Neural Network Potentials for Atomistic Simulations of Reactive Chemistry

A THESIS

SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL

OF THE UNIVERSITY OF MINNESOTA

BY

Adrian M. Gordon

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

Doctorate of Philosophy

Jason D. Goodpaster

May, 2024

Acknowledgements

I would like to acknowledge everyone who has helped me, including my advisor Jason Goodpaster, my thesis committee, and my many colleagues in the chemistry department at the University of Minnesota.

Dedication

I dedicate this to Robert Schumann for shining some moonlight upon my wanderings through the dull formulae and theorems of science.

Contents

List of Tables

List of Figures

Chapter 1

Constructing Datasets for the Training of Neural Network Potentials for Rate Constant Calculations

1.1 Introduction

Atomistic simulation is a highly effective tool for the study of chemistry. In complex systems, for which there exist no analytical solutions for properties of interest, molecular dynamics and Monte Carlo simulations provide a means of calculating these properties using the power of modern computers. Simulations have the ability to visit a vast number

of configurations, providing a route to useful properties such as equations of state, free energies, as well as detecting dynamical changes in molecular structure.[4] Atomistic simulations have been applied to a diverse range of chemical fields, including catalysis,[5–7] biochemical applications such as protein folding,[8,9] and polymer science.[10,11]

The accuracy of results obtained from atomistic simulations depends on the interatomic potential energy, often referred to as the potential energy surface. Ab initio quantum mechanical calculations provide the most accurate potential energy surfaces. However, the cost of quantum mechanical calculations is prohibitively expensive, especially for long simulations or systems with a large number of atoms. Therefore, force fields, analytical representations of the potential energy surface parametrized with experimental or ab initio data, have traditionally been used in the simulation community. While force fields have been applied with success to a wide range of systems, their accuracy is generally limited, and each new system they are applied to requires intensive development efforts in order to obtain a reasonable description of the potential energy surface. In addition force fields generally cannot model large changes in the electronic structure such as bond making and breaking. Reactive force fields have been developed that are able to model chemical reactivity,[12] but they still require intensive development efforts and approximations inherent in choosing a single functional form for the potential energy surface limit their accuracy.[13]

Another approach for the evaluation of interatomic potential energies is the use of highly flexible functions. This approach is sometimes termed "mathematical", because

the functional forms are designed without regard for the physics of the system under study.[13] Instead they contain a vast number of parameters that allow them to accurately model very complex high dimensional potential energy surfaces. Machine learning is the leading method in the "mathematical" approach. A variety of machine methods have been used for potential energy surfaces,[14–30] including forward neural networks,[31–40] graph neural networks,[41–45] Gaussian processes,[34,46–48] Kernel methods,[49,50] and the moment tensor potential.[51,52]

There are many advantages of machine learning potentials. They are fast to evaluate. For example, they have been shown to evaluate energies of organic molecules up to 5 orders of magnitude faster than density functional theory (DFT).[31] They also can be extremely accurate for atomic configurations similar to the training set. In contrast to force fields where approximations implicit in the functional form limit accuracy, machine learning methods employ extremely flexible functions, which can represent any high dimensional potential energy surface, provided the reference data used to train the model is accurate. In addition, neural networks promise significant reduction in the human effort needed to generate the potential energy surface. Because they do not require careful selection of a functional form, it is possible to imagine the training of neural networks becoming an automated, "black box" process. For these reasons, machine learning potentials are receiving more and more attention in the field of chemistry.

Of the machine learning techniques, neural networks have probably received the most attention for potential energy surfaces. In 2007, Behler and Parinello introduced the

high dimensional neural network, which decomposes the energy of a system into atomic contributions.[53] Since the development of high dimensional neural network potentials (HDNNP), they have been applied to a large variety of chemical systems, including proteins,[54,55] protonated water clusters,[56–58] and metal surfaces.[59–61] HDNNPs have been developed for some specific catalytic and organic reactions such as the dissociation of N_2 over a Ru(0001) surface,[62] and the Claisen rearrangement of allyl vinyl ether to 4-pentenal. Despite these studies, the application of neural network potentials to simulations that involve reactive chemistry has yet to be fully explored.

The generation of balanced datasets for training is critical to the success of PESs. Training data within the field of quantum chemistry is not inexpensive to obtain, particularly for chemical systems with a large number of atoms. Active learning is a commonly used method that iteratively selects data to be added to the training dataset.[59,63,64] It has been shown that active learning strategies can significantly reduce the size of training datasets while improving model accuracy.[65]

The ultimate test for a machine learning model is whether it can be used to calculate physical quantities of interest. Machine-learned PESs are well-suited for applications where the number of model evaluations needed for desired simulations far surpasses the quantum mechanical energy calculations required to create the training dataset.[66–71] Approximate quantum simulation techniques, which demand a high number of force calls, are thus an ideal application of machine learning potentials.[72–74]

Ring polymer molecular dynamics, is an approximate quantum simulation technique,

which employs classical molecular dynamics in an extended phase space. When combined with the Bennet-Chandler method for calculating rates it offers an efficient and accurate route for obtaining thermal rate coefficients.[75–79] Because RPMD simulations involve several copies of the physical system, a large number of force evaluations are required for each step of the simulation. Therefore, along with the development of new numerical integration schemes,[80,81] machine learning PESs are important for the widespread application of RPMD.

In this work, we investigate the construction of a machine learning potential for the reaction shown in Figure 1,

$$\mathrm{OH} + \mathrm{CH_4} \longrightarrow \mathrm{CH_3} + \mathrm{H_2O} \tag{1.1}$$

$$r = k[\mathrm{CH_4}][\mathrm{OH}] \tag{1.2}$$

This reaction is critical in both atmospheric and combustion chemistry, and its kinetics have been extensively studied using a variety of methods, including variational transition state theory with multi-dimensional tunneling,[2,82,83] quasi-classical trajectories method,[3] and RPMD.[81,82,84] Permutation invariant polynomials neural network (PIP-NN) potentials[2] have also been previously combined with RPMD[81] and have shown great accuracy in the prediction of the thermal rate constants. However, the PIP-NN was developed to obtain more than just the thermal rate constant, and the authors used

over 135,000 high-level ab initio points to fit the full 15-dimensional surface.[2] Therefore, it remains an open question as to how much data is really required to accurately predict thermal rate constants using RPMD.

We show that accurate neural network potentials (NNPs) can be trained on datasets of limited size and discuss active and transfer learning strategies for constructing small but representative training datasets. We validate our machine learning model through the calculations of the RPMD thermal rate constants. We choose feed-forward neural networks with the Behler-Parrinello symmetry functions as features for the majority of this study because of their efficiency and ability to scale to larger systems.[85] However, we also compare the Behler-Parinello model with the recently proposed NequIP model.[41]

1.2 Theory

1.2.1 Interatomic Potentials

Accurate evaluation of internuclear forces is critical for molecular dynamics. Traditionally, the internuclear forces are either determined directly from quantum mechanical calculations (ab initio molecular dynamics) or they are obtained from a parametrized form of the potential energy surface known as the force field. The variants of ab initio molecular dynamics can be derived from quantum first principles.[86] The time dependent

Schrödinger equation describing the time evolution of a molecular system is

$$i\hbar\frac{\partial}{\partial t}\Phi\left(\mathbf{r},\mathbf{R};\mathbf{t}\right)=\mathscr{H}\left(\mathbf{r},\mathbf{R}\right)\Phi\left(\mathbf{r},\mathbf{R};\mathbf{t}\right) \tag{1.3}$$

The molecular Hamiltonian $\mathscr{H}$ is given by

$$\begin{aligned}\mathscr{H}\left(\mathbf{r},\mathbf{R}\right)=&-\sum_{I}\frac{\hbar}{2M_I}\nabla_I^2-\sum_{i}\frac{\hbar}{2m_e}\nabla_i^2+\frac{1}{4\pi\epsilon_0}\sum_{i<j}\frac{e^2}{*r_j-r_i}\\&-\frac{1}{4\pi\epsilon_0}\sum_{i,I}\frac{e^2Z_I}{*R_I-r_i}+\frac{1}{4\pi\epsilon_0}\sum_{I<J}\frac{e^2Z_IZ_J}{*R_J-R_I}\\=&-\sum_{I}\frac{\hbar}{2M_I}\nabla_I^2+\mathscr{H}_e\left(\mathbf{r},\mathbf{R}\right)\end{aligned} \tag{1.4}$$

where r_i and R_I represent the electronic and nuclear coordinates M_I the nuclear mass, m_e the mass of an electron, Z_I the nuclear charge, and e the elementary charge. The Born-Oppenheimer (BO approximation allows certain terms in the full time-dependent Schrodinger equation to be ignored. Physically, the BO approximation is based on the fact that since the nuclei have much larger mass than the electrons, they move much more slowly. Using the BO approximation, the time-dependent Schrödinger equation may be reduced to

$$\left[-\sum_{I}\frac{\hbar}{2M_I}\nabla_I^2+E_k\left(\mathbf{R}\right)\right]\chi_k=i\hbar\frac{\partial}{\partial t}\chi_k \tag{1.5}$$

and the total wave function written as a product of electronic and nuclear functions

$$\Phi\left(\mathbf{r},\mathbf{R};\mathbf{t}\right)=\Psi_k\left(\mathbf{r};\mathbf{R}\right)\chi_k\left(\mathbf{R};\mathbf{t}\right) \tag{1.6}$$

where it is often assumed that state k is the ground state of the system. The BO approximation separates the task of solving the Schrödinger equation into two separate parts. One part involves the calculation of the "clamped nuclei" electronic wave function of equation,

$$\mathscr{H}_e \Psi_k \left(\mathbf{r};\mathbf{R}\right) = E_k \left(\mathbf{R}\right) \Psi_k \left(\mathbf{r};\mathbf{R}\right) \tag{1.7}$$

which depends only parametrically on the nuclei. The second part involves solving for the time evolution of the nuclei according to equation 1.5, The total electronic energy $E_k \left(\mathbf{R}\right)$ in equation 1.5 serves as the potential energy for the nuclei and is thus often called the potential energy surface.

To arrive at the usual formulation of Born-Oppenheimer molecular dynamics (BOMD), the nuclei are treated as classical particles, which is typically a good approximation except for the lightest nuclei. It can be shown that in the classical limit, the nuclei follow Newton's equations of motion

$$M_I \ \ddot{\mathbf{R}}_\mathbf{I} \left(t\right) = -\nabla_I E_k \left(\mathbf{R}_\mathbf{I} \left(t\right)\right) \tag{1.8}$$

In a BOMD simulation, E_k is found by solving the time independent Schrodinger equation for the electronic energy at each step of the simulation using an electronic structure method such as Hartree Fock or density functional theory.[86] Other approaches to ab initio molecular dynamics exist such as Ehrenfest dynamics, in which the time-dependent Schrödinger equation is solved simultaneously for the electrons.[87]

Instead of solving for the electronic energy "on-the-fly" during a simulation, an analytical functional form of the potential energy surface may be used to provide the forces. This analytical potential energy surface, often called a force field, is obtained by fitting a functional form with data from either experiment or ab initio calculations. Force fields typically break the total energy into various intra- and inter-molecular contributions. A basic force field usually includes bond stretching, angle bending, torsion, van der Waals, electrostatic and cross terms:[88]

$$E = E_{\text{str}} + E_{\text{bend}} + E_{\text{tors}} + E_{\text{vdw}} + E_{\text{el}} + E_{\text{cross}} \tag{1.9}$$

Each contribution has its own analytic form, for example bond stretches are often represented by Taylor series truncated at the quadratic or quartic terms, and van der Waals interactions are often represented by Lennard-Jones potentials. In general, terms may be added and the exact functional forms of each term changed in order to better describe the system of interest.

1.2.2 Neural Network Potentials

For the fitting of the potential energy surface, we use an NNP with inputs constructed from the symmetry functions first introduced by Behler and Parrinello.[53] In this NNP architecture, the total energy of the system is decomposed into atomic energies. Separate atomic neural network evaluations are performed to obtain atomic energies for each atom in the system, and these atomic energies are added together to obtain the total system

energy. The input for the atomic neural networks consists of vectors composed of the Behler-Parrinello radial and angular symmetry functions, describing the environment around each atom and ensuring invariance with respect to translation, rotation, and permutation of like atoms.[31]

The choice of atomic representations is extremely important for NNP performance. The obvious choice for NPPs, Cartesian coordinates, suffers from a number of issues.[89] Each additional atom leads to three more degrees of freedom included as input nodes, making the cost of training scale poorly with system size. In addition Cartesian coordinates are not translationally and rotationally invariant. Use of internal coordinates solves the problem of translational and rotational invariance, but variance with respect to the permutation of two atoms remains. In 2007, Behler and Parinello introduced high-dimensional neural networks that use set of functions called symmetry functions as inputs, overcoming the shortcomings of the atomic representations listed above.[53] A high-dimensional NNP consists of a separate NN for each atom in a system. The individual NNs return atomic energies which are summed to find the total energy. The total energy of a system is

$$E_s = \sum_{\nu=1}^{N_{\text{elem}}} \sum_{\mu=1}^{N_{\text{atoms},\nu}} E_\mu^\nu \tag{1.10}$$

where E_μ^{nu} is the atomic energy for the μth atom of the νth element.

For inputs, high-dimensional NNs use a set of symmetry functions which probe an atom's short range chemical environment within a cutoff radius R_c. The input vector

for each atomic NN consists of elements G_m^R and G_m^A describing the radial and angular environments respectively of the central atom. The radial symmetry functions are a sum of products between the Gaussian and the cutoff function:

$$G_m^R = \sum_j e^{-\eta(R_{ij}-Rs)^2} f_c\left(R_{ij}\right) \tag{1.11}$$

where $f_c\left(R_{ij}\right)$ is the cutoff function:

$$f_c\left(R_{ij}\right) = \begin{cases} 0.5\left[\cos\left(\frac{\pi R_{ij}}{R_c}\right)+1\right] & \text{for } R_{ij} \leq R_c \\ 0 & \text{for } R_{ij} > R_c \end{cases} \tag{1.12}$$

and m is an index over different values for the parameters R_s and η. η controls the width of the Gaussian functions and R_s the angular shell that is being probed. In addition the atomic environment is described by a set of angular symmetry functions

$$G_m^A = 2^{1-\xi} \sum_j \sum_{k \neq j} \left[1+\lambda\cos\left(\phi_{ijk}\right)\right]^\xi f_c\left(R_{ij}\right) f_c\left(R_{ik}\right) f_c\left(R_{jk}\right) \tag{1.13}$$

where ϕ_{ijk} is the angle formed between atoms i, j and k, and ξ is a parameter that allows the width of the peaks to be controlled. Variations on the standard Behler Parinello symmetry functions have been proposed.[13] Smith et al. found that modifying the angular symmetry functions to include additional parameters allows for greater resolution of the atomic environment.[31] They replaced the original angular symmetry function with

the following:

$$G_m^A = 2^{1-\xi} \sum_{j,k \neq i} \left[1 + \cos\left(\theta_{ijk} - \theta_s\right)\right]^\xi \times \exp\left[-\eta \frac{\left(R_{ij} - R_{ik}\right)^2}{2} - R_s\right] f_c\left(R_{ij}\right) f_c\left(R_{ik}\right) \tag{1.14}$$

Including ϕ_s allows specific regions of angular space to be probed. In addition, they found using input vectors that distinguish the atomic number of atoms in the environment lowered error and improved transferability of neural networks. It is important to note that NNPs are everywhere differentiable with respect to the symmetry functions and the Cartesian coordinates of the atoms. Therefore, forces may be easily calculated from the NNP, for use in geometry optimizations and molecular dynamics simulations.

1.2.3 Active and Transfer Learning

A commonly used active learning method is query by committee, in which the disagreement between an ensemble of models is used to identify the confidence of the model's predictions. Configurations with high ensemble disagreement (low confidence) are added to the training dataset, thus avoiding the inclusion of redundant configurations. To implement active learning, an ensemble of machine learning models is trained on the same training data. The ensemble is used to run molecular dynamics simulations, from which configurations with the highest ensemble variance are selected. Ab initio energies are calculated for the selected configurations, added to the training dataset, and the model

is retrained.

Transfer learning is a technique in which a model trained on one task is retrained on a different but related task. It has been shown that neural network potentials pretrained on DFT energies can be fine-tuned on a small targeted dataset of CCSD(T) energies to produce a network that predicts CCSD(T) energies with high accuracy.[90]

1.2.4 Ring Polymer Molecular Dynamics

Typical molecular dynamics simulations approximate nuclei as classical particles, ignoring their quantum nature. This approximation works well in many cases, but breaks down for lighter nuclei and lower temperatures, when quantum effects such as zero point energies and tunneling become important. In addition the quantum nature of nuclei is critical for studying equilibrium isotope effects, which form a basis for many experimental studies of reaction mechanisms.[91]

Ring polymer molecular dynamics (RPMD) is a commonly used method for including quantum nuclear effects in simulations. The basic method exploits a similarity between the quantum mechanical partition function and a classical partition function for a system of interacting particles.[92] The quantum mechanical partition function can be written as the trace of the density matrix:

$$Z = \sum_{n} \exp\left[-\beta E_n\right] = \mathrm{tr}(\rho) \tag{1.15}$$

where the density matrix is

$$\rho = \sum_n \exp(-\beta E_n)\phi_n\phi_n \tag{1.16}$$

and ϕ_n are eigen states of the Hamiltonian. In the coordinate representation, Z is

$$Z = \int \mathrm{d}x x \exp(-\beta H)x \tag{1.17}$$

For this derivation it is assumed that H is the Hamiltonian for a single particle system $H = \hat{K} + \hat{V}$. In order to evaluate expectation value in equation 1.15, the factorization $\exp(-\beta H) = [\exp(-(\beta/M)H)]^M$ is made and the completeness relation is inserted $M-1$ times to arrive at

$$Z = \int \prod_i^{M-1} \mathrm{d}x_i \left\langle x_{i+1} \middle| \exp(\frac{-\beta}{M}H) \middle| x_i \right\rangle \tag{1.18}$$

Now the partition function is written as a product of density matrix elements. Due to the factor of M, these are often called high temperature density matrices. Equation 1.18 is exact in the limit of $M \to \infty$, but is often a good approximation for finite M. When M is sufficiently large the Trotter approximation can be made

$$\exp\left(-\beta H\right) = \lim_{M\to\infty} [\exp(-\beta K)\exp(-\beta V)]^M \tag{1.19}$$

Using this factorization of the Hamiltonian it can be shown that the high temperature

density matrix elements evaluate to

$$\rho\left(x_{i+1};x_i;\epsilon\right)=\left(\frac{m}{2\pi\epsilon\hbar^2}\right)^{1/2}\exp\left(-\epsilon\sum_i\frac{m\left(x_i-x_{i+1}\right)^2}{2\epsilon\hbar^2}\right)\times\exp\left(-\epsilon V\left(x_i\right)\right) \qquad (1.20)$$

where $\epsilon=\beta/M$.

Now substituting these high temperature density matrix elements into the equation 1.18, a transformed expression for the density matrix is reached

$$\begin{aligned}Z=&\left(\frac{m}{2\pi\epsilon\hbar^2}\right)^{1/2}\int\left(\prod_i^{M-1}\mathrm{d}x_i\right)\\&\times\exp\left(-\frac{m}{2\epsilon\hbar^2}\sum_i\left(x_i-x_{i+1}\right)^2-\epsilon\sum_i V\left(x_i\right)\right)\end{aligned}$$

Since the partition function is the trace of the density matrix, the cyclic condition $x_0=x_M$ is enforced. This expression can be interpreted in the framework of the path integral formulation, as a sum over all the paths that in an interval of imaginary time interval 0 to $\beta\hbar$, wander away from and come back to x_0. Furthermore equation is isomorphic to the partition function for a system of classical particles interacting through harmonic potentials with force constants $m/\epsilon\hbar^2$ and subject to an external potential $V(x_i)$. Due to the cyclic condition these particles form a "ring polymer" and are often referred to as beads. To complete the analogy to classical statistical mechanics, M momenta are introduced with arbitrary mass, finally leading to a classical system of

particles governed by the Hamiltonian

$$H = \sum_i \frac{p_i^2}{2m_c} + \frac{1}{2} m\omega_n^2 \left(x_i - x_{i+1}\right)^2 + V\left(x_i\right) \tag{1.21}$$

where $\omega = \sqrt{M}/\left(\beta\hbar\right)$. This expression for the quantum mechanical density matrix is the at the heart of ring polymer molecular dynamics. Molecular dynamics simulations may be run on this fictitious ring polymer in order to obtain equilibrium properties of the system, using formulas to average over the beads. Quantum reaction rates have been calculated by considering the free energy surface along a reaction coordinate defined by the geometric center (centroid) of the ring polymer.[93,94] Approximate quantum dynamics may also be studied using quantum time correlation functions.[75,95]

The expression for the full RPMD rate coefficient is

$$k(T) = \kappa(t \to \infty; s_1) k^{\mathrm{QTST}}(T; s_1) \tag{1.22}$$

where $\kappa(t \to \infty; s_1)$ is the recrossing factor and $k^{\mathrm{QTST}}(T; s_1)$ is the static factor (centroid-density quantum transition state theory rate constant).[76,78] In order to make the calculation of the of the rate constant computationally feasible, two dividing surfaces are defined. The first s_0 is located in the reactant valley. The s_1 is located in the transition state region. A reaction coordinate is defined which interpolates between the two dividing surfaces,

$$\eta = \frac{s_0(\bar{\mathrm{x}})}{s_0(\bar{\mathrm{x}}) - s_1(\bar{\mathrm{x}})} \tag{1.23}$$

where $\bar{\mathrm{x}}$ represents the coordinates of the ring polymer centroid.

To evaluate the recrossing factor, independent configurations are collected from a RPMD simulation with the system constrained to the transition state dividing surface. Those configurations are then used to initialize unconstrained trajectories, in order to count how many trajectories end up on either side of the dividing surface.[78] The transition theory rate constant may be evaluated as

$$k^{QTST}(T; s_1) = 4\pi R_\infty^2 \left(\frac{m_\mathrm{A} + m_\mathrm{B}}{2\pi\beta m_\mathrm{A} m_\mathrm{B}} \right)^{1/2} e^{-\beta[W(\eta^\ddagger) - W(0)]} \tag{1.24}$$

where $W(\eta^\ddagger)$ is the potential of mean force when the value of the reaction coordinate is equal to s_1. The potential of mean force can be easily evaluated using a method such as umbrella integration.[78]

1.3 Computational Details

The NNPs were trained on energies calculated at the UCCSD(T)/aug-cc-pVTZ level of theory.[96] Ensembles of 10 NNPs were used throughout the training process. All electronic software calculations were performed in the Q-Chem software package.[97]

To create an initial training dataset, the minimum energy path (MEP) for the reaction was obtained from the nudged elastic band method at the ωB97X/6-31G(d) level of theory.[98] For each of the 40 images along the MEP, normal mode coordinates were obtained. The normal mode sampling procedure outlined previously[31] was applied to

each nudged elastic band geometry, except in our work, the random displacements were drawn from a uniform distribution between -0.3 and 0.3, rather than a Boltzmann-style temperature-dependent distribution. CCSD(T) energies were calculated for the geometries in this dataset, and an ensemble of differently initialized neural networks was trained on the initial dataset.

An active learning approach[64,90] was then used to expand the training dataset. Umbrella sampling simulations were run with the preliminary ensemble of NNPs. The umbrella bias potentials were placed along the reaction coordinate from reactants and products, ensuring that data was included for the entire reaction pathway. The umbrella sampling simulations were performed using i-PI[99] interfaced with the PLUMED library[100] version 2.7.5[101] for the application of bias potentials. For each configuration of the umbrella sampling trajectories, the standard deviation of the outputs of the NNP ensemble was calculated, and configurations with the highest standard deviations were added to the training dataset. A new ensemble of neural networks was then trained on this combined dataset. High-energy configurations (> -115.85 Hartree) were excluded since their inclusion led to fits with higher errors on our validation sets. These high-energy configurations (140 kcal/mol above the reaction barrier) would be rarely sampled during the reaction; therefore, omitting them leads to better NNPs in the energy range important for this reaction.

The final training dataset contained 11,440 configurations and CCSD(T) energies. During each training iteration, the training dataset was split into a train (80%) and

validation set (20%). To test the accuracy of the final ensemble of NNPs, a test set was constructed containing 976 configurations randomly selected from RPMD trajectories generated using the final ensemble of NNPs. An ensemble of three out of the 10 trained NNPs was chosen for the final RPMD calculations. This ensemble of NNP had a test RMSE of 0.39 kcal/mol. It should be noted that the test and training datasets include configurations along the entire reaction pathway, including the product channel. The training and test datasets are available in a publicly accessible GitHub repository [https://github.com/GoodpasterGroup/CH4_OH].[102]

For transfer learning, a neural network was trained on 167,196 DFT energies calculated using the ωB97X functional[98] and the 6-31G(D)[103] basis set, with the active learning approach described above. Then, layers of the neural network were held fixed, and it was retrained on a smaller selection of data points pulled from the CCSD(T) dataset. Architectures for the various NNPs constructed for this project are presented in Tables 1.1-1.5.

Table 1.1: NNP architectures for NNP-CCSD(T) (hydrogen).

	Layer1	Layer2
Nodes	128	1
Activation	CELU	Linear

Table 1.2: NNP architectures for NNP-CCSD(T) (carbon and oxygen).

	Layer1	Layer2
Nodes	112	1
Activation	CELU	Linear

Table 1.3: NNP architectures for NNP-DFT, Transfer-1 and Transfer-3 (hydrogen).

	Layer1	Layer2	Layer3	Layer4
Nodes	160	128	96	1
Activation	CELU	CELU	CELU	Linear

Table 1.4: NNP architectures for NNP-DFT, Transfer-1 and Transfer-3 (carbon and oxygen).

	Layer1	Layer2	Layer3	Layer4
Nodes	144	112	96	1
Activation	CELU	CELU	CELU	Linear

Table 1.5: NNP architectures for No-Transfer (hydrogen, carbon and oxygen)

	Layer1	Layer2
Nodes	96	1
Activation	CELU	Linear

All rate calculations were performed with the RPMD rate package.[104] For all RPMD simulations performed in this work, 128 beads were used. The contribution of the $^2\Pi_{1/2}$ excited state of OH (140 cm^{-1})[105] to the electronic partition function was taken into account. Further details about the RPMD calculations are included in the Supporting

Information section. The NNP was implemented in TorchANI, an open-source PyTorch-based code.[106] TorchANI implements the neural network architecture and input features first utilized for the construction of ANAKIN-ME, a transferable neural network potential for organic molecules.[31]

To calculate the static factor from umbrella sampling, 120 umbrella windows were set up along the reaction coordinate. At each umbrella window, 8 trajectories of 80 ps duration were run, following an equilibration of 20 ps. For calculation of the recrossing factor, 4 constrained MD trajectories were run at the dividing surface for 120 ps, after an equilibration period of 20 ps. Configurations were sampled every 2 ps and 100 trajectories were launched from each of these configurations. Initial momenta were sampled from the Maxwell distribution. An Anderson thermostat was used for all RPMD simulations.

For training the Behler-Parrinello nueral network potentials, the TorchANI package was employed.[106] The network architectures are shown in Tables 1.1-1.5. 18 radial shifting parameters R_s were used for the radial symmetry functions and 4 radial shifting parameters and 8 angular shifting parameters ϕ_{ijk} were used for the angular symmetry functions. The CELU (Continuously-Differentiable Exponential Linear Unit)[107] was used as the hidden layer activation function. The parameters were initialized using the Kaiming normal initialization method.[108]

The hyperparameters used for constructing the NequIP models are shown in table 1.6.

Table 1.6: NequIP model parameters

cutoff	5.0 Å
number of interaction layers	4
maximum rotation order l_{max}	3
number of features	32
parity	even and odd
number of radial basis functions	8
number of layers in radial network	3
number of radial features	64

MSE loss functions were used for the optimization of the Behler-Parrinello and the NequIP models.

1.4 Discussion

The thermal rate constants obtained from the NNPs show good agreement with the experimental values across the examined temperature range (Table 1). The percent errors in comparison to the experimental values are 10%, 15%, and 19% at 1000 K, 300 K, and 200 K, respectively. The close agreement with the experiment at 1000 K is expected since RPMD is known to be exact in the high-temperature limit. The small error is most likely a result of errors in the potential energy surface (PES). It has been demonstrated theoretically that RPMD overestimates the rate for asymmetric reactions in the low-temperature deep tunneling regime.[109] However, our results show excellent

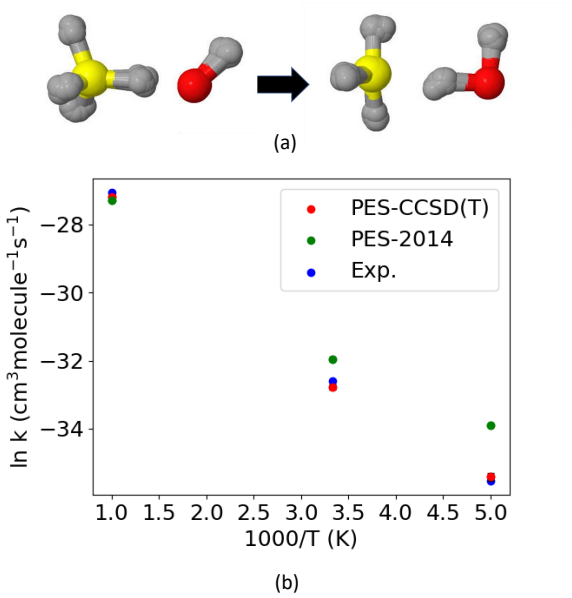

Figure 1.1: (a) The $OH + CH_4$ reaction (b) Arrehnius plot for the $OH + CH_4$ reaction, with rates calculated at 200 K, 300 K, and 1000 K using NNP-CCSD(T) and PES-2014.[3]

accuracy at 200 K and 300 K. These findings demonstrate the reliability of RPMD in the deep tunneling regime, although the close match between the calculated rates and the experiment may have been aided by the cancellation between errors in the PES and RPMD theory. The accuracy of RPMD for reaction (1) is in agreement with a recently published study, which also found that RPMD did not significantly overestimate the rate at low temperatures.[81]

The total statistical uncertainty in the rate constants was determined from the standard deviations of properties obtained from independent MD simulations. The statistical

Table 1.7: RPMD rate coefficients in $cm^3molecule^{-1}s^{-1}$ (and percent errors compared with experiment[1]) calculated using the neural network potential trained on coupled cluster data (NNP-CCSD(T)), the PES-2014 force field developed by Espinosa-García et al and the PIP-NN PES developed by Li et al.[2]

T (K)	NNP-CCSD(T)	PES-2014	PIP-NN	Experiment
200	4.39×10^{-16} (18.7)	1.91×10^{-15} (417)	5.05×10^{-16} (36.5)	3.70×10^{-16}
300	5.93×10^{-15} (-15.0)	1.33×10^{-14} (91.2)	8.37×10^{-15} (20.1)	6.97×10^{-15}
1000	1.55×10^{-12} (-10.7)	1.40×10^{-12} (-17.7)	1.70×10^{-12} (0.0)	1.70×10^{-12}

uncertainties were found to be within 11% of the values of the rates. The statistical uncertainty of the overall rate constant $k(T)$ (Eq. 1.22) was calculated from the propagation of error formula

$$\sigma^2_{k(T)} = \sigma^2_\kappa \left(\frac{\partial k}{\partial \kappa}\right)^2 + \sigma^2_{k^{\mathrm{QTST}}} \left(\frac{\partial k}{\partial k^{\mathrm{QTST}}}\right)^2 \tag{1.25}$$

where σ_x is the standard deviation of x obtained from four independent RPMD simulations. Using a 1-σ standard deviation, the percent uncertainties were 10.4%, 7.73% and 5.33% at 200 K, 300 K and 1000 K respectively.

The previously developed force field PES-2014[3] significantly overestimates the rates at 300 K and 200 K (Table 1). The PES-2014 force field does not adequately capture the reactant van der Waals energy well, leading to an underestimation of the free energy barrier.[81] In contrast, our NNP captures the van der Waals well and predicts a higher free energy at the transition state than does the PES-2014 (Figure 1.2). The accuracy of the NNP method was achieved while remaining relatively data-efficient. It required only 11,440 training data points to obtain an accurate global potential energy surface, including the product channel, suitable for molecular dynamics in the temperature range

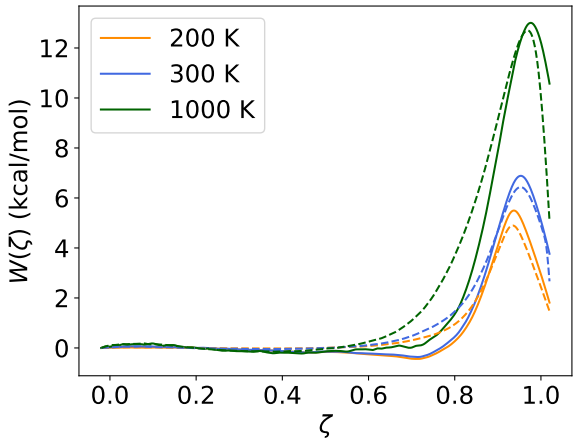

Figure 1.2: Potentials of mean force calculated with the neural network potential (solid line) and force field (dashed line) for the $OH + CH_4$ reaction at 200, 300, and 1000 K.

of 200 K - 1000 K.

In comparison with the previously developed PIP-NN PES,[2] our model was trained on significantly fewer training data points. This is expected because the PIP-NN PES was designed to be a global PES that performs well for molecular dynamics across a wider temperature range (not just 200-1000 K) and is suitable for the calculation of other dynamical properties such as cross sections.[2,110] However, the question of how much data is needed when targeting specific properties, such as the thermal rate coefficients, remains an open question. A more thorough study of data efficiency in training NNPs may be useful when approaching reactions for which a global PES is not already available and highly accurate calculations are desired.

To systematically examine the change in network accuracy with dataset size, we developed another training dataset using active learning. A fresh ensemble of networks

was trained on an initial dataset of 1211 normal mode sampling configurations. The active learning procedure was then performed as described in the methods section. Approximately 800 configurations were added in each round of active learning. Figure 1.3 shows the test RMSE and MAE of the NNPs at each round. The performance of the NNPs systematically improves before leveling off at an RMSE of 0.31 kcal/mol with > 7000 training data points. This may indicate that with over 7000 training examples that map the important configurational space, the accuracy is limited by aspects of the model used in this work, such as its input features and hyperparameters, rather than the extent of the training data set. Our original NNP, which was trained on 11440 configurations, had a similar accuracy of 0.39 kcal/mol, suggesting that the accuracy is limited by other aspects of the model and not the training data. We note that the active learning scheme was identical for this NNP and our original NNP, with the difference being that in the original NNP, we added more configurations at each round of active learning. This suggests that a smaller number of configurations added per round will lead to a smaller total number of configurations required; however, a more detailed analysis of this relationship is left to future work.

Graph neural networks have emerged as promising alternatives to traditional feed-forward neural networks for the construction of NNPs.[41–44] Rotational equivariant graph models have been shown to achieve high accuracy even when trained on small datasets. We compared the performance of the recently developed Neural Equivariant Interatomic

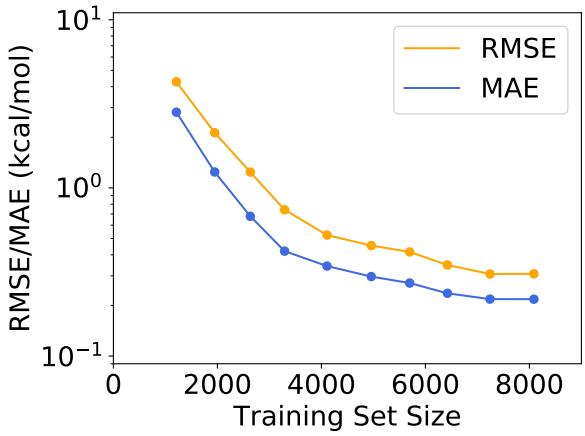

Figure 1.3: Log-log plot of test root mean square error (RMSE) and mean absolute error (MAE) of trained NNPs following each round of active learning.

Potentials (NequIP)[41] with the Behler Parrinello style networks (BP). The active learning procedure was performed with both models, selecting configurations from the previously created training dataset. Figure 4 shows the learning curves from the active learning process for both the NequIP and BP models. On small datasets, Nequip outperforms the BP model. However, on datasets larger than 3000 points, the Nequip model shows no advantage over the BP model. This suggests that for larger datasets, the BP model may be preferable since it is more computationally efficient, without a loss in accuracy. We note that the accuracy of both models could be improved with additional hyperparameter tuning.

To examine transfer learning, an ensemble of 10 NNPs (Table 4) was trained on a DFT dataset consisting of 167,196 configurations and energies. Then, the DFT NNPs

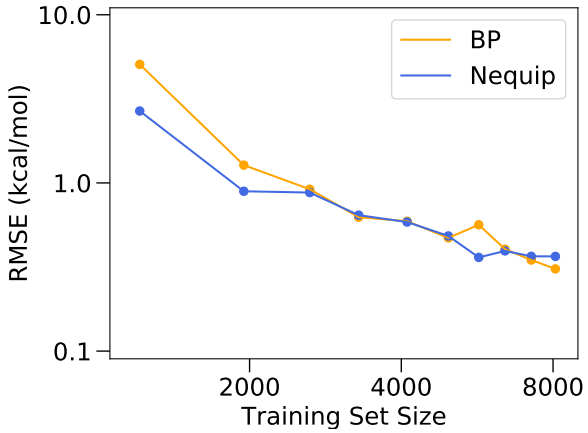

Figure 1.4: Log-log plot of the RMSE versus as a function of training dataset size for the NequIP and BP models.

were retrained on CCSD(T) data. This training was performed on two different ensembles of networks: one with three of the network's four hidden layers held fixed during training (Transfer-3), and the other with just one hidden layer held fixed (Transfer-1). For comparison, a smaller network (No-Transfer), without pre-training (parameters initialized using the Kaiming normal initialization method[108]), was trained on the same CCSD(T) data (Model architectures are displayed in Tables 1.1-1.5).

Figure 1.5 shows the test RMSEs of Transfer-1, Transfer-3, and No-Transfer, after training on CCSD(T) datasets of various sizes. For small training datasets, transfer learning greatly improved the NNP's performance on the test dataset, relative to the networks without previous training (No-Transfer). The No-Transfer NNPs overfit the data, leading to poor generalization on the validation dataset, while the transfer-learned networks were able to apply what they have learned about the DFT PES, allowing good

transferability for configurations not included in the small CCSD(T) dataset. However, as the training dataset becomes larger, No-Transfer begins to outperform the transfer-learned networks. The reason for the superior performance of No-Transfer when trained on larger CCSD(T) datasets is that it has greater flexibility to fit the data than the NNPs with fixed parameters. Even though Transfer-1 has a similar number of free optimizable parameters as No-Transfer, the fixed hidden layer encodes a particular transformation of the input features which may be unsuitable for fitting to the CCSD(T) target. These results show the advantages of a more flexible model when sufficient data is available and a less flexible pre-trained model when data is scarce. This trade-off is further demonstrated by comparing Transfer-1 and Transfer-3 NNPs. Transfer-1, which has only 1 hidden layer held fixed, performs worse than Transfer-3 on small training datasets (less than 500 data points) but better on large training datasets.

The improved accuracy of the rate calculation must be balanced with increases in computational time. We found that our CPU-based NNP is about two orders of magnitude slower than the CPU-based force field (PES-2014). The computational efficiency of NNPs and force fields could be improved through GPU parallelization and more efficient computer architectures. Despite this increase in required computational expense compared to traditional force fields, given that a couple trillion force calls were required to calculate one RPMD rate constant, this approach is still many orders of magnitude more efficient than direct CCSD(T) dynamics. Therefore, NNPs offer an appealing balance between accuracy and efficiency.

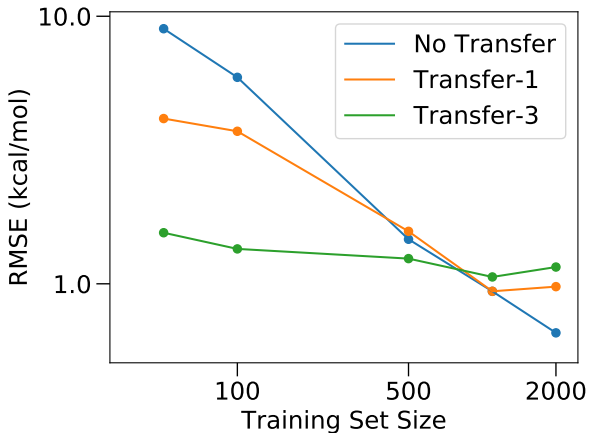

Figure 1.5: Test RMSEs of NNPs previously trained on DFT, then retrained on CCSD(T) datasets while one (Transfer-1) and three (Transfer-3) hidden layers were held fixed. In addition, RMSEs of a randomly initialized NNP (No-Transfer) trained on the same CCSD(T) datasets is shown.

Overall, the benefits of using neural network potentials over a force field include ease of construction and good accuracy throughout the PES. Highly accurate rates ($< 20\%$ error) were calculated using a PES trained on a very small dataset ($\approx 10,000$ points). Furthermore, BP NNPs have the advantage of being easily adapted to larger systems. Therefore, the remaining obstacle to tackling larger systems is the computational expense associated with dataset generation and neural network evaluations. The active learning strategies presented in this paper may be particularly advantageous for larger systems. The methodology outlined here paves the way for the application of RPMD rate theory to a wide range of chemical systems.

1.5 Conclusion

The combination of RPMD and neural network potential provides a powerful tool for the accurate calculation of gas-phase reaction rate coefficients. Importantly, we have shown that the total amount of training data for an accurate rate calculation is a relatively small number of calculations (10,000 CCSD(T) calculations) compared to the trillion of force calls required to perform the RPMD rate calculation. Despite the small number of training data points, we find that the error in the rate calculation using our NNPs to be around 20% at 200 K decreasing to 10% at 1000 K, demonstrating remarkable accuracy.

To further study what the minimum data limit is for these reactions, we explored transfer learning and active learning schemes. We found that for highly accurate NNPs, transfer learning where the NNPs were originally trained on DFT and then retrained on CCSD(T) was actually detrimental to the accuracy of the NNPs once the dataset size of the CCSD(T) calculations was larger than 1,000 data points. In contrast, we found that careful active learning can further reduce the amount of data required. Additionally, we compared Behler-Parrinello type models to NequIP. We found that in the small dataset regime, NequIP outperformed the BP model, but in the limit of large datasets, there was no significant difference between the two models' accuracies. These results suggest that when a sufficient amount of data is available, the choice of machine learning model type is not critically important. Therefore, we believe that future work in the field of machine learning potentials should focus on the development of high-quality training data.

Accurate rate calculations for significantly larger systems should be obtainable through a combination of NNPs, active learning, and RPMD. Both BP symmetry functions and graph-based models such as NequIP are known to scale to large systems, so if the number of required CCSD(T) calculations remains reasonable even for significantly larger systems than the one studied here, there should be no barrier for adoption of this approach for accurate rate calculations in small to medium-size systems. We expect that NNPs trained using active learning for specific reactions to have widespread adoption by the theoretical chemistry community for the accurate calculation of reaction rates.

Chapter 2

Machine Learning/Molecular Mechanics Methods for Condensed Phase Reactions

2.1 Ml/MM Schemes

In combined quantum mechanics/molecular mechanics (QM/MM) methods, a small part of a chemical system is simulated with quantum mechanics, while the rest of the system is simulated with molecular mechanics (MM) methods (a forcefield). This separation allows the critical part of the system, where electronic rearrangements or bond formation occur, to be modeled with quantum mechanics, and the effect of the environment can be included through molecular mechanics.[111]

The computation of the QM region remains a computational bottleneck in QM/MM calculations.[112] Recently the use of machine learning methods to describe the potential energy surfaces of the QM system has gained much attention.[55,113–117] Employing neural network potentials (NNPs) that predict the potential energy from descriptors of local atomic environments, we can replace the QM calculations with the NNP, allowing us to calculate forces at a fraction of the cost. Since the use of both NNPs and force-fields for molecular dynamics simulations is now wide-spread, the remaining challenge is developing methodologies for treating the interactions between the NNP and MM sub-systems.

A variety of ML/MM approaches have been proposed. Perhaps the simplest is the mechanical embedding strategy employed by Lahey et al.[55] In their method, the total energy of can be written as a sum of three terms,

$$V(\mathbf{r}) = V_{\mathrm{MM}}(\mathbf{r}_{\mathrm{MM}}) + V_{\mathrm{ML}}(\mathbf{r}_{\mathrm{ML}}) + V_{\mathrm{ML/MM}}(\mathbf{r}) \tag{2.1}$$

where $\mathbf{r}$ are the total coordinates, $\mathbf{r}_{\mathrm{MM}}$ is the coordinates of the MM region, $\mathbf{r}_{\mathrm{ML}}$ the coordinates of the ML region. V_{MM} is the energy provided by a force field, V_{ML} by a NNP, and $V_{\mathrm{ML/MM}}$ is a pair-wise interaction term between the MM and ML systems,

$$V_{\mathrm{ML/MM}}(\mathbf{r}) = \sum_{i}^{\mathrm{ML}} \sum_{j}^{\mathrm{MM}} \frac{q_i q_j}{4\pi\epsilon r_{ij}} + 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \tag{2.2}$$

This strategy was applied to compute binding poses and conformational free energies

for protein-ligand binding, where the ligand was modeled with a ML potential and the protein with a conventional forcefield.[55] Advantages of this method are its ease of implementation and that a general ML potential (ANI) could be employed, without modification. A limitation of this strategy is that the ML atoms are assigned fixed charges and Lennard-Jones parameters for the calculations of interactions between the ML and MM atoms. Hence, any polarization effects between the ML and MM regions are not taken into account.

In QM/MM calculations, the polarization of the QM region by the MM environment may be addressed by including interactions with MM atom charges as additional terms of the QM Hamiltonian, a strategy known as electrostatic embedding.[118] Electrostatic embedding schemes have been shown to generally be more accurate than mechanical embedding.[118,119] Böselt et al. implemented an electrostatic ML/MM strategy, introducing additional one-electron terms that incorporate the MM charges into the QM Hamiltonian.[115] They expressed the total energy as

$$V_{\mathrm{QM}}(\mathbf{r}) = V_{\mathrm{QM}}(\mathbf{r}_{\mathrm{QM}}) + V^{\mathrm{el}}_{\text{QM-MM}}(\mathbf{r}) + V_{\mathrm{MM}}(\mathbf{r}_{\mathrm{MM}}) + V^{\mathrm{vdW,SR}}_{\text{QM-MM}}(\mathbf{r}) \tag{2.3}$$

where $V^{\mathrm{el}}_{\text{QM-MM}}(\mathbf{r})$ term is the energy contribution due to the one-electron terms in the Hamiltonian describing the interaction of the QM charge density with MM atomic charges, and $V^{\mathrm{vdW,SR}}_{\text{QM-MM}}(\mathbf{r})$ describes the van der Waals interactions between QM and MM atoms through a classical Lennard-Jones potential. The first two terms in equation 2.3

are learned by the NNP. In order for the NNP to have information about the surrounding environment, MM atoms were included in the ML potential feature descriptors as an additional element type. It is important to note that the NNP is trained on QM/MM energies, i.e. the QM energies within a specific MM environment. The NNP cannot be trained on *in vacuo* QM energies, since the training examples must include the perturbation the QM region due to the environment. The need to train on QM/MM energies means that a generalized NNP trained on *in vacuo* energies cannot be used in this ML/MM embedding scheme.

Zinovjev proposed a method to electrostatically embed a NNP in an arbitrary MM environment.[120] This would allow any NNP to be used in the ML/MM computation, given that it is able to describe the QM region *in vacuo*. The polarization of the QM region in response to the MM environment is calculated using the Thole model.[121] The Thole model requires QM densities and molecular dipolar polarizabilities in order to compute the electrostatic embedding between ML and MM regions and these can be learned by an ML model. The model propsed by Zinovjev was employed to study protein-ligand interactions.

Another ML/MM scheme, known as the BuRNN method, was proposed by Lier et al.[113] The NNP learns the energy of the QM region in addition to the polarization energy of a buffer region, which is calculated as the difference between the energy of the buffer region in isolation and with the influence of the inner region. The buffer region essentially acts as a polarizable forcefield. Discontinuities that occur as molecules move

between buffer and outer regions largely cancel.

2.2 Adaptive QM regions

It is sometimes necessary to have a non-localized QM region, so that atoms can switch between QM and MM representations during the simulation. The QM region can be defined as all atoms within a given radius of a central point (defined, for example, as the center of mass of a reactive molecule). However, when atoms move between the QM and MM regions, there is a discontinuity in the potential energy surface and forces. One possible solution to the discontinuity problem, known as the hot spot method, is to define a buffer region as the atoms within inner and outer radii, r_1 and r_0 respectively.[122] A smoothing function is applied to the forces of atoms in the buffer region, so that they switch seamlessly between QM and MM representations,

$$f_i = S_m(r_i) f_{\mathrm{QM}} + (1 - S_m(r_i)) f_{\mathrm{MM}} \tag{2.4}$$

where f_i is the force on atom i and f_{QM} and f_{MM} are the forces calculated with quantum mechanics and molecular mechanics respectively, and S_m is a smoothing function defined as

$$f_c\left(R_{ij}\right) = \begin{cases} 1 & \text{for } r \leq r_1 \\ \frac{\left(r_0^2 - r^2\right)\left(r_0^2 + 2r^2 - 3r_1^2\right)}{\left(r_0^2 - r_1^2\right)^3} & \text{for } r_1 < r \leq r_0 \\ 0 & \text{for } r > r_0 \end{cases} \tag{2.5}$$

The hot spot method has been criticized because it lacks an expression for the potential energy.[123,124] An alternative ONIOM-XS method was has been proposed, which has a smooth energy function.[123] A function based on the average of switching functions for particles in the buffer region determines the relative contributions of the QM and MM energies in the buffer region. If a smooth potential energy surface is needed, for example in Monte Carlo simulations, the ONIOM-XS method is the preferred method.

2.3 Long-range Interactions

Standard neural network potentials are effectively short range, since atomic descriptors only include interactions with atoms that are within a given cutoff distance. They neglect long-range electrostatic and dispersion forces. For some chemical systems, long-range interactions are critical. For example, long-range interactions are typically not important for the simulation of bulk liquids; however, they are critical for interfaces between bulk liquids and their vapor phase.[125]

Several methods for including long-range interactions into neural network potentials have been proposed. One solution is to calculate the long-range electrostatic forces from neural network-predicted atomic charges.[126–128] The total energy is written as a sum of short range energies predicted by a standard NNP and long range electrostatic energy,

$$V_{\text{total}} = \sum_{i=1}^{N_{atom}} V_j + V_{elec} \tag{2.6}$$

The atomic charge NN is trained on reference atomic charges calculated using a charge petitioning method such as Mulliken charges,[129] Hirshfeld charges,[130] or Bader charges.[131] In order to avoid double-counting, the long range electrostatic energies predicted by the atomic charge NN are subtracted from the reference energies used to train the standard NNP.

Despite the successes of third generation NNPs, the training and evaluation of the additional atomic charge NN is a significant computational expense.[132] This expense can be spared by using a fixed set of charges for the calculation of the electrostatic interactions.

2.4 Proposed Methodology

We propose an additive embedding scheme that divides the energy into three terms,

$$V(\mathbf{r}) = V_{\mathrm{MM}}(\mathbf{r}_{\mathrm{MM}}) + V_{\mathrm{ML}}(\mathbf{r}_{\mathrm{ML}}) + V_{\mathrm{ML/MM}}(\mathbf{r}) \tag{2.7}$$

where V_{MM} is the forcefield energy of the MM region,

$$V_{\mathrm{MM}}(\mathbf{r}_{\mathrm{MM}}) = V_{\mathrm{Bonded}}(\mathbf{r}_{\mathrm{MM}}) + \sum_{i<j}^{\mathrm{MM}} \frac{q_i q_j}{4\pi\epsilon r_{ij}} + 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \tag{2.8}$$

V_{ML} is the energy of the ML region,

$$V_{\mathrm{ML}}(\mathbf{r}_{\mathrm{ML}}) = V_{\mathrm{ML}}^{SR}(\mathbf{r}_{\mathrm{ML}}) + \sum_{i<j}^{\mathrm{ML}} \frac{q_i q_j}{4\pi\epsilon r_{ij}} \mathrm{erf}(\alpha r_{ij}) \tag{2.9}$$

and $V_{\mathrm{ML/MM}}(\mathbf{r})$ is the interaction term between the ML and MM regions,

$$V_{\mathrm{ML/MM}}(\mathbf{r}) = \sum_{i}^{\mathrm{ML}} \sum_{j}^{\mathrm{MM}} \frac{q_i q_j}{4\pi\epsilon r_{ij}} + 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \tag{2.10}$$

V_{ML}^{SR} is the short-range ML potential that is trained on the reference energies with long range electrostatics subtracted.

$$V_{\mathrm{DFT}}^{\mathrm{SR}}(\mathbf{r}) = V_{\mathrm{DFT}}(\mathbf{r}) - \sum_{i<j}^{\mathrm{DFT}} \frac{q_i q_j}{4\pi\epsilon r_{ij}} \mathrm{erf}(\alpha r_{ij}) \tag{2.11}$$

The long-range electrostatic term in 2.11 can be readily obtained from the reciprocal space portion of an Ewald sum.[125]

In order to test our ML/MM methodology, we plan to study the dissociation of formic acid in water. In a previous study, free energy profiles of this dissociation were obtained through umbrella sampling simulations.[133] A NNP was employed to simulate a system of formic acid solvated with 128 water molecules. The dissociation was studied in bulk solution and at the air-water interface. We plan to reproduce their calculations of the free energy profiles; however we will employ an ML/MM method, replacing several water molecules which are described by the NNP in their work with a forcefield. It is expected

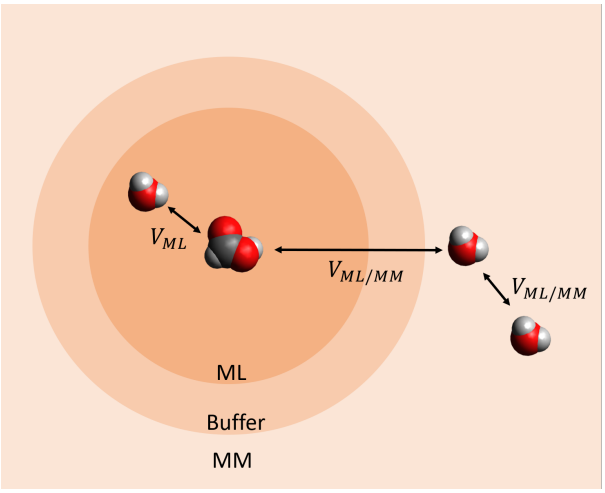

Figure 2.1: Illustration of proposed ML/MM method.

that the ML/MM method will lower the computational cost of the simulations.

Since the reaction involves a proton transfer to solvent, and any water molecule in the system can accept the proton, it is critical to have an adaptive QM(ML) region, which solvent molecules can enter and leave. We plan to define the ML region using the hot spot method, because it is straightforward to implement and does not require additional force calculations.[124] The center of the hot spot will be the center of mass of the formic acid molecule.

2.5 Initial Neural Network Potential Training

Neural network potentials were trained on a dataset of 6521 configurations and corresponding energies and forces, with each configuration consisting of formic acid and 128 water molecules. The energies and forces were calculated at the BLYP-D3/TZV2P level

Table 2.1: Energy and force RMSEs of NNPs constructed within TorchANI and DeePMD-kit.

	Energy (kcal/mol)		Force (kcal/(mol*Å)	
	Train	Test	Train	Test
TorchANI	6.82	19.30	3.35	2.99
DeePMD	4.16	6.24	3.07	3.43

of theory. Details about the electronic structure calculations and sampling methods are available in reference 133. In order to select the best framework for training the neural network potential, NNPs were trained using the TorchANI,[106] and DeePMD-kit.[134] The architecture and hyperparameters used to trained the model in DeePMD-kit are identical to those in reference 133. For the training of the TorchANI model, a radial cutoff distance of 6 Å was used. 21 radial shifting parameters R_s were used for the radial symmetry functions and 4 radial shifting parameters and 8 angular shifting parameters ϕ_{ijk} were used for the angular symmetry functions. The CELU (Continuously-Differentiable Exponential Linear Unit)[107] was used as the hidden layer activation function. The parameters were initialized using the Kaiming normal initialization method.[108] The model was trained on both energies and forces, the total loss function being the sum of the energy and force losses.

As shown in table 2.1, TorchANI are able to achieve good force RMSEs, comparable to those reported by Puenta et al.[133] However, the energy RMSE of the TorchANI model is significantly higher than that of the DeePMD model. It is not immediately clear why the DeePMD model outperforms the TorchANI model. The major difference between the two models is that DeePMD employs an embedding network, which learns a

featurization of the input coordinates, in contrast to TorchANI's pre-defined descriptors. Since it yields superior results, the DeePMD model will be used in further work.

Chapter 3

Contrasting Mechanisms of Aromatic and Aryl-Methyl Substituent Hydroxylation by the Rieske Monooxygenase Salicylate 5-Hydroxylase

3.1 Introduction

Catalysts play a key role in a wide range of industrial processes. The development of new catalysts has the power to make these processes more efficient and less toxic

and is therefore critical for environmental sustainability efforts.[135] Inspiration for the development of new catalysts may be derived from nature's catalysts: enzymes. Enzymes are "green"; they typically contain cheap metals such as iron, manganese and copper.[136] They are also highly selective.[137] A deep understanding of enzyme catalysts may inform the development of new sustainable, selective catalysts.

Rieske oxygenases are a class of enzymes best known for the degradation of aromatic compounds through dihydroxylation.[138] This function is practically useful for the removal of toxic aromatic compounds from the environment.[139] However their potential applications are varied, including roles in human health,[140–142] agriculture,[143,144] and chemoenzymatic synthesis.[132,145,146]

All ROs generally possess one of two quaternary structures: α_3 or $\alpha_3\beta_3$, where α is the catalytic subunit and β the structural subunit.[138] Each catalytic subunit contains a Rieske-type [2Fe-2S] iron-sulfur cluster and a mononuclear Fe(II) site. During catalysis, the mononuclear Fe(II) site performs the substrate oxidation, while electrons are transferred to and from the Rieske cluster. The electrons required for oxidation are donated to the Rieske center from NADH. There is a large distance (44 Å) between the Rieske cluster and the mononuclear Fe(II) within each subunit, making electron transfer between them unlikely. However, the subunits are arranged such that the Rieske cluster is approximately 12 Å away from the mononuclear Fe(II) site in the adjacent subunit, allowing electron transfer to occur between the Rieske and mononuclear Fe(II) sites in adjacent subunits.[138,147,148]

While much previous work has focused on Rieske dioxygenases, less attention has been paid to Rieske monooxygenases.[148] Salicylate 5-hydroxylase (S5H) is of particular interest because it has been structurally characterized and it exhibits great catalytic diversity, reacting with a wide range of substrates.[149] Previous work that investigated the reactivity of S5H with a variety of aromatic substrates found that it shares many mechanistic features with more widely studied cis-dihydroldiol forming dioxygenases.[148]

Kinetics studies and product analysis have been used to propose a mechanism of S5HH, the hydroxylase component of S5H. The O_2 binding does not occur until the substrate binds near the mononuclear Fe(II) site and the Rieske complex is reduced. The substrate binding leads to the loss of a water bound to the Fe(II) and also causes the Fe(II) to move away from the substrate, making room for the O_2 to bind. The O_2 binds to the mononuclear Fe site, to make an Fe(III)-superoxo species that attacks the substrate. A proton coupled electron transfer follows and then O-O bond cleavage occurs to form the product.[148]

Although this basic mechanism for S5H has been proposed, the mechanism may change when reacting with different substrates. In this work, the reactions of S5H with various substrates were investigated. Reaction mechanisms were inferred from experimental results that included kinetic and product isotope effects. In addition, density functional theory calculations were employed to provide information about the charge distributions and bond dissociation energies of substrates. With both the experimental and computational results, a general mechanism for S5H was proposed, as well as

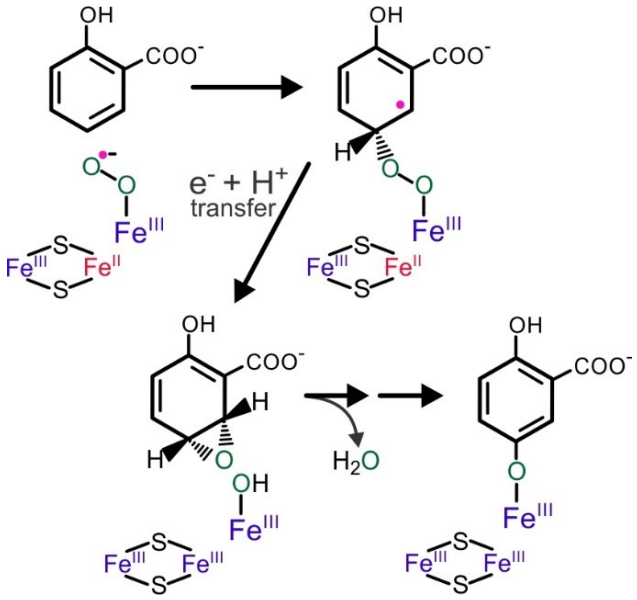

Figure 3.1: Proposed mechanism for Rieske monooxygenase S5H.

variations to this mechanism that depend on the identity of the substrate.

3.2 Computational Methods

All calculations were performed using the M06-2X functional[150] and the 6–311 + (2df,p) basis set[151] within the electronic structure modeling program Gaussian 16 version c01.[152] Geometries were optimized for salicylates and benzoates in their deprotonated conjugate base forms. The SMD (solvation model based on density) continuum solvation model using a dielectric constant of 5.6968, appropriate for the hydrophobic enzyme active site, was used.[153] CM5 (charge model 5) atomic charges were calculated with charges of hydrogens summed into those of heavy atoms.[154] Vibrational frequency calculations were performed for each optimized structure and thermal and entropic contributions

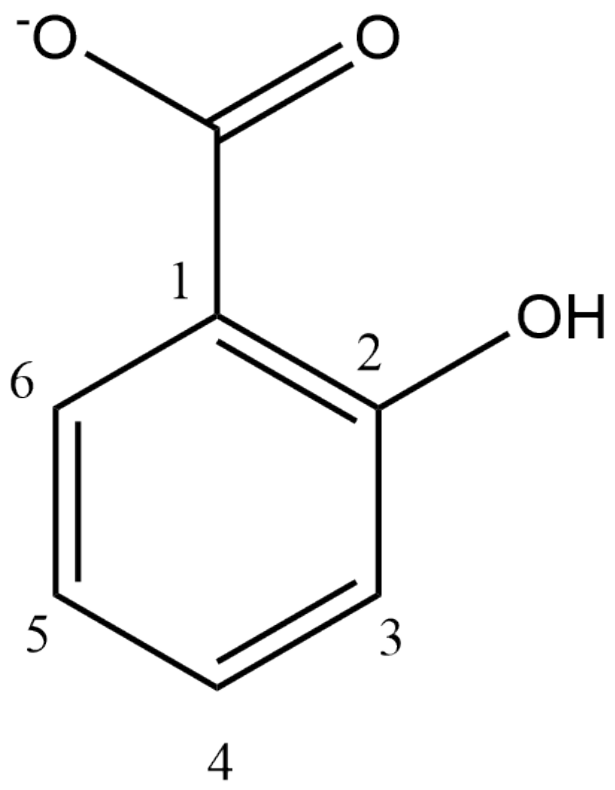

Figure 3.2: Carbon labeling on salicylate.

added to the self-consistent field energy. Bond dissociation free energies were calculated for bonds broken and formed in the enzyme catalyzed hydroxylation of the substrate.

Bond dissociation free energy (BDFE) of the bond R-X was calculated using eq 4.1 where $\mathrm{R-X}$ represents the salicylate or benzoate molecule, R-X represents the molecule after bond dissociation, and X^* represents the dissociated atom (either H or F).

$$\mathrm{BDFE(R-X)} = \Delta G(\mathrm{R}^*) + \Delta G(\mathrm{X}^*) - \Delta G(\mathrm{R-X}) \tag{3.1}$$

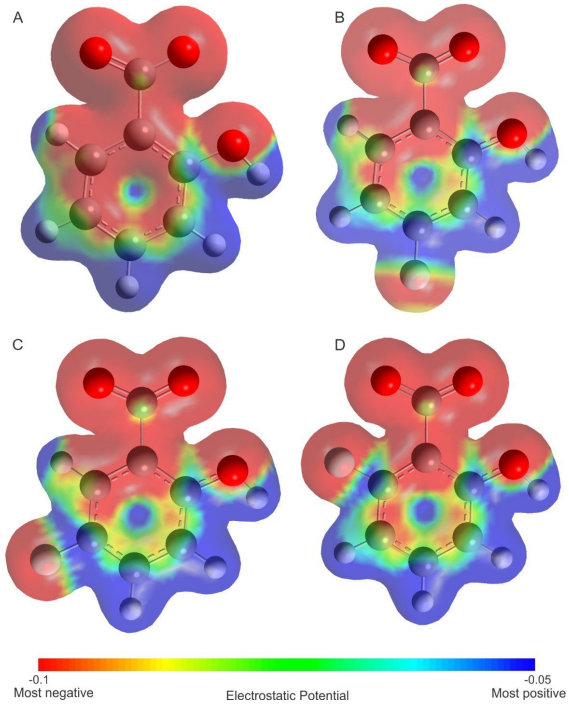

Figure 3.3: Electrostatic potential surface maps of (A) salicylate and (B) 4-F-salicylate, (C) 5-F-salicylate, and (D) 6-F-salicylate (computed with GaussView 6.1).

Here, we only report calculations using the M06-2X functional; however, additional computations were also performed using the ωB97XD functional,[98] and these computations showed that identical conclusions would be reached.

Electrostatic potential surface maps were produced on the $\rho = 0.02$ au contour of the molecular electronic density, from the optimized structures computed with the parameters described above. Electrostatic potential surface map images were made in GaussView 6.1.[155]

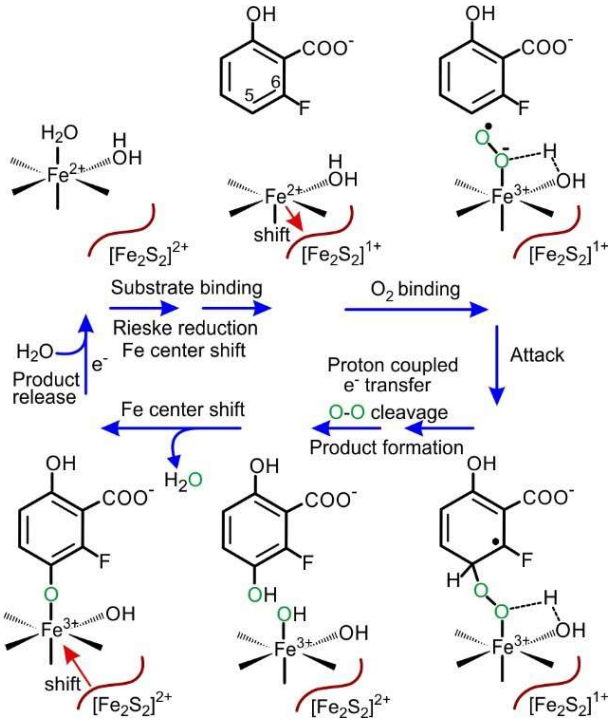

Figure 3.4: Proposed mechanism for S5H for fluorinated substrates.

3.3 Results and Discussion

The same general mechanism for the oxidation of salicylate can be drawn for salicylate with fluorine derivatives (Figure 3.4).[1] Two factors were proposed to determine which carbon on the aromatic ring hydroxylation occurs: the charge of the carbon and its placement relative to the attacking Fe(III) species. A previous study on cis-diol forming

[1]Figures 3.1,3.3,3.4, and 3.5 reprinted with permission from Melanie S. Rogers, Adrian M. Gordon, Todd M. Rappe, Jason D. Goodpaster, and John D. Lipscomb. Contrasting Mechanisms of Aromatic and Aryl-Methyl Sub- stituent Hydroxylation by the Rieske Monooxygenase Salicylate 5-Hydroxylase. Biochemistry, 62(2):507–523, jan 2023. Copyright 2023 American Chemical Society.

Table 3.1: Observed hydroxylation products and computed substrate atomic charges for salicylate and benzoate derivatives.

Substrate	Product	Atomic Charge		
		C4	C5	C6
Salicylate (Sal)	5-OH-Sal	-0.0169	-0.0451	-0.0217
3-F-Sal	3-F,5-OH-Sal	-0.0052	-0.0269	-0.0234
4-F-Sal	5-F,5-OH-Sal	+0.0823	-0.0310	+0.0016
5-F-Sal	5-F,6-OH-Sal	-0.0136	+0.0700	-0.0041
6-F-Sal	5-OH,6-F-Sal	-0.0088	-0.0313	+0.0882
benzoate (Ben)	3-OH-Ben	-0.0222	-0.0227	-0.0216
2-F-Ben	2-F,5-OH-Ben	-0.0043	-0.0245	-0.0098
3-CH_3-Ben	3-CH_2OH-Ben	-0.0281	-0.0303	-0.0242
5-CH_3-Sal	5-CH_2OH-Sal	-0.0317	-0.0436	-0.0163
5-CF_3-Sal	none detected	+0.0102	-0.0648	+0.0091

benzoate 1,2-dioxygenase showed that hydroxylation occurs on the ring carbon with the most negative charge.[147] In this work, 4-F-salicylate, 5-F-salicylate and 6-F salicylate exhibit hydroxylation at C5, the most negative carbon. However, when fluorine is placed at C5 hydroxylation occurs at C6. Accordingly the atomic charge computations show the C5 to be strongly deactivated, and hence hydroxylation takes place at C6. Owing to its high charge, hydroxylation could have occurred at C4. However, computational docking of the salicylate substrate to the active site show the C4 to be occluded from the oxygen binding site by a His229 ligand,[156] explaining why the hydroxylation occurs at C6, despite C4 having slightly more negative charge.

In the reaction of the methyl salicylate (5-CH3-Sal), hydroxylation occurs on the methyl

Table 3.2: Bond Dissociation Free Energies (BDFE) of Tested Substrates.

Substrate	Bond	BDFE
Salicylate	C5-H	101.01
Salicylate-d_6	C5-D	102.60
5-CH_3-salicylate	methyl C-H	81.42
5-CD_3-salicylate-d_8	methyl C-D	83.12
3-F-salicylate	C5-H	102.21
4-F-salicylate	C5-H	106.48
5-F-salicylate	C6-H	103.63
6-F-salicylate	C5-H	105.38
5-CH_3-salicylate	methyl C-5	111.81
benzoate	C3-H	101.05
	C5-H	101.07
3-CH_3-benzoate	methyl C-H	81.96
2-F-benzoate	C3-H	104.01
	C5-H	102.20

group instead of the aromatic ring. Surprisingly, an inverse kinetic isotope effect is observed in stopped-flow-single-turnover reactions.[149] The inverse KIE could be explained by a change in ring carbon hybridization from sp2 to sp3 when attacked by the superoxo intermediate. Therefore it is proposed that methyl-substituent hydroxylation reaction mechanism includes a Fe(III)-superoxo attack of the aromatic ring. The binding with the ring enables the oxidation of the mononuclear iron site and the subsequent transfer of an electron from the Rieske complex. The charge calculations, which show C4, C5 and C6 to be negatively charged, confirm that electrophilic attack of the aromatic ring carbons is likely.

A product isotope effect is also observed for the methyl substituent hydroxylation; an

excess of product from non-deuterated substrate is formed when equal amounts of deuterated and non-deuterated substrate are present.[149] The observation of the product isotope effect implies that substrate exchange occurs at the active site. Finally the C-H bond of the methyl substituent is quite strong (81.4 kcal/mol), requiring a powerful oxidizing intermediate. A mechanism consistent with these observations is shown in Figure 3.5. An Fe(III)-(hydro)peroxo intermediate not bound to the substrate is formed and then converted to a HO-FE(V)=O species capable of performing oxidation at the methyl substituent.

To further our understanding of the S5H mechanism, QM/MM computations may be performed on the substrate bound to the active site.[157] Computations could tell us, for example, how the active site can affect access of the reactive iron species to specific ring carbons on the substrate. Insights from active site and substrate calculations could provide insight into the observed kinetics. To explain the proposed substrate exchange for methyl-substituted reactions, computations could be performed to determine the stability of the reversibly formed Fe(III)-peroxo-substrate intermediate. The stability of this intermediate could be impacted by the particular orientation of the substrate within the active site, owing to the presence of a bulky methyl group. In addition QM calculations could investigate structural changes that result from changes in the Rieske cluster oxidation state. The structural insights afforded by QM/MM studies will advance our understanding of the reactivity of Rieske monooxygenases.

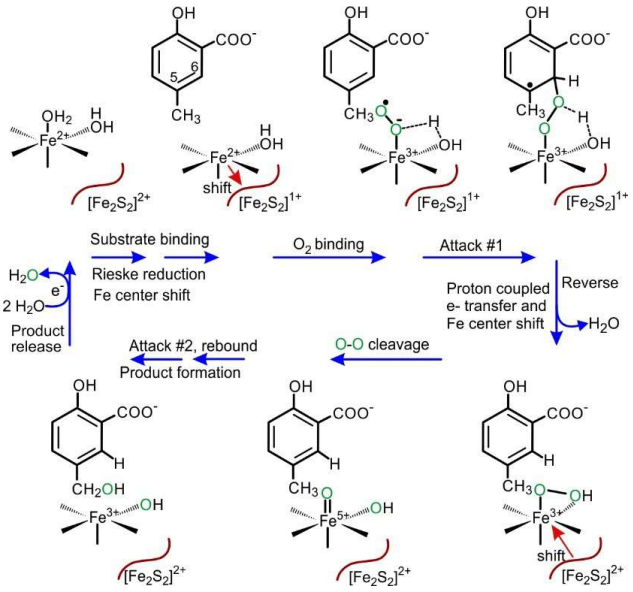

Figure 3.5: Proposed mechanism for S5H for methylated substrates.

Chapter 4

Conclusion

Machine learning for potential energy surfaces is a dynamic, fast-moving field of research. Every year new models are proposed that exhibit better accuracy and data efficiency. There are two main approaches to machine learning potentials. One approach is to train system-specific models. In this approach ab-initio molecular dynamics may be performed, and training configurations harvested from the trajectories to train a ML model. When the model has been trained on a training set sufficiently covering configurational space, the ab-initio simulation can switch to ML simulation. This allows longer simulation times to be achieved and avoids the need to perform redundant ab-initio calculations. The other approach is to train machine learning models generalized to a wide range of chemistry. Researchers could employ such a generalized model 'out of the box' without additional training data generation or training, much as researchers now use a DFT functional to calculate properties of interest. Both approaches to NNPs are

promising and should be pursued in future research efforts.

There are several exciting areas of further research in the field of NNPs. The generation of large training datasets is a critical area of work, especially for the development of generalizable NNPs. It is important that datasets cover a wide range of chemical and conformational space, while also avoiding redundant data.[158] Several large datasets have been constructed with the goal of training NNPs. QM9 is a dataset of 135k organic molecules in equilibrium configurations.[159] The ANI-1 dataset consists of a large set (20 million conformations) of off-equilibrium geometries and energies of organic molecules.[31] In the original ANI dataset, only molecules containing elements H, C, N, and O are represented, but the ANI-2 contains F, Cl, and S as well.[160] Limitations of the ANI datasets are that they do not include charged species, or high energy configurations in which covalent bonds are broken and formed. The SPICE dataset is another large dataset containing small molecules, dimers, dipeptides, and solvated amino acids.[161] It includes charged species and non-covalent interactions. However it does not include high energy configurations involved in chemical reactions. Transition1x was developed in order to address the lack of datasets containing reactive chemistry.[162] The configurations were generated by sampling along the Nudged Elastic Band pathways of 10k organic reactions.

The datasets mentioned above are good initial attempts at generating comprehensive quantum chemistry datasets for the training of ML potentials. But larger datasets containing more varied chemicals are still needed, as well as datasets at higher levels

of theory. As more datasets are developed, it is important that the methods for generating the data are well-documented. For example it is critical that the level of theory, and software used for ab initio calculations are documented, as well as the methods used to sample configurations. Guidelines have been proposed for the use of ML in chemistry.[134,163]

For both generalized or system-specific models, the algorithms for producing chemically varied, and balanced training data sets is a crucial area of research. Given the complexity of multi-dimensional potential energy surfaces, it seems that iterative methods, such as active learning, are needed to construct training datasets. Ensembles of models are widely used for uncertainty estimation in active learning algorithms. However, ensemble uncertainties have been shown to be unreliable estimators of model error.[164] We have seen this in our own work and it deserves further study. In addition, training multiple models in the ensemble is a significant computational expense. Better methods to estimate model uncertainties are needed. One approach is to use Bayesian neural networks that have internal uncertainty estimates based on the similarity between a configuration's atomic environments and environments already represented in the training set. Bayesian neural networks are often prohibitively expensive, but some success has been shown with low-dimensional Gaussian process models.[165] A fully automated active learning scheme was developed, in which the algorithm determined whether to add a new configuration into the training dataset based on the model uncertainty. An alternative approach to data selection within the active learning scheme is to select configurations

based on their geometric similarity with respect to configurations already in the training set. Those new configurations that are the least similar are likely the most important.

Finally, hybrid methods that combine physics-based molecular mechanics methods with NNPs is a promising method for the simulation of large chemical systems. Despite rapidly improving high performance computing resources, energy and force evaluations of NNPs, as well as generating training data for large systems, remain a challenge. Through hybrid methods, the components of the system requiring the flexibility of a neural network can be identified, and the remainder treated with cheaper methods.

References

[1] N. K. Srinivasan, M. C. Su, J. W. Sutherland, and J. V. Michael. Reflected shock tube studies of high-temperature rate constants for OH + CH 4 → CH 3 + H 2O and CH 3 + NO 2 → CH 3O + NO. *J. Phys. Chem. A*, 109(9):1857–1863, mar 2005.

[2] Jun Li and Hua Guo. Communication: An accurate full 15 dimensional permutationally invariant potential energy surface for the OH + CH4 → H2O + CH3 reaction. *J. Chem. Phys.*, 143(22):221103, dec 2015.

[3] Joaquin Espinosa-Garcia and Jose C Corchado. Qct dynamics study of the reaction of hydroxyl radical and methane using a new ab initio fitted full-dimensional analytical potential energy surface. *Theor Chem Acc*, 134:6, 2015.

[4] Michael P. Allen and Dominic J. Tildesley. *Computer simulation of liquids: Second edition.* OUP Oxford, 2017.

[5] Tao Cheng, Hai Xiao, and William A. Goddard. Full atomistic reaction mechanism

with kinetics for CO reduction on Cu(100) from ab initio molecular dynamics free-energy calculations at 298 K. *Proc. Natl. Acad. Sci. U. S. A.*, 114(8):1795–1800, feb 2017.

[6] Jon Fuller, Alessandro Fortunelli, William A. Goddard, and Qi An. Reaction mechanism and kinetics for ammonia synthesis on the Fe(211) reconstructed surface. *Phys. Chem. Chem. Phys.*, 21(21):11444–11454, may 2019.

[7] Jon Del Arco, Almudena Perona, Leticia González, Jesús Fernández-Lucas, Federico Gago, and Pedro A. Sánchez-Murcia. Reaction mechanism of nucleoside 2'-deoxyribosyltransferases: Free-energy landscape supports an oxocarbenium ion as the reaction intermediate. *Org. Biomol. Chem.*, 17(34):7891–7899, aug 2019.

[8] M. Karplus and J. Kuriyan. Molecular dynamics and protein function. *Proc. Natl. Acad. Sci. U. S. A.*, 102(19):6679–6685, may 2005.

[9] Jörg Gsponer and Amedeo Caflisch. Molecular dynamics simulations of protein folding from the transition state. *Proc. Natl. Acad. Sci. U. S. A.*, 99(10):6719–6724, may 2002.

[10] Kurt Kremer and Gary S. Grest. Dynamics of entangled linear polymer melts: A molecular-dynamics simulation. *J. Chem. Phys.*, 92(8):5057–5086, apr 1990.

[11] Mingjun Yang, Vasileios Koutsos, and Michael Zaiser. Interactions between polymers and carbon nanotubes: A molecular dynamics study. *J. Phys. Chem. B*,

109(20):10009–10014, may 2005.

[12] Thomas P. Senftle, Sungwook Hong, Md Mahbubul Islam, Sudhir B. Kylasa, Yuanxia Zheng, Yun Kyung Shin, Chad Junkermeier, Roman Engel-Herbert, Michael J. Janik, Hasan Metin Aktulga, Toon Verstraelen, Ananth Grama, and Adri C.T. Van Duin. The ReaxFF reactive force-field: Development, applications and future directions. *npj Comput. Mater.*, 2(1):1–14, mar 2016.

[13] Jörg Behler. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.*, 134(7):074106, feb 2011.

[14] Emir Kocer, Tsz Wai Ko, and Jörg Behler. Neural network potentials: A concise overview of methods. *Annual Review of Physical Chemistry*, 73(1):163–186, 2022, https://doi.org/10.1146/annurev-physchem-082720-034254. PMID: 34982580.

[15] Jörg Behler. Neural network potential-energy surfaces in chemistry: a tool for large-scale simulations. *Phys. Chem. Chem. Phys.*, 13:17930–17955, 2011.

[16] Jörg Behler. Perspective: Machine learning potentials for atomistic simulations. *The Journal of Chemical Physics*, 145(17):170901, 2016, https://doi.org/10.1063/1.4966192.

[17] Chris M. Handley and Paul L. A. Popelier. Potential energy surfaces fitted by artificial neural networks. *The Journal of Physical Chemistry A*, 114(10):3371–3383, 2010, https://doi.org/10.1021/jp9105585. PMID: 20131763.

[18] Pavlo O. Dral. Quantum chemistry in the age of machine learning. *The Journal of Physical Chemistry Letters*, 11(6):2336–2347, 2020, https://doi.org/10.1021/acs.jpclett.9b03664. PMID: 32125858.

[19] Volker L. Deringer, Miguel A. Caro, and Gábor Csányi. Machine learning interatomic potentials as emerging tools for materials science. *Advanced Materials*, 31(46):1902765, 2019, https://onlinelibrary.wiley.com/doi/pdf/10.1002/adma.201902765.

[20] Frank Noé, Alexandre Tkatchenko, Klaus-Robert Müller, and Cecilia Clementi. Machine learning for molecular simulation. *Annual Review of Physical Chemistry*, 71(1):361–390, 2020, https://doi.org/10.1146/annurev-physchem-042018-052331. PMID: 32092281.

[21] Tetiana Zubatiuk and Olexandr Isayev. Development of multimodal machine learning potentials: Toward a physics-aware artificial intelligence. *Accounts of Chemical Research*, 54(7):1575–1585, 2021, https://doi.org/10.1021/acs.accounts.0c00868. PMID: 33715355.

[22] Jun Zhang, Yao-Kun Lei, Zhen Zhang, Junhan Chang, Maodong Li, Xu Han, Lijiang Yang, Yi Isaac Yang, and Yi Qin Gao. A perspective on deep learning for molecular modeling and simulations. *The Journal of Physical Chemistry A*, 124(34):6745–6763, 2020, https://doi.org/10.1021/acs.jpca.0c04473. PMID: 32786668.

[23] Oliver T. Unke, Stefan Chmiela, Huziel E. Sauceda, Michael Gastegger, Igor Poltavsky, Kristof T. Schütt, Alexandre Tkatchenko, and Klaus-Robert Müller. Machine learning force fields. *Chemical Reviews*, 121(16):10142–10186, 2021, https://doi.org/10.1021/acs.chemrev.0c01111. PMID: 33705118.

[24] Pascal Friederich, Florian Häse, Jonny Proppe, and Alán Aspuru-Guzik. Machine-learned potentials for next-generation matter simulations. *Nature Materials*, 20(6):750–761, Jun 2021.

[25] Andrea Grisafi, Jigyasa Nigam, and Michele Ceriotti. Multi-scale approach for the prediction of atomic scale properties. *Chem. Sci.*, 12:2078–2090, 2021.

[26] Fenris Lu, Lixue Cheng, Ryan J. DiRisio, Jacob M. Finney, Mark A. Boyer, Pattarapon Moonkaen, Jiace Sun, Sebastian J. R. Lee, J. Emiliano Deustua, Thomas F. III Miller, and Anne B. McCoy. Fast near ab initio potential energy surfaces using machine learning. *The Journal of Physical Chemistry A*, 126(25):4013–4024, 2022, https://doi.org/10.1021/acs.jpca.2c02243. PMID: 35715227.

[27] Nikita Fedik, Roman Zubatyuk, Maksim Kulichenko, Nicholas Lubbers, Justin S. Smith, Benjamin Nebgen, Richard Messerly, Ying Wai Li, Alexander I. Boldyrev, Kipton Barros, Olexandr Isayev, and Sergei Tretiak. Extending machine learning beyond interatomic potentials for predicting molecular properties. *Nature Reviews Chemistry*, 6(9):653–672, Sep 2022.

[28] Hatice Gokcan and Olexandr Isayev. Learning molecular potentials with neural networks. *WIREs Computational Molecular Science*, 12(2):e1564, 2022, https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/wcms.1564.

[29] Michele Ceriotti, Cecilia Clementi, and O. Anatole von Lilienfeld. Introduction: Machine learning at the atomic scale. *Chemical Reviews*, 121(16):9719–9721, 2021, https://doi.org/10.1021/acs.chemrev.1c00598. PMID: 34428897.

[30] Bing Huang and O. Anatole von Lilienfeld. Ab initio machine learning in chemical compound space. *Chemical Reviews*, 121(16):10001–10036, 2021, https://doi.org/10.1021/acs.chemrev.0c01303. PMID: 34387476.

[31] J S Smith, O Isayev, and A E Roitberg. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost †. 2017.

[32] Yuzhi Zhang, Haidi Wang, Weijie Chen, Jinzhe Zeng, Linfeng Zhang, Han Wang, and Weinan E. DP-GEN: A concurrent learning platform for the generation of reliable deep learning based potential energy models. *Comput. Phys. Commun.*, 253:107206, aug 2020, 1910.12690.

[33] Nongnuch Artrith and Alexander Urban. An implementation of artificial neural-network potentials for atomistic materials simulations: Performance for TiO2. *Comput. Mater. Sci.*, 114:135–150, 2016.

[34] Lixue Cheng, Jiace Sun, and Thomas F. III Miller. Accurate molecular-orbital-based machine learning energies via unsupervised clustering of chemical space. *Journal of Chemical Theory and Computation*, 18(8):4826–4835, 2022, https://doi.org/10.1021/acs.jctc.2c00396. PMID: 35858242.

[35] Anders S. Christensen, Sai Krishna Sirumalla, Zhuoran Qiao, Michael B. O'Connor, Daniel G. A. Smith, Feizhi Ding, Peter J. Bygrave, Animashree Anandkumar, Matthew Welborn, Frederick R. Manby, and Thomas F. Miller. Orbnet denali: A machine learning potential for biological and organic chemistry with semi-empirical cost and dft accuracy. *The Journal of Chemical Physics*, 155(20):204103, 2021, https://doi.org/10.1063/5.0061990.

[36] Dylan M. Anstine and Olexandr Isayev. Machine learning interatomic potentials and long-range physics. *The Journal of Physical Chemistry A*, 127(11):2417–2431, 2023, https://doi.org/10.1021/acs.jpca.2c06778. PMID: 36802360.

[37] Michael S. Chen, Joonho Lee, Hong-Zhou Ye, Timothy C. Berkelbach, David R. Reichman, and Thomas E. Markland. Data-efficient machine learning potentials from transfer learning of periodic correlated electronic structure methods: Liquid water at afqmc, ccsd, and ccsd(t) accuracy. *Journal of Chemical Theory and Computation*, 0(0):null, 0, https://doi.org/10.1021/acs.jctc.2c01203. PMID: 36730728.

[38] SHUHAO ZHANG, Małgorzata Makoś, Ryan Jadrich, Elfi Kraka, Kipton Barros, Benjamin Nebgen, Sergei Tretiak, Olexandr Isayev, Nicholas Lubbers, Richard

Messerly, and et al. Exploring the frontiers of chemistry with a general reactive machine learning potential. *ChemRxiv*, 2022.

[39] Jinzhe Zeng, Liqun Cao, Mingyuan Xu, Tong Zhu, and John Z. H. Zhang. Complex reaction processes in combustion unraveled by neural network-based molecular dynamics simulation. *Nature Communications*, 11(1):5713, Nov 2020.

[40] Rongjun Chen, Kejie Shao, Bina Fu, and Dong H. Zhang. Fitting potential energy surfaces with fundamental invariant neural network. ii. generating fundamental invariants for molecular systems with up to ten atoms. *The Journal of Chemical Physics*, 152(20):204307, 2020, https://doi.org/10.1063/5.0010104.

[41] Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P. Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E. Smidt, and Boris Kozinsky. E(3)-Equivariant Graph Neural Networks for Data-Efficient and Accurate Interatomic Potentials. *Nat. Commun. 2022 131*, 13(1):1–11, jan 2021, 2101.03164v1.

[42] K. T. Schütt, H. E. Sauceda, P. J. Kindermans, A. Tkatchenko, and K. R. Müller. SchNet - A deep learning architecture for molecules and materials. *J. Chem. Phys.*, 148(24):241722, jun 2018, 1712.06113.

[43] Johannes Gasteiger, Janek Groß, and Stephan Günnemann. Directional Message Passing for Molecular Graphs. *8th Int. Conf. Learn. Represent. ICLR 2020*, mar 2020, 2003.03123.

[44] Oliver T. Unke, Stefan Chmiela, Michael Gastegger, Kristof T. Schütt, Huziel E. Sauceda, and Klaus Robert Müller. SpookyNet: Learning force fields with electronic degrees of freedom and nonlocal effects. *Nat. Commun. 2021 121*, 12(1):1–14, dec 2021, 2105.00304.

[45] Philipp Thölke and Gianni De Fabritiis. TORCHMD-NET: EQUIVARIANT TRANSFORMERS FOR NEURAL NETWORK BASED MOLECULAR POTENTIALS. In *ICLR 2022 - 10th Int. Conf. Learn. Represent.*, 2022, 2202.02541.

[46] Albert P. Bartók, Mike C. Payne, Risi Kondor, and Gábor Csányi. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.*, 104(13):136403, apr 2010, 0910.1019.

[47] Lixue Cheng, Jiace Sun, J. Emiliano Deustua, Vignesh C. Bhethanabotla, and Thomas F. Miller. Molecular-orbital-based machine learning for open-shell and multi-reference systems with kernel addition gaussian process regression. *The Journal of Chemical Physics*, 157(15):154105, 2022, https://doi.org/10.1063/5.0110886.

[48] Matthew Welborn, Lixue Cheng, and Thomas F. III Miller. Transferability in machine learning for electronic structure via the molecular orbital basis. *Journal of Chemical Theory and Computation*, 14(9):4772–4779, 2018, https://doi.org/10.1021/acs.jctc.8b00636. PMID: 30040892.

[49] Stefan Heinen, Guido Falk von Rudorff, and O. Anatole von Lilienfeld. Transition state search and geometry relaxation throughout chemical compound space with quantum machine learning. *The Journal of Chemical Physics*, 157(22):221102, 2022, https://doi.org/10.1063/5.0112856.

[50] Dominik Lemm, Guido Falk von Rudorff, and O. Anatole von Lilienfeld. Machine learning based energy-free structure predictions of molecules, transition states, and solids. *Nature Communications*, 12(1):4468, Jul 2021.

[51] Alexander V. Shapeev. Moment tensor potentials: A class of systematically improvable interatomic potentials. *Multiscale Model. Simul.*, 14(3):1153–1173, 2016, https://doi.org/10.1137/15M1054183.

[52] I. S. Novikov, Y. V. Suleimanov, and A. V. Shapeev. Automated calculation of thermal rate coefficients using ring polymer molecular dynamics and machine-learning interatomic potentials with active learning. *Phys. Chem. Chem. Phys.*, 20(46):29503–29512, nov 2018, 1805.11924.

[53] Jörg Behler and Michele Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.*, 98(14):146401, apr 2007.

[54] Zhilong Wang, Yanqiang Han, Jinjin Li, and Xiao He. Combining the Fragmentation Approach and Neural Network Potential Energy Surfaces of Fragments for

Accurate Calculation of Protein Energy. *J. Phys. Chem. B*, 124(15):3027–3035, apr 2020.

[55] Shae Lynn J. Lahey and Christopher N. Rowley. Simulating protein-ligand binding with neural network potentials. *Chem. Sci.*, 11(9):2362–2368, mar 2020.

[56] Christoph Schran, Jörg Behler, and Dominik Marx. Automated Fitting of Neural Network Potentials at Coupled Cluster Accuracy: Protonated Water Clusters as Testing Ground. *J. Chem. Theory Comput.*, 16(1):88–99, jan 2020, 1908.08734.

[57] Suresh Kondati Natarajan, Tobias Morawietz, and Jörg Behler. Representing the potential-energy surface of protonated water clusters by high-dimensional neural network potentials. *Phys. Chem. Chem. Phys.*, 17(13):8356–8371, apr 2015.

[58] Christoph Schran, Felix Uhl, Jörg Behler, and Dominik Marx. High-dimensional neural network potentials for solvation: The case of protonated water clusters in helium. *J. Chem. Phys.*, 148(10):102310, mar 2018.

[59] Nongnuch Artrith and Jörg Behler. High-dimensional neural network potentials for metal surfaces: A prototype study for copper. *Phys. Rev. B - Condens. Matter Mater. Phys.*, 85(4):045439, jan 2012.

[60] Nongnuch Artrith, Björn Hiller, and Jörg Behler. Neural network potentials for metals and oxides - First applications to copper clusters at zinc oxide. *Phys. status solidi*, 250(6):1191–1203, jun 2013.

[61] Brian Kolb, Levi C. Lentz, and Alexie M. Kolpak. Discovering charge density functionals and structure-property relationships with PROPhet: A general framework for coupling machine learning and first-principles methods. *Sci. Rep.*, 2017.

[62] N. Gerrits, Khosrow Shakouri, Jörg Behler, and Geert Jan Kroes. Accurate Probabilities for Highly Activated Reaction of Polyatomic Molecules on Surfaces Using a High-Dimensional Neural Network Potential: CHD 3 + Cu(111). *J. Phys. Chem. Lett.*, 10(8):1763–1768, apr 2019.

[63] Linfeng Zhang, De Ye Lin, Han Wang, Roberto Car, and E. Weinan. Active learning of uniformly accurate interatomic potentials for materials simulation. *Phys. Rev. Mater.*, 2019, 1810.11890.

[64] Qidong Lin, Liang Zhang, Yaolong Zhang, and Bin Jiang. Searching configurations in uncertainty space: Active learning of high-dimensional neural network reactive potentials. *Journal of Chemical Theory and Computation*, 17(5):2691–2701, May 2021.

[65] Justin S. Smith, Ben Nebgen, Nicholas Lubbers, Olexandr Isayev, and Adrian E. Roitberg. Less is more: Sampling chemical space with active learning. *J. Chem. Phys.*, 148(24):241733, jun 2018, 1801.09319.

[66] Khosrow Shakouri, Jörg Behler, Jörg Meyer, and Geert Jan Kroes. Accurate Neural Network Description of Surface Phonons in Reactive Gas-Surface Dynamics: N2 + Ru(0001). *J. Phys. Chem. Lett.*, 8(10):2131–2136, may 2017.

[67] Jun Chen, Xin Xu, and Dong H. Zhang. Communication: An accurate global potential energy surface for the OH + CO → H + CO2 reaction using neural networks. *J. Chem. Phys.*, 138(22):221104, jun 2013.

[68] Anyang Li and Hua Guo. A full-dimensional global potential energy surface of H3O+(ã3A) for the OH+(X3Σ-) + H2(X1Σg+) → H(2S) + H2O+(X2B1) reaction. *J. Phys. Chem. A*, 2014.

[69] Xi Chen and C. Franklin Goldsmith. Accelerating variational transition state theory via artificial neural networks. *The Journal of Physical Chemistry A*, 124(5):1038–1046, Feb 2020.

[70] Bin Jiang, Jun Li, and Hua Guo. High-fidelity potential energy surfaces for gas-phase and gas–surface scattering processes from machine learning. *The Journal of Physical Chemistry Letters*, 11(13):5120–5131, 2020, https://doi.org/10.1021/acs.jpclett.0c00989. PMID: 32517472.

[71] April M. Cooper, Philipp P. Hallmen, and Johannes Kästner. Potential energy surface interpolation with neural networks for instanton rate calculations. *The Journal of Chemical Physics*, 148(9):094106, 2018, https://doi.org/10.1063/1.5015950.

[72] Hsin Yu Ko, Linfeng Zhang, Biswajit Santra, Han Wang, E. Weinan, Robert A. DiStasio, and Roberto Car. Isotope effects in liquid water via deep potential molecular dynamics. *Mol. Phys.*, 2019.

[73] Hajime Kimizuka, Bo Thomsen, and Motoyuki Shiga. Artificial neural network-based path integral simulations of hydrogen isotope diffusion in palladium. *J. Phys. Energy*, 4(3):034004, jul 2022.

[74] Chenghan Li and Gregory A. Voth. Using machine learning to greatly accelerate path integral ab initio molecular dynamics. *Journal of Chemical Theory and Computation*, 18(2):599–604, Feb 2022.

[75] Ian R. Craig and David E. Manolopoulos. Quantum statistics and classical mechanics: Real time correlation functions from ring polymer molecular dynamics. *J. Chem. Phys.*, 121(8):3368–3373, aug 2004.

[76] Ian R. Craig and David E. Manolopoulos. Chemical reaction rates from ring polymer molecular dynamics. *J. Chem. Phys.*, 122(8):084106, feb 2005.

[77] Ian R. Craig and David E. Manolopoulos. A refined ring polymer molecular dynamics theory of chemical reaction rates. *J. Chem. Phys.*, 123(3):034102, jul 2005.

[78] Rosana Collepardo-Guevara, Yury V Suleimanov, and David E Manolopoulos. Bimolecular reaction rates from ring polymer molecular dynamics. *J. Chem. Phys.*, 130:174713, 2009.

[79] Yury V. Suleimanov, F. Javier Aoiz, and Hua Guo. Chemical Reaction Rate Coefficients from Ring Polymer Molecular Dynamics: Theory and Practical Applications. *J. Phys. Chem. A*, 120(43):8488–8502, nov 2016.

[80] Roman Korol, Nawaf Bou-Rabee, and Thomas F. Miller. Cayley modification for strongly stable path-integral and ring-polymer molecular dynamics. *J. Chem. Phys.*, 151(12):124103, sep 2019, 1907.07941.

[81] Xiongfei Gui, Wenbin Fan, Jiace Sun, and Yongle Li. New stable and fast ring-polymer molecular dynamics for calculating bimolecular rate coefficients with an example of oh + ch4. *Journal of Chemical Theory and Computation*, 18(9):5203–5212, Sep 2022.

[82] Yury V. Suleimanov and J. Espinosa-Garcia. Recrossing and Tunneling in the Kinetics Study of the OH + CH4 → H2O + CH3 Reaction. *J. Phys. Chem. B*, 120(8):1418–1428, mar 2015.

[83] Hai Lin, Yan Zhao, Benjamin A. Ellingson, Jingzhi Pu, and Donald G. Truhlar. Temperature dependence of carbon-13 kinetic isotope effects of importance to global climate change. *Journal of the American Chemical Society*, 127(9):2830–2831, Mar 2005.

[84] Joshua W. Allen, William H. Green, Yongle Li, Hua Guo, and Yury V. Suleimanov. Communication: Full dimensional quantum rate coefficients and kinetic isotope effects from ring polymer molecular dynamics for a seven-atom reaction oh + ch4 → ch3 + h2o. *Journal of Chemical Physics*, 138:221103, 6 2013.

[85] Jun Li, Kaisheng Song, and Jörg Behler. A critical comparison of neural network potentials for molecular reaction dynamics with exact permutation symmetry. *Phys. Chem. Chem. Phys.*, 21(19):9672–9682, may 2019.

[86] Dominik Marx and Jürg Hutter. *Ab initio molecular dynamics: Basic theory and advanced methods*, volume 9780521898638. Cambridge University Press, jan 2009.

[87] Xiaosong Li, John C. Tully, H. Bernhard Schlegel, and Michael J. Frisch. Ab initio Ehrenfest dynamics. *J. Chem. Phys.*, 2005.

[88] M.A. González. Force fields and molecular dynamics simulations. *École thématique la Société Française la Neutron.*, 12:169–200, 2011.

[89] Jörg Behler. First Principles Neural Network Potentials for Reactive Simulations of Large Molecular and Condensed Systems. 56(42):12828–12840, oct 2017.

[90] Justin S. Smith, Benjamin T. Nebgen, Roman Zubatyuk, Nicholas Lubbers, Christian Devereux, Kipton Barros, Sergei Tretiak, Olexandr Isayev, and Adrian E. Roitberg. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nat. Commun.*, 10(1):1–8, dec 2019.

[91] Thomas E. Markland and Michele Ceriotti. Nuclear quantum effects enter the mainstream. *Nat. Rev. Chem.*, 2(3), mar 2018, 1803.01037.

[92] Charusita Chakravarty. Path integral simulations of atomic and molecular systems. *Int. Rev. Phys. Chem.*, 1997.

[93] Konrad Hinsen and Benoît Roux. Potential of mean force and reaction rates for proton transfer in acetylacetone. *J. Chem. Phys.*, 106(9):3567–3577, mar 1997.

[94] Nicholas Boekelheide, Romelia Salomón-Ferrer, Thomas F. Miller, and III. Dynamics and dissipation in enzyme catalysis. *Proc. Natl. Acad. Sci. U. S. A.*, 108(39):16159, 2011.

[95] Scott Habershon, David E. Manolopoulos, Thomas E. Markland, and Thomas F. Miller. Ring-polymer molecular dynamics: Quantum effects in chemical dynamics from classical trajectories in an extended phase space. *Annu. Rev. Phys. Chem.*, 2013.

[96] Thom H. Dunning. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *J. Chem. Phys.*, 90:1007–1023, 1989.

[97] Yihan Shao, Zhengting Gan, Evgeny Epifanovsky, Andrew T.B. Gilbert, Michael Wormit, Joerg Kussmann, Adrian W. Lange, Andrew Behn, Jia Deng, Xintian Feng, Debashree Ghosh, Matthew Goldey, Paul R. Horn, Leif D. Jacobson, Ilya Kaliman, Rustam Z. Khaliullin, Tomasz Kus, Arie Landau, Jie Liu, Emil I. Proynov, Young Min Rhee, Ryan M. Richard, Mary A. Rohrdanz, Ryan P. Steele, Eric J. Sundstrom, H. Lee Woodcock, Paul M. Zimmerman, Dmitry Zuev, Ben Albrecht, Ethan Alguire, Brian Austin, Gregory J.O. Beran, Yves A. Bernard, Eric Berquist, Kai Brandhorst, Ksenia B. Bravaya, Shawn T. Brown, David Casanova,

Chun Min Chang, Yunqing Chen, Siu Hung Chien, Kristina D. Closser, Deborah L. Crittenden, Michael Diedenhofen, Robert A. Distasio, Hainam Do, Anthony D. Dutoi, Richard G. Edgar, Shervin Fatehi, Laszlo Fusti-Molnar, An Ghysels, Anna Golubeva-Zadorozhnaya, Joseph Gomes, Magnus W.D. Hanson-Heine, Philipp H.P. Harbach, Andreas W. Hauser, Edward G. Hohenstein, Zachary C. Holden, Thomas C. Jagau, Hyunjun Ji, Benjamin Kaduk, Kirill Khistyaev, Jaehoon Kim, Jihan Kim, Rollin A. King, Phil Klunzinger, Dmytro Kosenkov, Tim Kowalczyk, Caroline M. Krauter, Ka Un Lao, Adèle D. Laurent, Keith V. Lawler, Sergey V. Levchenko, Ching Yeh Lin, Fenglai Liu, Ester Livshits, Rohini C. Lochan, Arne Luenser, Prashant Manohar, Samuel F. Manzer, Shan Ping Mao, Narbe Mardirossian, Aleksandr V. Marenich, Simon A. Maurer, Nicholas J. Mayhall, Eric Neuscamman, C. Melania Oana, Roberto Olivares-Amaya, Darragh P. Oneill, John A. Parkhill, Trilisa M. Perrine, Roberto Peverati, Alexander Prociuk, Dirk R. Rehn, Edina Rosta, Nicholas J. Russ, Shaama M. Sharada, Sandeep Sharma, David W. Small, Alexander Sodt, Tamar Stein, David Stück, Yu Chuan Su, Alex J.W. Thom, Takashi Tsuchimochi, Vitalii Vanovschi, Leslie Vogt, Oleg Vydrov, Tao Wang, Mark A. Watson, Jan Wenzel, Alec White, Christopher F. Williams, Jun Yang, Sina Yeganeh, Shane R. Yost, Zhi Qiang You, Igor Ying Zhang, Xing Zhang, Yan Zhao, Bernard R. Brooks, Garnet K.L. Chan, Daniel M. Chipman, Christopher J. Cramer, William A. Goddard, Mark S. Gordon, Warren J. Hehre, Andreas Klamt, Henry F. Schaefer, Michael W. Schmidt, C. David

Sherrill, Donald G. Truhlar, Arieh Warshel, Xin Xu, Alán Aspuru-Guzik, Roi Baer, Alexis T. Bell, Nicholas A. Besley, Jeng Da Chai, Andreas Dreuw, Barry D. Dunietz, Thomas R. Furlani, Steven R. Gwaltney, Chao Ping Hsu, Yousung Jung, Jing Kong, Daniel S. Lambrecht, Wanzhen Liang, Christian Ochsenfeld, Vitaly A. Rassolov, Lyudmila V. Slipchenko, Joseph E. Subotnik, Troy Van Voorhis, John M. Herbert, Anna I. Krylov, Peter M.W. Gill, and Martin Head-Gordon. Advances in molecular quantum chemistry contained in the Q-Chem 4 program package. *Mol. Phys.*, 113(2):184–215, jan 2015.

[98] Jeng Da Chai and Martin Head-Gordon. Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections. *Phys. Chem. Chem. Phys.*, 10(44):6615–6620, nov 2008.

[99] Venkat Kapil, Mariana Rossi, Ondrej Marsalek, Riccardo Petraglia, Yair Litman, Thomas Spura, Bingqing Cheng, Alice Cuzzocrea, Robert H. Meißner, David M. Wilkins, Benjamin A. Helfrecht, Przemysław Juda, Sébastien P. Bienvenue, Wei Fang, Jan Kessler, Igor Poltavsky, Steven Vandenbrande, Jelle Wieme, Clemence Corminboeuf, Thomas D. Kühne, David E. Manolopoulos, Thomas E. Markland, Jeremy O. Richardson, Alexandre Tkatchenko, Gareth A. Tribello, Veronique Van Speybroeck, and Michele Ceriotti. i-PI 2.0: A universal force engine for advanced molecular simulations. *Comput. Phys. Commun.*, 236:214–223, mar 2019, 1808.03824.

[100] Massimiliano Bonomi, Giovanni Bussi, Carlo Camilloni, Gareth A. Tribello, Pavel Banáš, Alessandro Barducci, Mattia Bernetti, Peter G. Bolhuis, Sandro Bottaro, Davide Branduardi, Riccardo Capelli, Paolo Carloni, Michele Ceriotti, Andrea Cesari, Haochuan Chen, Wei Chen, Francesco Colizzi, Sandip De, Marco De La Pierre, Davide Donadio, Viktor Drobot, Bernd Ensing, Andrew L. Ferguson, Marta Filizola, James S. Fraser, Haohao Fu, Piero Gasparotto, Francesco Luigi Gervasio, Federico Giberti, Alejandro Gil-Ley, Toni Giorgino, Gabriella T. Heller, Glen M. Hocky, Marcella Iannuzzi, Michele Invernizzi, Kim E. Jelfs, Alexander Jussupow, Evgeny Kirilin, Alessandro Laio, Vittorio Limongelli, Kresten Lindorff-Larsen, Thomas Löhr, Fabrizio Marinelli, Layla Martin-Samos, Matteo Masetti, Ralf Meyer, Angelos Michaelides, Carla Molteni, Tetsuya Morishita, Marco Nava, Cristina Paissoni, Elena Papaleo, Michele Parrinello, Jim Pfaendtner, Pablo Piaggi, Giovanni Maria Piccini, Adriana Pietropaolo, Fabio Pietrucci, Silvio Pipolo, Davide Provasi, David Quigley, Paolo Raiteri, Stefano Raniolo, Jakub Rydzewski, Matteo Salvalaglio, Gabriele Cesare Sosso, Vojtěch Spiwok, Jiří Šponer, David W.H. Swenson, Pratyush Tiwary, Omar Valsson, Michele Vendruscolo, Gregory A. Voth, and Andrew White. Promoting transparency and reproducibility in enhanced molecular simulations. *Nat. Methods 2019 168*, 16(8):670–673, jul 2019.

[101] Gareth A. Tribello, Massimiliano Bonomi, Davide Branduardi, Carlo Camilloni, and Giovanni Bussi. PLUMED 2: New feathers for an old bird. *Comput. Phys. Commun.*, 185(2):604–613, feb 2014, 1310.0980.

[102] Adrian M. Gordon and Jason D. Goodpaster. Ch4-oh reaction dataset.

[103] P. C. Hariharan and J. A. Pople. The influence of polarization functions on molecular orbital hydrogenation energies. *Theor. Chim. Acta*, 28(3):213–222, sep 1973.

[104] Y. V. Suleimanov, J. W. Allen, and W. H. Green. RPMDrate: Bimolecular chemical reaction rates from ring polymer molecular dynamics. *Comput. Phys. Commun.*, 184:833–840, 2013.

[105] M Chase. NIST-JANAF Thermochemical Tables, 4th Edition, 1998.

[106] Xiang Gao, Farhad Ramezanghorbani, Olexandr Isayev, Justin S. Smith, and Adrian E. Roitberg. Torchani: A free and open source pytorch-based deep learning implementation of the ani neural network potentials. *J Chem Inf Model*, 60(7):3408–3415, Jul 2020.

[107] Jonathan T. Barron. Continuously Differentiable Exponential Linear Units. apr 2017, 1704.07483.

[108] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, 1502.01852.

[109] Jeremy O. Richardson and Stuart C. Althorpe. Ring-polymer molecular dynamics rate-theory in the deep-tunneling regime: Connection with semiclassical instanton theory. *J. Chem. Phys.*, 131:214106, 2009.

[110] Jun Li and Hua Guo. Thermal rate coefficients and kinetic isotope effects for the reaction oh + ch4 → h2o + ch3 on an ab initio-based potential energy surface. *The Journal of Physical Chemistry A*, 122(10):2645–2652, Mar 2018.

[111] Jorge Nochebuena, Sehr Naseem-Khan, and G. Andrés Cisneros. Development and application of quantum mechanics/molecular mechanics methods with advanced polarizable potentials. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 11(4):e1515, jul 2021.

[112] Frank Neese. Software update: the ORCA program system, version 4.0. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 8(1):e1327, jan 2018.

[113] Bettina Lier, Peter Poliak, Philipp Marquetand, Julia Westermayr, and Chris Oostenbrink. BuRNN: Buffer Region Neural Network Approach for Polarizable-Embedding Neural Network/Molecular Mechanics Simulations. *J. Phys. Chem. Lett.*, 13(17):3812–3818, may 2022.

[114] Lin Shen, Jingheng Wu, and Weitao Yang. Multiscale Quantum Mechanics/Molecular Mechanics Simulations with Neural Networks. *J. Chem. Theory Comput.*, 12(10):4934–4946, oct 2016.

[115] Lennard Böselt, Moritz Thürlemann, and Sereina Riniker. Machine Learning in QM/MM Molecular Dynamics Simulations of Condensed-Phase Systems. *J. Chem. Theory Comput.*, 17(5):2641–2658, may 2021, 2010.11610.

[116] Albert Hofstetter, Lennard Böselt, and Sereina Riniker. Graph-convolutional neural networks for (QM)ML/MM molecular dynamics simulations. *Phys. Chem. Chem. Phys.*, 24(37):22497–22512, sep 2022, 2206.14422.

[117] Raimondas Galvelis, Alejandro Varela-Rial, Stefan Doerr, Roberto Fino, Peter Eastman, Thomas E. Markland, John D. Chodera, and Gianni De Fabritiis. NNP/MM: Accelerating Molecular Dynamics Simulations with Machine Learning Potentials and Molecular Mechanics. *J. Chem. Inf. Model.*, 63(18):5701–5708, sep 2023, 2201.08110.

[118] Hai Lin and Donald G. Truhlar. QM/MM: What have we learned, where are we, and where do we go from here? *Theor. Chem. Acc.*, 117(2):185–199, feb 2007.

[119] Hans Martin Senn and Walter Thiel. QM/MM Methods for Biomolecular Systems. *Angew. Chemie Int. Ed.*, 48(7):1198–1229, feb 2009.

[120] Kirill Zinovjev. Electrostatic Embedding of Machine Learning Potentials. *J. Chem. Theory Comput.*, 19(6):1888–1897, mar 2023.

[121] B. T. Thole. Molecular polarizabilities calculated with a modified dipole interaction. *Chem. Phys.*, 59(3):341–350, aug 1981.

[122] Teerakiat Kerdcharoen, Klaus R. Liedl, and Bernd M. Rode. A QM/MM simulation method applied to the solution of Li+ in liquid ammonia. *Chem. Phys.*, 211(1-3):313–323, nov 1996.

[123] Teerakiat Kerdcharoen and Keiji Morokuma. ONIOM-XS: an extension of the ONIOM method for molecular simulation in condensed phase. *Chem. Phys. Lett.*, 355(3-4):257–262, apr 2002.

[124] Andreas Heyden, Hai Lin, and Donald G. Truhlar. Adaptive partitioning in combined quantum mechanical and molecular mechanical calculations of potential energy functions for multiscale simulations. *J. Phys. Chem. B*, 111(9):2231–2241, mar 2007.

[125] Samuel P. Niblett, Mirza Galib, and David T. Limmer. Learning intermolecular forces at liquid-vapor interfaces. *J. Chem. Phys.*, 155(16):164101, oct 2021, 2107.06208.

[126] Nongnuch Artrith, Tobias Morawietz, and Jörg Behler. High-dimensional neural-network potentials for multicomponent systems: Applications to zinc oxide. *Phys. Rev. B - Condens. Matter Mater. Phys.*, 83(15):153101, apr 2011.

[127] Tobias Morawietz, Vikas Sharma, and Jörg Behler. A neural network potential-energy surface for the water dimer based on environment-dependent atomic energies and charges. *J. Chem. Phys.*, 136(6), feb 2012.

[128] Jörg Behler. Four Generations of High-Dimensional Neural Network Potentials. *Chem. Rev.*, 121(16):10037–10072, aug 2021.

[129] R. S. Mulliken. Electronic Population Analysis on LCAO–MO Molecular Wave Functions. I. *J. Chem. Phys.*, 23(10):1833–1840, oct 1955.

[130] F. L. Hirshfeld. Bonded-atom fragments for describing molecular charge densities. *Theor. Chim. Acta*, 44(2):129–138, jun 1977.

[131] R. F. W. Bader. Atoms in Molecules: A Quantum Theory, Oxford University Press. page 438, 1990.

[132] Korey Bedard and Tomas Hudlicky. Enzymatic dihydroxylation of aromatic compounds: Nature's unique reaction and its impact on the synthesis of natural products. *Strateg. Tactics Org. Synth.*, 15:53–97, jan 2021.

[133] Miguel De La Puente, Rolf David, Axel Gomez, and Damien Laage. 2022, 144, 10524-10529 Downloaded via UNIV OF MINNESOTA on. *J. Am. Chem. Soc*, 144:10524–10529, 2022.

[134] Han Wang, Linfeng Zhang, Jiequn Han, and Weinan E. DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics. *Comput. Phys. Commun.*, 228:178–184, jul 2018, 1712.03641.

[135] T. Punniyamurthy, Subbarayan Velusamy, and Javed Iqbal. Recent advances in transition metal catalyzed oxidation of organic substrates with molecular oxygen. *Chem. Rev.*, 105(6):2329–2363, jun 2005.

[136] Lawrence Que and William B. Tolman. Biologically inspired oxidation catalysis. *Nat. 2008 4557211*, 455(7211):333–340, sep 2008.

[137] Miquel Costas, Mark P. Mehn, Michael P. Jensen, and Lawrence Que. Dioxygen Activation at Mononuclear Nonheme Iron Active Sites: Enzymes, Models, and Intermediates. *Chem. Rev.*, 104(2):939–986, feb 2004.

[138] Sarah M. Barry and Gregory L. Challis. Mechanism and Catalytic Diversity of Rieske Non-Heme Iron-Dependent Oxygenases. *ACS Catal.*, 3(10):2362–2370, oct 2013.

[139] Donald M. Jerina, John W. Daly, Alan M. Jeffrey, and David T. Gibson. Cis-1,2-dihydroxy-1,2-dihydronaphthalene: A bacterial metabolite from naphthalene. *Arch. Biochem. Biophys.*, 142(1):394–396, jan 1971.

[140] April L. Lukowski, Nicholas Denomme, Meagan E. Hinze, Sherwood Hall, Lori L. Isom, and Alison R.H. Narayan. Biocatalytic Detoxification of Paralytic Shellfish Toxins. *ACS Chem. Biol.*, 14(5):941–948, may 2019.

[141] Marco Massmig, Edward Reijerse, Joern Krausze, Christoph Laurich, Wolfgang Lubitz, Dieter Jahn, and Jürgen Moser. Carnitine metabolism in the human gut: characterization of the two-component carnitine monooxygenase CntAB from Acinetobacter baumannii. *J. Biol. Chem.*, 295(37):13065–13079, sep 2020.

[142] Mussa Quareshy, Muralidharan Shanmugam, Eleanor Townsend, Eleanor Jameson, Timothy D.H. Bugg, Alexander D. Cameron, and Yin Chen. Structural basis of carnitine monooxygenase CntA substrate specificity, inhibition, and intersubunit electron transfer. *J. Biol. Chem.*, 296:100038, jan 2021.

[143] Robert L. D'Ordine, Timothy J. Rydel, Michael J. Storek, Eric J. Sturman, Farhad Moshiri, Ryan K. Bartlett, Gregory R. Brown, Robert J. Eilers, Crystal Dart, Youlin Qi, Stanislaw Flasinski, and Sonya J. Franklin. Dicamba Monooxygenase: Structural Insights into a Dynamic Rieske Oxygenase that Catalyzes an Exocyclic Monooxygenation. *J. Mol. Biol.*, 392(2):481–497, sep 2009.

[144] Minggen Cheng, Dian Chen, Rebecca E. Parales, and Jiandong Jiang. Oxygenases as Powerful Weapons in the Microbial Degradation of Pesticides. *Annu. Rev. Microbiol.*, 76(Volume 76, 2022):325–348, sep 2022.

[145] Elizabeth A Osifalujo, Cristina Preston-Herrera, Phillip C Betts, Louis R Satterwhite, Jordan T Froese, J T Froese, E A Osifalujo, C Preston-Herrera, P C Betts, L R Satterwhite, and Preston-Herrera Cristina Preston-Herrera. Improving Toluene Dioxygenase Activity for Ester-Functionalized Substrates through Enzyme Engineering. *ChemistrySelect*, 7(11):e202200753, mar 2022.

[146] Julian L. Wissner, Wendy Escobedo-Hinojosa, Andreas Vogel, and Bernhard Hauer. An engineered toluene dioxygenase for a single step biocatalytical production of (-)-(1S,2R)-cis-1,2-dihydro-1,2-naphthalenediol. *J. Biotechnol.*, 326:37–39,

jan 2021.

[147] Brent S. Rivard, Melanie S. Rogers, Daniel J. Marell, Matthew B. Neibergall, Sarmistha Chakrabarty, Christopher J. Cramer, and John D. Lipscomb. Rate-Determining Attack on Substrate Precedes Rieske Cluster Oxidation during Cis-Dihydroxylation by Benzoate Dioxygenase. *Biochemistry*, 54(30):4652–4664, aug 2015.

[148] Melanie S. Rogers and John D. Lipscomb. Salicylate 5-Hydroxylase: Intermediates in Aromatic Hydroxylation by a Rieske Monooxygenase. *Biochemistry*, 58(52):5305–5319, dec 2019.

[149] Melanie S. Rogers, Adrian M. Gordon, Todd M. Rappe, Jason D. Goodpaster, and John D. Lipscomb. Contrasting Mechanisms of Aromatic and Aryl-Methyl Substituent Hydroxylation by the Rieske Monooxygenase Salicylate 5-Hydroxylase. *Biochemistry*, 62(2):507–523, jan 2023.

[150] Yan Zhao and Donald G. Truhlar. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theor. Chem. Acc.*, 120(1-3):215–241, may 2008.

[151] W Hehre, L Radom, P Schleyer, and J Pople. Ab initio molecular theory. page 576, 1986.

[152] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox. Gaussian˜16 Revision C.01, 2016. Gaussian Inc. Wallingford CT.

[153] Aleksandr V. Marenich, Christopher J. Cramer, and Donald G. Truhlar. Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions. *J. Phys. Chem. B*, 113(18):6378–6396, may 2009.

[154] Aleksandr V. Marenich, Steven V. Jerome, Christopher J. Cramer, and Donald G. Truhlar. Charge Model 5: An Extension of Hirshfeld Population Analysis for the

Accurate Description of Molecular Interactions in Gaseous and Condensed Phases. *J. Chem. Theory Comput.*, 8(2):527–541, feb 2012.

[155] Roy Dennington, Todd A. Keith, and John M. Millam. Gaussview Version 6, 2019. Semichem Inc. Shawnee Mission KS.

[156] Yan Jie Hou, Yuan Guo, De Feng Li, and Ning Yi Zhou. Structural and Biochemical Analysis Reveals a Distinct Catalytic Site of Salicylate 5-Monooxygenase NagGH from Rieske Dioxygenases. *Appl. Environ. Microbiol.*, 87(6):1–12, mar 2021.

[157] Hans Martin Senn and Walter Thiel. QM/MM studies of enzymes. *Curr. Opin. Chem. Biol.*, 11(2):182–187, apr 2007.

[158] Silvan Käser, Luis Itza Vazquez-Salazar, Markus Meuwly, and Kai Töpfer. Neural network potentials for chemistry: concepts, applications and prospects. *Digit. Discov.*, 2(1):28–58, feb 2023, 2209.11581.

[159] Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole Von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data 2014 11*, 1(1):1–7, aug 2014.

[160] Christian Devereux, Justin S. Smith, Kate K. Davis, Kipton Barros, Roman Zubatyuk, Olexandr Isayev, and Adrian E. Roitberg. Extending the Applicability

of the ANI Deep Learning Molecular Potential to Sulfur and Halogens. *J. Chem. Theory Comput.*, 16(7):4192–4202, jul 2020.

[161] Peter Eastman, Pavan Kumar Behara, David L. Dotson, Raimondas Galvelis, John E. Herr, Josh T. Horton, Yuezhi Mao, John D. Chodera, Benjamin P. Pritchard, Yuanqing Wang, Gianni De Fabritiis, and Thomas E. Markland. SPICE, A Dataset of Drug-like Molecules and Peptides for Training Machine Learning Potentials. *Sci. Data 2022 101*, 10(1):1–11, jan 2023, 2209.10702.

[162] Mathias Schreiner, Arghya Bhowmik, Tejs Vegge, Jonas Busk, and Ole Winther. Transition1x - a dataset for building generalizable reactive machine learning potentials. *Sci. Data 2022 91*, 9(1):1–9, dec 2022, 2207.12858.

[163] Nongnuch Artrith, Keith T. Butler, François Xavier Coudert, Seungwu Han, Olexandr Isayev, Anubhav Jain, and Aron Walsh. Best practices in machine learning for chemistry. *Nat. Chem. 2021 136*, 13(6):505–508, may 2021.

[164] Leonid Kahle and Federico Zipoli. Quality of uncertainty estimates from neural network potential ensembles. *Phys. Rev. E*, 105(1):015311, jan 2022, 2108.05748.

[165] Jonathan Vandermause, Steven B. Torrisi, Simon Batzner, Yu Xie, Lixin Sun, Alexie M. Kolpak, and Boris Kozinsky. On-the-fly active learning of interpretable Bayesian force fields for atomistic rare events. *npj Comput. Mater. 2020 61*, 6(1):1–11, mar 2020, 1904.02042.