Internal force corrections with machine learning for quantum mechanics/molecular mechanics simulations

Jingheng Wu, Lin Shen, and Weitao Yang

Citation: The Journal of Chemical Physics **147**, 161732 (2017);
View online: https://doi.org/10.1063/1.5006882
View Table of Contents: http://aip.scitation.org/toc/jcp/147/16
Published by the American Institute of Physics

Internal force corrections with machine learning for quantum mechanics/molecular mechanics simulations

Jingheng Wu,[1,2] Lin Shen,[1] and Weitao Yang[1,a)]
[1]*Department of Chemistry, Duke University, Durham, North Carolina 27708, USA*
[2]*School of Chemistry and Chemical Engineering, Sun Yat-sen University, Guangzhou 510275, People's Republic of China*

(Received 1 May 2017; accepted 27 September 2017; published online 12 October 2017)

Ab initio quantum mechanics/molecular mechanics (QM/MM) molecular dynamics simulation is a useful tool to calculate thermodynamic properties such as potential of mean force for chemical reactions but intensely time consuming. In this paper, we developed a new method using the internal force correction for low-level semiempirical QM/MM molecular dynamics samplings with a predefined reaction coordinate. As a correction term, the internal force was predicted with a machine learning scheme, which provides a sophisticated force field, and added to the atomic forces on the reaction coordinate related atoms at each integration step. We applied this method to two reactions in aqueous solution and reproduced potentials of mean force at the *ab initio* QM/MM level. The saving in computational cost is about 2 orders of magnitude. The present work reveals great potentials for machine learning in QM/MM simulations to study complex chemical processes. *Published by AIP Publishing.*
https://doi.org/10.1063/1.5006882

INTRODUCTION

The free energy change along the reaction coordinate (RC), or the potential of mean force (PMF), is a central quantity for describing the pathways and barriers of a chemical reaction.[1–5] Based on chemical intuition, one or a set of collective variables of the molecular structure is often defined as RC to capture the major geometric change in a chemical process. The PMF can be calculated from molecular dynamics (MD) simulations. In order to cross high-energy barriers efficiently, some RC-guided enhanced sampling techniques such as umbrella sampling[6,7] are necessary. However, MD simulations with the *ab initio* quantum mechanical or QM/MM method are still limited to relatively small molecules because of a large computational cost on electronic structure calculations.[8–12] A reparameterization or correction of a semiempirical QM (SQM) model to *ab initio* QM is an attractive way to achieve a higher accuracy with an affordable computational cost. One example is the force matching method, in which some specific parameters in the low-level model were modified with a fitness function to match the high-level calculations on the selected configurations along the reaction path.[13–15] Ruiz-Pernía *et al.* developed another dual-level QM/MM method in which the difference between SQM/MM and *ab initio* QM/MM potential energies was approximated as a spline interpolation function of RC.[16,17] These methods are as efficient as SQM/MM since only a few configurations are required for *ab initio* QM/MM calculations. The applications to complex reactions in the condensed phase and biochemical environment revealed their success, but there are still some nontrivial concerns in practice. First, an explicit fitness function has to be designed to map the correction term for each system, which restricts the accuracy and availability for these approaches. Second, the selection on the configurations for *ab initio* calculations is limited to the reaction path along RC. It may be largely affected by the fluctuation of MM environment, which is not easy to capture because of the complexity of a QM/MM energy landscape.

Machine learning (ML) is being increasingly used in the past decades to save the computational cost and overcome the limitation on conventional MD simulations. Botu and Ramprasad proposed an adaptive scheme combined with "big-data" techniques to predict atomic forces and accelerate materials simulations.[18] Li *et al.* developed a machine learning approach to perform QM-accurate MD simulations with a growing database.[19] Stecher *et al.* applied the Gaussian process regression to construct free energy surfaces after umbrella sampling simulations.[20] Recently, we developed a QM/MM neural network method to predict the SQM/MM and *ab initio* QM/MM potential energy differences based on the low-level MD simulations, and then the high-level PMF was produced using a reweighting scheme,[21] achieving a significant reduction of computational effort in free energy prediction compared with direct *ab initio* QM/MM calculations. However, a direct QM/MM MD algorithm guided with machine learning is still absent. In this paper, we present a machine learning method to predict the *ab initio* QM/MM forces with the information on RC. First, internal forces are introduced to correct the difference of atomic forces between SQM/MM and *ab initio* QM/MM models along RC directions. Their values are predicted using the k-nearest neighbor (kNN) algorithm[22] for any configuration during QM/MM MD. Second, MD simulations are performed using SQM/MM forces with the correction term, both of which can be obtained at a much lower computational cost than *ab initio* calculations. Finally, a highly

a)Author to whom correspondence should be addressed: weitao.yang@duke.edu

accurate PMF is achieved based on MD samplings with ML corrections.

The rest of this paper is organized as follows. We first describe the construction on the ML database. The central question is how to calculate the internal force from SQM/MM and *ab initio* QM/MM atomic forces. Then we will use the machine learning scheme to predict the high-level atomic forces for any configuration during SQM/MM MD and correct the low-level MD evolution, followed by computational validation, discussion, and conclusion.

THEORETICAL FOUNDATION

The key point of our method is that the dimensionality of machine learning can be reduced with some prior knowledge of reaction coordinate. Here we focus our discussions on the systems where a combination of bond lengths on the QM subsystem is defined as RC. Only the forces acting on the QM atoms involved in RC are corrected using ML, and the correction terms on the internal forces are added only along the direction of the corresponding bond length in RC. The deviation of SQM/MM forces from *ab initio* QM/MM forces should be reduced significantly with ML corrections. Consider RC as the distance between atoms i and j. Its value at time t, $r_{ij}(t)$, can be written as

$$r_{ij}(t)\,\hat{\mathbf{r}}_{ij}(t) = \mathbf{r}_j(t) - \mathbf{r}_i(t), \tag{1}$$

where $\mathbf{r}_i(t)$ and $\mathbf{r}_j(t)$ are the Cartesian coordinates of atoms i and j at time t, respectively, and $\hat{\mathbf{r}}_{ij}$ is the unit vector along the direction from i to j. The Cartesian coordinate and velocity of atom i at the next MD step can be expressed based on the common used velocity Verlet integration schemes[23] as follows:

$$\mathbf{r}_i(t+\Delta t) = \mathbf{r}_i(t) + \mathbf{v}_i(t)\,\Delta t + \frac{\mathbf{F}_i(t)}{2m_i}\Delta t^2 \tag{2}$$

and

$$\mathbf{v}_i(t+\Delta t) = \mathbf{v}_i(t) + \frac{\mathbf{F}_i(t) + \mathbf{F}_i(t+\Delta t)}{2m_i}\Delta t, \tag{3}$$

where Δt is the integration time step, m_i is the mass of atom i, and $\mathbf{v}_i(t)$ and $\mathbf{F}_i(t)$ are the velocity and force of atom i at time t, respectively. Given the same coordinates and velocities at time t, the deviation of $\mathbf{r}_i(t+\Delta t)$ between different potential energies (e.g., SQM/MM and *ab initio* QM/MM) only originates from the difference of $\mathbf{F}_i(t)$, leading to

$$\begin{aligned}&\left[\mathbf{r}_j^H(t+\Delta t) - \mathbf{r}_j^{LC}(t+\Delta t)\right] - \left[\mathbf{r}_i^H(t+\Delta t) - \mathbf{r}_i^{LC}(t+\Delta t)\right]\\ &\quad= \frac{\Delta t^2}{2m_j}\left[\mathbf{F}_j^H(t) - \mathbf{F}_j^L(t) - \mathbf{F}_j^{corr}(t)\right]\\ &\qquad- \frac{\Delta t^2}{2m_i}\left[\mathbf{F}_i^H(t) - \mathbf{F}_i^L(t) - \mathbf{F}_i^{corr}(t)\right].\end{aligned} \tag{4}$$

Here $\mathbf{r}_i^H(t+\Delta t)$ and $\mathbf{r}_i^{LC}(t+\Delta t)$ are the Cartesian coordinates of atom i at time $t+\Delta t$. The former one is calculated from $\mathbf{F}^H(t)$, and the latter one is calculated from $\mathbf{F}^L(t) + \mathbf{F}^{corr}(t)$. $\mathbf{F}_i^H(t)$ and $\mathbf{F}_i^L(t)$ are the high-level and low-level forces on atom i at time t, respectively, and $\mathbf{F}_i^{corr}(t)$ is the internal force correction on atom i at time t. The variables with the subscript j denote the corresponding coordinates or forces of atom j. $\mathbf{F}_i^{corr}(t)$ and $\mathbf{F}_j^{corr}(t)$ are added along the direction of $\hat{\mathbf{r}}_{ij}$ with the magnitudes as $F_i^{corr}(t)$ and $F_j^{corr}(t)$, respectively. They should be constrained by the condition

$$F_i^{corr}(t) + F_j^{corr}(t) = 0, \tag{5}$$

which ensures that the total correction force of the whole system is zero at any time. Note that we cannot obtain any prior knowledge about $\hat{\mathbf{r}}_{ij}(t+\Delta t)$ at time t, which means that we have to introduce an approximation as $\hat{\mathbf{r}}_{ij}(t+\Delta t) \approx \hat{\mathbf{r}}_{ij}(t)$. Both sides of Eq. (4) are projected onto $\hat{\mathbf{r}}_{ij}$ under this approximation, leading to

$$\begin{aligned}&\frac{2}{\Delta t^2}\left[r_{ij}^H(t+\Delta t) - r_{ij}^{LC}(t+\Delta t)\right]\\ &\quad\approx \frac{\left[\mathbf{F}_j^H(t) - \mathbf{F}_j^L(t)\right]\cdot\hat{\mathbf{r}}_{ij}(t) - F_j^{corr}(t)}{m_j}\\ &\qquad- \frac{\left[\mathbf{F}_i^H(t) - \mathbf{F}_i^L(t)\right]\cdot\hat{\mathbf{r}}_{ij}(t) - F_i^{corr}(t)}{m_i},\end{aligned} \tag{6}$$

where $r_{ij}^H(t+\Delta t)$ and $r_{ij}^{LC}(t+\Delta t)$ are the distances between atoms i and j at time $t+\Delta t$. The former is calculated from $\mathbf{F}^H(t)$, and the latter is calculated from $\mathbf{F}^L(t) + \mathbf{F}^{corr}(t)$. In order to get the same values of $r_{ij}^H(t+\Delta t)$ and $r_{ij}^{LC}(t+\Delta t)$, we have the relationship between the internal force corrections and the atomic forces calculated at two levels as

$$\begin{aligned}&\frac{\left[\mathbf{F}_j^H(t) - \mathbf{F}_j^L(t)\right]\cdot\hat{\mathbf{r}}_{ij}(t) - F_j^{corr}(t)}{m_j}\\ &\quad- \frac{\left[\mathbf{F}_i^H(t) - \mathbf{F}_i^L(t)\right]\cdot\hat{\mathbf{r}}_{ij}(t) - F_i^{corr}(t)}{m_i} = 0.\end{aligned} \tag{7}$$

Herein the values of the correction forces can be solved using the low-level and high-level atomic forces through Eqs. (5) and (7). Similarly, the internal force corrections at time $t+\Delta t$ can be derived from Eq. (3) in order to get the same values of $v_{ij}^H(t+\Delta t)$ and $v_{ij}^{LC}(t+\Delta t)$, leading to the same expressions as Eqs. (5) and (7). The projections of $\mathbf{r}^H$ and $\mathbf{r}^{LC}$ (and $\mathbf{v}^H$ and $\mathbf{v}^{LC}$) onto the direction of RC are expected to keep identical at sequential MD steps, so the internal force corrections at any time can be obtained recursively in the same way. The above derivation can be extended directly to any linear combination of bond lengths involving more than two atoms.

One motivation of machine learning is to reduce the expensive computational cost on the high-level forces in Eq. (7). Next we propose an ML scheme to predict the internal force corrections without *ab initio* calculations. For example, QM atom i is included in RC as a RC-related atom. First, its chemical environment is represented by the neighboring QM atoms within a cut-off distance r_{cut}. In other words, the position of atom i can be constructed with the positions of its neighbors, and the construction weight $w_j^{(i)}$ summarizes the contribution from atom j. Similar to the scheme used in the locally linear embedding algorithm,[24] the construction error ε_i is measured as

$$\varepsilon_i = |\mathbf{r}_i - \sum_{j\neq i} w_j^{(i)}\mathbf{r}_j|^2, \tag{8}$$

which is minimized with the constraint of $\sum_j w_j^{(i)} = 1$ to obtain the optimal weights as the input features. The weights are

further scaled to reduce the contribution from distant atoms with a cutoff function as

$$f_{\text{cut}}(r_{ij}) = \begin{cases} 1 - 2\left[e^{-r_{\text{cut}}(r_{ij}-r_{\text{cut}})} + 1\right]^{-1} & r_{ij} \le r_{\text{cut}} \\ 0 & r_{ij} > r_{\text{cut}} \end{cases}. \quad (9)$$

Second, the Mulliken atomic charges of the RC-related atoms, which have been calculated with the low-level electrostatic-embedding QM/MM model, are employed as another input feature to include the contributions from MM environment. As discussed in our previous work,[21] Mulliken atomic charges reflect external potentials of the MM environment with reduced degrees of freedom as small as the number of QM atoms. This feature also captures the polarization of the QM subsystem in response to MM electrostatic potentials. The input vector for atom i is thus expressed as

$$\mathbf{v}^{(i)} = \left(w_1^{(i)} f_{\text{cut}}(r_{i1}), \ \ldots, \ w_M^{(i)} f_{\text{cut}}(r_{iM}), \ q_i\right), \quad (10)$$

where q_i is the Mulliken charge of atom i, and M is the number of the neighbors of atom i that depends on the cutoff distance.

The k-nearest neighbor algorithm is utilized in this work. The predictions on the correction forces acting on different atoms are independent. For example, the correction term $\mathbf{F}_i^{corr}$ is only dependent on $\mathbf{v}^{(i)}$, where atom i is included in RC. The Euclidean distance between two configurations m and n is calculated as

$$d_{mn}^{(i)} = \|\mathbf{v}_m^{(i)} - \mathbf{v}_n^{(i)}\|, \quad (11)$$

where $\mathbf{v}_m^{(i)}$ and $\mathbf{v}_n^{(i)}$ are the input vectors of configurations m and n for atom i, respectively, which have been defined in Eq. (10). For a new configuration x, the k nearest samples in the database are selected based on the Euclidean distances, and the correction force on atom i of configuration x can be predicted in terms of a weighted average as

$$F_{i(x)}^{corr} = \frac{\sum_p^k \left[d_{xp}^{(i)}\right]^{-1} F_{i(p)}^{corr}}{\sum_p^k \left[d_{xp}^{(i)}\right]^{-1}}, \quad (12)$$

where p is one of the k nearest samples, and $F_{i(p)}^{corr}$ and $F_{i(x)}^{corr}$ are the correction forces on atom i of configuration p and x, respectively. The correction forces on other RC-related atoms of configuration x can be predicted in the same way.

The procedure of QM/MM MD simulation combined with our internal force machine learning correction is outlined as follows: (1) Define a set of geometric parameters that are involved in the reaction coordinate z and perform low-level SQM/MM MD simulations along z (e.g., umbrella sampling). (2) Choose a few configurations from MD trajectories in the whole range of z as data points in the ML database and calculate their input vectors for RC-related atoms using Eq. (10) and the *ab initio* QM/MM forces $\mathbf{F}^H$. (3) Apply $\mathbf{F}^L$ obtained in step 1 and $\mathbf{F}^H$ obtained in step 2 to Eq. (7) to calculate $\mathbf{F}^{corr}$ for all samples in the ML database using Eqs. (5) and (7) or the variant formulation (e.g., see the Appendix). (4) Perform SQM/MM MD simulations with the correction forces. At each MD step, the input vectors for RC-related atoms are calculated using Eq. (10). The ML correction forces are predicted using Eqs. (11) and (12) and added to the SQM/MM atomic forces to update atomic coordinates and velocities for the next integration step. (5) Calculate the free energy change along z based on the MD samplings obtained in step 4.

COMPUTATIONAL VALIDATION

To demonstrate the capability of this method, we calculated the PMFs for two aqueous systems, the Menshutkin reaction between methyl chloride and ammonia and the aliphatic Claisen rearrangement reaction of allyl vinyl ether (AVE), at different QM/MM levels. For both reactions, the solute was defined as the QM subsystem and solvated in a cubic water box with a 16 Å extended distance. The MM subsystem for the Menshutkin and Claisen rearrangement reaction contains 1686 and 1745 water molecules, respectively. The TIP3P water model was employed under periodic boundary condition.[25] The cutoff distance for nonbonded interactions was set as 14 Å. The QM/MM van der Waals (vdW) interactions were described with the Amber-ff14SB force field.[26] The Hartree-Fock (HF) method with the 6-31G(d) basis set and the self-consistent charge density functional tight binding (SCC-DFTB) method with the second-order formulation[27,28] were used, respectively, as the high-level and low-level QM models. Note that the two levels were selected based on the difference of the QM/MM MD simulation results, while their comparisons with experiments are not important for our purpose presently. The MD simulations for free energy calculations were carried out after geometry optimization on the solute in the gas phase and solvent equilibration. The integration time step was set as 1 fs, and the system temperature was maintained at 300 K using a Berendsen thermostat[29] with a time constant of 1.0 ps during all simulations. All the calculations were implemented in the modified Amber 14 program package[30] combined with the GAUSSIAN 03 program[31] for Hartree-Fock calculations.

The reaction coordinate for the Menshutkin reaction of methyl chloride and ammonia transfer was chosen as $z = r_{\text{C–Cl}} - r_{\text{C–N}}$. Umbrella samplings with 64 windows centering from $z = -1.6$ to 1.6 Å were applied. The SCC-DFTB/MM MD simulations were performed for 1 ns for each window, and the snapshots were saved every 5 ps from each window to build the ML database. Then 1 ns SCC-DFTB/MM MD simulations with ML corrections were performed for each window. In addition, the HF/6-31G(d)/MM MD simulations were performed for 20 ps for each window to obtain the high-level free energy change as our reference. The cutoff distance r_{cut} to construct the input vectors for ML was set as 3.8 Å, and the hyperparameter k in kNN was set as 5. The weighted histogram analysis method (WHAM)[32,33] was applied to calculate PMFs based on MD samplings. The PMFs obtained at different levels were shown in Fig. 1. The reaction free energies at the SCC-DFTB/MM and HF/6-31G(d)/MM levels are −10.9 and −25.9 kcal/mol, respectively. The values of free energy barriers at two levels are similar (20.8 and 21.2 kcal/mol), but the locations of transition state along RC are different. The low-level MD simulations with ML corrections reproduce the free energy profile in good agreement with the results obtained at the high-level, that is, −23.8 and 20.8 kcal/mol for the reaction free energy and barrier, respectively. A corrected position along RC for the transition state is also obtained. We also

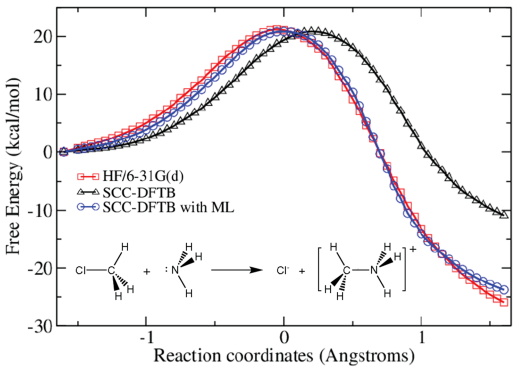

FIG. 1. Potential of mean force for the Menshutkin reaction. Different colors and shapes represent MD simulations with different methods (black triangle: SCC-DFTB/MM; red square: HF/6-31G(d)/MM; blue circle: SCC-DFTB/MM with ML corrections).

checked the convergence of the size of ML database using 50 snapshots from each window, getting similar reaction free energy and barrier with less than 2 kcal/mol difference from the previous results. It suggests that 200 snapshots per window, that is, 200 times of *ab initio* QM/MM calculations, have been sufficient to construct a database for ML training. On the other hand, a typical *ab initio* QM/MM MD needs at least tens of thousands times for *ab initio* calculations. The saving in the computational cost is about 2 orders of magnitude.

The reaction coordinate for the Claisen rearrangement of AVE was chosen as $z = r_{O3-C4} - r_{C1-C6}$. Umbrella samplings with 53 windows centering from $z = -4.2$ to 0.2 Å were applied. The simulation time at different levels, the size of the database for training, the hyperparameters in ML, and the method for free energy calculations were set as the same as those in the first system. As shown in Fig. 2, the free energy barriers at the SCC-DFTB/MM and HF/6-31G(d)/MM levels are 20.7 and 42.9 kcal/mol, respectively. The simulation with ML corrections achieves 42.4 kcal/mol, but the position of the transition state has a slight deviation. On account of the complication of the Claisen rearrangement of AVE such as the transition from extended to compact forms in the reactant region and the stereochemistry of transition states,[34,35] the identification of an accurate RC is challenging. It may cause some errors in our case because only the forces on RC-related atoms are

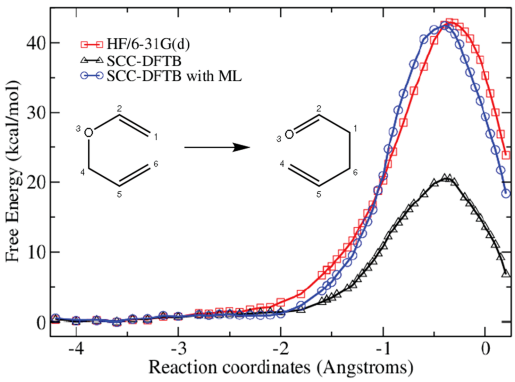

FIG. 2. Potential of mean force for the Claisen rearrangement reaction of AVE. Different colors and shapes represent MD simulations with different methods (black triangle: SCC-DFTB/MM; red square: HF/6-31G(d)/MM; blue circle: SCC-DFTB/MM with ML corrections).

corrected with ML. A preceding optimization on the transition path[36–39] or a further force correction along some additional selected degrees of freedom beyond RC may be useful to remedy this problem and overcome the RC-dependence as the main disadvantage of this method. The combination with RC-free enhanced sampling techniques[40,41] may be even beneficial to some more complex systems without obvious RCs.

Several improvements can be made to address the limitations observed in this method. One extension is the optimal choice of ML algorithms for force predictions. In order to measure the quality of kNN used in this paper, we constructed two testing sets for each system. The first one (set 1) consists of several configurations (200 snapshots from each window) that were sampled during original SCC-DFTB/MM simulations but excluded from the database of kNN. The second one (set 2) includes some other configurations (100 snapshots from each window) that were sampled during SCC-DFTB/MM simulations with ML corrections. The comparisons of predicted and reference correction forces were shown in Figs. S1 and S2 in the supplementary material. Compared with the reference values obtained from Eqs. (5) and (7), the mean absolute errors (MAEs) on the predicted atomic forces for set 1 and set 2 were 0.29 and 0.35 eV/Å for the system of the Menshutkin reaction, respectively. The larger error for set 2 reflects the deviation originated from the insufficient overlap between sampling spaces at two levels, which can be remedied with an adaptive ML database. For the system of the Claisen rearrangement reaction, the MAEs for set 1 and set 2 were 0.25 and 0.28 eV/Å, respectively. Although the errors on the calculated reaction free energies and free energy barriers for the two systems are as small as 1-2 kcal/mol, the intrinsic error of kNN on force predictions cannot be ignored. More recent ML algorithms based on nonlinear kernel, such as Gaussian process regression, have been reported to achieve higher accuracy within 0.1 eV/Å for the bulk silicon[19] and would be introduced in our future work. We further checked the energy conservation under the present ML-based model. The dynamic energy estimation scheme reported recently[42] was employed. The energy evolution of one representative trajectory during NVE simulations can be seen in Fig. S3 in the supplementary material. An energy drift of 1.8×10^{-4} kcal/mol per atom per picosecond was observed during a 10 ps evolution, showing that the energy conservation cannot be implied strictly because of the force-based predictions with ML.[19] Another extension is the optimal choice of MD algorithms for sampling, such as the thermostat applied to QM/MM systems. Here we compared two sets of simulation results on the Claisen rearrangement of AVE, that is, using the Berendsen thermostat with a time constant of 1.0 ps[29] and using Langevin dynamics with a friction coefficient of 1.0 ps^{-1}.[43] On the one hand, the free energy barriers using Langevin dynamics were calculated as 21.3 and 43.4 kcal/mol with original SCC-DFTB/MM and our ML-based model, respectively, indicating a small difference from the corresponding results using the Berendsen thermostat (within 1 kcal/mol, see above). On the other hand, the fluctuation of kinetic energy was reproduced with Langevin dynamics as the same as the analytical value but underestimated by 20% with the Berendsen thermostat, which means that the latter cannot give a canonical ensemble. On account of the significant role

of sampling not only on free energy calculations but also on ML database constructions, special care should be taken in the MD simulation techniques.

CONCLUSION

Compared with some recently reported ML methods combined with QM/MM calculations, in which the QM/MM reaction free-energy profiles or excitation energies were predicted satisfyingly,[21,44] this method has three additional outstanding features. The first one is its dimensionality reduction. As the same as our previous work,[21] thousands of MM degrees of freedom are reduced to the number of QM atoms using Mulliken charges. In this method, furthermore, only the input vectors for the atoms included in RC (2-4 atoms in general) are necessary, and the dimensionality of an input vector is determined by a cutoff distance. Therefore, the computational cost on ML would grow more slowly with the increase on QM degrees of freedom. The second feature is "truly force based,"[42] that is, atomic forces rather than a potential energy surface are predicted directly, avoiding the difficulty to calculate the gradients of some artificial input variables. Finally, compared with the neural network based models, the ML problem is moved from an input-feature space into a data-point space by kNN in our method, making it much easier to extend to a "learn-on-the-fly" scheme.[45]

In summary, we develop a machine learning internal force correction, which is guided by the information on a predefined reaction coordinate. It provides a pathway to improve the semiempirical QM/MM free energy calculations and achieves the accuracy of an *ab initio* QM/MM model. On the one hand, the use of RC can describe the dominant physical or chemical change in the reaction and capture the principal components contributing to a large portion of the free energy change. On the other hand, the use of ML can overcome the limitation of physical approximations or function forms in conventional RC-guided methods and decrease the computational cost on electronic structure calculations by about 2 orders of magnitude. The application to two systems reveals its reliability and efficiency. The limitations and some possible extensions are also discussed. This method makes a bridge between two active research topics, reaction coordinate guided sampling and machine learning based force field, and provides great potential for a broad range of interesting applications in molecular simulations.

SUPPLEMENTARY MATERIAL

See supplementary material for Figs. S1 and S2 for the comparisons of predicted and reference correction forces for two systems. Figure S3 for the energy evolution of one representative trajectory during NVE simulations using SCC-DFTB/MM with ML corrections.

ACKNOWLEDGMENTS

Financial support from the National Institute of Health (No. R01 GM061870-13) is gratefully appreciated. The support provided by the China Scholarship Council during a visit of J.W. to Duke University is also acknowledged.

APPENDIX: INTERNAL FORCES ON TESTING SYSTEMS

In order to construct the ML database, we need to calculate the internal force corrections based on SQM/MM and *ab initio* QM/MM atomic forces. The derivation in the main text can be extended to any linear combination of bond lengths involving more than two atoms. The RC for our second testing system is defined as $z = r_{\mathrm{O3-C4}} - r_{\mathrm{C1-C6}}$ with four atoms. The force corrections on O3 and C4 are decoupled from that on C1 and C6, and Eqs. (5) and (7) can be applied without any change.

The RC for our first testing system is defined as $z = r_{\mathrm{C-Cl}} - r_{\mathrm{C-N}}$. Note that one carbon atom is involved with two bond lengths, making it a little complicated. On the one hand, the correction force acting on C should be added along two directions, that is,

$$\mathbf{F}_{\mathrm{C}}^{corr} = \mathbf{F}_{\mathrm{C(Cl)}}^{corr} + \mathbf{F}_{\mathrm{C(N)}}^{corr}, \tag{A1}$$

where $\mathbf{F}_{\mathrm{C(Cl)}}^{corr}$ and $\mathbf{F}_{\mathrm{C(N)}}^{corr}$ are, respectively, along the direction of $\hat{\mathbf{r}}_{\mathrm{Cl-C}}$ and $\hat{\mathbf{r}}_{\mathrm{N-C}}$. On the other hand, the correction forces on Cl and N are only added along $\hat{\mathbf{r}}_{\mathrm{Cl-C}}$ and $\hat{\mathbf{r}}_{\mathrm{N-C}}$, respectively. Similar to Eq. (5), their components should satisfy

$$\mathbf{F}_{\mathrm{Cl}}^{corr}(t) + \mathbf{F}_{\mathrm{C(Cl)}}^{corr}(t) = 0 \tag{A2}$$

and

$$\mathbf{F}_{\mathrm{N}}^{corr}(t) + \mathbf{F}_{\mathrm{C(N)}}^{corr}(t) = 0, \tag{A3}$$

at any time t. Applying these correction terms to Eq. (4), we have

$$\begin{aligned}&\left[\mathbf{r}_{\mathrm{C}}^{H}(t+\Delta t) - \mathbf{r}_{\mathrm{C}}^{LC}(t+\Delta t)\right] - \left[\mathbf{r}_{\mathrm{Cl}}^{H}(t+\Delta t) - \mathbf{r}_{\mathrm{Cl}}^{LC}(t+\Delta t)\right] \\ &\quad = \frac{\Delta t^2}{2m_{\mathrm{C}}}\left[\mathbf{F}_{\mathrm{C}}^{H}(t) - \mathbf{F}_{\mathrm{C}}^{L}(t) - \mathbf{F}_{\mathrm{C}}^{corr}(t)\right] \\ &\quad\quad - \frac{\Delta t^2}{2m_{\mathrm{Cl}}}\left[\mathbf{F}_{\mathrm{Cl}}^{H}(t) - \mathbf{F}_{\mathrm{Cl}}^{L}(t) - \mathbf{F}_{\mathrm{Cl}}^{corr}(t)\right]\end{aligned} \tag{A4}$$

and

$$\begin{aligned}&\left[\mathbf{r}_{\mathrm{C}}^{H}(t+\Delta t) - \mathbf{r}_{\mathrm{C}}^{LC}(t+\Delta t)\right] - \left[\mathbf{r}_{\mathrm{N}}^{H}(t+\Delta t) - \mathbf{r}_{\mathrm{N}}^{LC}(t+\Delta t)\right] \\ &\quad = \frac{\Delta t^2}{2m_{\mathrm{C}}}\left[\mathbf{F}_{\mathrm{C}}^{H}(t) - \mathbf{F}_{\mathrm{C}}^{L}(t) - \mathbf{F}_{\mathrm{C}}^{corr}(t)\right] \\ &\quad\quad - \frac{\Delta t^2}{2m_{\mathrm{N}}}\left[\mathbf{F}_{\mathrm{N}}^{H}(t) - \mathbf{F}_{\mathrm{N}}^{L}(t) - \mathbf{F}_{\mathrm{N}}^{corr}(t)\right],\end{aligned} \tag{A5}$$

which leads to

$$\frac{\left[\mathbf{F}_{\mathrm{C}}^{H}(t) - \mathbf{F}_{\mathrm{C}}^{L}(t)\right]\cdot\hat{\mathbf{r}}_{\mathrm{Cl-C}}(t) - F_{\mathrm{C(Cl)}}^{corr}(t) - F_{\mathrm{C(N)}}^{corr}(t)\left[\hat{\mathbf{r}}_{\mathrm{N-C}}(t)\cdot\hat{\mathbf{r}}_{\mathrm{Cl-C}}(t)\right]}{m_{\mathrm{C}}} - \frac{\left[\mathbf{F}_{\mathrm{Cl}}^{H}(t) - \mathbf{F}_{\mathrm{Cl}}^{L}(t)\right]\cdot\hat{\mathbf{r}}_{\mathrm{Cl-C}}(t) - F_{\mathrm{Cl}}^{corr}(t)}{m_{\mathrm{Cl}}} = 0 \tag{A6}$$

and

$$\frac{\left[\mathbf{F}_{\mathrm{C}}^{H}(t)-\mathbf{F}_{\mathrm{C}}^{L}(t)\right]\cdot\hat{\mathbf{r}}_{\mathrm{N-C}}(t)-F_{\mathrm{C(N)}}^{corr}(t)-F_{\mathrm{C(Cl)}}^{corr}(t)\left[\hat{\mathbf{r}}_{\mathrm{N-C}}(t)\cdot\hat{\mathbf{r}}_{\mathrm{Cl-C}}(t)\right]}{m_{\mathrm{C}}}-\frac{\left[\mathbf{F}_{\mathrm{N}}^{H}(t)-\mathbf{F}_{\mathrm{N}}^{L}(t)\right]\cdot\hat{\mathbf{r}}_{\mathrm{N-C}}(t)-F_{\mathrm{N}}^{corr}(t)}{m_{\mathrm{N}}}=0. \quad \text{(A7)}$$

Combined with Eqs. (A2), (A3), (A6), and (A7), the internal force corrections on C, Cl, and N are solved using the low-level and high-level forces acting on the three atoms. The correction values at the following MD steps can be obtained in the same way based on the velocity Verlet integration schemes of Cartesian coordinates and velocities.

[1] P. G. Bolhuis, C. Dellago, and D. Chandler, Proc. Natl. Acad. Sci. U. S. A. **97**, 5877 (2000).
[2] *Free Energy Calculations: Theory and Applications in Chemistry and Biology*, Springer Series in Chemical Physics, edited by C. Chipot and A. Pohorille (Springer Berlin Heidelberg, 2007).
[3] E. Vanden-Eijnden, J. Comput. Chem. **30**, 1737 (2009).
[4] G. Fiorin, M. L. Klein, and J. Hénin, Mol. Phys. **111**, 3345 (2013).
[5] H. Vashisth, G. Skiniotis, and C. L. Brooks, Chem. Rev. **114**, 3353 (2014).
[6] G. M. Torrie and J. P. Valleau, Chem. Phys. Lett. **28**, 578 (1974).
[7] G. Torrie and J. Valleau, J. Comput. Phys. **23**, 187 (1977).
[8] P. Hu, S. Wang, and Y. Zhang, J. Am. Chem. Soc. **130**, 3806 (2008).
[9] P. Hu, S. Wang, and Y. Zhang, J. Am. Chem. Soc. **130**, 16721 (2008).
[10] H. Hu and W. Yang, Annu. Rev. Phys. Chem. **59**, 573 (2008).
[11] H. Hu and W. Yang, J. Mol. Struct.: THEOCHEM **898**, 17 (2009).
[12] X. Lu, D. Fang, S. Ito, Y. Okamoto, V. Ovchinnikov, and Q. Cui, Mol. Simul. **42**, 1056 (2016).
[13] F. Ercolessi and J. B. Adams, Europhys. Lett. **26**, 583 (1994).
[14] P. Maurer, A. Laio, H. W. Hugosson, M. C. Colombo, and U. Rothlisberger, J. Chem. Theory Comput. **3**, 628 (2007).
[15] Y. Zhou and J. Pu, J. Chem. Theory Comput. **10**, 3038 (2014).
[16] J. J. Ruiz-Pernía, E. Silla, I. n. Tuñón, S. Martí, and V. Moliner, J. Phys. Chem. B **108**, 8427 (2004).
[17] J. J. Ruiz-Pernía, E. Silla, I. n. Tuñón, and S. Martí, J. Phys. Chem. B **110**, 17663 (2006).
[18] V. Botu and R. Ramprasad, Phys. Rev. B **92**, 094306 (2015).
[19] Z. Li, J. R. Kermode, and A. De Vita, Phys. Rev. Lett. **114**, 096405 (2015).
[20] T. Stecher, N. Bernstein, and G. Csányi, J. Chem. Theory Comput. **10**, 4079 (2014).
[21] L. Shen, J. Wu, and W. Yang, J. Chem. Theory Comput. **12**, 4934 (2016).
[22] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Series in Statistics, 2nd ed. (Springer, 2009).
[23] D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications*, Computational Science Series (Elsevier Science, 2001).
[24] S. T. Roweis and L. K. Saul, Science **290**, 2323 (2000).
[25] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, J. Chem. Phys. **79**, 926 (1983).
[26] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling, J. Chem. Theory Comput. **11**, 3696 (2015).
[27] M. Elstner, D. Porezag, G. Jungnickel, J. Elsner, M. Haugk, T. Frauenheim, S. Suhai, and G. Seifert, Phys. Rev. B **58**, 7260 (1998).
[28] G. de M. Seabra, R. C. Walker, M. Elstner, D. A. Case, and A. E. Roitberg, J. Phys. Chem. A **111**, 5655 (2007).
[29] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak, J. Chem. Phys. **81**, 3684 (1984).
[30] D. A. Case, V. Babin, J. T. Berryman, R. M. Betz, Q. Cai, D. S. Cerutti, T. E. Cheatham III, T. A. Darden, R. E. Duke, H. Gohlke, A. W. Goetz, S. Gusarov, N. Homeyer, P. Janowski, J. Kaus, I. Kolossváry, A. Kovalenko, T. S. Lee, S. LeGrand, T. Luchko, R. Luo, B. Madej, K. M. Merz, F. Paesani, D. R. Roe, A. Roitberg, C. Sagui, R. Salomon-Ferrer, G. Seabra, C. L. Simmerling, W. Smith, J. Swails, R. C. Walker, J. Wang, R. M. Wolf, X. Wu, and P. A. Kollman, *AMBER 14* (University of California, San Francisco, 2014).
[31] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery, Jr., T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez, and J. A. Pople, GAUSSIAN 03, Revision D.02 (Gaussian, Inc., Wallingford, CT, 2004).
[32] A. M. Ferrenberg and R. H. Swendsen, Phys. Rev. Lett. **63**, 1195 (1989).
[33] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman, J. Comput. Chem. **13**, 1011 (1992).
[34] J. Zhang, Y. I. Yang, L. Yang, and Y. Q. Gao, J. Phys. Chem. B **119**, 5518 (2015).
[35] J. Zhang, Y. I. Yang, L. Yang, and Y. Q. Gao, J. Phys. Chem. B **119**, 14505 (2015).
[36] L. Xie, H. Liu, and W. Yang, J. Chem. Phys. **120**, 8039 (2004).
[37] S. K. Burger and W. Yang, J. Chem. Phys. **124**, 054109 (2006).
[38] H. Hu, Z. Lu, and W. Yang, J. Chem. Theory Comput. **3**, 390 (2007).
[39] H. Hu, Z. Lu, J. M. Parks, S. K. Burger, and W. Yang, J. Chem. Phys. **128**, 034105 (2008).
[40] M. Yang, L. Yang, Y. Gao, and H. Hu, J. Chem. Phys. **141**, 044108 (2014).
[41] L. Xie, L. Shen, Z.-N. Chen, and M. Yang, J. Chem. Phys. **146**, 024103 (2017).
[42] V. Botu, R. Batra, J. Chapman, and R. Ramprasad, J. Phys. Chem. C **121**, 511 (2017).
[43] W. F. van Gunsteren and H. J. C. Berendsen, Mol. Simul. **1**, 173 (1988).
[44] F. Häse, S. Valleau, E. Pyzer-Knapp, and A. Aspuru-Guzik, Chem. Sci. **7**, 5139 (2016).
[45] G. Csányi, T. Albaret, M. C. Payne, and A. De Vita, Phys. Rev. Lett. **93**, 175503 (2004).