Multiscale Quantum Mechanics/Molecular Mechanics Simulations with Neural Networks

Lin Shen[†], **Jingheng Wu**[†,‡], and **Weitao Yang**[*,†]

[†]Department of Chemistry, Duke University, Durham, North Carolina 27708, United States

[‡]School of Chemistry and Chemical Engineering, Sun Yat-sen University, Guangzhou 510275, P.R. China

Abstract

Molecular dynamics simulation with multiscale quantum mechanics/molecular mechanics (QM/MM) methods is a very powerful tool for understanding the mechanism of chemical and biological processes in solution or enzymes. However, its computational cost can be too high for many biochemical systems because of the large number of ab initio QM calculations. Semiempirical QM/MM simulations have much higher efficiency. Its accuracy can be improved with a correction to reach the ab initio QM/MM level. The computational cost on the ab initio calculation for the correction determines the efficiency. In this paper we developed a neural network method for QM/MM calculation as an extension of the neural-network representation reported by Behler and Parrinello. With this approach, the potential energy of any configuration along the reaction path for a given QM/MM system can be predicted at the ab initio QM/MM level based on the semiempirical QM/MM simulations. We further applied this method to three reactions in water to calculate the free energy changes. The free-energy profile obtained from the semiempirical QM/MM simulation is corrected to the ab initio QM/MM level with the potential energies predicted with the constructed neural network. The results are in excellent accordance with the reference data that are obtained from the ab initio QM/MM molecular dynamics simulation or corrected with direct ab initio QM/MM potential energies. Compared with the correction using direct ab initio QM/MM potential energies, our method shows a speed-up of 1 or 2 orders of magnitude. It demonstrates that the neural network method combined with the semiempirical QM/MM calculation can be an efficient and reliable strategy for chemical reaction simulations.

Graphical abstract

[*]**Corresponding Author:** weitao.yang@duke.edu.
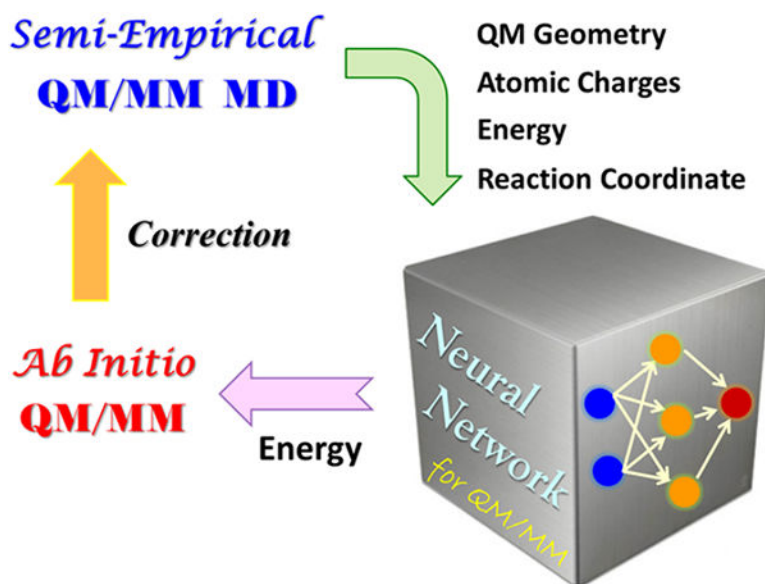
INTRODUCTION

Quantum mechanical phenomenon play an important role in fundamental chemical and biological processes, such as bond forming or breaking, proton or electron transfer, and electronic excitation. Most of these reactions take place in solution or enzymes rather than in gas phase. Because of the significant change in electronic structures of complex systems with thousands of degrees of freedom, the combined quantum mechanical and molecular mechanical (QM/MM) method, first proposed by Warshel and Levitt, is an accurate and computationally efficient tool toward QM descriptions on realistic chemical and biological systems.[1–5] In the QM/MM method, only a small number of atoms in the active site are selected for an accurate QM calculation, while the contribution of the rest of the system is described with an approximate, yet efficient MM force field. Furthermore, chemical and biological processes at room temperature are governed by changes on the free energy surface rather than potential energy surface. Therefore, molecular dynamics (MD) simulations from tens of picoseconds to hundreds of nanoseconds are usually required to achieve converged statistical sampling.[6–10] Since an electronic structure calculation is employed at each MD step, the application of direct QM/MM MD is still often limited to systems with relatively small active site for a short time simulation, such as several picoseconds.

Reducing the very demanding computational costs of QM calculations on the MD simulation is a major challenge. Focusing on the reaction path, Yang and co-workers developed a series of methods on reaction-path optimizations and free-energy calculations based on ab initio QM/MM approaches.[11–16] In the first method called QM/MM free energy perturbation (QM/MM-FEP), QM calculations were restricted to narrow regions such as stationary points on the reaction path. After an iterative optimization on the total QM/MM potential energy surface, the free energy difference between two fixed QM conformations was calculated with FEP to generate the free energy profile from reactant to product.[11] Hu et al. further developed the QM/MM minimum free-energy path (QM/MM-MFEP) method, in

which the influence of the environment on the structural properties of active site was well considered.[13,14] With the expression of free energy as a function of QM coordinates, the reaction path was minimized on the potential of mean force (PMF) surface of the QM degrees of freedom. Many enzymatic reactions including 4-oxalocrotonate tautomerase, orotidine 5′-monophosphate decarboxylase, and Cdc25B phosphatase have been studied successfully with the above approaches.[17–20] The efficiency is much higher than direct QM/MM MD simulations because the MD samplings on the QM subsystem are replaced by single-point calculations or iterative optimizations in an environment with a fixed ensemble of the MM subsystem. The dynamic contributions of the QM subsystem to the free energy are calculated within the reaction path potential method under the linear-response approximation on QM atomic charges or the vibrational frequency analysis under the harmonic approximation on stationary geometries,[12,21] requiring expensive QM calculations on the response kernel of QM charges or the Hessian matrix of the QM subsystem, respectively.

Another attractive way to speed up the direct QM/MM MD simulation is the semiempirical QM (SQM) models such as AM1, empirical valence bond (EVB), and the self-consistent charge density functional tight binding (SCC-DFTB) method in which the total energy was truncated at the second-order term.[22–26] The semiempirical QM calculations are so efficient that MD samplings on the biological processes for several nanoseconds are affordable. Compared with ab initio QM/MM calculations, however, the results obtained from SQM/MM simulations may be less accurate and less reliable because of some inherent deficiencies in SQM models. For example, the QM atomic charges from SCC-DFTB calculations were found to be much smaller than MM charges or charges derived from ab initio QM methods.[27] Many methods have been developed to account for the differences between semiempirical and ab initio QM/MM calculations, which can be further classified as "on-the-fly" and "reweighting" corrections. In the on-the-fly correction strategy, the potential of SQM model is reparametrized and adapted on-the-fly during MD simulations. The corresponding free energy profile is then obtained from the refined SQM/MM potential and finally converged to ab initio QM/MM results. This is in spirit analogous to the pioneering method developed by Gonzalez-Lafont et al., in which the specific reaction parameters for the NDDO molecular orbital theory were adjusted for individual reactions.[28] Plotnikov et al. reported a paradynamics approach, in which the EVB reference potential was reparametrized to the ab initio potential with an iterative and automated refinement procedure.[29] Zhou et al. developed a reaction path force matching method, in which the specific reaction parameters in the PM3 model were fitted iteratively to reproduce the atomic forces of the selected configurations along the reaction path at the Hartree–Fock (HF) level. [30] The application of these methods to reactions in condensed phase has demonstrated its success; but there are still some nontrivial concerns, such as the construction of the empirical potential and the converging process of the reaction path over iterations. In the reweighting correction strategy, an initial estimate of the free-energy profile is determined from SQM/MM MD simulations and then corrected after evaluating the free energy change from the approximate SQM/MM to the target ab initio QM/MM Hamiltonian. A general formulation of QM thermodynamic-cycle perturbation was proposed by Rod et al., in which the free energy difference along the reaction coordinate (RC) was calculated at a low level

such as MM or SQM/MM, while a vertical free energy change from low level to high level, e.g., MM → QM/MM or SQM/MM → ab initio QM/MM, was estimated with FEP.[31] Several more recent approaches share a similar spirit with different reweighting techniques. For example, König et al. applied the Bennett acceptance ratio estimator to connect MM samplings to QM/MM free energies.[32] Polyak et al. designed a dual Hamiltonian free energy perturbation method. After QM/MM MD samplings at the low level, the high-level QM/MM single-point calculations were performed at regular intervals while skipping a predefined number of MD steps, and then the perturbation energy difference between two levels was obtained.[33] Note that only the configurations saved each hundredth or thousandth step are used for free energy calculations, which reduces the ab initio computational cost significantly. In addition, the calculations on high-level gradients are unnecessary in the reweighting procedures compared to the on-the-fly corrections.

Ideally, one would like to perform MD simulations that achieve the accuracy of ab initio QM/MM yet as is efficient as a semiempirical QM/MM method. An essential challenge is how to reduce the number of required high-level QM calculations. For small QM systems in gas phase, neural network (NN) is a promising choice to directly provide a relation between molecular structure and potential energy.[34–39] After NN constructions from a set of reference data, the potential energy of the system with any configuration can be predicted with an accuracy comparable to ab initio methods at an MM computational cost. In the past decades, a number of NNs have been investigated for many types of systems, ranging from triatomic reactions to heterogeneous surface processes.[40,41] In 2007, Behler and Parrinello developed a high-dimensional neural network scheme, in which the total potential energy was represented as a sum of atomic energies.[35] The energy contribution of each atom depends on the atomic environment via a subset of standard NN, and the local environment of atoms can be collected through a set of symmetry functions as input vectors. The symmetry functions were designed to overcome several disadvantages of standard structure of network, including size limitation, transferability of parameters, and invariance with translations, rotations, and exchanges. High accuracy of the neural network at a few meV per atom has been demonstrated for a wide range of applications such as bulk silicon, zinc oxide, and water clusters.[35,42,43] Neural network also has a strong potential to improve the accuracy of existing QM models. Hu et al. developed a DFT-NN approach, in which an artificial neural network was built based on the results of first-principle calculations to reproduce several molecular properties such as heats of formation and absorption energies.[44,45] Dral et al. applied machine-learning techniques to the automatic tuning of parameters in the semiempirical OM2 model, which improves the accuracy without reducing transferability to individual molecules.[46] Ramakrishnan et al. introduced a -machine learning method that added machine learning corrections to quantum methods having less accuracy and less computational cost, showing that highly accurate predictions of high-level potential energies were possible based on low-level calculations.[47] This idea can be applied to correct the energies from SQM/MM to an ab initio QM/MM level, in which the configurations from SQM/MM MD samplings can be employed directly as the training set for machine learning. However, the description on systems in complex MM environments is more difficult than that in vacuum. Recently Häse et al. proposed a machine-learning

technique based on the Coulomb matrix to compute the excitation energies and spectral densities of BChls in the FMO complex.[48]

In our present work the neural network reweighting correction is employed to estimate free energy changes of chemical reactions at the ab initio QM/MM level with high accuracy based on the efficient SQM/MM MD simulations. Inspired by the high-dimensional NN reported by Behler et al.,[35] we developed a neural network method named QM/MM-NN for ab initio QM/MM potential energy calculations to reduce the expensive ab initio computational cost. Similar to the -machine learning method,[47] the potential energy difference between two levels was chosen as the output variable of our QM/MM-NN. We aim to develop an artificial neural network such that the NN potential energy predictions closely approximate the ab initio QM/MM calculations, and the free energy changes along the reaction coordinate through QM/MM-NN reweighting achieve similar free energy profiles. Beside the neural network error itself, a possible additional error in our current approach can be the poor overlap between the sampling spaces at two levels of theory. For our test reactions the final free energy difference after reweighting from the low-level PMF is very close to that obtained from direct ab initio QM/MM MD simulations. It is partially due to the similarity on the critical structures such as transition state and local minima along the reaction path at two levels.[10] The choice of reaction coordinates is also essential. In principle the error from reweighting can be remedied with additional samplings on the exact or NN-predicted high-level potential energy surface, but it is beyond our consideration in this paper.

The rest of this paper is organized as follows. We first give an outline of the high-dimensional neural network developed by Behler and Parrinello. Then we will describe our QM/MM-NN scheme to predict ab initio QM/MM potential energies and free energies, followed by Simulation Details and Results and Discussions.

THEORY

Neural Network Architecture.

Neural networks represent approximation of arbitrary functions with simple and highly interconnected processing elements, which process information by their dynamic state response to input variables. The structure of a standard neural network can be seen in some textbooks such as Figure 11.2 in ref 49. To predict the potential energy from a given configuration of a system, the information on molecular structures is provided to the nodes in input layer of NN, and the associated potential energy is produced from the node in output layer. There are one or more hidden layers between the input and output layers. Here we consider a simple network with one hidden layer and one node in output layer. The potential energy E is written as

$$E = \sum_{j=1}^{L} w_j f\left(\sum_{k=1}^{M} w_{jk} x_k + b_j \right) + b_0 \quad (1)$$

where x_k is the input variable in node k in the input layer, w_{jk} is the weight parameter that connects node k in the input layer with node j in the hidden layer, w_j is the weight parameter that connects node j in the hidden layer with the node in the output layer, b_j and b_0 are respectively the bias weights of hidden and output layers, M and L are respectively the number of nodes in input and hidden layers, and $f(x)$ is the nonlinear function that can be chosen as the sigmoid function, the hyperbolic function, or the Gaussian function. All of the weight parameters and bias weights are determined during a training process, in which the minimization of the error between the predicted energies obtained from eq 1 and the reference energies obtained from electronic structure calculations is implemented based on the data in the training set.

The structure of the high-dimensional neural network developed by Behler et al. was shown in Figure 2 in the original paper.[35] Here we provide a brief summary. The potential energy E is represented as

$$E = \sum_{i=1}^{N} E_i \quad (2)$$

where E_i is the atomic energy contribution of atom i, and N is the number of atoms in a molecule. The key is to divide the whole network into N subnets, and the atomic energy contribution E_i can be obtained from the i-th subnet as

$$E_i = \sum_{j=1}^{L} w_{ij} f\left(\sum_{k=1}^{M} w_{ijk} G_i^k + b_{ij} \right) + b_i \quad (3)$$

Here the subnet is assumed to have one hidden layer for simplicity, but two or more hidden layers are also used in the related works. The superscript i denotes the i-th subnet used to predict E_i. As the same as in eq 1, w_{ijk} and w_{ij} are the weight parameters coupling two nodes in neighbor layers, b_{ij} and b_i are respectively the bias weights of hidden and output layers, M and L are respectively the number of nodes in input and hidden layers of the i-th subnet, $f(x)$ is the nonlinear function, and G_i^k is the k-th generalized coordinate of atom i.

Since the individual energy E_i is dependent on the chemical environment of atom i, the relative positions of other atoms neighboring on atom i should be involved in the generalized coordinates explicitly or implicitly. Different types of symmetry functions have been designed.[50,51] In the current network, two generalized coordinates, a radial function and an angular function, are used for each atom for simplicity. The radial function of atom i in eq 3 is defined as

$$G_i^1 = \sum_{j \neq i}^{N} e^{-n\left(R_{ij} - R_s\right)^2} f_c\left(R_{ij}\right) \quad (4)$$

where R_{ij} is the distance between atom i and j, R_s and η are the predetermined parameters of NN that would be fixed during the training process, and $f_c(R_{ij})$ is the cutoff function as

$$
f_c(R_{ij}) = \begin{cases} \frac{1}{2}\left[\cos\left(\frac{\pi R_{ij}}{R_c}\right) + 1\right] & R_{ij} \leq R_c \\ 0 & R_{ij} > R_c \end{cases} \quad (5)
$$

where R_c is the predetermined cutoff radius. It means that the contribution of atom j to the generalized coordinates of atom i is neglected if the distance between two atoms is larger than R_c.

The angular function of atom i in eq 3 is defined as

$$
G_i^2 = 2^{1-\xi} \sum_{j,k \neq i}^{N} \left(1 \pm \cos\theta_{ijk}\right)^{\xi} e^{-n\left(R_{ij}^2 + R_{kj}^2 + R_{ik}^2\right)} f_c(R_{ij}) f_c(R_{jk}) f_c(R_{ik}) \quad (6)
$$

where ξ is the predetermined parameters of NN, θ_{ijk} is the angle that consists of atoms i, j, and k, and R_{ij}, R_{jk}, and R_{ik} are the distances between atoms i and j, j and k, and i and k, respectively. Other terms have been defined in eqs 4 and 5. More details on the high-dimensional neural network for potential energy predictions can be found in Behler's reviews and the related papers.[37–39]

QM/MM Neural Network.

The total potential energy of the whole system in the QM/MM model is written as

$$
E_{tol} = E_{QM} + E_{QM/MM} + E_{MM} \quad (7)
$$

Here E_{QM} is the quantum mechanical energy of the QM subsystem, E_{MM} is the standard molecular mechanical interactions involving exclusively atoms in the MM subsystem, and $E_{QM/MM}$ is the coupling term between QM and MM subsystems including electrostatic, van der Waals (vdW), and covalent interactions as

$$
E_{QM/MM} = E_{QM/MM}^{ele} + E_{QM/MM}^{vdW} + E_{QM/MM}^{cov} \quad (8)
$$

The QM/MM electrostatic interaction is the core of the QM/MM model. In the "mechanical-embedding" approach, all three terms in eq 8 are modeled classically, and E_{QM} in eq 7 is obtained from electronic structure calculations in gas phase. Thus, the neural network described above can be applied directly. In the more accurate and widely used "electrostatic-embedding" approach, however, the contribution of MM electrostatic potentials is involved in the QM self-consistent field (SCF) calculation. Therefore, the sum of E_{QM} and $E_{QM/MM}^{ele}$ can be obtained as the eigenvalue of an effective Hamiltonian as follows

$$E_{\text{QM}} + E_{\text{QM/MM}}^{\text{ele}} = \left\langle \Psi \middle| \hat{H}_0 + \sum_{l \in \text{MM}} q_1 v_{\text{MM}}(\mathbf{r}_l) \middle| \Psi \right\rangle \quad (9)$$

where

$$v_{\text{MM}}(\mathbf{r}_l) = \sum_{k \in \text{QM}} \frac{Z_k}{|\mathbf{r}_k - \mathbf{r}_l|} - \int d\mathbf{r}' \frac{\rho(\mathbf{r}')}{|\mathbf{r}' - \mathbf{r}_l|} \quad (10)$$

$\hat{H}_0$ is the Hamiltonian of the QM subsystem in vacuum that depends on the QM theory, $\rho(\text{r}')$ is the electron density of the QM subsystem, Z_k is the charge of the nuclei of the QM atom k, and q_l is the point charge of the MM atom l. The remaining terms in eqs 7 and 8 are calculated much more efficiently with classical force fields.

Consider QM/MM calculations on the same system under two models. The high-level total potential energy can be expressed as the low-level total potential energy with an energy correction term. Applying the same MM force field at two levels, the correction term of the total QM/MM potential energy can be obtained from the QM difference, that is

$$E_{tol}^H = E_{tol}^L + \left\langle \Psi^H \middle| \hat{H}_0^H + \sum_{l \in \text{MM}} q_l v_{\text{MM}}(\mathbf{r}_l) \middle| \Psi^H \right\rangle - \left\langle \Psi^L \middle| \hat{H}_0^L + \sum_{l \in \text{MM}} q_l v_{\text{MM}}(\mathbf{r}_l) \middle| \Psi^L \right\rangle \quad (11)$$

where E_{tol}^H and E_{tol}^L are the total QM/MM potential energies at high and low levels, e.g, ab initio QM/MM and semiempirical QM/MM, respectively, $\hat{H}_0^H$ and $\hat{H}_0^L$ are respectively the QM Hamiltonian in vacuum based on the high-level and low-level QM theory, and Ψ^H and Ψ^L are the wave functions of the QM subsystem corresponding to the high-level and low-level Hamiltonian, respectively, which can be obtained from ab initio and SQM calculations with the contribution of the MM environment as background charges.

To apply the neural network to QM/MM energy calculations, we perform three modifications on the high-dimensional NN mentioned in the previous section. The structure of our QM/MM-NN is shown in Figure 1. First, the potential energy difference between ab initio and semiempirical QM/MM models is predicted as the output of NN to approximate the ab initio QM/MM potential. Note that both the SQM and NN calculations are several orders of magnitude faster than ab initio approaches, so E_{tol}^H in eq 11 can be obtained with much less CPU times. As expressed in eq 11, the total QM/MM energy difference is

$$\Delta E = \left\langle \Psi^H \middle| \hat{H}_0^H + \sum_{l \in \text{MM}} q_l v_{\text{MM}}(\mathbf{r}_l) \middle| \Psi^H \right\rangle - \left\langle \Psi^L \middle| \hat{H}_0^L + \sum_{l \in \text{MM}} q_l v_{\text{MM}}(\mathbf{r}_l) \middle| \Psi^L \right\rangle \quad (12)$$

which is approximated as the output of our present QM/MM-NN, where H and L denote ab initio QM/MM and SQM/MM methods, respectively. Similar to eqs 2 and 3, E can be represented as

$$\Delta E = \sum_{i=1}^{N} \Delta E_i + \Delta E_{RC} \quad (13)$$

where N is the number of atoms in the QM subsystem, E_i is the atomic contribution of atom i predicted from the i-th subnet, and E_{RC} is predicted from the RC subnet, which depends on the reaction coordinate and will be discussed later. Note that in the previous NN predictions on homogeneous systems, one subnet was used for the same element on account of transferability and permutation invariance,[35,43] but the present reaction-specific QM/MM-NN is constructed with different subnets for different atoms. It is due to the fact that only the geometry of the QM subsystem needs to be described in NN. For example, in our test reactions each atom in the QM subsystem can be distinguished from others by its connectivity in the molecule except for the hydrogen atoms bonded to the same heavy atom, and these hydrogen atoms have been permutated before the prediction according to their bond lengths.[52].

The second and the most important feature of our NN model is that in order to capture the polarization of the QM subsystem in response to MM electrostatic potentials as much as possible, we use the Mulliken atomic charges from the low-level electrostatic-embedding QM/MM calculations as input variables. Mulliken atomic charges reflect not only the external potentials of the MM environment but also the polarizabilities of all atoms in different QM subsystems. E_i is thus represented as

$$\Delta E_i = \sum_{j=1}^{L} w_{ij}\tanh\left(w_{ij1}G_i^1 + w_{ij2}G_i^2 + w_{ij3}Q_i + b_{ij}\right) + b_i \quad (14)$$

where G_i^1 and G_i^2 are the generalized coordinates of atom i dependent on the QM coordinates, which have been defined in eqs 4 and 6, w_{ijk}, w_{ij}, b_{ij}, and b_i are respectively the weight parameters and bias weights as the same as in eq 3, and Q_i is the Mulliken charge of QM atom i obtained from SQM/MM calculations, accounting for the polarization response of the QM subsystem to the MM environment. Here three input nodes and one hidden layer are applied in one subnet, and the nonlinear hyperbolic function is used in the nodes in hidden layer. The current network can be extended to a more complex structure with more input nodes or more hidden layers.

The third feature of the present QM/MM-NN originates from the goal of our work, that is, to estimate the free energy change of chemical reactions at the ab initio QM/MM level based on SQM/MM MD simulations. Normally, a set of structural or energetic parameters is chosen as reaction coordinates to characterize the process as a low-dimensional free-energy curve from reactant to product. The properties directly corresponding to the reaction

coordinate are thus essential for the accuracy of QM/MM calculations. For example, it has been studied by Ruiz-Pernía et al. that a corrected energy as a function of the reaction coordinate can improve the quality of the results from a low-level method.[53] Here we introduce the term E_{RC} in eq 13 in order to include the information on the reaction coordinate in the node in input layer more explicitly. The value of E_{RC} is predicted with an additional subnet in NN, which is similar to E_i in eq 14 as

$$\Delta E_{RC} = \sum_{j=1}^{L} w'_j \tanh\left(w'_{j1}z + w'_{j2}A(z) + w'_{j3}\frac{\partial A(z)}{\partial z} + b'_j \right) + b' \quad (15)$$

where w'_{jk}, w'_j, b'_j, and b' are respectively the weight parameters and bias weights in the RC subnet, z is the one-dimensional reaction coordinate, $A(z)$ is the potential of mean force at the low level as a function of the reaction coordinate, and $\partial A(z)/\partial z$ is the first derivative with respect to z. Note that we put three nodes in the input layer in the RC subnet for a one-dimensional reaction coordinate, while more input variables in eq 15 could be used if two or more reaction coordinates were chosen. A cubic spline is used to interpolate $A(z)$ and $\partial A(z)/\partial z$ at any value of z based on the approximate free-energy profile obtained from SQM/MM MD simulations.

Procedure for NN Training and QM/MM Simulations.

Here we outline the procedure of the QM/MM simulation to calculate free energy change of reaction with QM/MM-NN as follows:

(1) Define a set of geometric parameters in the QM subsystem as reaction coordinate z. After SQM/MM MD simulations, the low-level free energy change along the reaction path as a function of z is calculated with classical free energy approaches, such as free energy perturbation, thermodynamic integration, and umbrella sampling processed with the weighted histogram analysis method (WHAM).[54,55]

(2) Select several configurations from MD trajectories in the entire range of reaction coordinates and calculate their QM/MM energies at the highly accurate ab initio QM/MM level. Some of the above configurations are chosen randomly to build the training set, while the remaining configurations belong to the testing set.

(3) Perform the training of QM/MM neural network based on the selected configurations, in which the root mean squared error (RMSE) defined as

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(E_i^{\text{pred}} - E_i^{\text{ref}}\right)^2} \quad (16)$$

for the training set is minimized. In eq 16, E_i^{pred} and E_i^{ref} are the QM/MM energies of the i-th configuration calculated with QM/MM-NN and ab initio QM/MM methods, respectively, and N is the total number of configurations in

the training or testing set. The procedure for the training of QM/MM-NN consists of two steps.[45,56] First, the genetic algorithm is used to optimize the weights of NN. The individual with the smallest RMSE for the training set is decoded as the initial values of the weights of NN in the next step. Second, the steep-descent optimization is applied to further train the NN. The "early-stopping" rule is employed to avoid overfitting by monitoring the RMSE for the testing set during the optimization. Once overfitting occurs, the RMSE for the testing set begins to increase even if the RMSE for the training set continues to decrease. At this point, we stop the training procedure and obtain the optimized weights of NN. To measure the accuracy of QM/MM-NN, we use the Q^2 value for the samples in the testing set, defined as

$$Q^2 = 1 - \frac{\sum_{i=1}^{N}\left(E_i^{\mathrm{pred}} - E_i^{\mathrm{ref}}\right)^2}{\sum_{i=1}^{N}\left(E_i^{\mathrm{ref}} - \bar{E}\right)^2} \quad (17)$$

where $\bar{E}$ is the average of E_i^{ref}. The Q^2 value corresponding to the QM/MM energy difference between two levels can be also obtained from eq 17, in which E_i^{pred}, E_i^{ref}, and $\bar{E}$ are replaced by $\Delta E_i^{\mathrm{pred}}$, $\Delta E_i^{\mathrm{ref}}$, and $\Delta\bar{E}$, respectively.

(4) Predict the ab initio QM/MM potential energies of other samplings from the low-level MD simulations using the constructed QM/MM-NN at a low computational cost.

(5) Calculate the free energy difference between two levels at reaction coordinate z as

$$\Delta A^{L \to H}(z) = -\beta^{-1}\ln\left\langle e^{-\beta\left(E^H - E^L\right)}\right\rangle_z \quad (18)$$

where E^L is the low-level QM/MM potential energy calculated in Step 1, E^H is the high-level QM/MM potential energy calculated in Step 4, $\beta = 1/k_B T$ is the inverse temperature, and the angular bracket denotes an average over the samplings from the low-level MD simulations. The free energy change along the reaction path at the high level is then obtained as

$$\Delta A_{z_1 \to z_2}^L = \Delta_{z_1 \to z_2}^L + \Delta A^{L \to H}(z_2) - \Delta A^{L \to H}(z_1) \quad (19)$$

where $\Delta A_{z_1 \to z_2}^L$ is the free energy change between two states on the reaction path at the SQM/MM level, which has been obtained in Step 1, and $\Delta A_{z_1 \to z_2}^H$ is the related free energy change at the ab initio QM/MM level. The Bennett

acceptance ratio estimator[57,58] or some variants of WHAM[59] can be also used to calculate the high-level free energy profile.

SIMULATION DETAILS

The simulations on three aqueous reactions were implemented to evaluate the performance of our method. These reactions include the S_N2 reaction of $CH_3Cl + Cl^- \rightarrow Cl^- + CH_3Cl$, the intramolecular proton transfer reaction for glycine, and the aliphatic Claisen rearrangement reaction of allyl vinyl ether (AVE) to 4-pentenal (see Figure 2). For the S_N2 reaction, the complex of CH_3Cl and Cl^- was defined as the QM subsystem and solvated in a cubic box of $48 \times 48 \times 48$ Å^3 containing 3,600 water molecules under periodic boundary condition. The cutoff distance for nonbonded interactions was set as 14 Å. The vdW interactions between the QM and MM subsystems were described with the CHARMM22 force field.[60] The DFT method with the B3LYP hybrid functional and the 6–31G(d) basis set was used as the high-level QM model,[61,62] and the SCC-DFTB method with the parameters for chlorine developed recently was used as the low-level QM model.[63] For the proton transfer reaction, the glycine molecule was defined as the QM subsystem and solvated in a cubic box of $64 \times 64 \times 64$ Å^3 containing 8,650 water molecules under periodic boundary condition. The cutoff distance for nonbonded interactions was set as 12 Å. The QM/MM vdW interactions were described with the CHARMM22 force field. The B3LYP/6–31G(d) and SCC-DFTB methods were used as the high-level and low-level QM models, respectively. For the Claisen rearrangement reaction, the solute was defined as the QM subsystem and solvated in a cubic box with a 16 Å extended distance, containing 1,745 water molecules under periodic boundary condition. The cutoff distance for nonbonded interactions was set as 14 Å. The QM/MM vdW interactions were described with the Amber-ff14SB force field.[64] The Hartree–Fock method with the 6–31G(d) basis set and SCCDFTB was respectively used as the high-level and low-level QM models. The TIP3P water model was employed for all reactions.[65] Because the motivation of our work is to develop a correction scheme for SQM/MM to ab initio QM/MM, the low-level and high-level QM methods used for three reactions were selected to distinguish ab initio QM/MM MD simulation results from that obtained from SQM/MM. In other words, the qualities of SQM/MM and ab initio QM/MM models compared with experiments are inconsequential. It should be also noted that the SCC-DFTB with the second-order formulation (DFTB2) was used as the low-level QM model in this work, while the newly issued versions of DFTB such as the third-order expansion of the DFT total energy (DFTB3) and the Klopman-Ohno functional form for QM/MM electrostatic interactions may provide more accurate results, especially for proton transfer reactions and the systems with highly charged QM regions.[66,67]

The initial geometries of all three systems were optimized with the SCC-DFTB/MM method, followed by MD simulations for 100 ps for reaching equilibrium at the same level. The reaction coordinate for the S_N2 reaction was chosen as $z = d_{CCl1} - d_{CCl2}$, where d_{ij} is the bond distance between atom i and j. Umbrella samplings with 37 windows centering from $z = -2.5$ to 2.5 Å were applied to calculate the low-level PMF with WHAM. Note that this reaction is symmetrical along the path, so only the samplings in the first 19 windows with $z \leq 0.0$ Å are necessary for the training of QM/MM-NN, and the potential energies of

other samplings with $z > 0.0$ Å can be predicted with NN after the exchange of two chlorine atoms. The reaction coordinate for the proton transfer reaction was chosen as $z = d_{NH} - d_{OH}$, where H denotes the transferred hydrogen atom. Umbrella samplings with 25 windows centering from $z = -1.5$ to 1.5 Å were applied. The reaction coordinate for the Claisen rearrangement was chosen as $z = d_{OC5} - d_{C2C3}$. Umbrella samplings with 28 windows centering from $z = -4.0$ to 0.2 Å were applied. For all three reactions, 1 ns of the SCCDFTB/MM MD simulation was performed for each window, and the samplings were saved every 500 steps for each trajectory to calculate the QM/MM potential energies at the ab initio QM/MM level. In addition, 20 ps of the ab initio QM/MM MD simulation for each window was performed to obtain the free energy change at the high level for comparison. During all simulations the integration time step was set as 1 fs, and the system temperature was maintained at 300 K with a Berendsen thermostat.[68] The simulations for the S_N2 reaction were implemented in an in-house QM4D program package[69] combined with the GAUSSIAN 03 program[70] for DFT calculations and the Amber SQM (version 14) program[71,72] for SCC-DFTB calculations. The simulations for glycine were performed in the QM4D program package combined with the GAUSSIAN 03 program for DFT calculations. The simulations for AVE were carried in the AmberTools15 program package[73] combined with the GAUSSIAN 03 program for Hartree–Fock calculations.

RESULTS AND DISCUSSION

S_N2 Reaction in Water.

First the QM/MM-NN was constructed to predict the potential energy of any configuration along the S_N2 reaction path at the B3LYP/6–31(d)/MM level based on SCC-DFTB/MM calculations. The data in the training and testing sets for QM/MM-NN were randomly selected from the snapshots of the MD trajectories in Step 1 of the procedure. Since the efficiency of our method depends on the computational cost of ab initio QM/MM calculations on all configurations in the training set, we explored the performance of QM/MM-NN with different sizes of training sets. Five training sets were built with 20, 30, 40, 60, and 80 samples from each window, i.e., with the total number of data as 380, 570, 760, 1,140, and 1,520, respectively. Another 200 samples were selected from each window to build the testing set to further check the convergence of the size of training sets. The target high-level QM/MM potentials vary from −80 to 30 kcal/mol, while the energy differences between two levels as the direct output variables of NN are distributed as narrow as from −5 to 10 kcal/mol. The parameters of radial and angular functions in eqs 4, 5, and 6 were set as follows: $R_c = 6.0$ Å for all elements, $R_s = 0.0$ Å for all elements, $\eta = 1.80, 1.20,$ and 0.09 bohr^{-2} for C, H, and Cl, respectively, and $\xi = 1.80, 1.20,$ and 0.09 for C, H, and Cl, respectively. The weights for NN were optimized with the training method in Step 3. The RMSEs of energies over the training and testing sets were given in Table 1. The error on the training set of larger size decreases slightly from 1.15 to 1.09 kcal/mol. The RMSE on the testing set obtained from the smallest training set is 1.16 kcal/mol (0.0084 eV per atom) while that obtained from the largest training set is 1.15 kcal/mol, indicating that only 20 samples from each window are sufficient for training. The comparison of predicted and reference potential energies (E^{pred} and E^{ref}) for all samples in the testing set was shown in

Figure 3(a). The Q^2 value for E is calculated as 0.72, indicating the reliability of the present NN.

The reaction coordinate is essential for calculating free energy changes during a reaction. Here we explore the role of RC on the QM/MM-NN in two ways. First, the distribution of errors along RC was shown in Figure 3(b). The error for the reactant is somewhat larger than that for the transition state, because the chloride ion in the complex of the reactant is almost free in solution, leading to more flexible degree of freedoms. Second, an alternative QM/MM-NN without the subnet of RC, that is, E_{RC} in eq 13 is a constant of zero rather than that in eq 15, was constructed with the same database and training procedure. The RMSEs on the training and testing sets were 1.18 and 1.27 kcal/mol, respectively. The subnet of RC makes only minor improvement of about 0.1 kcal/mol in this system.

The free energy barrier for this reaction was estimated as 26.9 kcal/mol at the SCC-DFTB/MM level. The QM/MM potential energies of all snapshots (2,000 samples from each window) were then calculated at the B3LYP/6–31G(d)/MM level with DFT and QM/MM-NN, respectively. Following the procedure in Step 5, the free energy barrier reweighted with direct B3LYP/6–31G(d)/MM energies is 22.3 kcal/mol and with QM/MM-NN is 22.2 kcal/mol. The value obtained from B3LYP/MM MD simulations is 22.4 kcal/mol. As shown in Figure 4, QM/MM-NN reproduces the free-energy profile of the S_N2 reaction with a good agreement with the results obtained from ab initio QM/MM calculations.

In our method the high-level computation has been restricted to the training set, and thus most of the high-level potentials in the summation in eq 18 are predicted from QM/MM-NN much more efficiently. If all of the exact high-level energies were applied to eq 18 for reweighting, however, hundreds or thousands of samples from each window should be recalculated with expensive ab initio QM/MM methods. The computational saving of QM/MM-NN depends on the size of the training set relative to the required number of points of high-level potentials for reweighting. In other words, the number of configurations used in the training set of NN ($N_{training}$) compared with that used for the reweighting procedure ($N_{reweight}$) reflects the efficiency of our approach. In this system we employed $N_{training}$ = 380 from 19 windows and $N_{reweight}$ = 74,000 from 37 windows along the entire reaction path. The saving in computational cost is about 2 orders of magnitude. The symmetry of this reaction along the path makes our method more efficient.

Proton Transfer Reaction of Glycine in Water.

Another QM/MM-NN was constructed to predict the potential energy of glycine with different conformations from zwitterion to neutral form during the proton transfer reaction, where B3LYP/6–31G(d)/MM and SCC-DFTB/MM methods were employed as two levels. We set up three training sets with 20, 40, and 80 configurations from each window, i.e., with the total number of data as 500, 1,000 and 2,000, respectively. Another 160 configurations from each window were chosen for the testing set. The high-level QM/MM potentials vary from −40 to 90 kcal/mol, and the energy differences between two levels are distributed from −5 to 25 kcal/mol. Both of them are broader than that in the S_N2 reaction. The parameters of radial and angular functions were set as follows: R_c = 6.0 Å for all elements, R_s = 0.0 Å for all elements, η = 0.8, 0.2, 0.8, and 0.2 bohr^{-2} for C, O, N, and H, respectively, and ξ = 0.8,

0.9, 1.0, and 0.6 for C, O, N, and H, respectively. It can be seen from Table 2 that the training set with 20 samples from each window can achieve the accuracy comparable to other training sets with 40 or 80 samples from each window. Thus, the smallest training set was employed to build NN, obtaining the RMSE as 1.22 kcal/mol for the training set and 1.25 kcal/mol (0.0054 eV per atom) for the testing set. The comparison of E^{pred} and E^{ref} for all samples in the testing set was shown in Figure 5(a). The Q^2 value for ΔE is 0.97, which is somewhat larger than that in the S_N2 reaction because of the broader energy difference distribution. As depicted in Figure 5(b), the neutral form of glycine has the largest RMSE, and the error of zwitterion is a little larger than that of transition state. The RMSEs obtained from the alternative NN without the RC subnet were calculated as 1.38 and 1.39 kcal/mol for the training and testing sets, respectively. Although the accuracy of two types of NNs is similar in this system, the use of RC as an additional subnet accelerated the NN optimization in Step 3.

Interestingly, the free energy calculations based on the B3LYP/6–31G(d)/MM and SCC-DFTB/MM models provide very different conclusions. The free energy difference and barrier were estimated from SCC-DFTB/MM MD simulations as −7.8 and 5.0 kcal/mol, respectively, while those obtained from B3LYP/6–31G(d)/MM MD simulations were respectively 7.7 and 10.0 kcal/mol. The SCC-DFTB/MM MD simulations give an incorrect result because the zwitterion stabilized by water solvents is the predominant form of aqueous glycine. The reweighting correction from DFTB to DFT model works well. As shown in Figure 6, the free energy differences reweighted with the potentials obtained from B3LYP/6–31G(d)/MM calculations and QM/MM-NN predictions are 8.0 and 8.1 kcal/mol, respectively, and the related free energy barriers are 10.3 and 9.6 kcal/mol, respectively. The error originates from QM/MM-NN is less than 1 kcal/mol. The deviation increases in the product region, which is consistent with the distribution of errors with the largest RMSE for the neutral form during QM/MM-NN predictions. The use of NN is very efficient in this case ($N_{training} = 500$ and $N_{reweight} = 50,000$) compared with the reweighting procedure with direct QM/MM potentials, again showing a saving of 2 orders of magnitude.

Claisen Rearrangement Reaction of AVE in Water.

We applied the QM/MM-NN to the Claisen rearrangement reaction of AVE to predict the potential energies at the HF/6–31(d)/MM level based on SCC-DFTB/MM calculations. The data in the training and testing sets for QM/MM-NN were randomly selected from the snapshots. Four training sets were generated with 20, 30, 40, and 50 configurations from each window, i.e., with the total number of data as 560, 840, 1,120 and 1,400, respectively. The testing set consists of other 100 configurations from each window. The high-level QM/MM potentials are distributed from −30 to 60 kcal/mol, while the range of energy differences between two levels is much broader than that in the S_N2 reaction, varying from −30 to 25 kcal/mol. The parameters of radial and angular functions were set as follows: $R_c =$ 6.0 Å for all elements, $R_s = 0.0$ Å for all elements, $\eta = 0.4, 0.4,$ and 0.1 bohr^{-2} for C, O, and H, respectively, and $\xi = 0.2, 0.8,$ and 0.4 for C, O, and H, respectively. As shown in Table 3, the RMSE on the testing set is reduced from 2.36 to 2.21 kcal/mol when the total number of samples in the training set increases from 560 to 840. However, the training set with larger size is not useful for further improving the accuracy of predictions on the testing set.

Therefore, 30 configurations were selected from each window, giving the RMSEs as 2.05 kcal/mol for the training set and 2.21 kcal/mol (0.0068 eV per atom) for the testing set. The comparison of E^{pred} and E^{ref} for all samples in the testing set and the distribution of errors along RC were shown in Figure 7. The Q^2 value for E is 0.95. The RMSEs obtained from the alternative NN without the RC subnet were calculated as 3.08 and 3.24 kcal/mol for the training and testing sets, respectively. Here the difference is as large as 1 kcal/mol, indicating that the RC subnet is more important for larger molecules in improving the predictive ability of QM/MM-NN.

The free-energy barrier for this reaction is much larger at the HF/6–31G(d)/MM level than that obtained from SCC-DFTB/MM simulations, i.e., 42.6 and 20.9 kcal/mol, respectively. The former is overestimated compared with the experimental measurement,[74] while the SCC-DFTB/MM method combined with enhanced sampling has been successfully applied to the dynamics and kinetics study on this reaction.[75] In the present work, however, the low-level and high-level QM methods were chosen based on the difference of the QM/MM MD simulation results at two levels rather than their qualities. Thus, the HF/6–31G(d)/MM model was still employed as the reference. Here we adopted $N_{training}$ = 840 and $N_{reweight}$ = 56,000 during the whole procedure. As shown in Figure 8, the free energy barrier reweighted with direct HF/6–31G(d)/MM energies is 46.4 kcal/mol and with QM/MM-NN energies is 45.1 kcal/mol.

Two factors should be responsible for the error of QM/MMNN. On one hand, there is a difference of −1.3 kcal/mol between the reweighted results from direct QM/MM and QM/MM-NN potential energies, reflecting the error of NN predictions. It can be remedied with further improvements on the quality of NN in different ways, for example, the use of advanced training algorithms,[76] the Morte Carlo sampling of symmetry functions,[77] or the interpolation of gradients.[78] On the other hand, even if we applied direct HF/6–31G(d)/MM potential energies, the low-level free-energy profile cannot be reweighted to the accurate PMF obtained from ab initio QM/MM MD simulations. The error on the free energy barrier is 3.8 kcal/mol, indicating that the overlap between sampling spaces at two levels is not as sufficient as that in the previous systems. It is consistent with the chemical view that the change in molecular structure is much larger in the Claisen rearrangement reaction than that in the identity S_N2 or proton transfer reactions. One possible improvement is to calculate the free energy difference in Step 5 using a linear response approximation after a short-time MD sampling at the high level.[8,10] For the long-time dynamic simulations on larger biochemical systems, the "learn-on-the-fly" simulation on the high-level potential energy surface predicted with QM/MM-NN is a good candidate to reduce the error from reweighting.[79–81] Compared with some existing correction schemes with reparametrization on SQM models, [28–30] the combination with neural network make it more attractive in two aspects. First, these other approaches are restricted by some physical approximations and functional forms used in the SQM model, while the neural network method is based on generic mathematical potentials to overcome the limitations of physical considerations and capture the high-level data on the entire reaction path more accurately and flexibly. Second, the existing "on-the-fly" corrections rely on a semiempirical QM model with predefined parameters, while the QM/MM-NN corrections can be generalized to other multiscale or multistate simulations. For example, we can correct the QM/MM potential energy surface from the HF level to the

MP2 level to consider dynamical correlations or perform MD samplings on the excited state at the cost of calculations on the ground state. The "learn-on-the-fly" QM/MM-NN method would be addressed in our future work.

CONCLUSIONS

In summary, we developed a neural network method to predict the ab initio QM/MM potential energies for chemical reaction systems. On the basis of the high-dimensional neural network developed by Behler and Parrinello, three extensions have been developed for the potential energy predictions of QM/MM systems with a complex environment. (1) The semiempirical QM/MM simulations are performed initially, and then the energy difference between SQM/MM and ab initio QM/MM, rather than the absolute high-level potential energy, is predicted with NN. (2) The QM atomic charges at the SQM/MM level are introduced to NN in order to capture the polarization of the QM subsystem induced by the MM environment. (3) The information about the reaction coordinate is included in NN explicitly to enhance the overall accuracy on the entire reaction path. The improvement of the RC subnet is more important for larger systems.

Three reactions in water, the S_N2 reaction of $CH_3Cl + Cl^-$, the proton transfer reaction of glycine, and the Claisen rearrangement reaction of AVE, were studied. The ab initio QM/MM potential energies were predicted with QM/MM-NN with the RMSE as 1.16, 1.25, and 2.21 kcal/mol, i.e., 8.4, 5.4, and 6.8 meV per atom, respectively. The free-energy profile along the reaction coordinate estimated from SQM/MM MD simulations was then reweighted with NN predicted potential energies. The results are consistent with the free energy difference that is reweighted with direct high-level calculations or obtained from ab initio QM/MM MD simulations. It was found that tens of configurations from each window along the reaction path are sufficient for NN optimizations, while hundreds or thousands of snapshots in each MD trajectory are required for reweighting, showing about 2 orders of magnitude speed-up of the present QM/MM-NN computations. This work opens up the possibility to combine the machine learning method with QM/MM calculations for long-time dynamic simulations on large-scale biochemical systems.

ACKNOWLEDGMENTS

REFERENCES

(1). Warshel A; Levitt M Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. J. Mol. Biol 1976, 103, 227–249. [PubMed: 985660]

(2). Senn HM; Thiel W QM/MM Methods for Biomolecular Systems. Angew. Chem., Int. Ed 2009, 48, 1198–1229.

(3). Acevedo O; Jorgensen WL Advances in Quantum and Molecular Mechanical (QM/MM) Simulations for Organic and Enzymatic Reactions. Acc. Chem. Res 2010, 43, 142–151. [PubMed: 19728702]

(4). Brunk E; Rothlisberger U Mixed Quantum Mechanical/Molecular Mechanical Molecular Dynamics Simulations of Biological Systems in Ground and Electronically Excited States. Chem. Rev 2015, 115, 6217–6263. [PubMed: 25880693]

(5). Pezeshki S; Lin H Recent developments in QM/MM methods towards open-boundary multi-scale simulations. Mol. Simul 2015, 41, 168–189.

(6). Carloni P; Rothlisberger U; Parrinello M The Role and Perspective of Ab Initio Molecular Dynamics in the Study of Biological Systems. Acc. Chem. Res 2002, 35, 455–464. [PubMed: 12069631]

(7). Röhrig UF; Frank I; Hutter J; Laio A; VandeVondele J; Rothlisberger U QM/MM Car-Parrinello Molecular Dynamics Study of the Solvent Effects on the Ground State and on the First Excited Singlet State of Acetone in Water. ChemPhysChem 2003, 4, 1177–1182. [PubMed: 14652995]

(8). Kamerlin SCL; Haranczyk M; Warshel A Progress in Ab Initio QM/MM Free-Energy Simulations of Electrostatic Energies in Proteins: Accelerated QM/MM Studies of pKa, Redox Reactions and Solvation Free Energies. J. Phys. Chem. B 2009, 113, 1253–1272. [PubMed: 19055405]

(9). Isborn CM; Götz AW; Clark MA; Walker RC; Martínez TJ Electronic Absorption Spectra from MM and ab Initio QM/MM Molecular Dynamics: Environmental Effects on the Absorption Spectrum of Photoactive Yellow Protein. J. Chem. Theory Comput. 2012, 8, 5092–5106. [PubMed: 23476156]

(10). Lu X; Fang D; Ito S; Okamoto Y; Ovchinnikov V; Cui Q QM/MM free energy simulations: recent progress and challenges. Mol. Simul 2016, 42, 1056–1078. [PubMed: 27563170]

(11). Zhang Y; Liu H; Yang W Free energy calculation on enzyme reactions with an efficient iterative procedure to determine minimum energy paths on a combined ab initio QM/MM potential energy surface. J. Chem. Phys 2000, 112, 3483–3492.

(12). Lu Z; Yang W Reaction path potential for complex systems derived from combined ab initio quantum mechanical and molecular mechanical calculations. J. Chem. Phys 2004, 121, 89–100. [PubMed: 15260525]

(13). Hu H; Lu Z; Yang W QM/MM Minimum Free-Energy Path: Methodology and Application to Triosephosphate Isomerase. J. Chem. Theory Comput. 2007, 3, 390–406. [PubMed: 19079734]

(14). Hu H; Lu Z; Parks JM; Burger SK; Yang W Quantum mechanics/molecular mechanics minimum free-energy path for accurate reaction energetics in solution and enzymes: Sequential sampling and optimization on the potential of mean force surface. J. Chem. Phys 2008, 128, 034105. [PubMed: 18205486]

(15). Hu H; Yang W Free Energies of Chemical Reactions in Solution and in Enzymes with Ab Initio Quantum Mechanics/Molecular Mechanics Methods. Annu. Rev. Phys. Chem 2008, 59, 573–601. [PubMed: 18393679]

(16). Hu H; Yang W Development and application of ab initio QM/MM methods for mechanistic simulation of reactions in solution and in enzymes. J. Mol. Struct.: THEOCHEM 2009, 898, 17–30.

(17). Cisneros GA; Liu H; Zhang Y; Yang W Ab Initio QM/MM Study Shows There Is No General Acid in the Reaction Catalyzed by 4-Oxalocrotonate Tautomerase. J. Am. Chem. Soc 2003, 125, 10384–10393. [PubMed: 12926963]

(18). Hu H; Boone A; Yang W Mechanism of OMP Decarboxylation in Orotidine 5′-Monophosphate Decarboxylase. J. Am. Chem. Soc 2008, 130, 14493–14503. [PubMed: 18839943]

(19). Parks JM; Hu H; Rudolph J; Yang W Mechanism of Cdc25B Phosphatase with the Small Molecule Substrate p-Nitrophenyl Phosphate from QM/MM-MFEP Calculations. J. Phys. Chem. B 2009, 113, 5217–5224. [PubMed: 19301836]

(20). Wu P; Cisneros GA; Hu H; Chaudret R; Hu X; Yang W Catalytic Mechanism of 4-Oxalocrotonate Tautomerase: Significances of Protein-Protein Interactions on Proton Transfer Pathways. J. Phys. Chem. B 2012, 116, 6889–6897. [PubMed: 22417185]

(21). Ghysels A; Woodcock HL; Larkin JD; Miller BT; Shao Y; Kong J; Neck DV; Speybroeck VV; Waroquier M; Brooks BR Efficient Calculation of QM/MM Frequencies with the Mobile Block Hessian. J. Chem. Theory Comput 2011, 7, 496–514. [PubMed: 26596169]

(22). Dewar MJS; Storch DM Development and use of quantum molecular models. 75. Comparative tests of theoretical procedures for studying chemical reactions. J. Am. Chem. Soc 1985, 107, 3898–3902.

(23). Åqvist J; Warshel A Simulation of enzyme reactions using valence bond force fields and other hybrid quantum/classical approaches. Chem. Rev 1993, 93, 2523–2544.

(24). Elstner M; Porezag D; Jungnickel G; Elsner J; Haugk M; Frauenheim T; Suhai S; Seifert G Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties. Phys. Rev. B: Condens. Matter Mater. Phys 1998, 58, 7260–7268.

(25). Cui Q; Elstner M; Kaxiras E; Frauenheim T; Karplus M A QM/MM Implementation of the Self-Consistent Charge Density Functional Tight Binding (SCC-DFTB) Method. J. Phys. Chem. B 2001, 105, 569–585.

(26). Akimov AV; Prezhdo OV Large-Scale Computations in Chemistry: A Bird's Eye View of a Vibrant Field. Chem. Rev 2015, 115, 5797–5890. [PubMed: 25851499]

(27). Elstner M; Frauenheim T; Suhai S An approximate DFT method for QM/MM simulations of biological structures and processes. J. Mol. Struct.: THEOCHEM 2003, 632, 29–41.

(28). Gonzalez-Lafont A; Truong TN; Truhlar DG Direct dynamics calculations with NDDO (neglect of diatomic differential overlap) molecular orbital theory with specific reaction parameters. J. Phys. Chem 1991, 95, 4618–4627.

(29). Plotnikov NV; Kamerlin SCL; Warshel A Paradynamics: An Effective and Reliable Model for Ab Initio QM/MM Free-Energy Calculations and Related Tasks. J. Phys. Chem. B 2011, 115, 7950–7962. [PubMed: 21618985]

(30). Zhou Y; Pu J Reaction Path Force Matching: A New Strategy of Fitting Specific Reaction Parameters for Semiempirical Methods in Combined QM/MM Simulations. J. Chem. Theory Comput 2014, 10, 3038–3054. [PubMed: 26588275]

(31). Rod TH; Ryde U Quantum Mechanical Free Energy Barrier for an Enzymatic Reaction. Phys. Rev. Lett 2005, 94, 138302. [PubMed: 15904045]

(32). König G; Hudson PS; Boresch S; Woodcock HL Multiscale Free Energy Simulations: An Efficient Method for Connecting Classical MD Simulations to QM or QM/MM Free Energies Using Non-Boltzmann Bennett Reweighting Schemes. J. Chem. Theory Comput 2014, 10, 1406–1419. [PubMed: 24803863]

(33). Polyak I; Benighaus T; Boulanger E; Thiel W Quantum mechanics/molecular mechanics dual Hamiltonian free energy perturbation. J. Chem. Phys 2013, 139, 064105. [PubMed: 23947841]

(34). Blank TB; Brown SD; Calhoun AW; Doren DJ Neural network models of potential energy surfaces. J. Chem. Phys 1995, 103, 4129–4137.

(35). Behler J; Parrinello M Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. Phys. Rev. Lett 2007, 98, 146401. [PubMed: 17501293]

(36). Handley CM; Popelier PLA Potential Energy Surfaces Fitted by Artificial Neural Networks. J. Phys. Chem. A 2010, 114, 3371–3383. [PubMed: 20131763]

(37). Behler J Neural network potential-energy surfaces in chemistry: a tool for large-scale simulations. Phys. Chem. Chem. Phys 2011, 13, 17930–17955. [PubMed: 21915403]

(38). Behler J Representing potential energy surfaces by high-dimensional neural network potentials. J. Phys.: Condens. Matter 2014, 26, 183001. [PubMed: 24758952]

(39). Behler J Constructing high-dimensional neural network potentials: A tutorial review. Int. J. Quantum Chem. 2015, 115, 1032–1050.

(40). Jiang B; Guo H Permutation invariant polynomial neural network approach to fitting potential energy surfaces. J. Chem. Phys 2013, 139, 054112. [PubMed: 23927248]

(41). Shen X; Chen J; Zhang Z; Shao K; Zhang DH Methane dissociation on Ni(111): A fifteen-dimensional potential energy surface using neural network method. J. Chem. Phys 2015, 143, 144701. [PubMed: 26472389]

(42). Artrith N; Morawietz T; Behler J High-dimensional neural-network potentials for multicomponent systems: Applications to zinc oxide. Phys. Rev. B: Condens. Matter Mater. Phys 2011, 83, 153101.

(43). Kondati Natarajan S; Morawietz T; Behler J Representing the potential-energy surface of protonated water clusters by high-dimensional neural network potentials. Phys. Chem. Chem. Phys 2015, 17, 8356–8371. [PubMed: 25436835]

(44). Hu L; Wang X; Wong L; Chen G Combined first-principles calculation and neural-network correction approach for heat of formation. J. Chem. Phys 2003, 119, 11501–11507.

(45). Li H; Shi L; Zhang M; Su Z; Wang X; Hu L; Chen G Improving the accuracy of density-functional theory calculation: The genetic algorithm and neural network approach. J. Chem. Phys 2007, 126, 144101. [PubMed: 17444695]

(46). Dral PO; von Lilienfeld OA; Thiel W Machine Learning of Parameters for Accurate Semiempirical Quantum Chemical Calculations. J. Chem. Theory Comput. 2015, 11, 2120–2125. [PubMed: 26146493]

(47). Ramakrishnan R; Dral PO; Rupp M; von Lilienfeld OA Big Data Meets Quantum Chemistry Approximations: The -Machine Learning Approach. J. Chem. Theory Comput 2015, 11, 2087–2096. [PubMed: 26574412]

(48). Häse F; Valleau S; Pyzer-Knapp E; Aspuru-Guzik A Machine learning exciton dynamics. Chem. Sci 2016, 7, 5139–5147. [PubMed: 30155164]

(49). Hastie T; Tibshirani R; Friedman J The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed.; Springer Series in Statistics Springer: 2009.

(50). Behler J Atom-centered symmetry functions for constructing high-dimensional neural network potentials. J. Chem. Phys 2011, 134, 074106. [PubMed: 21341827]

(51). Jose KVJ; Artrith N; Behler J Construction of high-dimensional neural network potentials using environment-dependent atom pairs. J. Chem. Phys 2012, 136, 194111. [PubMed: 22612084]

(52). Chen J; Xu X; Xu X; Zhang DH A global potential energy surface for the $H_2 + OH \leftrightarrow H_2O + H$ reaction using neural networks. J.Chem. Phys 2013, 138, 154301. [PubMed: 23614417]

(53). Ruiz-Pernía JJ; Silla E; Tuñon I. n.; Martí S; Moliner V Hybrid QM/MM potentials of mean force with interpolated corrections. J. Phys. Chem. B 2004, 108, 8427–8433.

(54). Ferrenberg AM; Swendsen RH Optimized Monte Carlo data analysis. Phys. Rev. Lett 1989, 63, 1195–1198. [PubMed: 10040500]

(55). Kumar S; Rosenberg JM; Bouzida D; Swendsen RH; Kollman PA The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. J. Comput. Chem 1992, 13, 1011–1021.

(56). Wu J; Mei J; Wen S; Liao S; Chen J; Shen Y A self-adaptive genetic algorithm-artificial neural network algorithm with leave-one-out cross validation for descriptor selection in QSAR study. J. Comput. Chem 2010, 31, 1956–1968. [PubMed: 20512843]

(57). Bennett CH Efficient estimation of free energy differences from Monte Carlo data. J. Comput. Phys 1976, 22, 245–268.

(58). Shirts MR; Chodera JD Statistically optimal analysis of samples from multiple equilibrium states. J. Chem. Phys 2008, 129, 124105. [PubMed: 19045004]

(59). Yang M; Yang L; Gao Y; Hu H Combine umbrella sampling with integrated tempering method for efficient and accurate calculation of free energy changes of complex energy surface. J. Chem. Phys 2014, 141, 044108. [PubMed: 25084882]

(60). MacKerell AD, Jr.; Bashford D; Bellott M; Dunbrack RL; Evanseck JD; Field MJ; Fischer S; Gao J; Guo H; Ha S; Joseph-McCarthy D; Kuchnir L; Kuczera K; Lau FTK; Mattos C; Michnick S; Ngo T; Nguyen DT; Prodhom B; Reiher WE; Roux B; Schlenkrich M; Smith JC; Stote R; Straub J; Watanabe M; Wiórkiewicz-Kuczera J; Yin D; Karplus M All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. J. Phys. Chem. B 1998, 102, 3586–3616. [PubMed: 24889800]

(61). Lee C; Yang W; Parr RG Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. Phys. Rev. B: Condens. Matter Mater. Phys 1988, 37, 785–789.

(62). Becke AD Density functional thermochemistry. III. The role of exact exchange. J. Chem. Phys 1993, 98, 5648–5652.

(63). Kubar T; Bodrog Z; Gaus M; Köhler C; Aradi B; Frauenheim T; Elstner M Parametrization of the SCC-DFTB Method for Halogens. J. Chem. Theory Comput 2013, 9, 2939–2949. [PubMed: 26583977]

(64). Maier JA; Martinez C; Kasavajhala K; Wickstrom L; Hauser KE; Simmerling C ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. J. Chem. Theory Comput 2015, 11, 3696–3713. [PubMed: 26574453]

(65). Jorgensen WL; Chandrasekhar J; Madura JD; Impey RW; Klein ML Comparison of simple potential functions for simulating liquid water. J. Chem. Phys 1983, 79, 926–935.

(66). Gaus M; Cui Q; Elstner M DFTB3: Extension of the Self-Consistent-Charge Density-Functional Tight-Binding Method (SCC-DFTB). J. Chem. Theory Comput 2011, 7, 931–948.

(67). Hou G; Zhu X; Elstner M; Cui Q A Modified QM/MM Hamiltonian with the Self-Consistent-Charge Density-Functional-Tight-Binding Theory for Highly Charged QM Regions. J. Chem. Theory Comput. 2012, 8, 4293–4304. [PubMed: 23275762]

(68). Berendsen HJC; Postma JPM; van Gunsteren WF; DiNola A; Haak JR Molecular dynamics with coupling to an external bath. J. Chem. Phys 1984, 81, 3684–3690.

(69). Hu X; Hu H; Yang W QM4D: An integrated and versatile quantum mechanical/molecular mechanical simulation package. http://www.qm4d.info/ (accessed 2016).

(70). Frisch MJ; Trucks GW; Schlegel HB; Scuseria GE; Robb MA; Cheeseman JR; Montgomery JA, Jr.; Vreven T; Kudin KN; Burant JC; Millam JM; Iyengar SS; Tomasi J; Barone V; Mennucci B; Cossi M; Scalmani G; Rega N; Petersson GA, Nakatsuji H; Hada M; Ehara M; Toyota K; Fukuda R; Hasegawa J; Ishida M; Nakajima T; Honda Y; Kitao O; Nakai H; Klene M; Li X; Knox JE; Hratchian HP; Cross JB; Bakken V; Adamo C; Jaramillo J; Gomperts R; Stratmann RE; Yazyev O; Austin AJ; Cammi R; Pomelli C; Ochterski JW; Ayala PY; Morokuma K; Voth GA; Salvador P; Dannenberg JJ; Zakrzewski VG; Dapprich S; Daniels AD; Strain MC; Farkas O; Malick DK; Rabuck AD; Raghavachari K; Foresman JB; Ortiz JV; Cui Q; Baboul AG; Clifford S; Cioslowski J; Stefanov BB; Liu G; Liashenko A; Piskorz P; Komaromi I; Martin RL; Fox DJ; Keith T; Al-Laham MA; Peng CY; Nanayakkara A; Challacombe M; Gill PMW; Johnson B; Chen W; Wong MW; Gonzalez C; Pople JA Gaussian 03, Revision D.02; Gaussian, Inc.: Wallingford, CT, 2004.

(71). Seabra G. d. M.; Walker RC; Elstner M; Case DA; Roitberg AE Implementation of the SCC-DFTB Method for Hybrid QM/MM Simulations within the Amber Molecular Dynamics Package. J. Phys. Chem. A 2007, 111, 5655–5664. [PubMed: 17521173]

(72). Walker RC; Crowley MF; Case DA The implementation of a fast and accurate QM/MM potential method in Amber. J. Comput. Chem 2008, 29, 1019–1031. [PubMed: 18072177]

(73). Case D; Berryman J; Betz R; Cerutti D; Cheatham T; Darden T; Duke R; Giese T; Gohlke H; Goetz A; Homeyer N; Izadi S; Janowski P; Kaus J; Kovalenko A; Lee T; LeGrand S; Li P; Luchko T; Luo R; Madej B; Merz K; Monard G; Needham P; Nguyen H; Nguyen H; Omelyan I; Onufriev A; Roe D; Roitberg A; Salomon-Ferrer R; Simmerling C; Smith W; Swails J; Walker R; Wang J; Wolf R; Wu X; York D; Kollman P AMBER 2015; University of California: San Francisco, 2015.

(74). Davidson MM; Hillier IH; Vincent MA The Claisen rearrangement of allyl vinyl ether in the gas phase and aqueous solution. Structures and energies predicted by high-level ab initio calculations. Chem. Phys. Lett 1995, 246, 536–540.

(75). Zhang J; Yang YI; Yang L; Gao YQ Dynamics and Kinetics Study of "In-Water" Chemical Reactions by Enhanced Sampling of Reactive Trajectories. J. Phys. Chem. B 2015, 119, 14505–14514. [PubMed: 26485567]

(76). 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. [PubMed: 26574419]

(77). Kolb B; Zhao B; Li J; Jiang B; Guo H Permutation invariant potential energy surfaces for polyatomic reactions using atomistic neural networks. J. Chem. Phys 2016, 144, 224103. [PubMed: 27305992]

(78). Pukrittayakamee A; Malshe M; Hagan M; Raff LM; Narulkar R; Bukkapatnum S; Komanduri R Simultaneous fitting of a potential-energy surface and its corresponding force fields using feedforward neural networks. J. Chem. Phys 2009, 130, 134101. [PubMed: 19355711]

(79). Csányi G; Albaret T; Payne MC; De Vita A "Learn on the Fly": A Hybrid Classical and Quantum-Mechanical Molecular Dynamics Simulation. Phys. Rev. Lett 2004, 93, 175503. [PubMed: 15525089]

(80). Botu V; Ramprasad R Learning scheme to predict atomic forces and accelerate materials simulations. Phys. Rev. B: Condens. Matter Mater. Phys 2015, 92, 094306.

(81). Li Z; Kermode JR; De Vita A Molecular Dynamics with On the-Fly Machine Learning of Quantum-Mechanical Forces. Phys. Rev. Lett 2015, 114, 096405. [PubMed: 25793835]
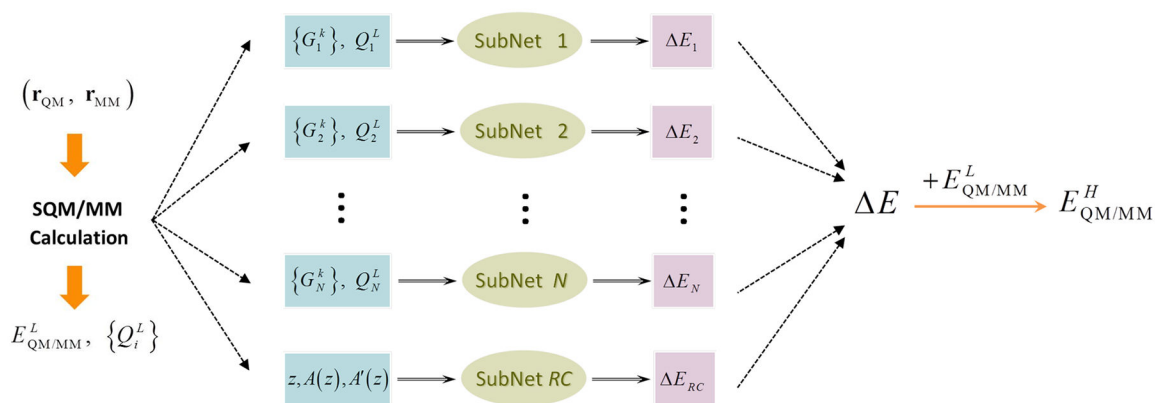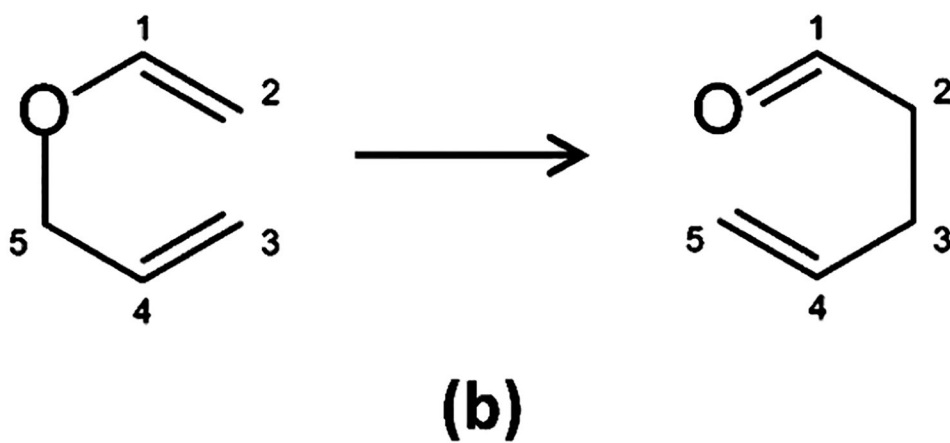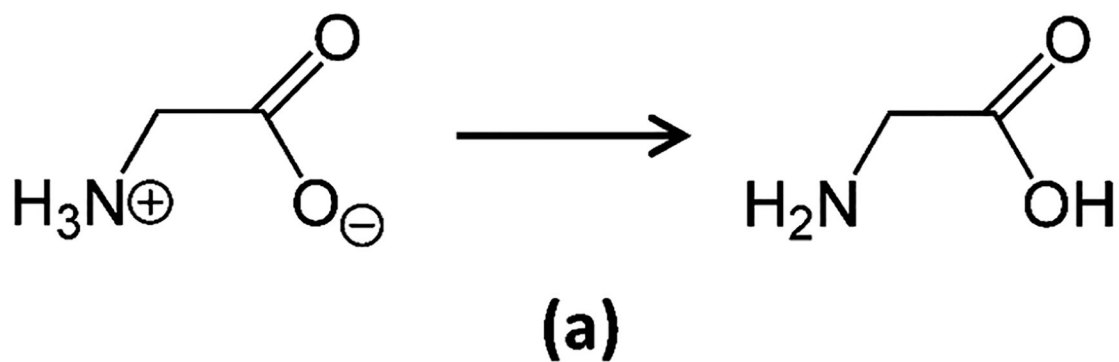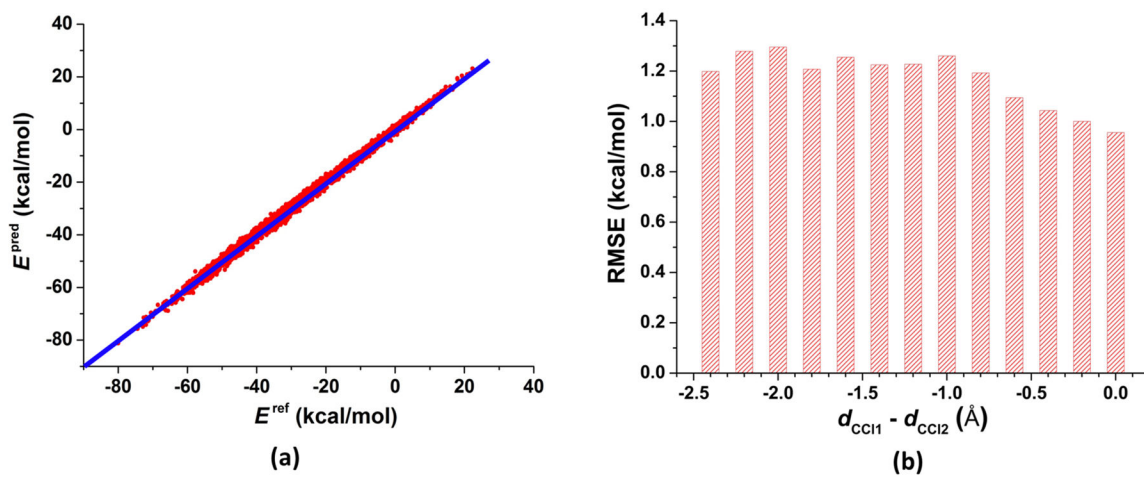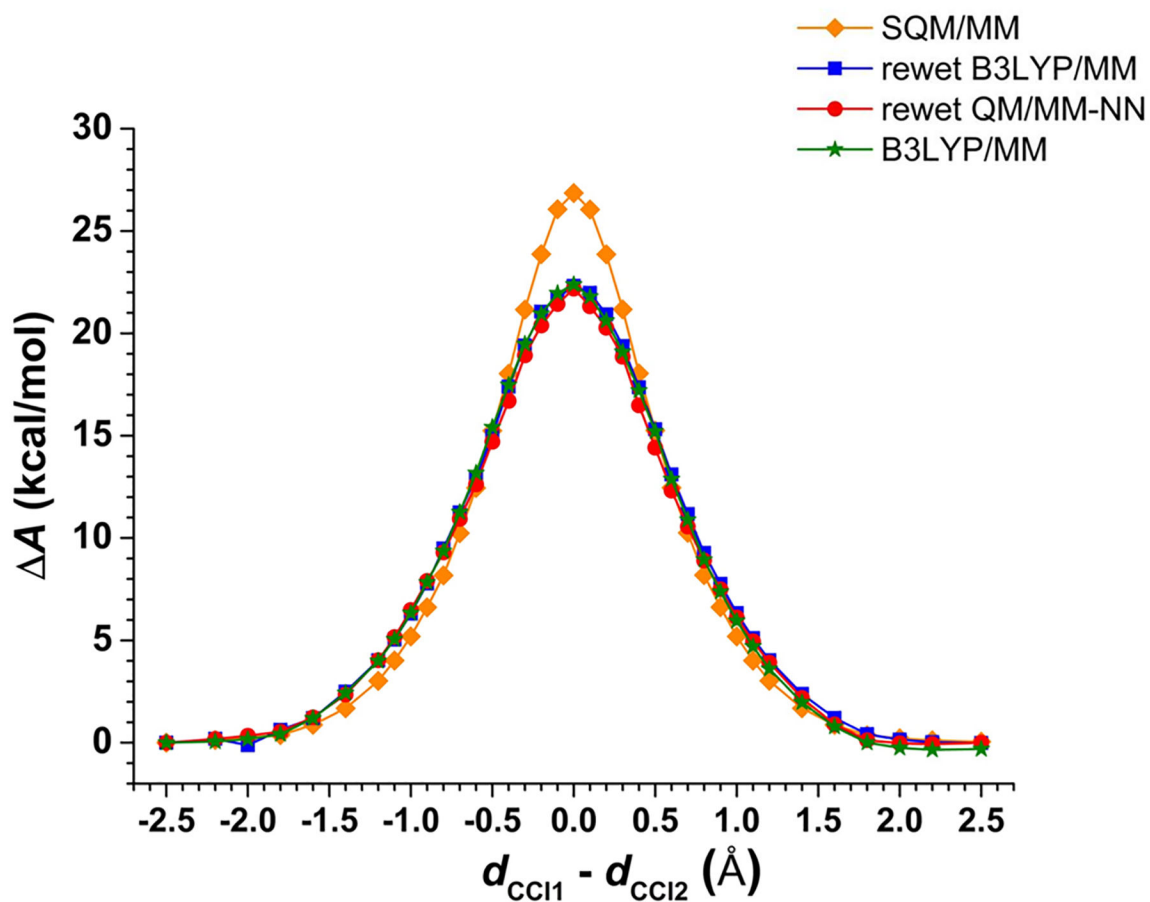
Figure 1.
Schematic structure of a QM/MM-NN for a system containing N atoms in the QM subsystem. Here $\mathbf{r}_{QM}$ and $\mathbf{r}_{MM}$ are respectively the Cartesian coordinates of the atoms in the QM and MM subsystem, $\{G_i^k\}$ is the symmetry function that depends on $\mathbf{r}_{QM}$, and z is the reaction coordinate as a function of $\mathbf{r}_{QM}$. After semiempirical QM/MM calculations with $\mathbf{r}_{QM}$ and $\mathbf{r}_{MM}$, the total QM/MM potential energy $E_{QM/MM}^L$ at the low level and the Mulliken atomic charges Q_i^L that have been polarized by the MM environment are known. The approximate free-energy profile as $A(z)$ and its first derivative with respect to z are also provided from the low-level QM/MM MD simulations. Then the energy difference between two levels as E is predicted with QM/MM-NN. Finally, the total QM/MM potential energy $E_{QM/MM}^H$ at the high level is obtained.
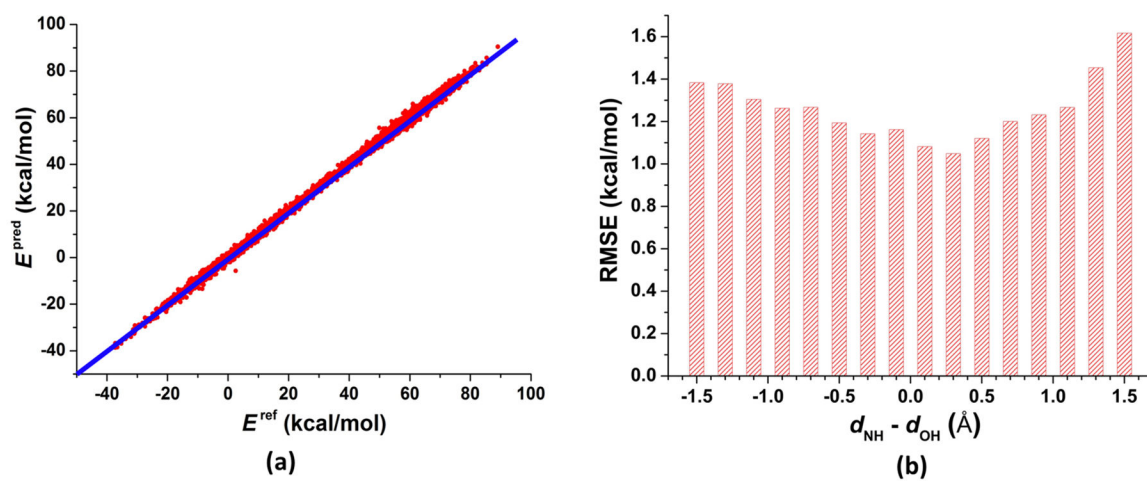
Figure 2.
(a) Glycine intramolecular proton transfer reaction from zwitterion form (left) to neutral form (right). (b) Aliphatic Claisen rearrangement reaction of allyl vinyl ether (AVE).
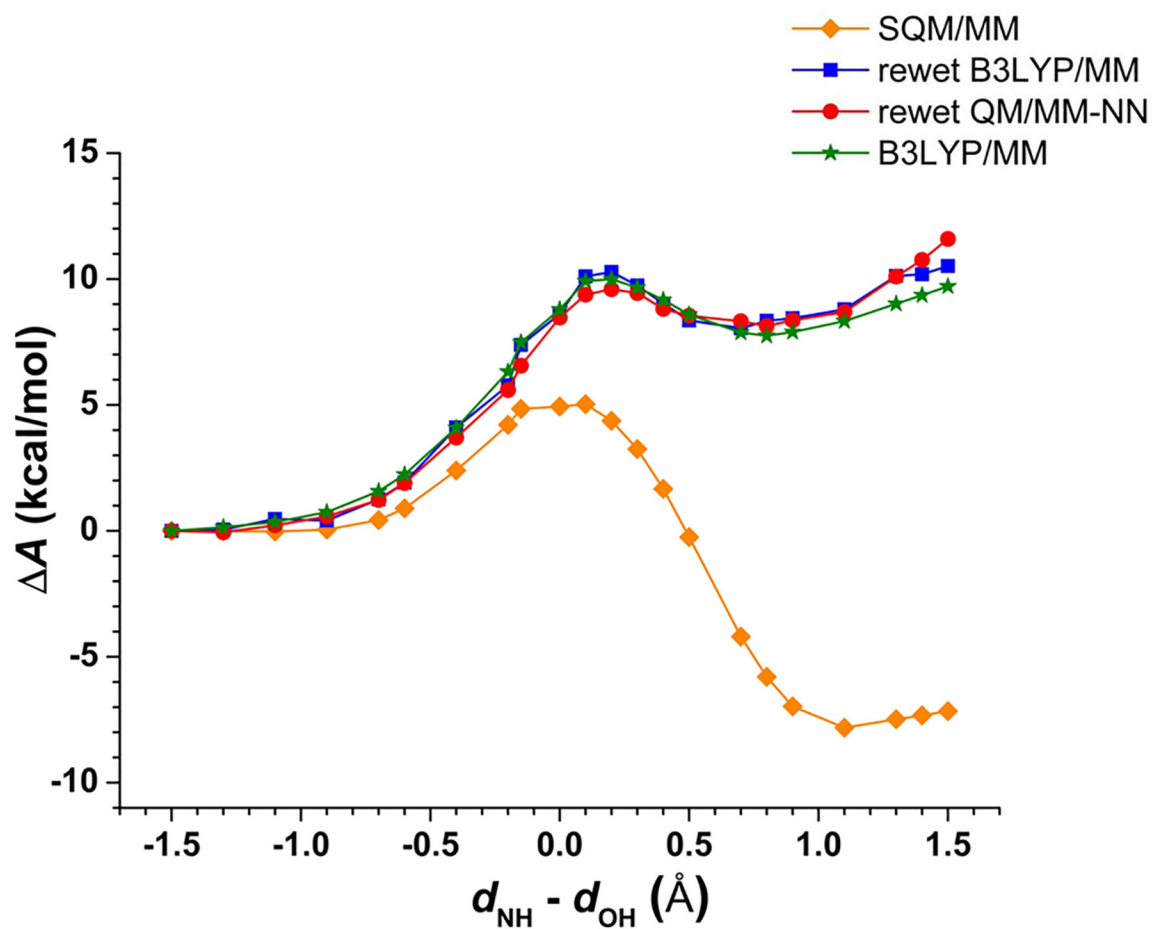
Figure 3.
Accuracy of QM/MM-NN for the S_N2 reaction. (a) Comparison of QM/MM-NN predicted potential energies (E^{pred}) with that obtained from B3LYP/6–31G(d)/MM calculations (E^{ref}). (b) Distribution of RMSEs along the reaction coordinate.
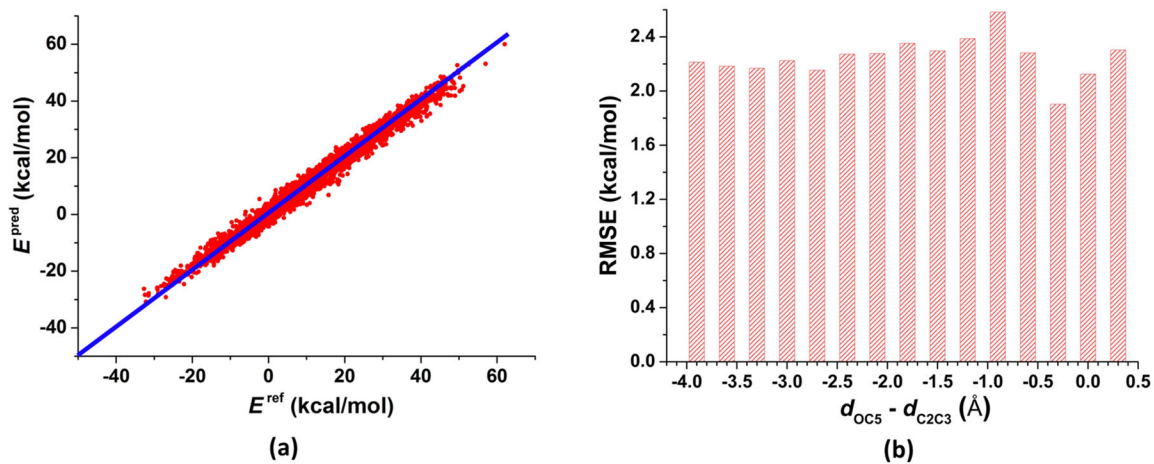
Figure 4.
Potential of mean force for the S_N2 reaction. Different colors and shapes represent different methods (orange diamond: direct SCC-DFTB/MM MD; blue square: reweighted with B3LYP/6–31G(d)/MM potentials; red circle: reweighted with QM/MM-NN predicted potentials; green star: direct B3LYP/6–31G(d)/MM MD).

Figure 5.
Accuracy of QM/MM-NN for the proton transfer reaction of glycine. (a) Comparison of QM/MM-NN predicted potential energies (E^{pred}) with that obtained from B3LYP/6–31G(d)/MM calculations (E^{ref}). (b) Distribution of RMSEs along the reaction coordinate.
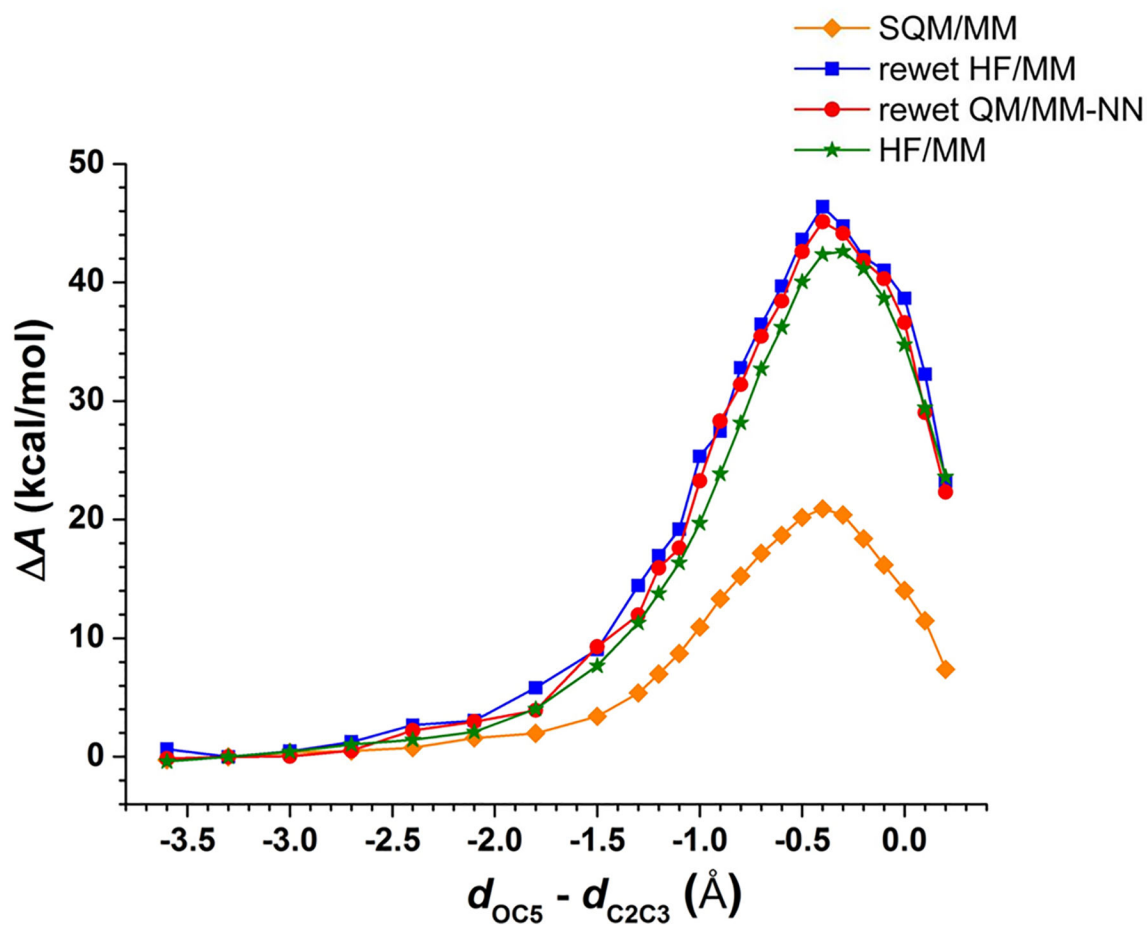
Figure 6.
Potential of mean force for the proton transfer reaction of glycine. Different colors and shapes represent different methods (orange diamond: direct SCC-DFTB/MM MD; blue square: reweighted with B3LYP/6–31G(d)/MM potentials; red circle: reweighted with QM/MM-NN predicted potentials; green star: direct B3LYP/6–31G(d)/MM MD).

Figure 7.
Accuracy of QM/MM-NN for the Claisen rearrangement reaction of AVE. (a) Comparison of QM/MM-NN predicted potential energies (E^{pred}) with that obtained from HF/6–31G(d)/MM calculations (E^{ref}). (b) Distribution of RMSEs along the reaction coordinate.

Figure 8.
Potential of mean force for the Claisen rearrangement reaction of AVE. Different colors and shapes represent different methods (orange diamond: direct SCC-DFTB/MM MD; blue square: reweighted with HF/6–31G(d)/MM potentials; red circle: reweighted with QM/MM-NN predicted potentials; green star: direct HF/6–31G(d)/MM MD).

Table 1.

Root Mean Squared Errors (kcal/mol) of Training and Testing Sets for S_N2 Reaction with Q^2 Values (in Parentheses)[b]

	training set	testing set
20[a]	1.15	1.16 (0.717)
30	1.13	1.17 (0.714)
40	1.12	1.18 (0.713)
60	1.11	1.14 (0.728)
80	1.09	1.15 (0.727)

[a]Number of configurations selected from each window for training. For the S_N2 reaction the total number of samplings in the training set is the product of 19 and this number.

[b]Ab initio QM/MM potential energies for the S_N2 reaction are predicted with QM/MM-NN. Five training sets with different sizes and one testing set are used.

Table 2.

Root Mean Squared Errors (kcal/mol) of Training and Testing Sets for Proton Transfer Reaction of Glycine with Q^2 Values (in Parentheses)[b]

	training set	testing set
20[a]	1.22	1.25 (0.971)
40	1.20	1.30 (0.969)
80	1.28	1.29 (0.970)

[a]Number of configurations selected from each window for training. For the proton reaction of glycine the total number of samplings in the training set is the product of 25 and this number.

[b]Ab initio QM/MM potential energies for the proton transfer reaction of glycine are predicted with QM/MM-NN. Three training sets with different sizes and one testing set are used.

Table 3.

Root Mean Squared Errors (kcal/mol) of Training and Testing Sets for Claisen Rearrangement Reaction of AVE with Q^2 Values (in Parentheses)[b]

	training set	testing set
20[a]	2.08	2.36 (0.950)
30	2.05	2.21 (0.955)
40	2.00	2.22 (0.955)
50	1.95	2.21 (0.955)

[a] Number of configurations selected from each window for training. For the Claisen rearrangement reaction of AVE the total number of samplings in the training set is the product of 28 and this number.

[b] Ab initio QM/MM potential energies for the Claisen rearrangement reaction of AVE are predicted with QM/MM-NN. Four training sets with different sizes and one testing set are used.