Active Machine Learning for Chemical Dynamics Simulations. I. Estimating the Energy Gradient

Kazuumi Fujioka, Yuheng Luo, Rui Sun

April 2021

Abstract

Ab initio molecular dymamics (AIMD) simulation studies are a direct way to visualize chemical reactions and help elucidate non-statistical dynamics that does not follow the intrinsic reaction coordinate. However, due to the enormous amount of the *ab initio* energy gradient calculations needed for AIMD, it has been largely restrained to limited sampling and low level of theory (i.e., density functional theory with small basis sets). To overcome this issue, a number of machine learning (ML) methods have been employed to predict the energy gradient of the system of interest. In this manuscript, we outline the theoretical foundations of a novel ML method which trains from a varying set of atomic positions and their energy gradients, called interpolating moving ridge regression (IMRR), and directly predicts the energy gradient of a new set of atomic positions. Several key theoretical findings are presented regarding the inputs used to train IMRR and the predicted energy gradient. A hyperparameter used to guide IMRR is rigorously examined as well. The method is then applied to three bimolecular reactions studied with AIMD, including $HBr^+ + CO_2$, $H_2S + CH$, and $C_4H_2 + CH$, to demonstrate IMRR's performance on different chemical systems of different sizes. This manuscript also compares the computational cost of the energy gradient calculation with IMRR vs. *ab initio*, and the results highlight IMRR as a viable option to greatly increase the efficiency of AIMD.

1 Introduction

Ab initio molecular dynamics (AIMD) simulations of chemical reactions have shown great success in revealing their complicated dynamics at an atomistic level, elucidating discoveries from experiments that are nonintuitive, and predicting behaviors of chemical reactions whose conditions are difficult to realize.[1, 2, 3, 4, 5, 6, 7, 8] In AIMD, the interaction between atoms (i.e. energy gradient, corresponding to forces acting on atoms) is directly calculated on-the-fly with *ab initio* methods and their positions (referred to as "configurations") are propagated iteratively by solving the classical equations of motion over a small time interval.[2, 9, 10] In this way, the time-evolution of the coordinates of the system

(referred to as "trajectories") is collected. To ensure the conservation of the total energy of the system, the time interval between updating the coordinates of the atoms of a trajectory is usually of the order of one-tenth of a femtosecond. A chemical reaction in the gas phase takes place on the scale of picoseconds, as a result, there are usually a few thousand to tens of thousands *ab initio* energy gradient calculations involved in simulating each reactive trajectory.

Further, to accurately model reactions in real life, AIMD simulations of chemical reactions need to sample a statistical ensemble corresponding to the conditions of the experiments.[11, 12] For example, AIMD simulations of crossed-beam experiments (i.e. bimolecular collisions) should sample all possible impact parameters (b) and orientations of the collision (θ). Practically, this is done by first detecting bmax, the largest b in which a reactive trajectory can be observed, and then sampling trajectories with random orientations within $b_{\max}$. To account for the collision probability, the number of trajectories sampled at each b value should be proportional to $2\pi b$. For a gas phase bimolecular collision of small molecules, the $b_{\max}$ is usually a few (4.0 6.0) Å when the collision energy is less than 1.0 eV.[4, 6, 7, 8] Assuming b is sampled with 0.5 Å intervals and 100 trajectories are simulated at $b = 1.0$ Å, the smallest sampled impact parameter. Such a simulation study will contain a total of 3,600-7,800 trajectories. Multiplying the number of trajectories with the number of *ab initio* energy gradient calculations per trajectory leads to enormous computation cost for a simulation study of one chemical reaction under just one condition (a certain collision energy, temperature, vibrational excitation, etc.).

The millions of *ab initio* energy gradient calculations take up the overwhelming majority the computation involved in AIMD and present an obvious dilemma: there is an inevitable tradeoff between the accuracy of the *ab initio* method and the ergodicity of the sampling. For example, coupled cluster theories with large basis sets (e.g. CCSD(T)[13]/aug-cc-pVTZ[14]) can be expected to accurately model the ground state potential energy of a gas phase system. However, this level of theory is of no practical use in AIMD: even for systems with less than 10 atoms, a single *ab initio* energy gradient calculation of such method may take hours on one computer node with twenty processors. Millions of such calculations demanded by one simulation study would drain the capacity of a medium-size supercomputer for years. On the other hand, insufficient sampling inevitably compromises the reliability of AIMD, as the chance of observing some minor reaction pathways could be as low as 1%.[4, 6, 7, 8] The balance between accuracy and ergodicity usually limits AIMD to single reference *ab initio* methods, such as density functional theory (DFT[15]) or Moller-Plesset perturbation theory to the second order (MP2[16]) with basis sets of limited sizes (e.g. cc-pVDZ or 6-31G*[17]). Selecting a feasible yet accurate single reference *ab initio* method is laborious: the potential energy of a chemical reaction calculated from various candidate methods are compared against experimental benchmarks and/or results from a high-level *ab inito* method (e.g CCSD(T) extrapolated to the complete basis set limit[18]).

The daring burden of computation has greatly limited the application of AIMD, therefore, an on-the-fly and efficient algorithm that is able to predict

the energy gradient of configurations that replaces the expensive ab initio calculation is highly desirable. Over the last decade, various methods have been developed for this purpose and one popular approach is to estimate the energy gradient from a large database of ab initio calculations with machine learning (ML). For a more in-depth overview, see reviews by Hansen et al.[19, 20] and Faber et al.[21], as well as a more general review by Noe et al.[22] One broad class of ML methods treats the atoms in the system individually and predicts the energy gradient of each atom according to its surroundings. Several research have successfully demonstrated this, employing neural networks[23, 24, 25, 26, 27], kernel ridge regression[28, 29, 30], or Gaussian process regression.[31, 32] Another broad class of ML methods instead looks at the configuration of the entire system and predicts the energy gradients of all the atoms in the system at once. These types of ML often use linear interpolation[33, 34, 35, 36, 37], reproducing kernel interpolation[38, 39, 40, 41, 42, 43], or kernel ridge regression.[44, 45] The molecular systems that serve as proof of concepts in those ML developments are mostly close-to-optimal structures, for example, oscillations of atoms in metal or vibrations of molecules at a relatively high temperature (up to 500 K). In spite of the success of the aforementioned ML methods, they have not been well-adapted for AIMD simulations of chemical reactions, as a majority of which involve separated molecules, e.g. bimolecular collisions and unimolecular dissociations.

Take the former as an example, the simulations start with two reactant molecules separated at a relatively large distance (e.g. 10+ Å to minimize the interaction between them) which then collide to form a complex and dissociate to form product(s). Some trajectories could even be trapped by one or more potential energy minima (i.e. intermediates) and experience multiple barrier (re-)crossings. And as opposed to the regular thermal excitation used to train most ML methods, the excitation immediately after collision is much more localized and, as a result, certain vibrational modes of the newly-formed complex could be highly excited before the intramolecular vibrational redistribution is able to disperse the excess energy in those modes to the entire molecule. In conclusion, not only many more unique molecules, including reactants, intermediates, transition states, and products, needs to be sampled for AIMD of chemical reaction, but also their vibrations need to be sampled at a higher excitation. Figure 1 demonstrates this effect with an example reaction of $HBr^+ + CO_2 \longrightarrow HOCO^+ + Br\cdot$, which is roughly captured by the collective variables (CVs) of two distances: the distance between H and C and the distance between Br and O. The area of the phase space sampled from 10 reactive trajectories of collision energy of 8 kcal/mol, one of the lowest collision energies explored in the experiments, is much larger than 10 trajectories of oscillations starting from an intermediate at 500 K, a temperature commonly used for proof of concept in other ML methods. The extended phase space demanded by AIMD simulations of chemical reactions presents a challenge to the aforementioned ML methods. Once a trajectory either visits different intermediates than it has trained with or dissociates to separated molecules, these ML methods would have to employ new regiments of *ab initio* calculations to train these specific areas in the phase

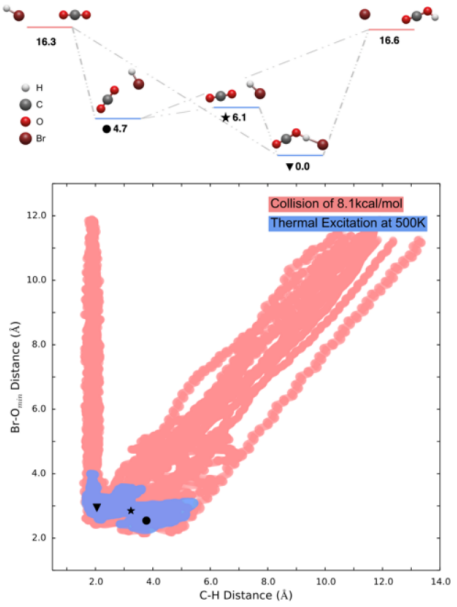

Figure 1: Top panel: the potential energy (in kcal/mol) of the $HBr^+ + CO_2 \longrightarrow HOCO^+ + Br$ reaction. Bottom panel: the blue area is sampled from 10 AIMD trajectories of a thermal excitation at 500 K (initiated at the triangle, an intermediate, see top panel). This temperature is commonly examined by other ML works of small molecules. The coral area is sampled from 10 AIMD bimolecular collision trajectories initiated from reactants (top right area of the figure) of a collision energy of 8.1 kcal/mol. This collision energy is among the lowest employed by the crossed-beam experiment.[46, 47] The triangle, star, and circle in the bottom panel correspond to the structures in the top panel. This figure shows that although the chemical space of interest in other ML works is able to sample the isomerization of two intermediates separated by a low transition state (star), it is not nearly enough for chemical reactions involving separated molecules. The potential energy and the AIMD are both generated at the MP2/cc-pVDZ/lanl08d level of theory.[48, 49]

space[35, 36] and/or the global ML potential energy surface (PES) would have to be iteratively retrained.[41, 42, 43, 37] To our best knowledge, there has not been a ML method that has been developed to tackle the complexity of the phase space involved in a chemical reaction.

In this manuscript, a novel ML algorithm, "Interpolating Moving Ridge Regression" (IMRR), that is specifically designed for estimating energy gradients for AIMD simulations, is introduced. The training set for IMRR, which is referred to as the "input of IMRR", is the energy gradients ($g(q_i)$) of configurations (q_i) that are geometrically close to the configuration of interest (q_0). The outcome of IMRR is $Z(q_0)$, the estimated energy gradient of q_0, which is

necessary to propagate the trajectory. IMRR is able to assess the risk of $Z(q_0)$, the likelihood of its deviation from the *ab initio* energy gradient, $(g(q_0))$, being larger than a user-defined threshold, which could be used to refine and/or reject $Z(q_0)$. It is important to note that the risk-assessment of IMRR is done without computing the *ab initio* gradient of the configuration of interest. As a summary, IMRR highlights a few characteristics that are attractive to AIMD simulations of chemical reactions: a) In theory, the training set of IMRR could be cost-free, as they are made from traditional AIMD simulations, e.g., the first 100 trajectories. In other words, all of the *ab initio* calculations involved in AIMD simulations directly contribute to the propagation of the trajectories – unlike other ML methods, there is not a number of *ab initio* calculations set aside purely for the purpose of ML training. b) IMRR's risk-assessing capability features the flexibility of referring back to the *ab initio* energy gradient when necessary. Combining with its nature of local regression, whenever an IMRR gradient is deemed risky (e.g., trajectory traverses through a poorly-learned regions in the phase space), the AIMD trajectory is not forced to adapt a potentially high error (i.e. high risk) energy gradient that would have negatively impacted its validity. This feature is not obvious in many ML methods, where their energy gradients are estimated from a global function. And c), IMRR is highly efficient – as shown later in this manuscript, its computational cost (both wall-time and CPU-time) is only a fraction of the *ab initio* energy gradient calculation. As a result, trajectories propagated with a mix of IMRR (when deemed low risk) / *ab initio* (when deemed high risk) could be expected to be much more efficient as compared to traditional AIMD trajectories. In this manuscript, the theory and performance of IMRR will be laid out in great detail, while its implementation with trajectory propagation will be introduced in a separate manuscript.

The rest of the manuscript is organized as the following. The theory of IMRR and the numerical protocol of minimizing the deviation between energy gradient from IMRR and *ab initio* are provided in the Methodology section. The dependance of this deviation on the input of IMRR and the hyperparameter is provided in the Result section. The computational cost of IMRR is also reported. The manuscript concludes with discussions on the IMRR's risk-assessing capability and how the chemistry and size of the system impact IMRR's performance.

2 Methods

2.1 The Upper Bound of the Error

Consider a chemical system of N atoms with configuration q and energy gradient $g(q)$. While the system may be fully described by a number of different schemes (distance matrix, smooth overlap of atomic positions (SOAP), etc.), consider the scheme defined by $3N$ coordinates (x, y, z for each atom), i.e. $q \in \mathbb{R}^{3N}$ and $g(q) \in \mathbb{R}^{3N}$. Assume for the configuration of interest at a certain step q_0, to propagate the system to the next time step, AIMD demands $g(q_0) \in \mathbb{R}^{3N}$, the

forces acting on the atoms, which is calculated from an *ab initio* method. The goal of IMRR is to estimate the energy gradient of q_0, named $Z(q_0) \in \mathbb{R}^{3N}$ with a training set made of *ab initio* energy gradients $g(q_i)$, calculated from previous simulations. In IMRR, $Z(q_0)$ is computed as the weighted average of the energy gradients of K configurations with w_i as the weight:

$$Z(q_0) = \sum_{i=1}^{k} w_i g(q_i) \qquad 1 \leq i \leq K \tag{1}$$

A successful IMRR execution would replace an expensive *ab initio* energy gradient calculation. IMRR optimizes w_i in order to minimize the difference (referred to as the "error of IMRR") between $Z(q_0)$ and $g(q_0)$, which is expressed as:

$$\left|g(q_0) - Z(q_0)\right| = \sqrt{\frac{1}{N}\sum_{j=1}^{3N}\Big(g_j(q_0) - Z_j(q_0)\Big)^2} \qquad 1 \leq j \leq 3N \tag{2}$$

Intuitively, those q_i that are geometrically close to q_0 should be prioritized in making up the training set. This closeness depends on the coordinates scheme; for the set of $3N$ x,y,z coordinates, the geometrical closeness between q_i and q_0 can be assessed by the root mean square displacement (RMSD, t_i) after q_i has been properly translated and rotated to maximize its overlap with q_0.[50] It is important to note that permutation should be allowed for chemically identical atoms if it increases the overlap. The RMSD is computed as:

$$t_i = \mathrm{RMSD}(q_0, q_i) = \sqrt{\frac{1}{N}\sum_{j=1}^{3N}\Big(q_{0,j} - q_{i,j}\Big)^2} \leq t_{cut} \qquad 1 \leq j \leq 3N \tag{3}$$

in which t_{cut} is a user-defined parameter that controls the geometrical closeness of the configurations in the training set with respect to the target. Using the notation of t_i, q_i is related to q_0 by:

$$q_i = q_0 + \sqrt{N} t_i \hat{h_i} \tag{4}$$

in which $\hat{h_i}$ is the unit vector of h_i, the displacement between q_i and q_0, i.e. $h_i = q_i - q_0 \in \mathbb{R}^{3N}$. Using this notation, the *ab initio* energy gradient of q_i can be rewritten as a single-variable function of t_i, $g(q_i) = f_i(t_i) \in \mathbb{R}^{3N}$. Assume the chemical system stays in the same electronic state (adiabatic process), its energy gradient can be assumed to be continuous and infinitely differentiable in each of its $3N$ components. Therefore, the j^{th} component of function f_i can be expanded with Taylor's theorem:

$$f_{i,j}(t) = f_{i,j}(0) + t f'_{i,j}(0) + \frac{1}{2}t^2 f''_{i,j}(0) + \frac{1}{3!}t^3 f'''_{i,j}(0) + \ldots \quad 1 \leq i \leq K, 1 \leq j \leq 3N \tag{5}$$

The first subscript of $f_{i,j}(t)$ denotes that it is the function for q_i and the second subscript of $f_{i,j}(t)$ deontes the it is the j^{th} component. As defined, the first term $f_{i,j}(0)$ is the j^{th} component of $g(q_0)$, i.e. $f_{i,j}(0) = g_j(q_0)$. The second term can be shown as the derivative of g_j with respect to the i^{th} direction:

$$\begin{aligned} f'_{i,j}(t_i) &= \frac{\mathrm{d}}{\mathrm{d}t_i} g_j\big(q_0 + \sqrt{N} t_i \hat{h_i}\big) \\ &= g'_j\big(q_0 + \sqrt{N} t_i \hat{h_i}\big) \cdot \sqrt{N}\hat{h_i} \\ f'_{i,j}(0) &= g'_j(q_0) \cdot \sqrt{N}\hat{h_i} \qquad g'_j(q_0) \in \mathbb{R}^{3N} \qquad 1 \le i \le K, 1 \le j \le 3N \end{aligned} \tag{6}$$

We first establish that the error of IMRR (Eq. 2) is bounded above the considereing the l^{th} component (among $3N$ components) of $Z(q_0)$ where it deviates the most from $g(q_0)$, e.g.

$$\big|g(q_0) - Z(q_0)\big| \le \sqrt{3} \cdot \big|g_l(q_0) - Z_l(q_0)\big| \qquad l = \underset{1 \le j \le 3N}{\operatorname{argmax}} \big|g_j(q_0) - Z_j(q_0)\big|$$

The inequality can be further derived as

$$\begin{aligned} \big|g(q_0) - Z(q_0)\big| &\le \sqrt{3} \cdot \Big|g_l(q_0) - \sum_{i=1}^{K} w_i g_l(q_i)\Big| \\ &= \sqrt{3} \cdot \Big|g_l(q_0) - \sum_{i=1}^{K} w_i f_{i,l}\big(t_i\big)\Big| \\ &= \sqrt{3} \cdot \Big|g_l(q_0) - \sum_{i=1}^{K} w_i \big(f_{i,l}(0) + t_i f'_{i,l}(0) + \frac{1}{2} t_i^2 f''_{i,l}(0) + \\ &\qquad \frac{1}{3!} t_i^3 f'''_{i,l}(0) + ...\big)\Big| \end{aligned}$$

Distribute the summation and apply the triangle inequality:

$$\begin{aligned} \big|g(q_0) - Z(q_0)\big| &\le \sqrt{3} \cdot \Big|\Big(g_l(q_0) - \sum_{i=1}^{K} w_i f_{i,l}(0)\Big) - \sum_{i=1}^{K} w_i t_i f'_{i,l}(0) - \\ &\qquad \sum_{i=1}^{K} \Big(w_i \frac{1}{2} t_i^2 f''_{i,l}(0) + w_i \frac{1}{3!} t_i^3 f'''_{i,l}(0) + \dots\Big)\Big| \\ &\le \sqrt{3} \cdot \Big|g_l(q_0) - \sum_{i=1}^{K} w_i f_{i,l}(0)\Big| + \sqrt{3} \cdot \Big|\sum_{i=1}^{K} w_i t_i f'_{i,l}(0)\Big| + \\ &\qquad \sqrt{3} \cdot \Big|\sum_{i=1}^{K} \Big(w_i \frac{1}{2} t_i^2 f''_{i,l}(0) + w_i \frac{1}{3!} t_i^3 f'''_{i,l}(0) + \dots\Big)\Big| \end{aligned} \tag{7}$$

The first term can be rewritten as:

$$\begin{aligned}\sqrt{3}\cdot\Big|g_l(q_0)-\sum_{i=1}^{K}w_i f_{i,l}(0)\Big| &= \sqrt{3}\cdot\big|g_l(q_0)\big|\cdot\Big|1-\sum_{i=1}^{K}w_i\Big| \\ &= \underbrace{\sqrt{3}\cdot\big|g_l(q_0)\big|}_{C_0}\cdot\underbrace{\Big|\sum_{i=1}^{K}w_i-1\Big|}_{R_0} \end{aligned} \tag{8}$$

in which C_0 depends only on the nature of the potential energy surface (specifically, its derivative), i.e. $C_0=\sqrt{3}\cdot\big|q_l(q_0)\big|$, and R_0 depends only on the weights of q_i, i.e. $R_0=\sum_{i=1}^{K}w_i-1$. Similarly, with results from Eq. 4 and Eq. 6, the second term in Eq. 7 could be derived as an inequality with a single C_1 term that depends only on the derivatives of $g(q_0)$ and a single R_1 term that depends only on q_i and their weights w_i:

$$\begin{aligned}\sqrt{3}\cdot\Big|\sum_{i=1}^{K}w_i t_i f'_{i,l}(0)\Big| &= \sqrt{3}\cdot\Big|\sum_{i=1}^{K}g'_l(q_0)\cdot w_i\sqrt{N}\hat{h_i}\Big| \\ &\le \sqrt{3}\cdot\Big|g'_l(q_0)\Big|\cdot\Big|\sum_{i=1}^{K}w_i\sqrt{N}t_i\hat{h_i}\Big| \\ &= \underbrace{\sqrt{3}\cdot\big|g'_l(q_0)\big|}_{C_1}\cdot\underbrace{\Big|\sum_{i=1}^{K}w_i h_i-\mathbf{0}\Big|}_{R_1} \end{aligned} \tag{9}$$

in which $\big|g'_l(q_0)\big|=\Big|\Big|g'(q_0)\Big|\Big|$ is defined as the magnitude of the largest value among the $3N\times 3N$ elements of the matrix $g'(q_0)$, the Hessian of q_0. In regard to the third term in Eq. 7:

$$\begin{aligned}&\sqrt{3}\cdot\Big|\sum_{i=1}^{K}\Big(w_i\frac{1}{2}t_i^2 f''_{i,l}(0)+w_i\frac{1}{3!}t_i^3 f'''_{i,l}(0)+\dots\Big)\Big| \\ &\qquad= \sqrt{3}\cdot\Big|\sum_{i=1}^{K}w_i t_i^2\Big(\frac{1}{2}f''_{i,l}(0)+\frac{1}{3!}t_i f'''_{i,l}(0)+\dots\Big)\Big| \\ &\qquad\le \sqrt{3}\cdot\sum_{i=1}^{K}\Big|w_i t_i^2\Big(\frac{1}{2}f''_{i,l}(0)+\frac{1}{3!}t_i f'''_{i,l}(0)+\dots\Big)\Big|\end{aligned}$$

Recall that t_i is bounded by t_{cut}, which is chosen to be a small value in practice, thus according to Taylor's theorem, the summation of higher terms in the Taylor expansion of an analytical function is bounded:

$$\left|t_i^2\left(\frac{1}{2}f''_{i,l}(0)+\frac{1}{3!}t_i f'''_{i,l}(0)+\dots\right)\right| \leq t_i^2 C_{i,l}$$

in which $C_{i,l}$ is a constant characterized by the nature of the potential energy for the system (i,e, higher order potential energy derivatives). Similarly, the third term in Eq. 7 can be epxressed as the product of two terms: the C_2 term that depends only on the nature of the potential energy surface and the R_2 term that depends only on q_i and their weights w_i. The inequality from above becomes:

$$\begin{aligned}
\sqrt{3}\cdot\sum_{i=1}^{K}\left|w_i t_i^2\left(\frac{1}{2}f''_{i,l}(0)+\frac{1}{3!}t_i f'''_{i,l}(0)+\dots\right)\right| &\leq \sqrt{3}\cdot\sum_{i=1}^{K}\left|w_i t_i^2 C_{i,l}\right| \\
&\leq \sqrt{3}\cdot\sum_{i=1}^{K}\left(\left|w_i t_i^2\right|\cdot\left|C_{i,l}\right|\right) \\
&\leq \sqrt{3}\cdot C_{i,l}\cdot\sum_{i=1}^{K}\left|w_i t_i^2\right| \\
&= \underbrace{\sqrt{3}\cdot C_{i,l}}_{C_2}\cdot\underbrace{\left|\sum_{i=1}^{K}|w_i|t_i^2-0\right|}_{R_2} \quad (10)
\end{aligned}$$

Substituting Eq. 8, 9, and 10 into Eq. 7 establishes that the error in energy gradients between the interpolated configuration and the target configuration is bounded above as:

$$\left|g(q_0)-Z(q_0)\right| \leq C_0R_0+C_1R_1+C_2R_2 \quad (11)$$

To summarize, the C terms depend only on the nature of the potential energy surface of the system (C_2 also depends on the threshold t_{cut} of selecting q_i) and the R terms depend only on the structural closeness between q_i and q_0. The geometric representations of these R terms for $K=2$ are summarized in Figure 2.

2.2 Minimize the Upper Bound of the Error

IMRR pursues to minimize the error between the estimated energy gradient and the *ab initio* energy gradient of the configuration of interest (Eq. 2) by minimizing its upper bound. As shown in Eq. 11, the source of the error has been categorized into two components: The C terms that depend on the nature of the potential energy surface and the R terms that do not. Clearly, the nature of the potential energy surface varies from system to system, therefore, IMRR focuses on minimizing Eq. 11 through the R terms. We first rewrite this upper bound into the form of a linear equation:

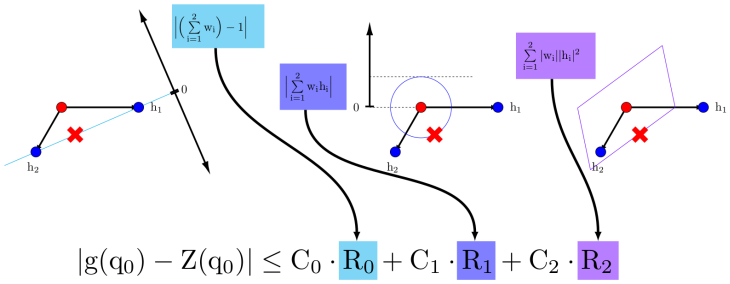

Figure 2: The R_0 distance measures the difference between the sum of the weights and unity whereas the R_1 distance measures the actual distance in space between the interpolated configuration (red cross) and the target configuration (red circle), as illustrated in the left panel. The R_2 distance measures the weighted sum of the magnitudes of the input configurations (blue circles). Cyan, dark blue, and purple lines indicate interpolation configurations which have the same value of R_0, R_1,and R_2, respectively.

$$
\begin{aligned}
|g(q_0) - Z(q_0)| &\leq C_0 R_0 + C_1 R_1 + C_2 R_2 \\
&= C_0 \left| \sum_{i=1}^{K} w_i - 1 \right| + C_1 \left| \sum_{i=1}^{K} w_i h_i - 0 \right| + C_2 \left| \sum_{i=1}^{K} w_i t_i^2 - 0 \right| \\
&= C_0 \left| \mathbf{U}^T w - \mathbf{1} \right| + C_1 \left| \mathbf{A} w - \mathbf{0} \right| + C_2 \left| \mathbf{B} w - \mathbf{0} \right| \\
&= \left| C_0 \mathbf{U}^T w - C_0 \mathbf{1} \right| + \left| C_1 \mathbf{A} w - C_1 \mathbf{0} \right| + \left| C_2 \mathbf{B} w - C_2 \mathbf{0} \right| \\
&\leq \sqrt{3} \cdot \sqrt{\left| C_0 \mathbf{U}^T w - C_0 \mathbf{1} \right|^2 + \left| C_1 \mathbf{A} w - C_1 \mathbf{0} \right|^2 + \left| C_2 \mathbf{B} w - C_2 \mathbf{0} \right|^2} \\
&= \sqrt{3} \left| \begin{bmatrix} C_2 \mathbf{B} \\ C_1 \mathbf{A} \\ C_0 \mathbf{U^T} \end{bmatrix} w - \begin{bmatrix} C_2 \mathbf{0} \\ C_1 \mathbf{0} \\ C_0 \mathbf{1} \end{bmatrix} \right| \qquad (12)
\end{aligned}
$$

$\mathbf{A} \in R^{K \times 3N}$ is defined as the matrix of the input configurations' deviations $h_i \in R^{3N}, 1 \leq i \leq K$:

$$
\mathbf{A} = \begin{bmatrix} | & | & & | \\ h_1 & h_2 & \dots & h_K \\ | & | & & | \end{bmatrix} = \begin{bmatrix} h_{1,1} & h_{2,1} & & h_{K,1} \\ h_{1,2} & h_{2,2} & & h_{K,2} \\ \vdots & \vdots & \dots & \vdots \\ h_{1,3N} & h_{2,3N} & & h_{K,3N} \end{bmatrix}
$$

$\mathbf{B} \in R^{K \times K}$ is the diagonal matrix of the magnitude of the displacement of the input configurations from the target configuration:

$$\mathbf{B} = \begin{bmatrix} t_1^2 & 0 & 0 & \dots & 0 \\ 0 & t_2^2 & 0 & \dots & 0 \\ 0 & 0 & t_3^2 & & \vdots \\ \vdots & \vdots & & \ddots & \\ 0 & 0 & \dots & & t_K^2 \end{bmatrix}$$

$\mathbf{U} \in R^K$ is a vector with elements of 1 and its transpose is:

$$\mathbf{U^T} = \begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix}$$

$w \in R^K$ is the weights of the K input configurations:

$$w = \begin{bmatrix} w_1 & w_2 & \dots & w_K \end{bmatrix}^T$$

Finally, $\mathbf{0} = \begin{bmatrix} 0 & 0 & \dots & 0 \end{bmatrix}^T \in \mathbb{R}^K$ and $\mathbf{1}$ is just the scalar 1. For simplicity, this can be further simplified as:

$$|g(q_0) - Z(q_0)| \leq |\mathbf{H}w - \mathbf{h}^*| = \sqrt{(\mathbf{H}w - \mathbf{h}^*)^T \cdot (\mathbf{H}w - \mathbf{h}^*)} = r(w) \tag{13}$$

in which $\mathbf{H} \in R^{(3N+K+1)\times K}$, $\mathbf{H} = \begin{bmatrix} C_2\mathbf{B} & C_1\mathbf{A} & C_0\mathbf{U^T} \end{bmatrix}^T$ and $\mathbf{h}^* \in R^{3N+K+1}$, $\mathbf{h}^* = \begin{bmatrix} C_2\mathbf{0} & C_1\mathbf{0} & C_0\mathbf{1} \end{bmatrix}^T$. IMRR solves for the optimal $w \in \mathbb{R}^K$ that minimizes $r(w)$ by setting the derivative of $r(w)^2$ to be zero. The optimal w that minimizes the upper bound of the error is:

$$w = \left(\mathbf{H}^T\mathbf{h}^*\right)^{-1}\mathbf{H}^T\mathbf{h}^* \tag{14}$$

This equation can be proven to have a non-trivial answer, as shown in the Supplementary Information.

2.3 Solving for the optimal w with restrictions

The derivation above details that the error of IMRR (Eq. 2) is bounded and how IMRR minimizes this bound. Nonetheless, it is important to note that the C terms used in constructing $\mathbf{H}$ and $\mathbf{h}^*$ (e.g. C_0, C_1, C_2 in Eq. 13) depend on the nature of the potential energy surface and they are not known *a priori*. To minimize the number of unknowns in $\mathbf{H}$ and $\mathbf{h}^*$, one of the R terms could be eliminated by setting it to zero as a constraint. The R_0 term is only a single row in the matrix $\mathbf{H}$ and would not dramatically shrink the number of solutions as compared to the R_1 term. Consequently, this can be eliminated by imposing the constraint:

$$\mathbf{U}^T w - \mathbf{1} = 0$$

Following the same procedure as shown in the previous section, the upper bound could be rewritten with the linear equation:

$$
\begin{aligned}
|g(q_0) - Z(q_0)| &\leq \left| \begin{bmatrix} C_2\mathbf{B} \\ C_1\mathbf{A} \end{bmatrix} w - \begin{bmatrix} C_2\mathbf{0} \\ C_1\mathbf{0} \end{bmatrix} \right| \\
&= C_1 \left| \begin{bmatrix} \alpha\mathbf{B} \\ \mathbf{A} \end{bmatrix} w - \begin{bmatrix} \alpha\mathbf{0} \\ \mathbf{0} \end{bmatrix} \right| \\
&= \mathbf{H_r} - \mathbf{h_r}^*
\end{aligned}
$$

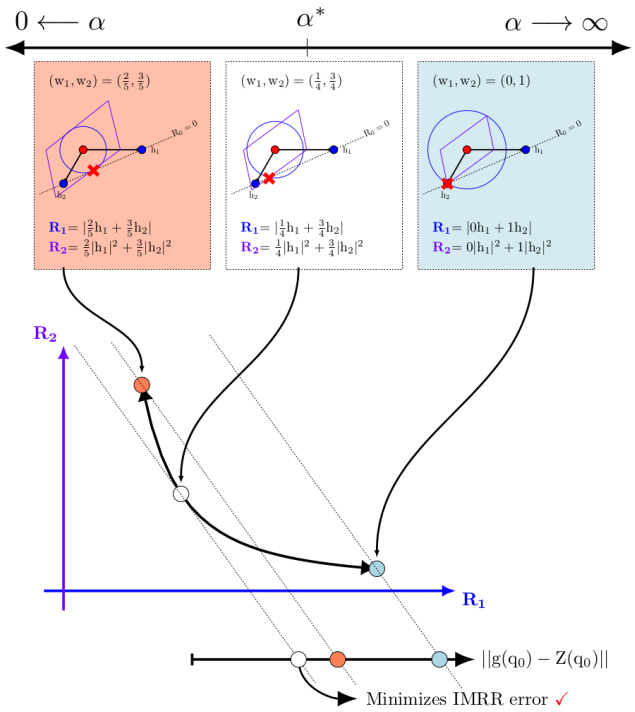

Figure 3: By constraining $R_0 = 0$, in an example where there are two input configurations ($K = 2$), all solutions lie on a one-dimensional line. Three possible solutions are presented. In each, the interpolated configuration is denoted by a red cross and the corresponding values of R_1 and R_2 vary. The hyperparameter α^* would lead the algorithm to pick the middle interpolation in this case, minimizing the error function defined by $R_1 + \alpha^* R_2$. In general (for all K), α describes the slope of the hyperplane tangent to the R_1-R_2 surface.

The terms in this equation are defined as before and a hyperparameter α is introduced to denote the ratio between C_2 and C_1, i.e. $\alpha = C_2/C_1$. The true value of α depends on the potential energy surface of the system (as well as q_0) and is not known either. As is customary in a machine learning method, this is left as a user-controlled hyperparameter. Details of estimating α that controls

the error are provided in the Results section. As seen in Figure 3, generally, the IMRR has two limiting behaviours for very small and large α. In the small α case, minimizing R_1 is preferred so the weights are chosen to minimize the distance between the interpolated frame and the target, or having an "accurate" interpolation. In the large α case, minimizing R_2 is preferred so the weights are chosen to minimize the distance between the interpolated frame and the closest possible input, or having a "precise" interpolation.

Define $\mathbf{H_r} = \begin{bmatrix}\alpha\mathbf{B} & \mathbf{A}\end{bmatrix}^T$ and $\mathbf{h_r}^* = \begin{bmatrix}\alpha\mathbf{0} & \mathbf{0}\end{bmatrix}^T$. This constrained optimization of w that minimizes the bound of the error could be solved by constructing a Lagrangian, $L(\lambda, w)$:

$$L(\lambda, w) = (\mathbf{H}w - \mathbf{h}^*)^T \cdot (\mathbf{H}w - \mathbf{h}^*) + \lambda(\mathbf{U}^T w - \mathbf{1})$$

Setting the gradient of $L(\lambda, w)$ with respect to λ and w to zero, the equation above could be rearranged as a set of linear equations:

$$\begin{bmatrix} w \\ \lambda \end{bmatrix} = \begin{bmatrix} 2{\mathbf{H_r}}^T\mathbf{H_r} & \mathbf{U} \\ \mathbf{U}^T & \mathbf{0} \end{bmatrix}^{-1} \begin{bmatrix} 2{\mathbf{H_r}}^T\mathbf{h_r^*} \\ \mathbf{U}^T \end{bmatrix}$$

It has been proved elsewhere that the matrix is invertible if $\begin{bmatrix}{\mathbf{H_r}}^T & \mathbf{U}\end{bmatrix}^T$ has linearly dependent columns.[51]

3 Results

The first representative system employed to demonstrate the performance of IMRR is the bimolecular collision of $HBr^+ + CO_2$, which after collision, forms the proton-transfer product $HOCO^+ + Br$, or goes back to the reactant molecules (i.e., non-reactive trajectories). This system has been studied extensively with the guided-beam experiments, quantum calculations, and AIMD. The IMRR employs *ab initio* energy gradients computed with MP2/cc-pVDZ by NWChem that made up the previous AIMD trajectories as its targets (q_0, $g(q_i)$) and inputs (q_i, $g(q_i)$) when applicable. As noted in Methods, the error of IMRR depends on various factors, such as the geometrical closeness of inputs (t_{cut} in Eq. 3), the number of inputs (K in Eq. 1), the hyperparameter (α in Eq. 15), etc. Thus, in this section, we treat these factors as independent variables – studying the dependence of the error of IMRR with respect to one factor while keeping the rest at fixed, reasonable values, with justifications provided in each respective section below. This setting is to have 15 input frames (q_i and its corresponding $g(q_i)$), each of which has an RMSD less than 0.15 A ($t_{cut} = 0.15$ A) to the target q_0, and the hyperparameter α is held at 1.0 x 10−1. Unless noted otherwise, these values are adapted as default for the rest of the manuscript.

3.1 Accuracy of IMRR vs. the Geometrical Closeness and the Number of Inputs

As defined in Eq. 3, t_i represents the geometrical closeness between the input configuration, q_i, $1 \leq i \leq K$ (the number of inputs for the IMRR) and the target, q_0, after their overlap has been maximized through center of mass translation and rotation. t_{cut} is the upper bound of t_i and the previous section has shown the error (i.e., Eq. 2) of IMRR decreases as ti gets smaller (see R_1 term in Eq. 9 and R_2 term in Eq. 10). Therefore, controlling t_{cut}, is the first proof of concept in this section. Since IMRR aims to predict the energy gradient that could be applied to simulate chemical reactions, it is important to ensure the targets employed in the test represent all relevant phase space of the reaction. An illustration of the phase space of this reaction, characterized by two collective variables (CVs), the distance between H-O and the shorter distance between two Br-O, is provided in Fig. S1 of the Supplementary Information. The configurations of 4,000 AIMD trajectories are binned into the CV-space, and 68 1A x 1A cells are populated. These cells are determined to be relevant to the reaction and one configuration from each cell is selected as the target to assess the performance of IMRR.

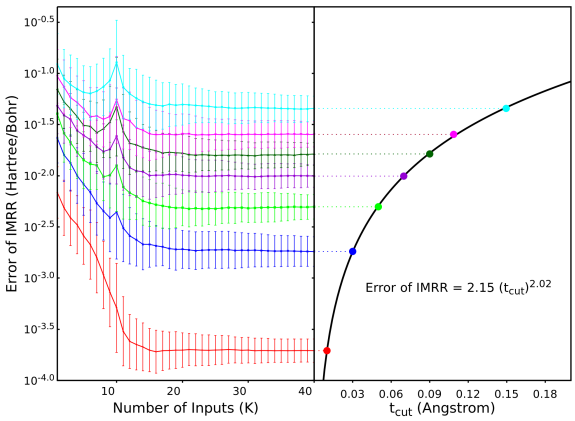

Figure 4: Left panel: the error of IMRR with vs. the number of inputs at various t_{cut} values. Right panel: the error of IMRR vs. t_{cut} with $K = 40$ inputs. The targets of these figures uniformly sample the phase space of the $HBr^+ + CO_2$ reaction.

The energy gradient of these targets is estimated by IMRR at various t_{cut} and compared with its *ab initio* counterpart, whose difference is defined as the error of IMRR (Eq. 2) and plotted in the right panel of Fig. 4. For each target (q_0), up to 40 inputs (q_i) are randomly generated by displacing atoms from q_0, while enforcing its t_i (geometrical closeness to q_0) to be between 90-100% of each t_{cut}. The energy gradient of the inputs, $g(q_i)$, are computed at

MP2/cc-pVDZ level of. As Fig. 4 demonstrates, the error of IMRR decreases monotonically as the input configurations, q_i, get more geometrically similar to the target, q_0. In other words, the energy gradient predicted by IMRR, $Z(q_0)$, approaches $g(q_0)$, the energy gradient of the target from MP2/cc-pVDZ, as q_i approaches q_0 (i.e., smaller t_{cut}). Fig. 4 also illustrates that the error of IMRR demonstrates a strong logarithmic relation with respect to t_{cut}. Note the x-axis is linear and the y-axis is logarithmic – when the y-axis is linearized, the black curve becomes a power relation with the general form of ax^b, where a and b are constants. The fitted line (black solid curve) closely resembles a quadratic (ax^2), indicating that with the optimized weights in Eq. 2, the error is largely bounded by C_2R_2, since R_2 is proportional to the sum of the squared t_i of the input. Overall, the right panel of Fig. 4 suggests that the accuracy of IMRR is improvable as more inputs closer to the target become available – a scenario that is at least achievable in theory with more sampling of AIMD trajectories.

Fig. 4 (left panel) also illustrates the correlation between the error of IMRR and the number of inputs (K). For each target, their inputs are sorted before being fed into the IMRR. For example, $K = 2$ means the IMRR is carried out with the two inputs closest to the target, and $K = 3$ includes the three inputs closest to the target. Although the newly included inputs (as a result of increasing K) are geometrically further away from the target, the IMRR, across all values of t_{cut}, is able to estimate an energy gradient much closer to *ab initio* value (i.e., smaller error) with a larger number of inputs. The gain in IMRR accuracy by including more inputs is significant when K is less than 15 and becomes marginal after $K > 15$. The convergence of the error of IMRR after 15 inputs has led to the usage $K = 15$ as a default number of inputs in the rest of the manuscript. It is interesting to note, the error of IMRR demonstrates a local maximum around $K = 10$ for almost all t_{cut}, which is particularly obvious for larger t_{cut}. The origin of this counterintuitive maximum is debatable, while one explanation could be that IMRR is more 'fragile' when inputs are geometrically further away from the target, and is not well-behaved until there is at least 10 inputs to work with.

It is important to confirm that the behavior of IMRR with respect to the geometrical closeness and number of inputs is not a result of the artificial method of generating inputs, i.e., by displacing atoms of the target. Herein, a more large-scale assessment of IMRR is carried out, in which 1000 inputs uniformly sampling the feasible CV-space are selected from 150 AIMD trajectories, thus roughly 15 inputs are selected from each 1A x 1A cell shown in Fig. S1. The energy gradients of these inputs are estimated by IMRR with inputs (15 each) selected from another 4000 AIMD trajectories (a total of 28.1 million energy gradient calculations). We note that these inputs are not restricted by a lower bound (i.e. t_i in this test does not have to be larger than 90% of the t_{cut}) as imposed in the previous test. This setting closely mimics how IMRR would be applied to an AIMD simulation – estimating the energy gradient with inputs from *ab initio* calculations of already-computed trajectories. The results are summarized in the Supplementary Information (Fig. S2) and shows very good agreement with Fig. 4, demonstrating the potential of IMRR in producing

low-error energy gradient given enough inputs that are close to the target.

3.2 Accuracy of IMRR vs. the Hyperparameter α

As shown in the previous section, IMRR demands a hyperparameter α in solving for the optimal weights (w_i) that minimize the upper bound of the error. Defined Section II.c, α is expressed as the ratio between C_2, a term depending on the derivatives of the energy gradient of the target, and C_1, a term depending on the energy gradient of the target, thus at least in theory computable with an *ab initio* method. However, it would be highly unwise to evaluate α for each IMRR, since it would be more expensive than just computing the *ab initio* energy gradient itself. Therefore, like many other ML methods, the hyperparameter α is tested and chosen over a range of values determined empirically to reliably produce minimal IMRR error.

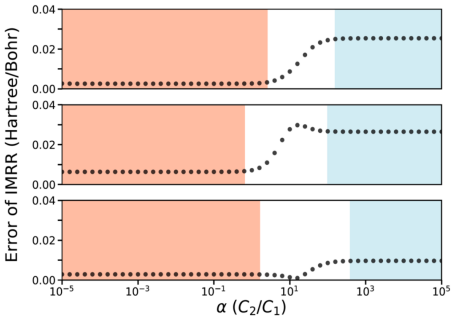

Figure 5: Three representative behaviors of the error of IMRR vs. the value of the hyperparameter. The small α, large α, and intermediate α regions are colored with coral, blue, and white, respectively.

Intuitively, when a variety of inputs are available to IMRR (while controlling the inputs' geometrical closeness to the target by enforcing $t_i < t_{cut}$), the hyperparameter α balances the relative importance between the inputs that are relatively far from the target (i.e., larger t_i) and that are relatively close (i.e., smaller t_i). As Eq. 15 suggests, a large α (blue region in Fig. 5) will minimize the C_2R_2 term of the upper bound of the error; while in contrast, a small α (coral region in Fig. 5) will minimize the C_1R_1 term of the upper bound of the error. Herein, the 1000 targets (q_0) that uniformly sample the CV-space of the reaction are employed as the targets again to investigate the behavior of the error of IMRR with respect to different α, whose value ranges between 10^{-5} and 10^{+5}. Although the exact curve of the error of IMRR vs. α curve varies from target to target (see in Fig. 5), it can be divided into three regions: large α (blue), small α (coral), and intermediate α (white). The first two regions are detected when the error of IMRR becomes independent of α, although each region is associated with a different error. The error of IMRR in the intermediate region heavily depends on α and smoothly connects the other two regions.

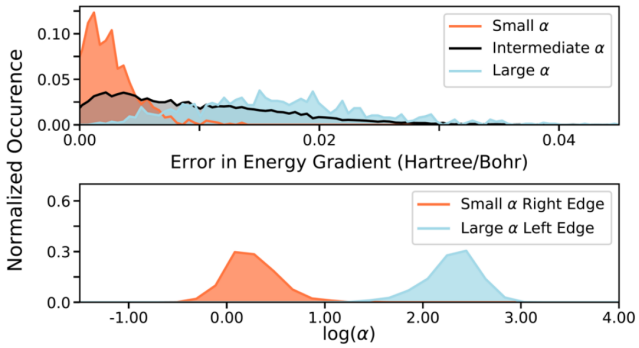

Figure 6: Top panel: the error of IMRR when the hyperparameter α is in different regions in Figure 5. Bottom panel: the histogram of the position of the edges of the small α and large α regions.

To identify the optimal α that empirically minimizes the error of IMRR, a histogram of the errors of IMRR from the aforementioned three regions is depicted in the top panel of Fig. 6. The data show that when an α value is small, the error of IMRR is overwhelmingly smaller than those in the large α region (93%, area under the blue curve). It is true that the possible minimal error of IMRR could correspond to an α value within the intermediate region, as the bottom panel of Fig. 5 shows, nevertheless, the intermediate region still statistically (61%, area under the black curve in the top panel of Fig. 6) has a larger error than those in the coral region. As a result, the optimal α is empirically set to be in the coral region in Fig. 5, whose position is detected by consolidating its upper bound (i.e., the right edge). A histogram of the position of these upper bounds is depicted in the bottom panel of Fig. 6, and a value of 1.0 x 10^{-1} is chosen as the default for the IMRR.

As a brief summary of the error of IMRR reported so far, it is worth noting that the theory in the Methods section only deals with the upper bound of the error. To demonstrate the behavior of the upper bound, not only is an enormous amount of sampling (i.e., targets) required, it is also of little use to the actual dynamics simulation. Nonetheless, the empirical data provided so far demonstrate that by optimizing the weight of the inputs, which are controlled over K, t_{cut}, and α, the error of IMRR is well-behaved and the IMRR energy gradient approaches its *ab initio* counterpart. This can further be seen in Fig. S3 of the Supporting Information, where the average error of IMRR of various t_{cut} is plotted against $R_1 + \alpha R_2$ – as $R_1 + \alpha R_2$ gets smaller, the error of IMRR monotonically decreases.

3.3 The Computational Cost of IMRR

The computational cost for one time step (i.e., update the configuration of the system once) in molecular dynamics simulations can be decomposed into two

parts: the generation of the energy gradient (e.g., *ab initio* calculation, force field evaluation, etc.) and all other overhead cost (e.g., propagate the system, evaluation trajectories, etc.). As discussed in the Introduction, the former makes up the overwhelming majority of the computational cost in AIMD. With a ML method like IMRR that aims to replace a majority of the *ab initio* energy gradient calculations, the overhead cost could possibly become the rate-limiting step in the simulation. This would be the case if searching through previous trajectories' energy gradients for satisfactory IMRR inputs takes an excessive amount of time. It is obvious that the speed of identifying inputs of IMRR (configurations that are geometrically close to the target) among an enormous number of configurations depends heavily on the data structure of those configurations, the searching algorithm, the hardware of the computer, etc. A thorough discussion on that front is beyond the scope of this manuscript, nonetheless, here we provide a preliminary computational cost of IMRR, including its overhead cost, with a bare bone protocol that is subject to further improvement.

Consider the set of all *ab initio* energy gradients that are computed in the early phase of the AIMD study as the "library" of available inputs for IMRR. Clearly, the overhead cost would be unmanageable should the entire library be searched through for the aforenoted inputs for each IMRR. To address this concern, the same pair of CVs described earlier are employed to construct the library. Before an *ab initio* energy gradient (i.e., $g(q)$) is put into the library, its corresponding CVs are calculated, and $g(q)$ would be written/appended into a file that is indexed by the CVs. With this setting, the library contains numerous files and each file stores only those $g(q)$ that share similar CVs. When searching for inputs of IMRR, only those files sharing CVs similar to the target of IMRR are loaded in the memory. The premise is that the $g(q)$ stored in these files are likely to be similar to the energy gradient of the target, and thus likely to be selected as IMRR inputs. As a result, only a small subset of the entire library is relevant to the IMRR input search and the overhead cost is kept at a manageable level. Further, to minimize the I/O of the computer system, a buffer is designed and implemented to store the $g(q)$ from the aforementioned files in the memory. This buffer is updated only when the target has moved significantly away from the previous one. The computational cost of the IMRR is tested with a library of 4,000 AIMD trajectories (about 25 million ab initio energy gradients, 1836285 files, 17.5 GB size) of the $HBr^+ + CO_2 \longrightarrow HOCO^+ + Br$ reaction.

Eight trajectories that are not part of the library are simulated, and their energy gradients (about 50,000) are computed with an *ab initio* calculation (these trajectories are still propagated with ab initio energy gradient) followed by an IMRR. Their timings are compared in Fig. 7. As shown, the wall time of IMRR (0.51 seconds per step on average) is less than a quarter of the wall time of the AIMD (2.16 seconds per step on average) to propagate the system by one step. The most populated bar of IMRR corresponds to those IMRRs that do not require an update of the buffer, and the larger the portion of the buffer needs to be updated, the longer IMRR takes. Further, as the pie chart shows, IMRR spends a majority (almost 90%) of the time on the overhead cost, while

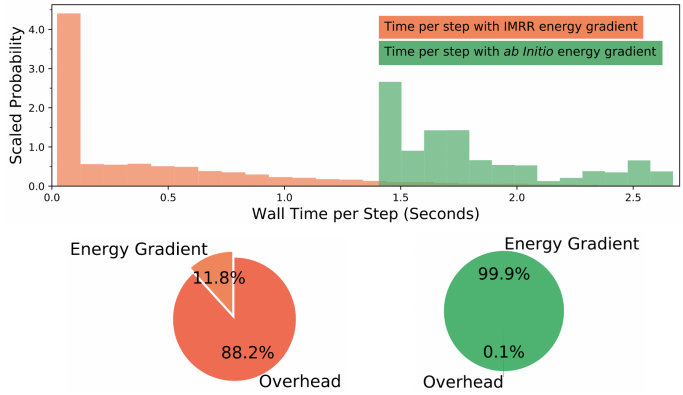

Figure 7: A histogram of the wall time it takes for one MD step if the energy gradient is generated from IMRR (coral) and *ab initio* (green). The heights of the bars are scaled so that they integrate to 1. The average wall times are further broken down in the inserted pie-charts to gradient generation and overhead.

almost all of the wall time of AIMD is spent on generating the energy gradients. It is also important to note that, the timing is measured with IMRR occupying only 1 CPU and ab initio occupying 20 CPUs. The preliminary timing results of IMRR, even though carried out with a bare minimal data structure, bespeaks its great potential efficiency as compared to *ab initio* energy gradient calculations.

4 Discussion

4.1 The Efficacy of IMRR

The previous section presented numerical results on the error of IMRR, which describes how close the estimated energy gradient ($Z(q_0)$, Eq. 1) is to the *ab initio* energy gradient ($g(q_0)$, Eq. 2). The error of IMRR was demonstrated to be affected by (and can be tuned by) the number of inputs (K, left panel of Fig. 4), the geometrical closeness of inputs (t_{cut}, right panel of Fig. 4), and the hyperparameter (α, see Fig. 5). Since the difference in energy gradient between any given pair of configurations should generally decrease as the configurations get geometrically close, without IMRR, one would expect the energy gradient of q_m, the input that is geometrically the closest to the target, to be the closest to the energy gradient of the target. Therefore the efficacy of IMRR (r) is defined as how much the IMRR energy gradient has improved upon $g(q_m)$ in faithfully representing the *ab initio* energy gradient of the target, i.e.,

$$r = \frac{|g(q_0) - g(q_m)|}{|g(q_0) - Z(q_0)|}, \quad m = \underset{1 \le i \le K}{\operatorname{argmin}} |q_0 - q_i| \tag{15}$$

As defined, r is non-negative and the larger its value, the more effective the IMRR is in reproducing the *ab initio* energy gradient of the target as compared to simply the input closest to the target. The same sets of targets as Fig. 4 (83 targets randomly selected from AIMD trajectories that distributed evenly in the CV space) are employed to probe into the efficacy of IMRR with inputs of various t_{cut} values. Fig. 8 demonstrates the results of r with 15 inputs ($K = 15$), which indicates that comparing to the energy gradient of the input (q_m) that is geometrically closest to the target, the IMRR energy gradients are 50 times closer to the target energy gradient. The inserted panels of Fig. 8 show that the efficacy of IMRR gradually increases when t_{cut} gets smaller, indicating that even when the input(s) are very close to the target, IMRR takes advantages of these inputs and estimates a much more accurate energy gradient for the target.

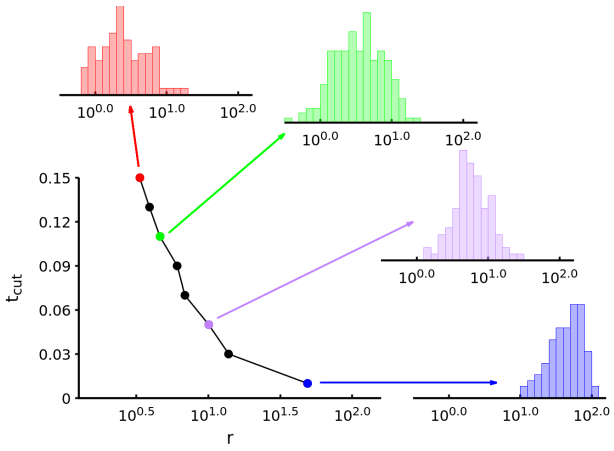

Figure 8: The average efficacy of IMRR (r) is plotted across the range of t_{cut} used in the dataset for Figure 4. For each t_{cut} the distribution of efficacies observed are plotted above the corresponding average.

The efficacy of IMRR at small t_{cut} indicates that it can be closely coupled with AIMD as a time interval (δt, time between updating configurations of the system) multiplier in addition to the active learning discussed in the Introduction. In AIMD, should be chosen as large as possible while conserving the total energy of the system and an *ab initio* energy gradient is calculated every δt to propagate the trajectory. δt is sub-femtosecond and the configurations of consecutive steps are geometrically very close. Therefore, if IMRR could estimate an energy gradient better representing the target than the energy gradient from the previous step does, one can propose a small integer n (e.g., $n = 2, 3, 4, ...$) such that for n steps, the trajectory is propagated with the *ab initio* energy gradient only once and the rest of the $(n-1)$ steps are propagated with IMRR energy gradients. In such applications, the inputs of IMRR could fall into two categories: the "history", those (h) inputs that are the previous h steps of the same trajectory, and the "library", those $(K-h)$ inputs that are geometrically close to the target from previous trajectories. The motivation is to have IMRR build upon the "history" with information from the "library" to effectively reduce the number of *ab initio* calculations (i.e., $n = 2$ will make the simulation almost twice as fast).

The error of IMRR with $h = 1$ and 2 are provided in Fig. 9. The targets of these IMRR are from an AIMD trajectory of the $HBr^+ + CO_2$ reaction ($\sim$4000 targets), and $(15 - h)$ inputs of various t_{cut} are selected from the library of 4000 trajectories. The results show that including just one or two "history" inputs can dramatically decrease the error of IMRR – for example, the error of IMRR with 2 inputs from the previous steps + 13 inputs of a t_{cut} of 0.15 A is on the same level as the error of IMRR with 15 inputs of t_{cut} of 0.01 A. It is also important to note that the error of IMRR does not further decrease with more than 2 "history" points which could be allotted to several factors. First, as more "history" inputs are included, inputs that are geometrically further and further away from the target (i.e., larger t_{cut}) are included. Previous results (Fig. 4) have shown that when the number of inputs is fixed, the error of IMRR increases with respect to t_{cut}. Second, "history" inputs may be nearly linearly dependent over short periods of time where the momentum changes little. As the IMRR linearly combines coordinates to determine the weights, having more than two of these nearly linearly dependent inputs contributes little. Nonetheless, it is important to point out that the inclusion of "history" inputs aligns well with the nature of local interpolation of IMRR and the level of the error as a result is on par or superior to other state of the art ML methods.

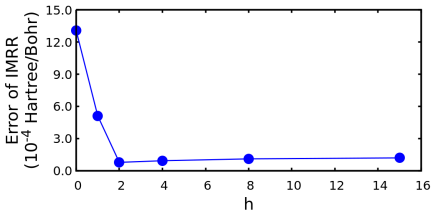

Figure 9: The average error of IMRR is plotted against several values of h.

4.2 The Risk of IMRR

As suggested in the Method section, IMRR utilizes local information, thus in principle avoids the risk of over- and under-fitting certain regions of the phase space that is inherited with those MLs utilizing a globally-fitted function. It has also been emphasized that, unlike other MLs, IMRR has the potential of providing a means to assess its risk associated its energy gradient. This risk is defined as the likelihood of the error of IMRR being more than the level desired by the user or required to maintain a stable trajectory – whichever is smaller. Recall that the upper limit of the error of IMRR is governed by the nature of the PES of the target (C_1 and C_2 terms in Eq. 13) and the terms related to the relative position of the inputs with respect to the target (R_1 and R_2 terms in Eq. 13) – IMRR does not provide any information on C_1 or C_2, but once the weights of the inputs are determined, the values of R_1 and R_2 become available. The theory of IMRR suggests that the upper bound of its error decreases monotonically if the values of R_1 and R_2 decrease. Therefore, when everything else (the geometrical closeness of the inputs, number of inputs, and hyperparameter) is equal, the error of those IMRR associated with relatively small R_1 and R_2 is expected to be small. To verify, a series of IMRR with 23,000

targets randomly selected from 160 AIMD trajectories and their corresponding R_1 and R_2 values are summarized in Fig. 10. The inputs of the IMRR are obtained from a library containing 4,000 AIMD trajectories and no "history" inputs are included. The left panel of Fig. 10 depicts the distribution of R_1 and R_2 values from these IMRR calculations and the region of small R_1 and R_2 values (bottom left) is highly populated. As the theorem suggests, the right panel of Fig. 10, the average error of IMRR with respect to R_1 and R_2, demonstrates that this region corresponds to low average error of IMRR.

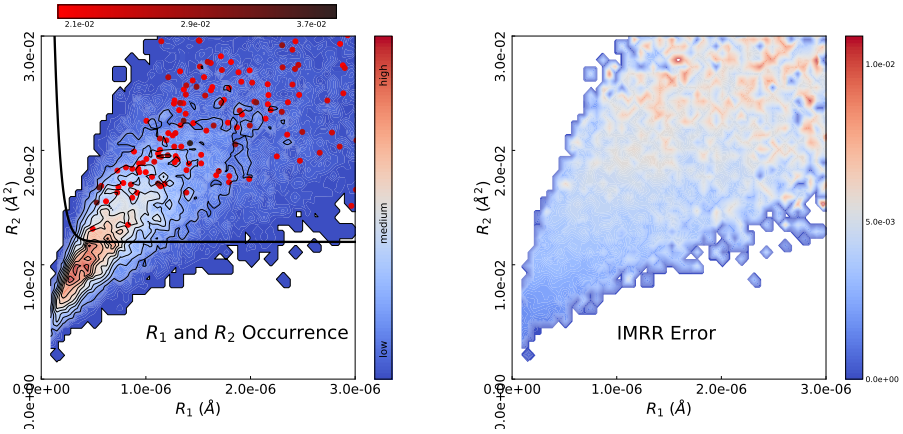

Figure 10: Left panel: the R_1 and R_2 values of 23,000 IMRR calculations are binned and plotted; the occurence of each R_1 and R_2 value is colored by the blue-white-red palette. The 0.1% of targets with the highest errors of IMRR have their R_1 and R_2 plotted on top of this with their errors colored by the red-black palette. Right panel: after binning the targets by R_1 and R_2, the average error of IMRR can also be computed and is colored by the blue-white-red palette.

The right panel of Figure 10 verifies another ideal behavior of IMRR – the error of IMRR generally increases as R_1 and R_2 from IMRR move from the bottom left to the top right of the figure. Therefore, they could be used as indicators to assess the risk of IMRR, e.g., almost certainly, an IMRR of both large R_1 and R_2 implies a large error, thus high risk, while an IMRR of both small R_1 and R_2 implies a small error, thus low risk. It is important to emphasize that all the IMRR in Figure 10 is prepared as suggested (number and geometrically closeness of the inputs, hyperparameter) in the Results section – filtering those IMRR with respect to R_1 and R_2 value is to add a safety net to get rid of those of high error. Several R-related thresholds have been proposed and their efficacy in eliminating those IMRR of high error in energy gradient can be seen in the additional panel of Figure 10.

Being able to assess the risk of the predicted energy gradient of a ML method locally is a valuable means to precisely control its usage – only low-risk results

from the ML method are adapted for the simulation. This feature is critical for the application of IMRR in active learning AIMD: With access to the risk of IMRR on-the-fly, the trajectory is not obligated to propagate with $Z(q_0)$ if it is deemed to be of high risk, but instead, calls an *ab initio* calculation to evaluate its energy gradient, which will be used to propagate the trajectory. Although the number of energy gradient from IMRR eliminated by the R-related thresholds could be small compared to total number of IMRR, we note that the system is physically defined as a chaotic system and the dynamics could be thrown off by just one single configuration with an energy gradient of high error. This feature also goes hand in hand with IMRR's nature of local interpolation – since the risk is local (i.e., related to the amount of the information stored in the library that can be used for the target), the decision of trusting IMRR or referring back to *ab initio* should only take into account the local information. Therefore, at least in theory, the performance of IMRR is irrelevant to the number of unique molecules the system experiences, whether it is just the thermal vibration of a single molecules or a bimolecular collision that visits several intermediates.

References

[1] S. Pratihar, X. Ma, Z. Homayoon, G. L. Barnes, and W. L. Hase, "Direct chemical dynamics simulations," *Journal of the American Chemical Society*, vol. 139, no. 10, pp. 3570–3590, 2017.

[2] J. M. Bowman and P. L. Houston, "Theories and simulations of roaming," *Chemical Society Reviews*, vol. 46, no. 24, pp. 7615–7624, 2017.

[3] J. Xie and W. L. Hase, "Rethinking the sn2 reaction," *Science*, vol. 352, no. 6281, pp. 32–33, 2016.

[4] J. Mikosch, S. Trippel, C. Eichhorn, R. Otto, U. Lourderaj, J. Zhang, W. Hase, M. Weidemuller, and R. Wester, "Imaging nucleophilic substitution dynamics," *Science*, vol. 319, no. 5860, pp. 183–186, 2008.

[5] J. Mikosch, J. Zhang, S. Trippel, C. Eichhorn, R. Otto, R. Sun, W. A. De Jong, M. Weidemuller, W. L. Hase, and R. Wester, "Indirect dynamics in a highly exoergic substitution reaction," *Journal of the American Chemical Society*, vol. 135, no. 11, pp. 4250–4259, 2013.

[6] R. Sun, C. J. Davda, J. Zhang, and W. L. Hase, "Comparison of direct dynamics simulations with different electronic structure methods. f-+ ch 3 i with mp2 and dft/b97-1," *Physical Chemistry Chemical Physics*, vol. 17, no. 4, pp. 2589–2597, 2015.

[7] J. Zhang, U. Lourderaj, R. Sun, J. Mikosch, R. Wester, and W. L. Hase, "Simulation studies of the cl-+ ch3i sn2 nucleophilic substitution reaction: comparison with ion imaging experiments," *The Journal of chemical physics*, vol. 138, no. 11, p. 114309, 2013.

[8] Y. T. Lee and Y. R. Shen, "Molecular beam studies of ir laser induced multiphoton dissociation and vibrational predissociation," Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States), Tech. Rep., 1980.

[9] M. E. Tuckerman, "Ab initio molecular dynamics: basic concepts, current trends and novel applications," *Journal of Physics: Condensed Matter*, vol. 14, no. 50, p. R1297, 2002.

[10] R. Iftimie, P. Minary, and M. E. Tuckerman, "Ab initio molecular dynamics: Concepts, recent developments, and future trends," *Proceedings of the National Academy of Sciences*, vol. 102, no. 19, pp. 6654–6659, 2005.

[11] M. Paranjothy, R. Sun, Y. Zhuang, and W. L. Hase, "Direct chemical dynamics simulations: coupling of classical and quasiclassical trajectories with electronic structure theory," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 3, no. 3, pp. 296–316, 2013.

[12] U. Lourderaj, K. Park, and W. L. Hase, "Classical trajectory simulations of post-transition state dynamics," *International Reviews in Physical Chemistry*, vol. 27, no. 3, pp. 361–403, 2008.

[13] K. Raghavachari, G. W. Trucks, J. A. Pople, and M. Head-Gordon, "A fifth-order perturbation comparison of electron correlation theories," *Chemical Physics Letters*, vol. 157, no. 6, pp. 479–483, 1989.

[14] T. H. Dunning Jr, "Gaussian basis sets for use in correlated molecular calculations. i. the atoms boron through neon and hydrogen," *The Journal of chemical physics*, vol. 90, no. 2, pp. 1007–1023, 1989.

[15] A. D. Beck, "Density-functional thermochemistry. iii. the role of exact exchange," *J. Chem. Phys*, vol. 98, no. 7, pp. 5648–6, 1993.

[16] C. M. Aikens, S. P. Webb, R. L. Bell, G. D. Fletcher, M. W. Schmidt, and M. S. Gordon, "A derivation of the frozen-orbital unrestricted open-shell and restricted closed-shell second-order perturbation theory analytic gradient expressions," *Theoretical Chemistry Accounts*, vol. 110, no. 4, pp. 233–253, 2003.

[17] R. Ditchfield, W. J. Hehre, and J. A. Pople, "Self-consistent molecular-orbital methods. ix. an extended gaussian-type basis for molecular-orbital studies of organic molecules," *The Journal of Chemical Physics*, vol. 54, no. 2, pp. 724–728, 1971.

[18] K. A. Peterson, D. E. Woon, and T. H. Dunning Jr, "Benchmark calculations with correlated molecular wave functions. iv. the classical barrier height of the h+ h2 h2+ h reaction," *The Journal of chemical physics*, vol. 100, no. 10, pp. 7410–7415, 1994.

[19] K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. Von Lilienfeld, A. Tkatchenko, and K.-R. Muller, "Assessment and validation of machine learning methods for predicting molecular atomization energies," *Journal of Chemical Theory and Computation*, vol. 9, no. 8, pp. 3404–3419, 2013.

[20] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. Von Lilienfeld, K.-R. Muller, and A. Tkatchenko, "Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space," *The journal of physical chemistry letters*, vol. 6, no. 12, pp. 2326–2331, 2015.

[21] F. A. Faber, L. Hutchison, B. Huang, J. Gilmer, S. S. Schoenholz, G. E. Dahl, O. Vinyals, S. Kearnes, P. F. Riley, and O. A. Von Lilienfeld, "Prediction errors of molecular machine learning models lower than hybrid dft error," *Journal of chemical theory and computation*, vol. 13, no. 11, pp. 5255–5264, 2017.

[22] F. Noe, A. Tkatchenko, K.-R. Muller, and C. Clementi, "Machine learning for molecular simulation," *Annual review of physical chemistry*, vol. 71, pp. 361–390, 2020. [Online]. Available: https://www.annualreviews.org/doi/full/10.1146/annurev-physchem-042018-052331

[23] J. Behler, "First principles neural network potentials for reactive simulations of large molecular and condensed systems," *Angewandte Chemie International Edition*, vol. 56, no. 42, pp. 12 828–12 840, 2017.

[24] J. Behler and M. Parrinello, "Generalized neural-network representation of high-dimensional potential-energy surfaces," *Physical review letters*, vol. 98, no. 14, p. 146401, 2007.

[25] K. J. Jose, N. Artrith, and J. Behler, "Construction of high-dimensional neural network potentials using environment-dependent atom pairs," *The Journal of chemical physics*, vol. 136, no. 19, p. 194111, 2012.

[26] M. Gastegger, J. Behler, and P. Marquetand, "Machine learning molecular dynamics for the simulation of infrared spectra," *Chemical science*, vol. 8, no. 10, pp. 6924–6935, 2017.

[27] J. S. Smith, O. Isayev, and A. E. Roitberg, "Ani-1: an extensible neural network potential with dft accuracy at force field computational cost," *Chemical science*, vol. 8, no. 4, pp. 3192–3203, 2017.

[28] V. Botu and R. Ramprasad, "Learning scheme to predict atomic forces and accelerate materials simulations," *Physical Review B*, vol. 92, no. 9, p. 094306, 2015.

[29] F. Brockherde, L. Vogt, L. Li, M. E. Tuckerman, K. Burke, and K.-R. Müller, "Bypassing the kohn-sham equations with machine learning," *Nature communications*, vol. 8, no. 1, pp. 1–10, 2017. [Online]. Available: https://www.nature.com/articles/s41467-017-00839-3

[30] V. Botu, R. Batra, J. Chapman, and R. Ramprasad, "Machine learning force fields: construction, validation, and outlook," *The Journal of Physical Chemistry C*, vol. 121, no. 1, pp. 511–522, 2017.

[31] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, "Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons," *Physical review letters*, vol. 104, no. 13, p. 136403, 2010.

[32] A. P. Bartók, R. Kondor, and G. Csányi, "On representing chemical environments," *Physical Review B*, vol. 87, no. 18, p. 184115, 2013.

[33] J. Ischtwan and M. A. Collins, "Molecular potential energy surfaces by interpolation," *The Journal of chemical physics*, vol. 100, no. 11, pp. 8080–8088, 1994.

[34] M. A. Collins, "Molecular potential energy surfaces by interpolation," in *International Conference on Computational Science*. Springer, 2003, pp. 159–167.

[35] G. G. Maisuradze, A. Kawano, D. L. Thompson, A. F. Wagner, and M. Minkoff, "Interpolating moving least-squares methods for fitting potential energy surfaces: Analysis of an application to a six-dimensional system," *The Journal of chemical physics*, vol. 121, no. 21, pp. 10 329–10 338, 2004.

[36] Y. Guo, A. Kawano, D. L. Thompson, A. F. Wagner, and M. Minkoff, "Interpolating moving least-squares methods for fitting potential energy surfaces: Applications to classical dynamics calculations," *The Journal of chemical physics*, vol. 121, no. 11, pp. 5091–5097, 2004.

[37] B. J. Braams and J. M. Bowman, "Permutationally invariant potential energy surfaces in high dimensionality," *International Reviews in Physical Chemistry*, vol. 28, no. 4, pp. 577–606, 2009.

[38] T.-S. Ho and H. Rabitz, "A general method for constructing multidimensional molecular potential energy surfaces from ab initio calculations," *The Journal of chemical physics*, vol. 104, no. 7, pp. 2584–2597, 1996.

[39] T.-S. Ho, T. Hollebeek, H. Rabitz, L. B. Harding, and G. C. Schatz, "A global h2o potential energy surface for the reaction o (1 d)+ h2 $\longrightarrow$ oh+ h," *The Journal of chemical physics*, vol. 105, no. 23, pp. 10 472–10 486, 1996.

[40] T. Hollebeek, T.-S. Ho, and H. Rabitz, "A fast algorithm for evaluating multidimensional potential energy surfaces," *The Journal of chemical physics*, vol. 106, no. 17, pp. 7223–7227, 1997.

[41] K. C. Thompson, M. J. Jordan, and M. A. Collins, "Polyatomic molecular potential energy surfaces by interpolation in local internal coordinates," *The Journal of chemical physics*, vol. 108, no. 20, pp. 8302–8316, 1998.

[42] T. Hollebeek, T.-S. Ho, and H. Rabitz, "Constructing multidimensional molecular potential energy surfaces from ab initio data," *Annual review of physical chemistry*, vol. 50, no. 1, pp. 537–570, 1999.

[43] T.-S. Ho, H. Rabitz, F. J. Aoiz, L. Banares, S. A. Vázquez, and L. B. Harding, "Implementation of a fast analytic ground state potential energy surface for the n (2 d)+ h 2 reaction," *The Journal of chemical physics*, vol. 119, no. 6, pp. 3063–3070, 2003.

[44] S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, and K.-R. Müller, "Machine learning of accurate energy-conserving molecular force fields," *Science advances*, vol. 3, no. 5, p. e1603015, 2017.

[45] H. E. Sauceda, S. Chmiela, I. Poltavsky, K.-R. Müller, and A. Tkatchenko, "Molecular force fields with gradient-domain machine learning: Construction and application to dynamics of small molecules with coupled cluster forces," *The Journal of chemical physics*, vol. 150, no. 11, p. 114102, 2019.

[46] L. Paetow, F. Unger, W. Beichel, G. Frenking, and K.-M. Weitzel, "Rotational dependence of the proton-transfer reaction $hbr^+ + co_2 \longrightarrow hoco^+ +$ br. i. energy versus angular momentum effects," *The Journal of chemical physics*, vol. 132, no. 17, p. 174305, 2010.

[47] L. Paetow, F. Unger, B. Beutel, and K.-M. Weitzel, "Rotational dependence of the proton-transfer reaction $hbr^+ + co_2 \longrightarrow hoco^+ +$ br. ii. comparison of hbr^+ $(2\pi 3/2)$ and hbr^+ $(2\pi 1/2)$," *The Journal of chemical physics*, vol. 133, no. 23, p. 234301, 2010.

[48] A. Shoji, D. Schanzenbach, R. Merrill, J. Zhang, L. Yang, and R. Sun, "Theoretical study of the potential energy profile of the $hbr^+ + co_2 \longrightarrow hoco^+$ br. reaction," *The Journal of Physical Chemistry A*, vol. 123, no. 45, pp. 9791–9799, 2019.

[49] Y. Luo, T. Kreuscher, C. Kang, W. L. Hase, K.-M. Weitzel, and R. Sun, "A chemical dynamics study of the hcl+ hcl^+ reaction," *International Journal of Mass Spectrometry*, vol. 462, p. 116515, 2021.

[50] E. A. Coutsias, C. Seok, and K. A. Dill, "Using quaternions to calculate rmsd," *Journal of computational chemistry*, vol. 25, no. 15, pp. 1849–1857, 2004.

[51] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.