqs 7 and 8 are trained together via a compact matrix-vector form: $\mathbf{y}^\alpha = \boldsymbol{\phi}^\alpha \mathbf{w}$, where $\{\mathbf{y}^\alpha \mid \alpha = 1, \cdots, N_{st}\}$ denotes column vectors containing the dimensionless QM potential energy and the forces for α in the

reference structure dataset, with $m^\alpha = (1 + 3N^\alpha)$ components for N^α atoms. $\boldsymbol{\phi}^\alpha$ is a $m^\alpha \times N_B$ matrix, in which the first line is made up by $\Sigma_i K(\mathbf{X}_i^\alpha, \mathbf{X}_{i_B})$ and the partial derivatives of the function K (with respect to the coordinates in the structure α) in the subsequent rows.

After fitting, one can obtain both the coefficient **w** and the uncertainty from the GP regression model. In Gaussian process, the coefficient **w** is assumed as a Gaussian distribution, the posterior distribution is express as

$$p(\mathbf{w} \mid \mathbf{Y}) = N(\overline{\mathbf{w}}, \boldsymbol{\Sigma}) \tag{11}$$

$$\overline{\mathbf{w}} = \frac{1}{\sigma_v^2} \boldsymbol{\Sigma} \boldsymbol{\Phi}^{\mathbf{T}} \mathbf{Y} \tag{12}$$

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{\sigma_{\mathrm{w}}^2} \mathbf{I} + \frac{1}{\sigma_v^2} \boldsymbol{\Phi}^{\mathbf{T}} \boldsymbol{\Phi} \tag{13}$$

Here, **w** is not a constant vector after fitting, but considered as a multi-dimensional Gaussian distribution with mean vector $\overline{\mathbf{w}}$ and covariance matrix **Σ**. **Y** is a supervector (with its dimension as $M = \Sigma_\alpha m^\alpha$) to collect all QM energies and forces in the reference structure datasets $\{\mathbf{y}^\alpha \mid \alpha = 1, \cdots, N_{st}\}$. The matrix **Φ** with the size of $M \times N_B$ is a collection of all matrices $\boldsymbol{\phi}^\alpha$ on the reference datasets, and **I** is a unit matrix. The symbols σ_v^2 and σ_w^2 are optimized iteratively by the evidence approximation[41,42] to balance the accuracy and robustness of the machine learning force field.

For a new structure, the energy, forces, and uncertainty can be obtained by

$$\mathbf{y} = \boldsymbol{\phi} \overline{\mathbf{w}} \tag{14}$$

$$\boldsymbol{\delta} = \boldsymbol{\phi} \boldsymbol{\Sigma} \boldsymbol{\phi}^T \tag{15}$$

Here, $\mathbf{X}_i$ is the descriptor of atom i, and **ϕ** comprises $\Sigma_i K(\mathbf{X}_i, \mathbf{X}_{i_B})$ in the first row and the partial derivatives of the function K (with respect to the coordinates) in the subsequent rows. The uncertainty is express as Bayes error **δ**, which is used to decide whether the QM calculations are needed or not during the on-the-fly force field generation.[36,38]

2.2. GEBF-ML.

In the GEBF approach,[37,43] the ground-state energy of a target system can be obtained from a series of small "electrostatically embedded" subsystems as,

$$\mathrm{E}_{tot} = \sum_m^M C_m \tilde{E}_m - (\sum_m^M C_m - 1) \sum_A \sum_{B>A} \frac{Q_A Q_B}{|\mathbf{r}_A - \mathbf{r}_B|} \tag{16}$$

Here, $\tilde{E}_m$ and C_m are the energy (including self-energy of point charges) and coefficient of the mth subsystems, M is the number of subsystems, and $\mathbf{r}_A$ and Q_A are the coordinates of atom A and the net point charge located on atom A, respectively. The details of the fragmentation scheme are described in the Supporting Information (SI). For alkanes under study, there are no polar groups so that the net point charge on each atom can be approximately taken as zero. Thus, the total energy in eq 16 can be simplified as

$$E_{tot} = \sum_m^M C_m E_m \tag{17}$$

To verify whether eq 17 is a good approximation for normal alkanes, 10 randomly chosen conformers of $C_{60}H_{122}$ are calculated with eq17 at the ωB97X-D/6-31G(d,p) level. Our calculations show that the mean absolute errors (MAEs) of energies and forces are only 0.0025 kcal/(mol atom) and 0.01 kcal/(mol ·Å), respectively, relative to their corresponding conventional results. Thus, eq 17 works well for alkanes under study.

In the GEBF-ML scheme, different types of subsystems are predicted by their own Gaussian processes. Then, the energy of the whole system can be expressed as:

$$E_{tot} = \sum_m^{M_s} \sum_j^{S_m} C_j^m E_j^m \tag{18}$$

where M_s is the number of different types of subsystems and S_m is the number of subsystems for the mth type. C_j^m and E_j^m are the coefficient and energy of the jth subsystem in the mth type. The forces of the target system can also be obtained from all subsystems as,

$$\frac{\partial E_{tot}}{\partial \mathbf{r}_A} = \sum_m^{M_s} \sum_j^{S_m} C_j^m \frac{\partial E_j^m}{\partial \mathbf{r}_A} \tag{19}$$

where $\mathbf{r}_A$ is the coordinates of atom A. In addition, the molecular properties, such as dipole moments, can be evaluated as

$$\Omega_{tot} = \sum_m^{M_s} \sum_j^{S_m} C_j^m \Omega_j^m \tag{20}$$

where the symbol Ω denotes the molecular properties.

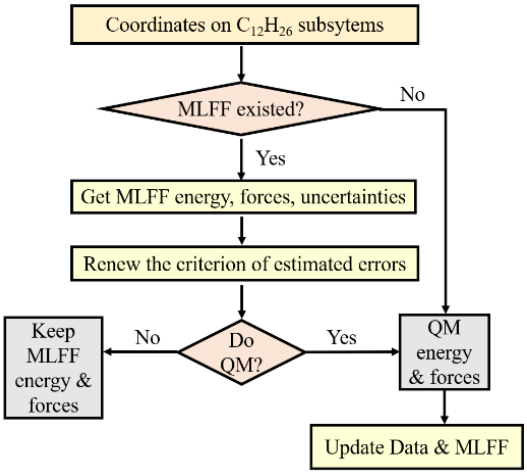

Figure 3. Flowchart of subsystem force field generation by taking $C_{12}H_{26}$ as an example.

2.3. Outline of the on-the-fly Force Field Generation.

To avoid a large amount of machine learning operations on subsystems generated by the GEBF approach, a nonparametric Gaussian process and a robust on-the-fly data generation scheme is chosen to efficiently generate the force field on the

fly.[36,38] Take a normal alkane $C_{60}H_{122}$ as an example. The flowchart of our GEBF-ML scheme is shown in Figure 2, in which data sets are automatically generated from subsystems rather than the entire molecule. Details of subsystems force field generation are displayed in Figure 3 by taking $C_{12}H_{26}$ as an example. The key steps in the scheme are outlined as below:

(1) The subsystems are reconstructed (or constructed at the very beginning) by the GEBF approach when 1000 MD steps passed after the last construction. The differences between the GEBF and full QM results are expected to be small if the subsystem types are fixed in such a short time, which makes the criterion for Bayesian on each subsystem being robust enough.

(2) The GEBF subsystems of $C_{60}H_{122}$ are classified into 7 types, $C_{16}H_{36}$ ($C_{12}H_{26}$+C_4H_{10}), $C_{16}H_{34}$, $C_{12}H_{28}$ (C_8H_{18}+C_4H_{10}), $C_{12}H_{26}$, C_8H_{20} (C_4H_{10}+C_4H_{10}), C_8H_{18}, and C_4H_{10}, according to their bonded connection. The (re)construction of the fragmentations is described in the Supporting Information (SI 4). Subsystems are constructed with different fragments, so we can easily classify the subsystem with the prior knowledge of fragments. In each type of subsystems, the energy, forces, and uncertainties are predicted by its own Gaussian process model.

(3) For the Gaussian process model of each set, if a criterion based on the uncertainties, the history of previous sampling, and the history of previous subsystem construction is met, QM calculations at the ωB97X-D/6-31G(d,p) level will be performed on the subsystems individually, otherwise, skip to step 5.

(4) For any type of Gaussian process model, if the number of the newly collected subsystems reaches a certain threshold, or if the uncertainties becomes too large, the set of reference structure datasets and local reference configurations belonging to this machine learning model are updated and the machine learning model is retrained.

(5) Predict the energies and forces of subsystems by the machine learning model.

(6) Combine the energies and forces predicted by subsystems Gaussian process model to obtain the total energy and forces of $C_{60}H_{122}$ by the GEBF formulas in eqs 18 and 19.

(7) Update the atomic positions and velocities by solving equation of motion, and then return to step 1 until the finalization of MD simulation ($I = N_{MD}$ in Figure 2).

The details of the decision on whether to do QM calculation or not at the I-th MD step are shown in Figure 4. Taking the $C_{12}H_{26}$ set as an example, the Gaussian process model gives the energies, forces, and uncertainties of all the $C_{12}H_{26}$ subsystems. The criterion setting is shown in the first dotted box, where the threshold (ε_{Bayes}) for the Bayes error (δ_{Bayes}) is automatically determined on the fly. Here, the threshold is set to zero at the beginning. To measure the lowest currently attainable Bayesian error, at the MD step I just after the refinement of the force field, the maximum value of the Bayes errors of the forces predicted for all the $C_{12}H_{26}$ is stored as $\delta_{Bayes}^{\max,I}$. The threshold is updated to be the average of the last ten $\delta_{Bayes}^{\max}$ if their relative standard deviation is < 0.2. The decision of whether to do QM calculation or not is shown in the second dotted square. First, if the maximum Bayesian error in all subsystems is larger than twice of the threshold, the QM calculations will be performed. It avoids instabilities in the MD simulation caused by less accurate forces. Next, our program examines the previous subsystem construction step. If the current step is within 20 MD steps from the previous subsystem construction step and new subsystems are generated, the maximum Bayes error is examined directly. If the maximum Bayes error is larger than the threshold, QM calculations is performed. This operation avoids the new structure, which are significantly different from the structures in the training set, to be predicted by the machine learning model. Next, the program checks the previous data sampling step. The QM calculations will always be skipped if the current step is within 10 MD steps from the previous sampling step. Otherwise, if the maximum Bayes error is larger than the threshold, the QM calculation is performed, this operator avoids too dense sampling during the MD simulation.

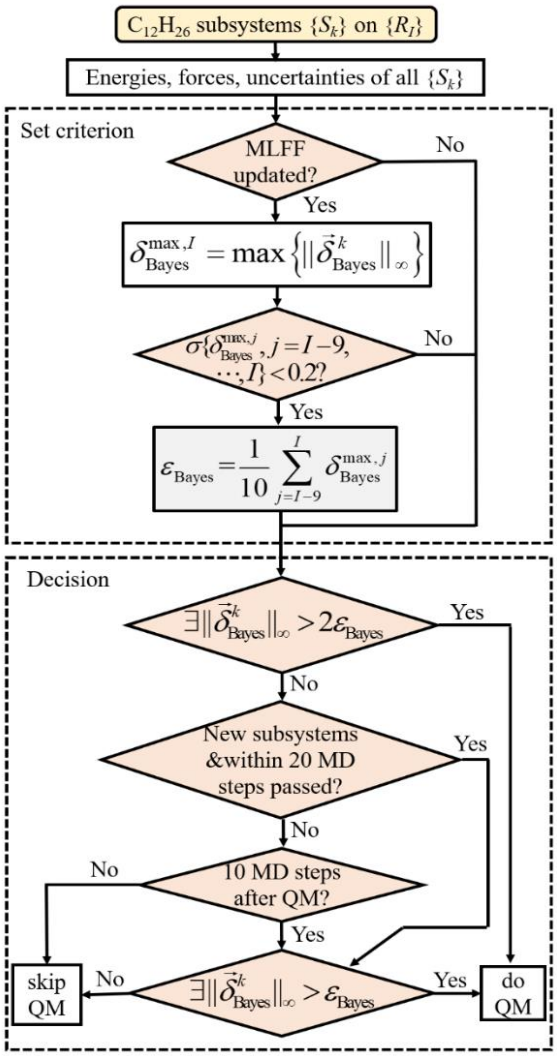

Figure 4. Flowchart of the decision step whether to perform subsystem QM calculation or not by taking $C_{12}H_{26}$ as an example. The symbols $\vec{\delta}_{Bayes}^{k}$ denotes the Bayesian forces error of kth subsystem in $C_{12}H_{26}$ set. $\|\vec{x}\|_\infty$ denotes the infinity norm, ε_{Bayes} donates the criterion for the Bayesian error, $\sigma(x)$ refer to the variance of the data x. At the beginning of training, the criterion ε_{Bayes} is set to zero.

For each type of subsystem force field generation, dataset and the ML force field is updated if five newly QM calculations are performed or when the estimated errors are twice larger than the

determined criteria. The local configuration choice and reference dataset sparsification are performed individually, which are the same as the previous operation on periodic systems.[36,38] Here, the main purpose is to reduce the computational costs and the memory requirement.

2.4. Machine learning dipole moments and IR spectra.

Since the GEBF method is also applicable for calculating the molecular properties, we also try to predict the dipole moments with the GEBF and machine learning model in this work. Moreover, IR spectra depending on the molecular dipole moments are obtained to check the accuracy of the GEBF-ML model.

In AIMD, vibrational spectra are computed via the Fourier transformation of time autocorrelation functions.[44] IR spectra depend on the molecular dipole moments as:

$$I_{IR} \propto \int_{-\infty}^{+\infty} \left\langle \dot{\mu}(\tau)\dot{\mu}(\tau+t) \right\rangle_{\tau} e^{-iwt} dt \qquad (21)$$

where $\dot{\mu}$ is the time derivative of the molecular dipole moment, ω is the vibrational frequency, τ is a time lag, and t is the time.

Here we also use the Gaussian process to predict the dipole moments, similar to the approach used in neural network.[45–47] The molecular dipole moments are expressed as:.

$$\boldsymbol{\mu} = \sum_{i_B=1}^{N_B} \mathrm{w}_{i_B}^{q} \sum_{i=1}^{N^{\alpha}} K(\mathbf{X}_i^{\alpha}, \mathbf{X}_{i_B}) \mathbf{r}_i \qquad (22)$$

Here, $\{\mathrm{w}_{i_B}^{q} \mid i_B = 1,....N_B\}$ are the coefficients in dipole moment machine learning model after fitting, $\mathbf{r}_i$ is the distance vector of the atom i from the molecular center of mass, and K is the kernel function used to measure the similarity between a local configuration of interest $\rho_i(\mathbf{r})$ and the reference configurations $\rho_{i_B}(\mathbf{r})$. In the GEBF-ML scheme, we first construct dipole machine learning models for all types of subsystems, and obtain the dipole moment of the target system by eq 20.

Table 1. The Root Mean Squared Errors (RMSEs) of the Energies [in kcal/(mol atom)], Forces [in kcal/(mol ·Å)], and Dipole Moments (in Debye) (with respect to the Conventional DFT Results) for the Test Set of Subsystems Obtained with the GEBF-ML Force Field.

Subsystem Type	Energy	Force	Dipole Moment
C_4H_{10}	0.076	1.38	0.023
C_8H_{18}	0.039	1.76	0.027
$(C_4H_{10})_2$	0.045	1.70	0.038
$C_{12}H_{26}$	0.040	1.95	0.044
C_8H_{18}+C_4H_{10}	0.033	1.46	0.078
$C_{16}H_{34}$	0.043	1.69	0.070
$C_{12}H_{26}$+C_4H_{10}	0.036	1.84	0.076

To construct the force field for normal alkanes, on-the-fly machine learning MD simulations were performed on $C_{60}H_{122}$ to generate different types of subsystems and their respective machine learning force fields are obtained automatically. The energies, forces and molecular properties of any other alkanes can also be obtained from the force fields of various subsystems using the GEBF method.

During the on-the-fly force field generation, more than 99% of the QM calculations are skipped, which reduces the computational time by a factor > 200. The details of the skipping ratio and the acceleration on each subsystem are summarized in Table S1.

On-the-fly force field generation has shown efficient sampling on the liquids, solids, and interfaces in previous studies.[36,38] In the GEBF-ML method, the number of structures in the reference structure dataset of each type of subsystem is typically < 1000, and the number of local reference configurations is < 1500 as shown in Table S2. Both also show the high efficiency sampling, and our method can be easily extended to more advanced electronic structure methods due to the computational linear scaling of these fragmentation methods.[48–50]

In addition to the significant acceleration of the computations, the accuracy of the generated force field should also be evaluated. A total of 300 $C_{60}H_{122}$ structures were randomly chosen from the trajectory at 500 K, different types of GEBF subsystems were generated with the LSQC program.[43] The electronic structure calculations of those subsystems were carried out at the ωB97X-D/6-31G(d,p) level with the Gaussian 16 package.[51] Parameters of SOAP for different subsystems are listed in Table S3. In Table 1, the root mean squared errors (RMSEs) of energy, forces and dipole moment obtained with the GEBF-ML method, relative to the conventional ωB97X-D/6-31G(d,p) results, on each type of subsystem are shown. One can see that our force field can accurately predict the potential energy surface with the RMSEs of energies and forces < 0.04 kcal/(mol·atom) and 2.0 kcal/(mol·Å), respectively. Here for each type of subsystem, the dipole moment is also predicted based on the reference structure dataset and local configurations. The RMSEs of dipole moments are typically < 0.07 Debye (over a range of 1.436 Debye), which is small enough for predicting the IR spectra.

3. Results and Discussion

3.1. Molecular Dynamics Simulation and Infrared Spectra of $C_{60}H_{122}$.

To evaluate the performance of our force field on the target molecule ($C_{60}H_{122}$), we have randomly chosen 300 structures from the GEBF-ML MD trajectory at 500 K. Table S4 shows that the RMSEs of the energies and forces for $C_{60}H_{122}$ are 0.033 kcal/(mol atom) and 2.37 kcal/(mol ·Å), respectively. The distributions of the energy and force errors between the ML force field and the conventional ωB97X-D reference data are displayed in Figure S1. It can be seen that almost all the force errors are < 10 kcal/(mol ·Å). Although the RMSE of forces on the target molecule is slightly larger than those of subsystems, it is still small enough to perform the MD simulations.

With the GEBF-ML force field, MD simulations using a Langevin thermostat[52] have been performed directly at 500 K with a timestep of 0.5 fs. To quantitatively describe the conformational changes, the RMSDs with respect to the initial structure of the $C_{60}H_{122}$ during the simulation are shown in Figure 5a. The RMSD increases rapidly in the first 5 ps and reaches the maximum value (10 Å) at 40 ps. It is consistent with the evolution of the structure in Figure 5b, which shows that the structure of the alkane is gradually changed from the straight chain to the folded one. Figure S2b depicts total energy fluctuations obtained from microcanonical (NVE) simulations whose initial

velocities are consistent with T = 300 K. In our GEBF-ML method, atomic energy of added Hydrogen for valence saturation of subsystem (subsystem construction is described in SI) is ignored as their net number in target system is zero. During the NVE simulation, the energy drift was found to be 0.03 meV/atom/ps, which may be caused by the discontinuities during the subsystem construction. In the AIMD simulations of sodium-ion batteries[53] and molten salt,[54] an energy drift of less than 1 meV/atom/ps is promised to ensure the NVT simulations. During the eReaxFF (a reactive force field) based MD simulations, the energy drift of about 0.4 meV/atom/ps is considered to be small.[55] Because the forces are analytically predicted by the machine learning force field during the simulations, our energy drift is much lower than that of direct GEBF-AIMD simulations, 1.2 meV/atom/ps.[56] Thus, our GEBF-ML method could be used to perform the NVT MD simulations of large alkanes to investigate their conformational changes.

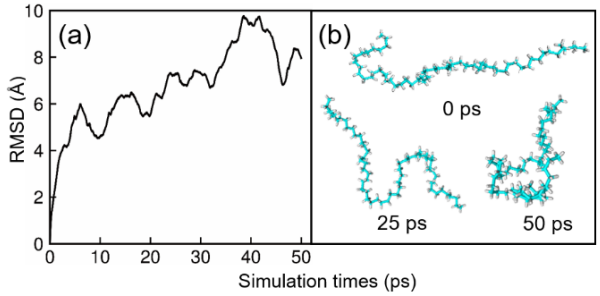

Figure 5. Time evolution of the total energy, potential energy (a) and RMSD with respect to the initial structure (b) in the machine learning molecular dynamics simulation of $C_{60}H_{122}$. (c) Conformational changes for $C_{60}H_{122}$ during the 50 ps machine learning molecular dynamic simulations.

After the force field has been trained, IR spectra of $C_{60}H_{122}$ were obtained with MD simulations in the gas phase employing a timestep 0.5 fs. After a short initial equilibration period (5 ps), constant temperature MD simulations were run for 50 ps. As shown in Figure 6a, the IR spectrum of $C_{60}H_{122}$ exhibits all of the spectroscopic features typical for simple hydrocarbons: the C-H scissoring (1505 cm^{-1}), methyl rock (1383 cm^{-1}), long-chain methyl rock (732 cm^{-1}), and the strong band in the 3100-3000 cm^{-1} region due to the C-H symmetric and asymmetric stretching. In comparison with the experimental frequencies, some peak positions predicted by the machine learning force field deviate from the experimental values to some extent. There is a blue shift from the typical experimental value of 2950 cm^{-1} to 3050 cm^{-1} for the C-H stretching vibrations. For $C_{60}H_{122}$, direct AIMD simulation are not available for comparison, due to very expensive *ab initio* calculations.

Instead, we performed direct AIMD simulations for $C_{12}H_{26}$ and compared the corresponding IR spectra of this system with the GEBF-ML MD results in Figure 6b. Figure 6b shows that the two IR spectra are almost coincident and have similar blue shifts for the C-H stretching vibrations. It indicates that the GEBF-ML MD could reproduce the IR spectra of the direct AIMD, and the blue shift may be caused by the underlying electronic structure method or the classical description of the nuclear dynamics.[57] Therefore, our machine learning force field can well reproduce the direct AIMD trajectory by skipping more than 99% of the QM calculations.

3.2. Force Field Application to Different Sizes of Alkanes.

Although the ML force fields of various subsystems are only trained from $C_{60}H_{122}$, other alkanes with different sizes may also be well predicted by the current force field, because the energies, forces and properties of any long-chain normal alkanes can be obtained with the GEBF method. Here, $C_{40}H_{82}$ and $C_{80}H_{162}$ are employed as two examples to test the accuracy of our force field on those alkanes which are not used during the on-the-fly force field generation. The initial (0 ps), middle (25 ps), and final (50 ps) snapshots for $C_{40}H_{82}$ and $C_{80}H_{162}$ during the GEBF-ML MD simulations are displayed in Figure S3. Both the $C_{40}H_{82}$ and $C_{80}H_{162}$ have large conformational changes from the straight structure to the folded one.

The RMSDs of $C_{40}H_{82}$ and $C_{80}H_{162}$ with respect to their initial structures are plotted in Figure S4a and b to quantitatively describe their conformational changes. For $C_{40}H_{80}$, the RMSDs increase to the maximum value at 25 ps and then decrease, which is consistent with the conformational changes in Figure S3. While for $C_{80}H_{162}$, the RMSDs exhibit a continuous increase. For $C_{40}H_{82,}$ energy drift is not observed in the NVE simulations, as shown in Figure S2a. Figure S2c shows that the energy drift for $C_{80}H_{162}$ is only about 0.03 meV/atom/ps, which is also small enough.

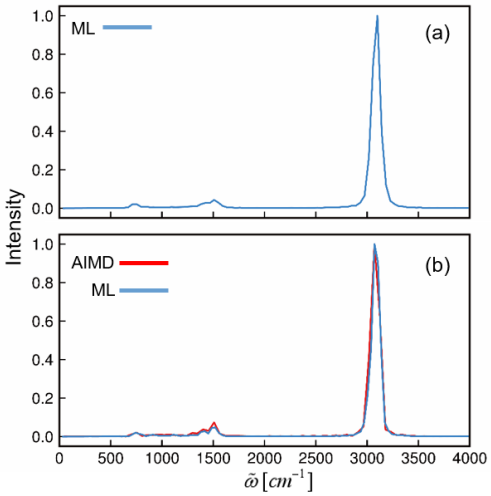

Figure 6. Infrared spectra of (a) $C_{60}H_{122}$ predicted by the ML model, and (b) $C_{12}H_{26}$ predicted by direct AIMD simulations (red) and GEBF-ML MD simulations (blue)

To evaluate the accuracy of the GEBF-ML force field on $C_{40}H_{82}$ and $C_{80}H_{162}$, 300 configurations for each alkane were randomly sampled from the GEBF-ML MD trajectories at 500 K. Table S4 shows that the RMSEs in energies and forces are only < 0.04 kcal/(mol atom) and 2.9 kcal/(mol ·Å), respectively. The distributions of the energy and forces errors between the GEBF-ML force field and the ωB97X-D reference data are shown in Figure S1, which suggests that almost all the force errors are < 10 kcal/(mol ·Å). The force errors of $C_{40}H_{82}$ and $C_{80}H_{162}$ are slightly larger than that of $C_{60}H_{122}$, but are still small enough for describing their potential energy surfaces. The energy errors per atom for these two alkanes are similar to that in $C_{60}H_{122}$, indicating that the GEBF-ML force field is applicable even for larger systems. Based on the GEBF-ML MD trajectory at 300 K, the IR spectra of $C_{40}H_{82}$ and $C_{80}H_{162}$ are also shown in Figure S5. The IR spectra also exhibit all of the typical spectroscopic features for simple hydrocarbons: the C-H scissoring (1505 cm^{-1}), methyl rock (1383 cm^{-1}), long-chain methyl rock

(732 cm^{-1}), and the strong band in the 3100-3000 cm^{-1} region due to the C-H symmetric and asymmetric stretching.

Finally, based on the existing machine learning force field trained by $C_{60}H_{122}$, the acceleration factors *x*2 (the ratio between the number of machine learning predictions and the number of QM calculations) on $C_{40}H_{82}$ and $C_{80}H_{162}$ are much larger than that on $C_{40}H_{82}$ during the on-the-fly force field generation. Although the initial $C_{40}H_{82}$ and $C_{80}H_{162}$ are very different from the $C_{60}H_{122}$, the QM calculations are rarely performed during the MD simulations. The acceleration factor *x*2 for $C_{40}H_{82}$ and $C_{80}H_{162}$ are about 2700 and 1400, respectively, while that for $C_{60}H_{122}$ is about 200 during the on-the-fly force field generation. With the scheme, the machine learning force field will be continuously improved during the MD simulations, so that no QM calculation is required in the final.

4. Conclusion

In summary, we have developed a GEBF-ML method via a Gaussian process to construct machine learning force fields for complex systems without data selection and parameter optimization. Only small subsystems of the target system generated from the GEBF method are used to automatically and efficiently construct the ML force fields of various types of subsystems and the ML force field of the target system. With this approach, long-time GEBF-ML MD simulations were performed on alkanes with different sizes. Our results show that the accuracies of energies, forces, and dipole moments of systems under study predicted with the GEBF-ML method are comparable with those from full QM calculations. Furthermore, infrared spectra of those alkanes could be accurately obtained from the GEBF-ML MD simulations, which are consistent with those from the corresponding direct AIMD or experimental results. Our GEBF-ML scheme provides an automatic and efficient way of building machine learning force fields for a broad range of complex systems such as biomolecules and supramolecular systems.

ASSOCIATED CONTENT

Supporting Information. Computational efficiency, additional ML results, additional MD results, fragmentation scheme, the construction of GEBF subsystems, and full citation of reference 51. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Authors

*Email: wli@nju.edu.cn (W. Li).
*Email: shuhua@nju.edu.cn (S. Li).

ORCID

Dongbo Zhao: 0000-0002-0927-4361
Jing Ma: 0000-0001-5848-9775
Wei Li: 0000-0001-7801-3643
Shuhua Li: 0000-0001-6756-057X

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (Grants Nos. 21833002, 21873046, 21873045, and 21673110) and China Postdoctoral Science Foundation (Grant No. 2019M651773). Part of the calculations were performed using computational resources on an IBM Blade cluster system from the High Performance Computing Center (HPCC) of Nanjing University.

REFERENCES

(1) Li, S.; Li, W.; Ma, J. Generalized Energy-Based Fragmentation Approach and Its Applications to Macromolecules and Molecular Aggregates. *Acc. Chem. Res*. **2014**, *47*, 2712−2720.

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

(3) Ganesh, V.; Dongare, R. K.; Balanarayan, P.; Gadre, S. R. Molecular Tailoring Approach for Geometry Optimization of Large Molecules: Energy Evaluation and Parallelization Strategies. *J. Chem. Phys.* **2006**, *125*, 104109.

(4) Dahlke, E. E.; Truhlar, D. G. Electrostatically Embedded Many-Body Expansion for Large Systems, with Applications to Water Clusters. *J. Chem. Theory Comput*. **2007**, *3*, 46−53.

(5) Gordon, M. S.; Fedorov, D. G.; Pruitt, S. R.; Slipchenko, L. V. Fragmentation Methods: A Route to Accurate Calculations on Large Systems. *Chem. Rev*. **2012**, *112*, 632−672.

(6) He, X; Zhang, J. Z. H. The generalized molecular fractionation with conjugate caps/molecular mechanics method for direct calculation of protein energy. *J. Chem. Phys*. **2006**, *124*, 184703.

(7) Battens, R. P. A.; Lee, A. M. A New Algorithm for Molecular Fragmentation in Quantum Chemical Calculations. *J. Phys. Chem. A* **2006**, *110*, 8777–8785.

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

(9) Richard, R. M.; Herbert, J. M. A generalized many-body expansion and a unified view of fragment-based methods in electronic structure theory. *J. Chem. Phys.* **2012**, *137*, 064113.

(10) Mayhall, N. J.; Raghavachari, K. Many-Overlapping-Body (MOB) Expansion: A Generalized Many Body Expansion for Nondisjoint Monomers in Molecular Fragmentation Calculations of Covalent Molecules. *J. Chem. Theory Comput*. **2012**, *8*, 2669–2675.

(11) Hu, D.; Xie, Y.; Li, X.; Li, L.; Lan, Z. Inclusion of Machine Learning Kernel Ridge Regression Potential Energy Surfaces in On-the-Fly Nonadiabatic Molecular Dynamics Simulation. *J. Phys. Chem. Lett*. **2018**, *9*, 2725-2732.

(12) Faber, F. A.; Hutchison, L.; Huang, B.; Gilmer, J.; Schoenholz, S. S.; Dahl, G. E.; Vinyals, O.; Kearnes, S.; Riley, P. F.; von Lilienfeld, O. A. Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error. *J. Chem. Theory Comput*. **2017**, *13*, 5255−5264.

(13) Wei, J. N.; Duvenaud, D.; Aspuru-Guzik, A. Neural Networks for the Prediction of Organic Chemistry Reactions. *ACS Cent. Sci*. **2016**, *2*, 725−732.

(14) Gómez-Bombarelli, R.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Duvenaud, D.; Maclaurin, D.; Blood-Forsythe, M. A.; Chae, H. S.; Einzinger, M.; Ha, D.-G.; Wu, T.; Markopoulos, G.; Jeon, S.; Kang, H.; Miyazaki, H.; Numata, M.; Kim, S.; Huang, W.; Hong, S. I.; Baldo, M.; Adams, R. P.; Aspuru-Guzik, A. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nat. Mater*. **2016**, *15*, 1120−1127.

(15) Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett*. **2007**, *98*, 146401.

(16) Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons. *Phys. Rev. Lett*. **2010**, *104*, 136403.

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

(18) Morawietz, T.; Behler, J. A Density-Functional Theory-Based Neural Network Potential for Water Clusters Including van der Waals Corrections. *J. Phys. Chem. A* **2013**, *117*, 7356–7366.

(19) Chiriki, S.; Bulusu, S. S. Modeling of DFT quality neural network potential for sodium clusters: Application to melting of sodium clusters (Na_{20} to Na_{40}). *Chem. Phys. Lett*. **2016**, *652*, 130–135.

(20) Chiriki, S.; Jindal, S.; Bulusu, S. S. Neural network potentials for dynamics and thermodynamics of gold nanoparticles. *J. Chem. Phys.* **2017**, *146*, 084314.

(21) Szlachta, W. J.; Bartók, A. P.; Csányi, G. Accuracy and transferability of Gaussian approximation potential models for tungsten. *Phys. Rev. B* **2014**, *90*, 104108.

(22) Dragoni, D.; Daff, T. D.; Csányi, G.; Marzari, N. Achieving DFT accuracy with a machine-learning interatomic potential: Thermomechanics and defects in bcc ferromagnetic iron. *Phys. Rev. M* **2018**, *2*, 013808.

(23) Behler, J.; Martoňák, R.; Donadio, D.; Parrinello, M. Metadynamics Simulations of the High-Pressure Phases of Silicon Employing a High-Dimensional Neural Network Potential. *Phys. Rev. Lett.* **2008**, *100*, 185501.

(24) Bartók, A. P.; Kermode, J.; Bernstein, N.; Csányi, G. Machine Learning a General-Purpose Interatomic Potential for Silicon. *Phys. Rev. X* **2018**, *8*, 041048.

(25) Artrith, N.; Behler, J. High-dimensional neural network potentials for metal surfaces: A prototype study for copper. *Phys. Rev. B* **2008**, *85*, 045439.

(26) Morawietz, T.; Singraber, A.; Dellago, C.; Behler, J. How van der Waals interactions determine the unique properties of water. *Proc. Natl. Acad. Sci. U.S.A.* **2016**, *113*, 8368–8373.

(27) Cheng, B.; Behler, J.; Ceriotti, M. Nuclear Quantum Effects in Water at the Triple Point: Using Theory as a Link Between Experiments. *J. Phys. Chem. Lett.* **2016**, *7*, 2210–2215.

(28) Veit, M.; Jain, S. K.; Bonakala, S.; Rudra, I.; Hohl, D.; Csányi, G. Equation of State of Fluid Methane from First Principles with Machine Learning Potentials. *J. Chem. Theory Comput.* **2019**, *15*, 2574–2586.

(29) Hellström, M.; Behler, J. Structure of aqueous NaOH solutions: insights from neural-network-based molecular dynamics simulations. *Phys. Chem. Chem. Phys.* **2017**, *19*, 82–96.

(30) Natarajan, S. K.; Behler, J. Neural network molecular dynamics simulations of solid-liquid interfaces: water at low-index copper surfaces. *Phys. Chem. Chem. Phys.* **2016**, *18*, 28704–28725.

(31) Yao, K.; Herr, J. E.; Parkhill, J. The Many-Body Expansion Combined with Neural Networks. *J. Chem. Phys.* **2017**, *146*, 014106.

(32) Wang, H.; Yang, W. Force Field for Water Based on Neural Network. *J. Phys. Chem. Lett.* **2018**, *9*, 3232–3240.

(33) Wang, H.; Yang, W. Toward Building Protein Force Fields by Residue-Based Systematic Molecular Fragmentation and Neural Network. *J. Chem. Theory Comput.* **2019**, *15*, 1409–1417.

(34) Wang, Z.; Han, Y.; Li, J.; He, X. Combining the Fragmentation Approach and Neural Network Potential Energy Surfaces of Fragments for Accurate Calculation of Protein Energy. *J. Phys. Chem. B* **2020**, *124*, 3027-3035.

(35) Chen, W.; Fang, W.; Cui, G. Integrating Machine Learning with the Multilayer Energy-Based Fragment Method for Excited States of Large Systems. *J. Phys. Chem. Lett.* **2019**, *10*, 7836–7841.

(36) Jinnouchi, R. Karsai, F.; Kresse, G. On-the-fly machine learning force field generation: Application to melting points. *Phys. Rev. B* **2019**, *100*, 014105.

(37) Li, W.; Li, S.; Jiang, Y. Generalized Energy-Based Fragmentation Approach for Computing the Ground-State Energies and Properties of Large Molecules. *J. Phys. Chem. A* **2007**, *111*, 2193–2199.

(38) Jinnouchi, R.; Lahnsteiner, J.; Karsai, F.; Kresse, G.; Bokdam, M. Phase Transitions of Hybrid Perovskites Simulated by Machine-Learning Force Fields Trained on the Fly with Bayesian Inference. *Phys. Rev. Lett.* **2019**, *122*, 225701.

(39) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Phys. Rev. B* **2013**, *87*, 184115.

(40) Bartók, A. P.; Csányi, G. Gaussian approximation potentials: A brief tutorial introduction. *Int. J. Quantum Chem.* **2015**, *115*, 1051–1057.

(41) MacKay, D. J. C. Bayesian interpolation. *Neural Comput.* **1992**, *4*, 415–447.

(42) Jinnouchi, R.; Asahi, R. Predicting catalytic activity of nanoparticles by a DFT-aided machine-learning algorithm. *J. Phys. Chem. Lett.* **2017**, *8*, 4279–4283.

(43) Li. W.; Chen, C.; Zhao, D.; Li, S. LSQC: Low scaling quantum chemistry program. *Int. J. Quantum Chem.* **2015**, *115*, 641–646.

(44) Thomas, M.; Brehm, M.; Fligg, R.; Vöhringer, P. Kirchner, B. Computing vibrational spectra from ab initio molecular dynamics. *Phys. Chem. Chem. Phys.* **2013**, *15*, 6608–6622.

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

(46) Sifain, A. E.; Lubbers, N.; Nebgen, B. T.; Smith, J. S.; Lokhov, A. Y.; Isayev, O.; Roitberg, A. E.; Barrors. K.; Tretiak, S. Discovering a Transferable Charge Assignment Model Using Machine Learning. *J. Phys. Chem. Lett.* **2018**, *9*, 4495–4501.

(47) Unke, O. T.; Meuwly, M. PhysNet: A Neural Network for Predicting Energies, Forces, Dipole Moments, and Partial Charges. *J. Chem. Theory Comput.* **2019**, *15*, 3678–3693.

(48) Wang, K.; Li, W.; Li, S. Generalized energy-based fragmentation CCSD(T)-F12a method and application to the relative energies of water clusters $(H_2O)_{20}$. *J. Chem. Theory Comput*. **2014**, *10*, 1546–1553.

(49) Zhang, L.; Li, W.; Fang, T.; Li, S. Accurate Relative Energies and Binding Energies of Large Ice-Liquid Water Clusters and Periodic Structures. *J. Phys. Chem. A* **2017**, *121*, 4030–4038.

(50) Yuan, D.; Li, Y.; Ni, Z.; Paulay, P.; Li, W.; Li, S. Benchmark Relative Energies for Large Water Clusters with the Generalized Energy-Based Fragmentation Method. *J. Chem. Theory Comput.* **2017**, *13*, 2696–2704.

(51) Frisch, M. J. et al. Gaussian 16, Revision A.03; Gaussian Inc.: Wallingford, CT, 2016.

(52) Biswas, R.; Hamann, D. R. Simulated annealing of silicon atom clusters in Langevin molecular dynamics. *Phys. Rev. B* **1986**, *34*, 895–901.

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

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

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

(56) Zhang, L.; Li, W.; Fang, T.; Li, S. Ab initio molecular dynamics with intramolecular noncovalent interactions for unsolvated polypeptides. *Theor. Chem. Acc.* **2016**, *135*, 34.

(57) Fischer, S. A.; Ueltschi, T. W.; El-Khoury, P. Z.; Mifflin, A. L.; Hess, W. P.; Wang, H.; Cramer, C. J.; Govind, N. Infrared and Raman spectroscopy from ab initio molecular dynamics and static normal mode analysis: The C-H region of DMSO as a case study. *J. Phys. Chem. B* **2016**, *120*, 1429–1436.

Supporting information for:

An On-the-fly Approach to Construct Generalized Energy-Based Fragmentation Machine Learning Force Fields of Complex Systems

Zheng Cheng[†], Dongbo Zhao[†,‡], Jing Ma[†], Wei Li[†,*], and Shuhua Li[†,*]

[†]Institute of Theoretical and Computational Chemistry, School of Chemistry and Chemical Engineering, Nanjing University, Nanjing 210023, People's Republic of China

[‡]Kuang Yaming Honors School, Nanjing University, 210023, People's Republic of China

E-mail: wli@nju.edu.cn; shuhua@nju.eud.cn

Contents

S1 Computational efficiency

The data showing the computational efficiency of the on-the-fly force field generation is summarized in Table S1. The numbers of structures providing the reference structure datasets and the numbers of the local reference configurations are listed in Table S2.

Table S1. Fraction x_1 (%) of the MD steps, where QM calculations were bypassed, and acceleration factor x_2 of the 50-ps MD simulation by the on-the-fly scheme on each type of subsystem.

subsystem	x_1	x_2
C_4H_{10}	98.2	57
C_8H_{18}	99.6	240
$(C_4H_{10})_2$	97.8	45
$C_{12}H_{26}$	99.6	250
C_8H_{18}+C_4H_{10}	98.9	92
$C_{16}H_{34}$	99.3	148
$C_{12}H_{26}$+C_4H_{10}	99.3	138

Table S2. The numbers of structures providing the reference structure datasets (N_{st}) and the numbers of local reference configurations (N_B).

subsystem	N_{st}	$N_{\boldsymbol{B}}$
C_4H_{10}	278	461
C_8H_{18}	1351	1599
$(C_4H_{10})_2$	299	996
$C_{12}H_{26}$	1179	1578
C_8H_{18}+C_4H_{10}	406	2197
$C_{16}H_{34}$	412	1413
$C_{12}H_{26}$+C_4H_{10}	429	1497

S2 Additional ML Results

The ML force fields are accurate for different size of alkanes. Parameters in descriptor are collected in Table S3. Table S4 collects root mean squared errors (RMSEs) of the energies and forces of $C_{40}H_{82}$, $C_{60}H_{122}$, and $C_{80}H_{162}$ (relative to the ωB97X-D reference data). The distributions of errors between the ML force field and the ωB97X-D reference method are displayed in Figure S1.

Table S3. Parameters in descriptor

subsystem	R_c	σ_{atom}	ζ	N_R^l	$L_{\max}$
C_4H_{10}	2.5	0.4	2.0	6	4
C8H18	3.0	0.4	2.0	6	5
(C4H10)2	2.5	0.4	2.0	6	5
C12H26	3.0	0.4	3.0	6	5
C8H18+C4H10	3.0	0.35	2.5	6	6
C16H34	2.7	0.4	2.5	6	5
C12H26+C4H10	3.0	0.4	3.0	6	6

Table S4. The RMSEs of the energies (E) [in kcal/(mol·atom)] and forces (F) [in kcal/(mol·Å)] (relative to the conventional DFT results) on the test set.

system	RMSE E	RMSE F
$C_{40}H_{82}$	0.040	2.84
$C_{60}H_{122}$	0.033	2.37
$C_{80}H_{162}$	0.031	2.61

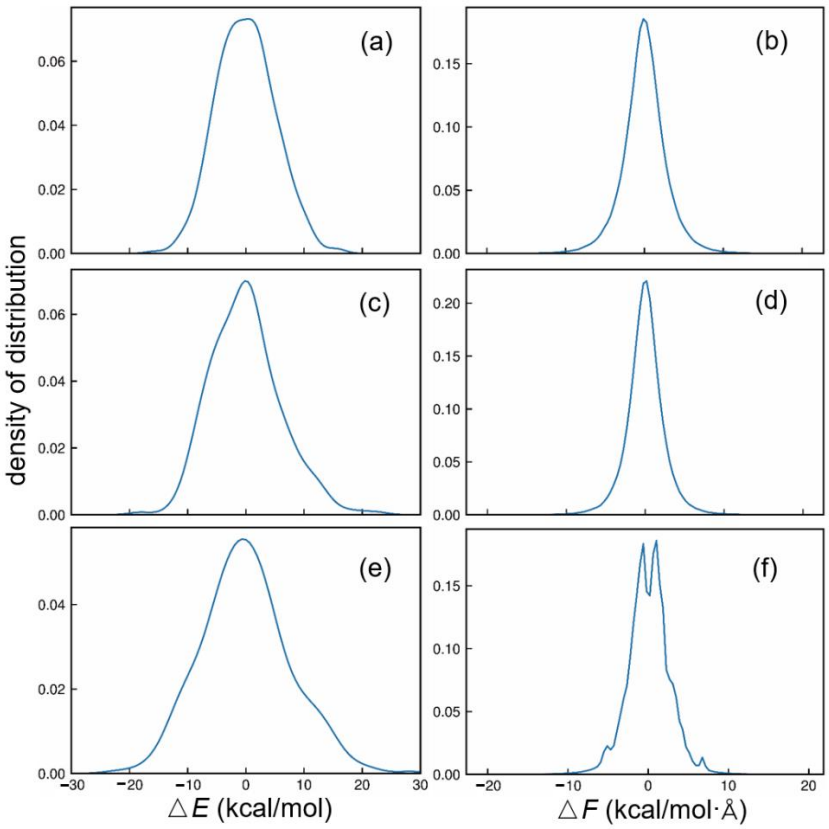

Figure S1. Distributions of the energy and force errors between the ML force field and the ωB97X-D reference using three test data sets: $C_{40}H_{82}$, $C_{60}H_{122}$, and $C_{80}H_{162}$ (from top to bottom). Left and right panels show the distributions of the errors of energies and forces, respectively.

S3 Additional MD Results

The time evolutions of the energies of $C_{40}H_{82}$, $C_{60}H_{122}$, and $C_{80}H_{162}$ at the microcanonical (NVE) ensemble are shown in Figure S2. The conformational changes and RMSDs (with respect to their initial structures) of $C_{40}H_{82}$ and $C_{80}H_{162}$ at the NVT ensemble are shown in Figures S3 and S4, respectively. The IR spectra of $C_{40}H_{82}$ and $C_{80}H_{162}$ are displayed in Figure S5.

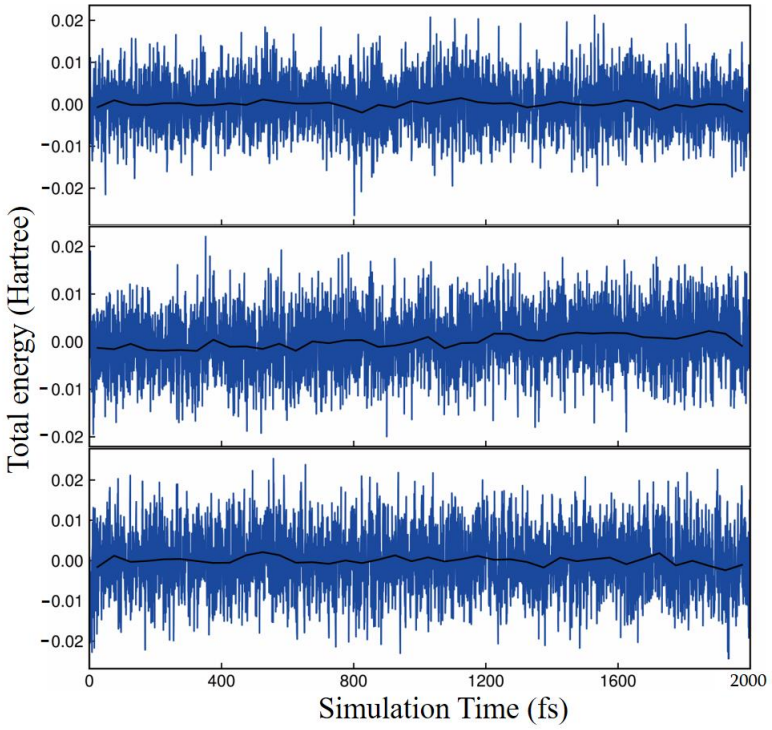

Figure S2. The total energy changes as functions of time in GEBF-ML MD simulations of (a) $C_{40}H_{82}$, (b) $C_{60}H_{122}$, and (c) $C_{80}H_{162}$ at the NVE ensemble with a time step of 0.5 fs. Black line denotes the average of total energy per 50 fs. The zero energy is chosen to be the average energy of total energy. The GEBF-ML MD simulations were performed using the force field without any retraining.

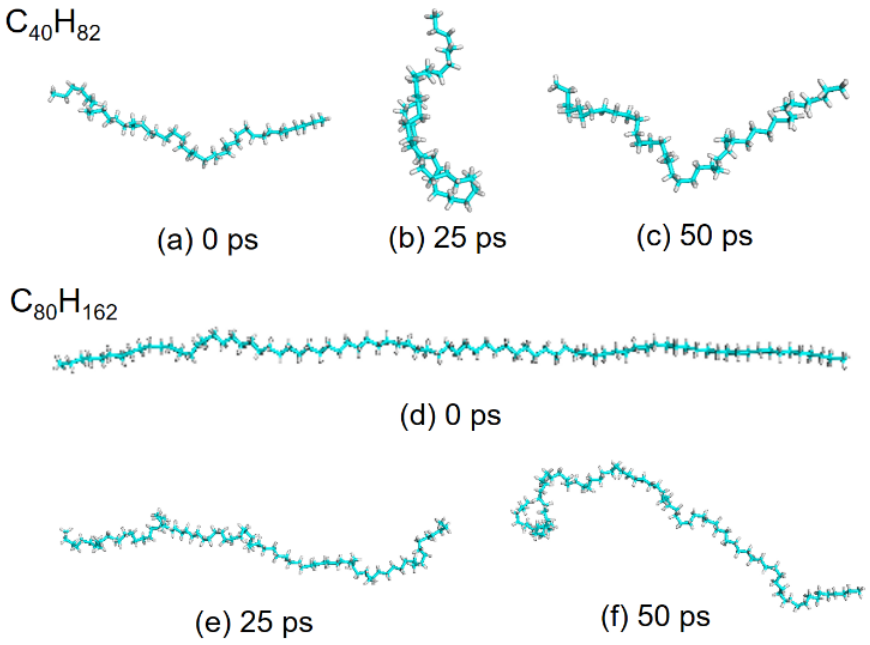

Figure S3. The conformational changes for $C_{40}H_{82}$ and $C_{80}H_{162}$ during the 50-ps GEBF-ML MD simulations.

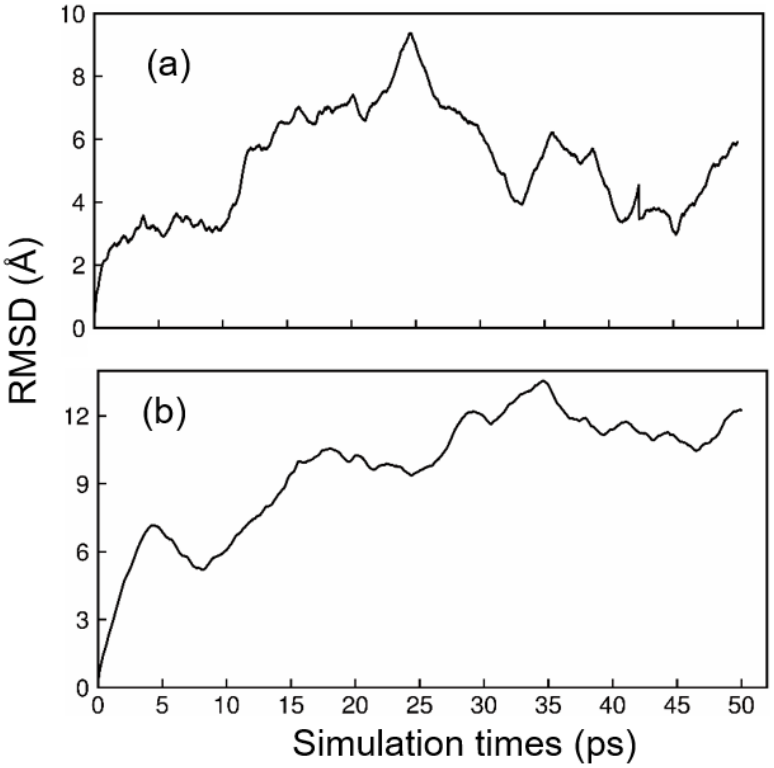

Figure S4. The RMSDs with respect to the initial structures of (a) $C_{40}H_{82}$ and (b) $C_{80}H_{162}$ in the GEBF-ML MD simulations.

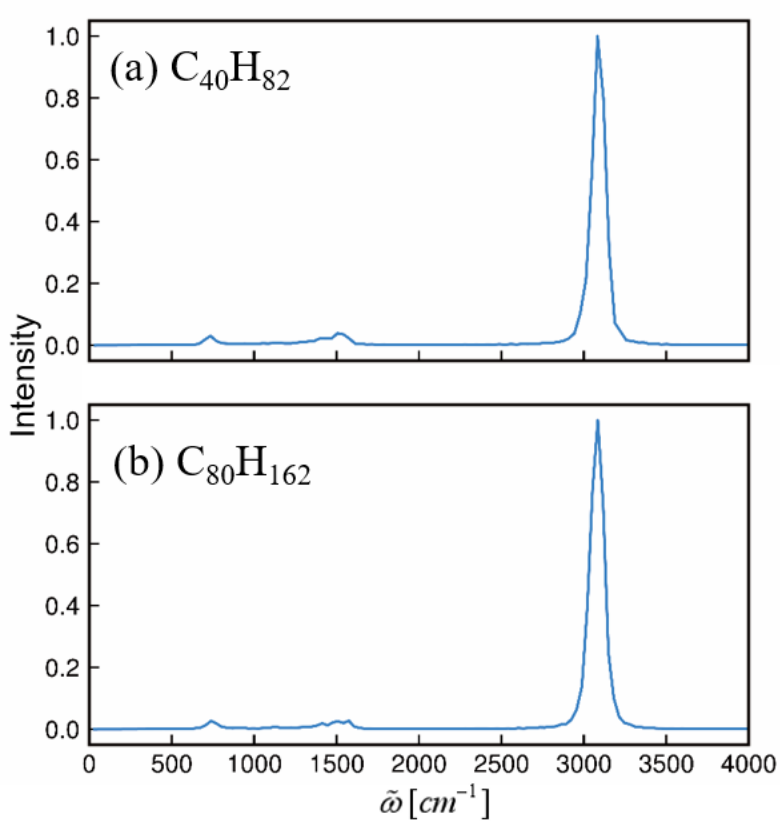

Figure S5. IR spectra of (a) $C_{40}H_{82}$ and (b) $C_{80}H_{162}$ predicted by the ML model.

S4 Fragmentation scheme and the construction of GEBF subsystems

The main procedures in a GEBF calculation in this work are summarized as follows. (1) Divide total system into various fragments. (2) For each fragment, construct a primitive subsystem by adding its neighboring environmental fragments within a distance threshold ζ and limit the maximum number of environmental fragments as η. Hydrogen atoms are added to subsystems for valence saturation to avoid dangling bonds. (3) Once all primitive subsystems are obtained, derivative subsystems with their coefficients are constructed with the inclusion-exclusion principle, to cancel the overlapping of primitive subsystems. The GEBF calculation is denoted as GEBF(ζ, η). Here, $C_{20}H_{42}$ was used as an example to illustrate our fragmentation scheme and the construction of GEBF subsystems.

(1) As shown in Figure S6(a), the $C_{20}H_{42}$ was first divided into five fragments, the box model of the fragmentation scheme is shown in Figure S6(b).

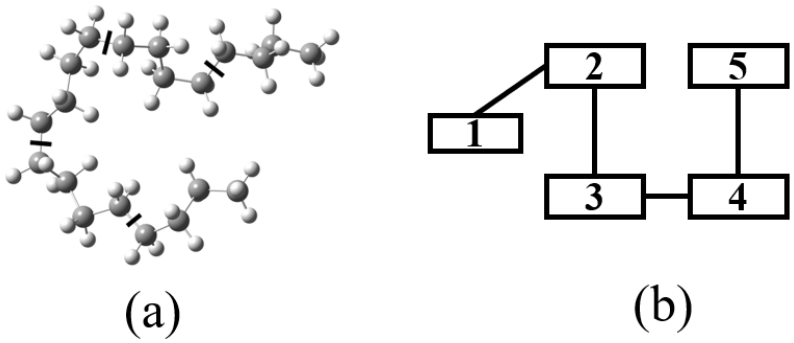

Figure S6. Fragmentation scheme of $C_{20}H_{42}$: (a) molecular structure of $C_{20}H_{42}$ and four C-C bonds (denoted in solid line) are cut to generate five fragments; (b) box model of five fragments. The solid lines represent covalent single bonds.

(2) For each fragment (denoted as central fragment), several neighboring (environmental) fragments were added to construct its primitive subsystem with ξ and η being 3.0 Å and 4, respectively. Hydrogen atoms are added for valence saturation. All the primitive subsystems are listed in Figure S7.

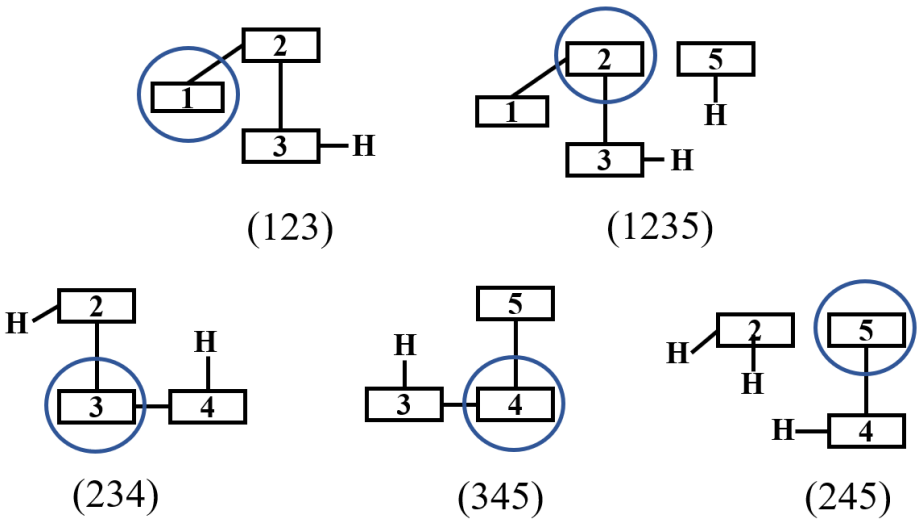

Figure S7. Primitive subsystems of the $C_{20}H_{42}$, each of which contains a central fragment (inside the circle) and its environmental fragments. The fragment indices in each subsystem are listed in parentheses.

(3) Delete the redundant small primitive subsystems, which are included in larger ones. For the $C_{20}H_{42}$, subsystem (123) is deleted as it is included in subsystem (1235). The retained primitive subsystems and their coefficients are shown in Figure S8.

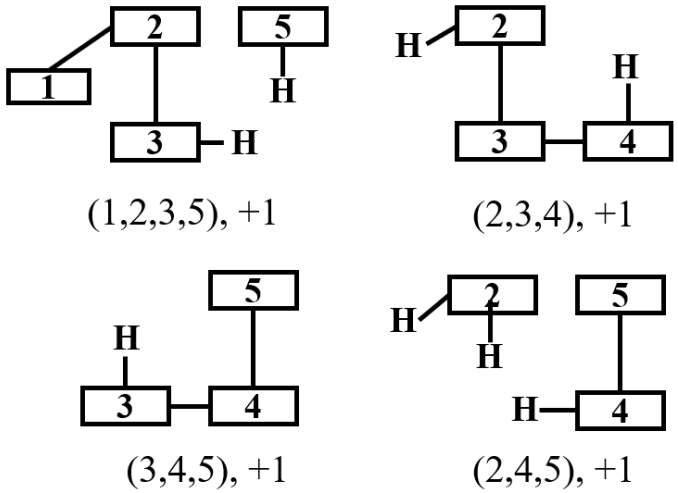

Figure S8. The retained primitive subsystems. Fragment indices in each subsystem are listed in parentheses, and the coefficients are denoted after the parentheses.

(4) 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 Figure S9.

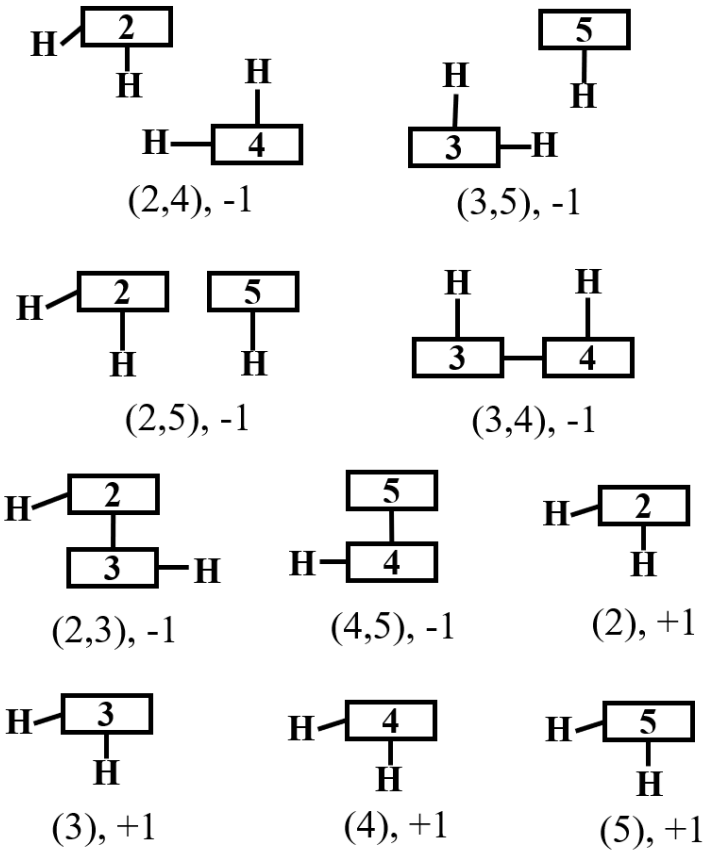

Figure S9. The derivative subsystems generated with the inclusion-exclusion principle. Fragment indices in each subsystem are listed in parentheses, and the coefficients are denoted after the parentheses.

In this work, ten $C_{60}H_{122}$ conformers are calculated with GEBF(3,4)-ωB97X-D, with their errors respective to the conventional ωB97X-D results collected in Table S5. The mean absolute error (MAE) is only 0.0025 kcal/(mol atom) and 0.01 kcal/(mol ·Å) for energy and forces, respectively. Thus, the parameters ξ and η in GEBF-ML method are chosen to be 3.0 Å and 4, respectively.

Table S5. Deviation of the energies (E) [in kcal/(mol atom)] and mean absolute error (MAE) of the forces (F) [in kcal/(mol ·Å)] from GEBF(3,4)-ωB97X-D computations for ten $C_{60}H_{122}$ conformers at the 6-31G** basis set level, compared to those obtained from conventional QM methods.

conformer	E	F
1	-0.0013	0.0093
2	-0.0006	0.0148
3	-0.0011	0.0074
4	-0.0020	0.0101
5	-0.0026	0.0103
6	-0.0035	0.0105
7	-0.0034	0.0080
8	-0.0016	0.0070
9	-0.0040	0.0100
10	-0.0044	0.0117
MAE	0.0025	0.0100

Full reference 51: Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. Gaussian 16, Revision A.03; Gaussian Inc.: Wallingford, CT, 2016.