

# Interpolating Moving Ridge Regression (IMRR): A machine learning algorithm to predict energy gradients for *ab initio* molecular dynamics simulations

Kazuumi Fujioka, Rui Sun<sup>\*</sup>

University of Hawai'i at Manoa, 2500 Campus Rd, Honolulu, HI 96822, United States

## ARTICLE INFO

### Keywords:

Ab initio molecular dynamics  
Direct dynamics  
Quantum chemistry  
Machine learning

## ABSTRACT

*Ab initio* molecular dynamics (AIMD) simulations 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 *ab initio* energy gradient calculations needed, 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 developed 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 its predicted energy gradient. A hyper-parameter used to guide IMRR is rigorously examined as well. IMRR is then applied to three bimolecular reactions studied with AIMD, including  $\text{HBr}^+ + \text{CO}_2$ ,  $\text{H}_2\text{S} + \text{CH}$ , and  $\text{C}_4\text{H}_2 + \text{CH}$ , to demonstrate its 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 [1–8]. 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 a "trajectory") is collected. To ensure the conservation of the physical properties of the system (e.g., total energy, momentum, etc.), the time interval between updating the coordinates of the atoms of a trajectory is usually on 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 trajectory.

Further, to accurately model reactions in real life, AIMD simulations 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 ( $\theta$ ). Practically, this is done by first detecting  $b_{\text{max}}$ , the largest  $b$  in which a reactive trajectory can be observed, and then sampling trajectories with random orientations within  $b_{\text{max}}$ . 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,  $b_{\text{max}}$  is usually a few (4.0–6.0) Å when the collision energy is less than 1.0 eV [4,6–8]. 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 an enormous

<sup>\*</sup> Corresponding author.

E-mail address: [ruisun@hawaii.edu](mailto:ruisun@hawaii.edu) (R. Sun).

<https://doi.org/10.1016/j.chemphys.2022.111482>

Received 21 September 2021; Received in revised form 29 December 2021; Accepted 9 February 2022

Available online 23 February 2022

0301-0104/© 2022 Elsevier B.V. All rights reserved.

computation cost for a simulation study of one chemical reaction under just one condition (e.g., a certain collision energy, temperature, vibrational excitation, etc.).

The millions of *ab initio* energy gradient calculations take up the overwhelming majority of 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. On one hand, for example, coupled cluster theories with triple-zeta 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% [4,6–8]. 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 [17] or 6-31G\* [18,19]). Selecting a feasible yet accurate combination of *ab initio* method and basis set is laborious: the potential energy of a chemical reaction calculated from various combinations are compared against experimental heats of reaction and/or results from a high-level *ab initio* method (e.g., CCSD(T) extrapolated to the complete basis set limit [20]).

The large 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 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 Hansen et al. [21,22] and Faber et al. [23], as well as a more general review by Noe et al. [24]. 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 [25–29], kernel ridge regression [30–33], or Gaussian process regression [34,35]. 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 [36–40], reproducing kernel interpolation [41–46], or kernel ridge regression [47,48].

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 ( $q_0$ ) which is referred to as the “target of IMRR”. The outcome of IMRR is  $Z(q_0)$ , the estimated energy gradient of  $q_0$ , which is necessary to propagate the trajectory. IMRR also assesses the risk of  $Z(q_0)$ , which is defined as the likelihood of  $Z(q_0)$  deviating from the true *ab initio* energy gradient more than some user-defined threshold. Targets with large risk may reject  $Z(q_0)$  and instead fall back to the *ab initio* energy gradient to propagate forward. It is important to note that the risk-assessment of IMRR is done without computing the *ab initio* gradient of the target. This type of risk or uncertainty prediction has often been used in ML methods in conjunction with *active learning* algorithms [24,49,50]. Actively learning through repeated sampling and retraining has been found to model potential energy surfaces to the level of chemical accuracy often by adding configurations to the training set which have high uncertainty [51–54]. While many of these involve having multiple neural networks that assess each others’ risks, [51–54] others construct theoretical probabilistic uncertainties [55–58].

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. b) IMRR’s risk-assessing capability features the flexibility of referring back to the *ab initio* energy gradient when necessary. Combined 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. And c), IMRR is highly efficient—as shown later in this manuscript, its computational cost 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 AIMD 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 upper bound of 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 practically minimizing the deviation, 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  $g(q)$ , which can be described 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  $q_0$ , referred to as the “target” of IMRR, 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 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:

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

IMRR optimizes  $w_i$  in order to minimize the deviation (referred to as the “error of IMRR”) between  $Z(q_0)$  and  $g(q_0)$ , which is expressed as:

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

Intuitively, those  $q_i$  that are geometrically close to  $q_0$  should be prioritized in making up the training set. With Cartesian coordinates, the “geometrical closeness” between  $q_i$  and  $q_0$  is assessed by the root mean square displacement (RMSD,  $t_i$ ) after  $q_i$  has been properly translated and rotated to maximize its overlap with  $q_0$  [59]. It is important to note that permutation should be allowed for chemically identical atoms if it increases the overlap. The RMSD between  $q_i$  and  $q_0$  is computed as:

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

in which  $t_{cut}$  is a user-defined parameter that enforces the geometrical closeness between  $q_0$  and  $q_i$ , the relation between which is:

$$q_i = q_0 + \sqrt{N} t_i \hat{h}_i \quad (4)$$

in which  $\hat{h}_i$  is the unit vector of  $h_i$ , the displacement between  $q_i$  and  $q_0$ , i.e.,  $h_i = q_i - q_0 \in \mathbb{R}^{3N}$ . An example geometric interpretation for these vectors can be seen for the  $K = 2$  case in Fig. 1.  $f_i(t)$  is defined as the energy gradient function, i.e.,  $f_i(t_i) = g(q_0 + \sqrt{N} t_i \hat{h}_i) = g(q_i) \in \mathbb{R}^{3N}$ . Assume the chemical system stays in the same electronic state (adiabatic process); the  $f_i(t_i)$  is continuous and infinitely differentiable in each of its  $3N$  components. Therefore, the  $j^{\text{th}}$  component of function  $f_i$  can be expanded with Taylor's theorem:

$$f_{i,j}(t_i) = f_{i,j}(0) + t_i f'_{i,j}(0) + \frac{1}{2} t_i^2 f''_{i,j}(0) + \frac{1}{3!} t_i^3 f'''_{i,j}(0) + \dots \quad 1 \leq i \leq K, 1 \leq j \leq 3N \quad (5)$$

As defined, the first term  $f_{i,j}(0)$  is the  $j^{\text{th}}$  component of  $g(q_0)$ , i.e.,  $f_{i,j}(0) = g_j(q_0)$ . The error of IMRR (Eq. 2) is bounded above by considering the  $l^{\text{th}}$  component ( $1 \leq l \leq 3N$ ) of  $Z(q_0)$  where it deviates the most from  $g(q_0)$ , i.e.,

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

The inequality can be further derived as

$$\begin{aligned} & |g(q_0) - Z(q_0)| \leq \sqrt{3} \cdot \left| g_l(q_0) - \sum_{i=1}^K w_i g_l(q_i) \right| \\ & \leq \sqrt{3} \cdot \left| g_l(q_0) - \sum_{i=1}^K w_i f_{i,l}(0) \right| + \sqrt{3} \cdot \left| \sum_{i=1}^K w_i t_i f'_{i,l}(0) \right| + \\ & \quad \sqrt{3} \cdot \left| \sum_{i=1}^K \left( w_i \frac{1}{2} t_i^2 f''_{i,l}(0) + w_i \frac{1}{3!} t_i^3 f'''_{i,l}(0) + \dots \right) \right| \end{aligned} \quad (6)$$

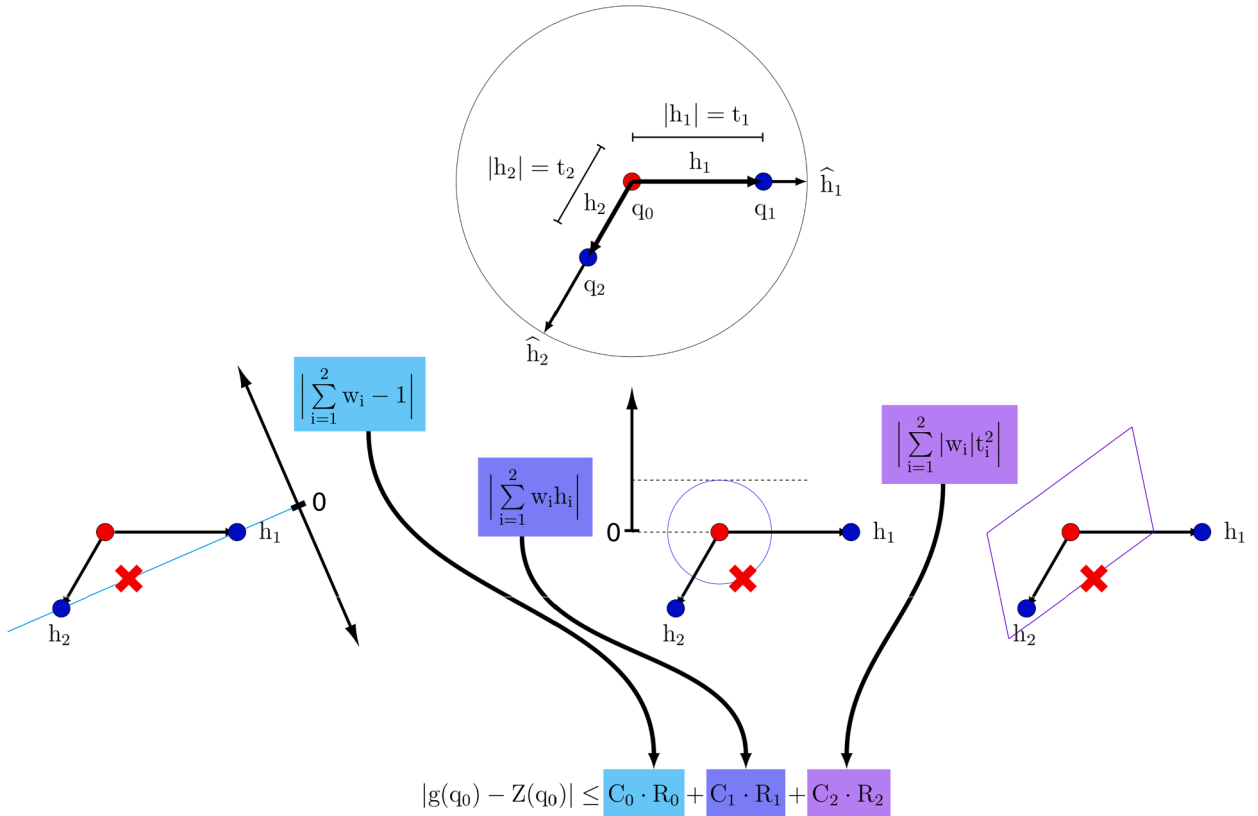
The first term can be rewritten as:

$$\begin{aligned} \sqrt{3} \cdot \left| g_l(q_0) - \sum_{i=1}^K w_i f_{i,l}(0) \right| &= \sqrt{3} \cdot |g_l(q_0)| \cdot \left| 1 - \sum_{i=1}^K w_i \right| \\ &= \underbrace{\sqrt{3} \cdot |g_l(q_0)|}_{C_0} \cdot \underbrace{\left| \sum_{i=1}^K w_i - 1 \right|}_{R_0} \end{aligned} \quad (7)$$

in which  $C_0$  depends only on the nature of the potential energy surface (specifically, its derivative) and  $R_0$  depends only on the weights of  $q_i$ . Similarly, the second term in Eq. 6 can 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  $w_i$ :

$$\begin{aligned} \sqrt{3} \cdot \left| \sum_{i=1}^K w_i t_i f'_{i,l}(0) \right| &= \sqrt{3} \cdot \left| \sum_{i=1}^K g'_l(q_0) \cdot w_i \sqrt{N} \hat{h}_i \right| \\ &\leq \underbrace{\sqrt{3} \cdot |g'_l(q_0)|}_{C_1} \cdot \underbrace{\left| \sum_{i=1}^K w_i h_i - \mathbf{0} \right|}_{R_1} \end{aligned} \quad (8)$$

in which  $|g'_l(q_0)|$  is the magnitude of the largest value among the  $3N \times$



**Fig. 1.** An example IMRR interpolation for a target (red circle) given two inputs (blue circles) to produce an interpolated configuration (red cross) is given.  $R_0$  is the difference between the sum of the weights and unity whereas  $R_1$  is the distance in space between the interpolated configuration and the target.  $R_2$  is the weighted sum of the magnitudes of the inputs. Cyan, dark blue, and purple lines indicate possible interpolation configurations which have the same value of  $R_0$ ,  $R_1$ , and  $R_2$ , respectively.

$3N$  elements of the matrix  $g'(q_0)$ , the Hessian of  $q_0$ . Finally, 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 higher order potential energy derivatives. Thus, the third term in Eq. 6 can be expressed as the product of two terms: a  $C_2$  term that depends only on the nature of the potential energy surface and a  $R_2$  term that depends only on  $q_i$  and their weights  $w_i$ , e.g.,

$$\begin{aligned} \sqrt{3} \cdot \sum_{i=1}^K \left| w_i t_i^2 \left( \frac{1}{2} f''_{i,l} (0) + \frac{1}{3!} t_i f'''_{i,l} (0) + \dots \right) \right| &\leq \sqrt{3} \cdot \sum_{i=1}^K |w_i t_i^2 C_{i,l}| \\ &= \underbrace{\sqrt{3} \cdot C_{i,l}}_{C_2} \cdot \underbrace{\left| \sum_{i=1}^K w_i t_i^2 - 0 \right|}_{R_2} \end{aligned} \quad (9)$$

Substituting Eq. 7, 8 into Eq. 6 establishes that the error in energy gradients between the interpolated configuration and the target configuration is bounded above as:

$$|g(q_0) - Z(q_0)| \leq C_0 R_0 + C_1 R_1 + C_2 R_2 \quad (10)$$

The geometric representations of these  $R$  terms for a model system of two inputs ( $K = 2$ ) are demonstrated in Fig. 1.

## 2.2. Minimize the upper bound of the error

IMRR minimizes its error (Eq. 2) by minimizing the upper bound of the error. Eq. 10 demonstrates that the error is determined by 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. 10 through the  $R$  terms. Eq. 10 can be rewritten into the form of a linear equation:

$$\begin{aligned} |g(q_0) - Z(q_0)| &\leq C_0 R_0 + C_1 R_1 + C_2 R_2 \\ &= |C_0 \mathbf{U}^T \mathbf{w} - C_0 \mathbf{1}| + |C_1 \mathbf{A} \mathbf{w} - C_1 \mathbf{0}| + |C_2 \mathbf{B} \mathbf{w} - C_2 \mathbf{0}| \\ &\leq \sqrt{3} \cdot \sqrt{|C_0 \mathbf{U}^T \mathbf{w} - C_0 \mathbf{1}|^2 + |C_1 \mathbf{A} \mathbf{w} - C_1 \mathbf{0}|^2 + |C_2 \mathbf{B} \mathbf{w} - C_2 \mathbf{0}|^2} \\ &= \sqrt{3} \left\| \begin{bmatrix} C_2 \mathbf{B} \\ C_1 \mathbf{A} \\ C_0 \mathbf{U}^T \end{bmatrix} \mathbf{w} - \begin{bmatrix} C_2 \mathbf{0} \\ C_1 \mathbf{0} \\ C_0 \mathbf{1} \end{bmatrix} \right\| \end{aligned} \quad (11)$$

$\mathbf{A} \in \mathbb{R}^{K \times 3N}$  is the matrix of the displacement between  $q_0$  and  $q_i$ ,  $h_i \in \mathbb{R}^{3N}$ ,  $1 \leq i \leq K$ :

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

$\mathbf{B} \in \mathbb{R}^{K \times K}$  is the diagonal matrix of the magnitude of the displacement between  $q_0$  and  $q_i$ :

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

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

$$\mathbf{U}^T = [1 \quad 1 \quad \dots \quad 1]$$

$\mathbf{w} \in \mathbb{R}^K$  is the weights of the  $K$  input configurations:

$$\mathbf{w} = [w_1 \quad w_2 \quad \dots \quad w_K]^T$$

Finally,  $\mathbf{0} = [0 \quad 0 \quad \dots \quad 0]^T \in \mathbb{R}^K$  and  $\mathbf{1}$  is just the scalar 1. Eq. 11 can be further simplified as:

$$\left| g(q_0) - Z(q_0) \right| \leq \left| \mathbf{H} \mathbf{w} - \mathbf{h}^* \right| = \sqrt{(\mathbf{H} \mathbf{w} - \mathbf{h}^*)^T \cdot (\mathbf{H} \mathbf{w} - \mathbf{h}^*)} = r(\mathbf{w}) \quad (12)$$

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

$$\mathbf{w} = (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{h}^* \quad (13)$$

Eq. 13 can be proven to have a solution  $\mathbf{w}$ , as shown in the Supporting Information.

## 2.3. Solving for the optimal $\mathbf{w}$ with restrictions

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. 10) 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. Therefore,  $R_0$  is set to be zero by imposing the constraint that  $\mathbf{U}^T \mathbf{w} - \mathbf{1} = 0$ . Following the same procedure as shown in the previous section, the upper bound could be rewritten with the linear equation:

$$\begin{aligned} \left| g(q_0) - Z(q_0) \right| &\leq \left\| \begin{bmatrix} C_2 \mathbf{B} \\ C_1 \mathbf{A} \end{bmatrix} \mathbf{w} - \begin{bmatrix} C_2 \mathbf{0} \\ C_1 \mathbf{0} \end{bmatrix} \right\| \\ &= C_1 \left\| \begin{bmatrix} \alpha \mathbf{B} \\ \mathbf{A} \end{bmatrix} \mathbf{w} - \begin{bmatrix} \alpha \mathbf{0} \\ \mathbf{0} \end{bmatrix} \right\| \\ &= \mathbf{H}_r \mathbf{w} - \mathbf{h}_r^* \end{aligned} \quad (14)$$

To solve for this equation, a hyperparameter  $\alpha$  is introduced as the ratio between  $C_2$  and  $C_1$ , i.e.,  $\alpha = C_2/C_1$ . The true value of  $\alpha$  depends on the potential energy surface of the system and is not known *a priori*. As is customary in ML methods, this is left as a user-controlled hyperparameter. As seen in Fig. 2, generally, the IMRR has two limiting behaviours for very small and large  $\alpha$ . In the small  $\alpha$  case, minimizing  $R_1$  is preferred so the weights are chosen to minimize the distance between the interpolated configuration and the target. In the large  $\alpha$  case, minimizing  $R_2$  is preferred so the weights are chosen to minimize the distance between the interpolated frame and the closest possible input.

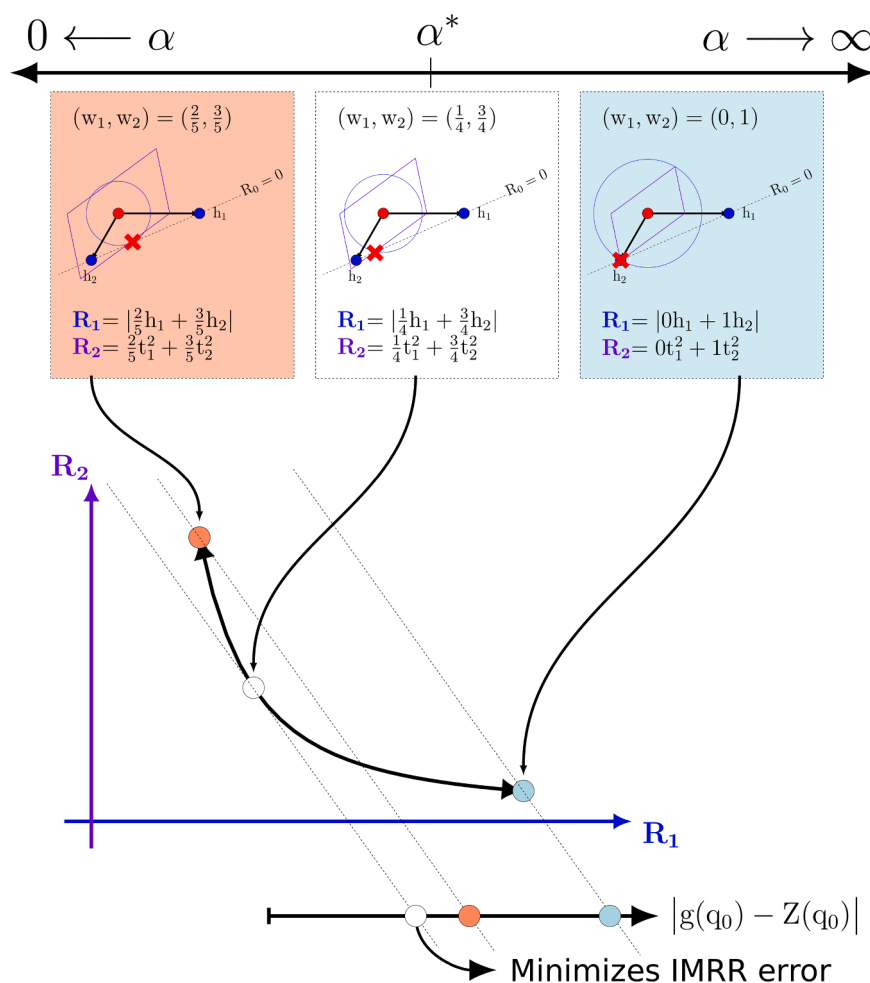
Define  $\mathbf{H}_r = [\alpha \mathbf{B} \quad \mathbf{A}]^T$  and  $\mathbf{h}_r^* = [\alpha \mathbf{0} \quad \mathbf{0}]^T$ . This constrained optimization of  $\mathbf{w}$  that minimizes the upper bound of the error could be solved by constructing a Lagrangian,  $L(\lambda, \mathbf{w})$ :

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

and setting the gradient to zero.

## 3. Results

The first representative system employed to demonstrate the performance of IMRR is the bimolecular collision of  $\text{HBr}^+ + \text{CO}_2$ , which after collision, forms the proton-transfer product  $\text{HOCO}^+ + \text{Br}$ , 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 [60–63]. The IMRR employs *ab initio* energy gradients computed with MP2/cc-pVDZ by NWChem that made



**Fig. 2.** By constraining  $R_0 = 0$ , in the same example from Fig. 1 where there are two inputs ( $K = 2$ ), 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  $\alpha^*$  would lead the algorithm to minimize the error function defined by  $R_1 + \alpha^* R_2$ .

up the previous AIMD trajectories [62,63] as its targets  $(q_0, g(q_0))$  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 ( $\alpha$  in Eq. 14), 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  $g(q_i)$ ,  $K = 15$ ), each of which has an RMSD less than  $0.15 \text{ \AA}$ . ( $t_{cut} = 0.15 \text{ \AA}$ ) to the target  $q_0$ , and the hyperparameter  $\alpha$  is held at  $1.0 \times 10^{-1}$ . Unless noted otherwise, these values are adapted as the default for the rest of the manuscript.

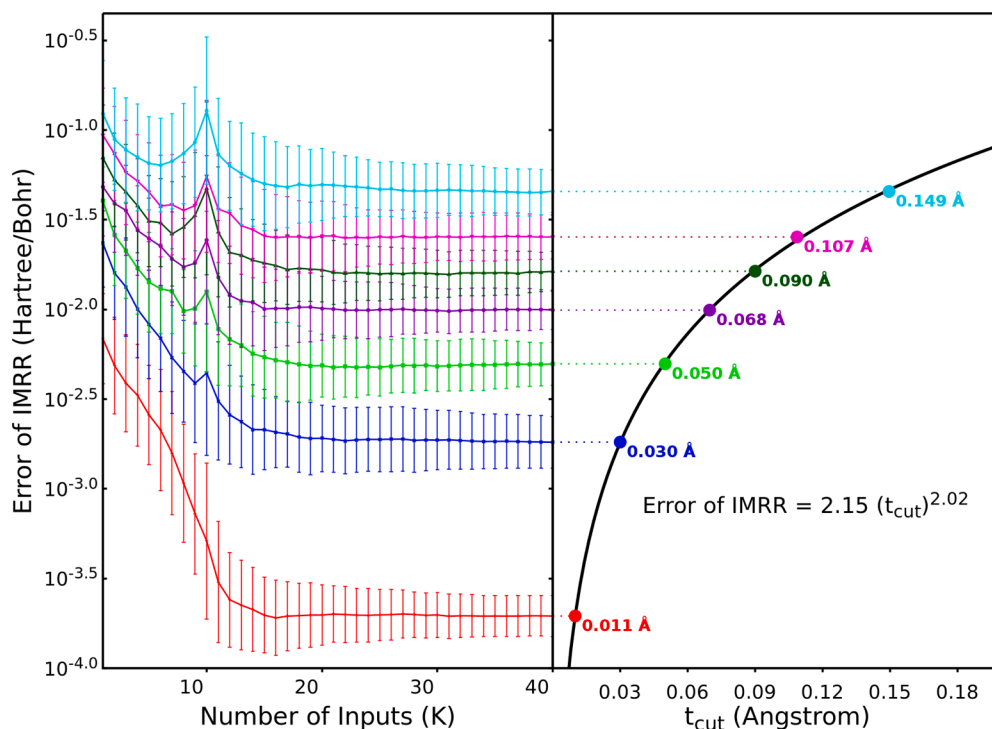
### 3.1. Accuracy of IMRR vs. the geometrical closeness and the number of inputs

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  $t_i$  gets smaller (see  $R_1$  term in Eq. 9 and  $R_2$  term in Eq. 10). Therefore, controlling  $t_{cut}$  is the first trial 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-C and the shorter distance between two Br-O, is provided in Fig. S1 of the Supporting Information. The configurations of 4,000 AIMD trajectories are binned into the CV-space, and  $68 \text{ 1 \AA} \times 1 \text{ \AA}$  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.

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. 3. For each target ( $q_0$ ), up to 40 inputs ( $q_i$ ) are randomly generated by displacing atom(s) from  $q_0$ , while enforcing its  $t_i$  (geometrical closeness to  $q_0$ ) to be between 90% and 100% of each  $t_{cut}$ . The energy gradient of the inputs,  $g(q_i)$ , are computed at MP2/cc-pVDZ level of theory. As Fig. 3 demonstrates, the error of IMRR decreases monotonically as the input configurations,  $q_i$ , get geometrically closer 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. 3 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 function ( $ax^2$ ), indicating that with the optimized weights in Eq. 10, the error is largely bounded by  $C_2 R_2$ , since  $R_2$  is proportional to the sum of the squared  $t_i$  of the input. Overall,





**Fig. 3.** Left panel: the error of IMRR with vs. the number of inputs at various  $t_{\text{cut}}$  values. Right panel: the error of IMRR vs.  $t_{\text{cut}}$  with  $K = 40$  inputs. The targets of these figures uniformly sample the phase space of the  $\text{HBr}^+ + \text{CO}_2$  reaction (see Fig. S1).

the right panel of Fig. 3 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. 3 (left panel) also illustrates the correlation between the error of IMRR and the number of inputs ( $K$ ). For each target, their inputs are sorted with respect to  $t_i$  before being fed into the IMRR. For example,  $K = 2$  means the IMRR is carried out with the two inputs geometrically closest to the target, and  $K = 3$  includes the three inputs geometrically 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  $t_{\text{cut}}$ , is able to estimate an energy gradient closer to the *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 for the  $\text{HBr}^+ + \text{CO}_2$  system. It is interesting to note, the error of IMRR demonstrates a local maximum around  $K = 10$  for almost all  $t_{\text{cut}}$ , which is particularly obvious for larger  $t_{\text{cut}}$ . The origin of this counterintuitive maximum is debatable, while one explanation could be that IMRR is more 'fragile' when inputs are overall geometrically further away from the target (i.e., larger  $t_{\text{cut}}$ ). This behaviour will be elucidated further in the Discussion.

It is important to confirm that the aforementioned 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 atom(s) 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, [64,65] thus roughly 15 inputs are selected from each  $1 \text{ \AA} \times 1 \text{ \AA}$  cell shown in Fig. S1. The energy gradients of these targets are estimated by IMRR with inputs selected from another 4000 AIMD trajectories (a total of 28.1 million energy gradients) with  $t_i$  chosen such that  $0.9t_{\text{cut}} < t_i < t_{\text{cut}}$ , as in the previous test. The results are summarized in the Supporting Information (Fig. S2) and show very good agreement with Fig. 3, 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 $\alpha$

As shown in the previous section, IMRR demands a hyperparameter  $\alpha$  in solving for the optimal weights ( $\mathbf{w}$ ) that minimize the upper bound of the error.

$\alpha$  for each IMRR, since it would be more expensive than just computing the *ab initio* energy gradient itself. Defined in Section 2.3,  $\alpha$  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. Therefore, like many other ML methods, the hyperparameter  $\alpha$  is tested over a range of values and determined empirically to reliably produce minimal IMRR error.

Intuitively, when a variety of inputs are available to IMRR (while controlling the inputs' geometrical closeness to the target by enforcing  $t_i < t_{\text{cut}}$ ), the hyperparameter  $\alpha$  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. 14 suggests, a large  $\alpha$  (blue region in Fig. 4) will minimize the  $C_2R_2$  term of the upper bound of the error; while in contrast, a small  $\alpha$  (coral region in Fig. 4) 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  $\alpha$ , whose value ranges between  $10^{-5}$  and  $10^{+5}$ . Although the exact curve of the error of IMRR vs.  $\alpha$  curve varies from target to target (see Fig. 4), it can be divided into three regions: large  $\alpha$  (blue), small  $\alpha$  (coral), and intermediate  $\alpha$  (white). The first two regions are detected when the error of IMRR becomes independent of  $\alpha$ , although each region could be associated with a different error. The error of IMRR in the intermediate region heavily depends on  $\alpha$  and smoothly connects the other two regions.

To identify the optimal  $\alpha$  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. 5. The data show that when  $\alpha$  is small, the error of IMRR is overwhelmingly smaller than those in the

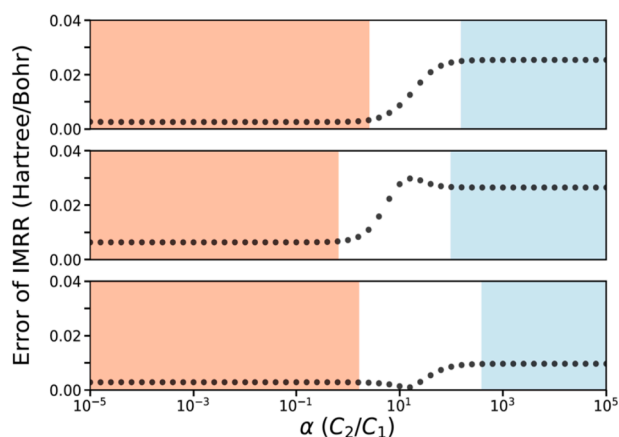


Fig. 4. Three representative behaviors of the error of IMRR vs. the value of the hyperparameter. The small  $\alpha$ , large  $\alpha$ , and intermediate  $\alpha$  regions are colored with coral, blue, and white, respectively.

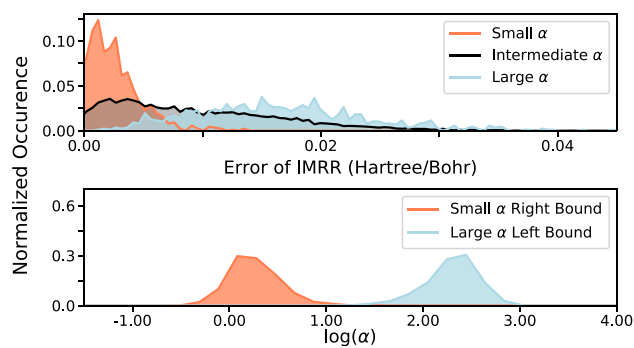


Fig. 5. Top panel: the error of IMRR when the hyperparameter  $\alpha$  is in different regions in Fig. 4. Bottom panel: the histogram of the position of the edges of the small  $\alpha$  and large  $\alpha$  regions.

large  $\alpha$  region (93%, area under the blue curve). It is true that the possible minimal error of IMRR could correspond to an  $\alpha$  value within the intermediate region, as the bottom panel of Fig. 4 shows, nevertheless, the intermediate region still statistically (61%, area under the black curve in the top panel of Fig. 5) has a larger error than those in the coral region. As a result, the optimal  $\alpha$  is empirically set to be in the coral region in Fig. 4, 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. 5 value of  $1.0 \times 10^{-1}$  is chosen as the default for the IMRR.

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 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  $\alpha$ , the error of IMRR is well-behaved and the IMRR energy gradient approaches its *ab initio* counterpart.

### 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 an ML method like

IMRR that aims to replace a majority of the *ab initio* energy gradient calculations, the overhead cost could possibly become rate-limiting 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) from an enormous number of configurations depends heavily on the data structure, 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 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 aforementioned 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 are loaded in the memory. The premise is that the  $g(q)$  stored in these files are likely to share geometrically close configurations to 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  $\text{HBr}^+ + \text{CO}_2 \rightarrow \text{HOCO}^+ + \text{Br}$  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. 6. As shown, the wall time of IMRR (0.51 s per step on average) is less than a quarter of the wall time of the AIMD (2.16 s 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 almost all of the wall time of AIMD is spent on generating the energy gradients. It is also

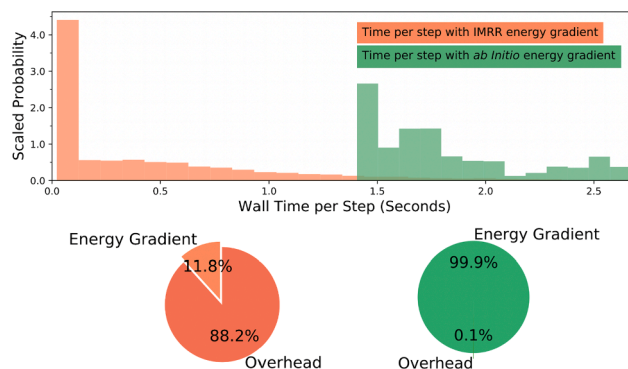


Fig. 6. A histogram of the wall time it takes for one MD step if the energy gradient is generated from IMRR (coral) and *ab initio* (green). The heights of the bars are scaled so that they integrate to 1. The average wall times are further broken down in the pie-charts to gradient generation and overhead.

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. 3), the geometrical closeness of inputs ( $t_{cut}$ , right panel of Fig. 3), and the hyperparameter ( $\alpha$ , Fig. 4). Since the difference in energy gradient between any given pair of configurations generally decreases 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 accurately representing the *ab initio* energy gradient of the target, i.e.,

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

As defined,  $r$  is non-negative and the larger its value, the more effective IMRR is in predicting the energy gradient of the target as compared to the input that is geometrically closest to the target. The same sets of targets as Fig. 3 (68 targets randomly selected from AIMD trajectories that distributed uniformly in the CV space) are employed to probe into the efficacy of IMRR with inputs of various  $t_{cut}$  values.

Fig. 7 demonstrates several histograms of  $r$  from IMRR on these 68 targets, each with 15 inputs ( $K = 15$ ). Compared to the energy gradient of the input ( $q_m$ ) that is geometrically closest to the target, the IMRR energy gradients are on average 50 times closer to the target energy gradient from the *ab initio* calculation. The inserted panels of Fig. 7 show that the efficacy of IMRR gradually increases when  $t_{cut}$  gets smaller, indicating that even when the input(s) are geometrically very close to the target, IMRR still takes advantages of these inputs and estimates a much more accurate energy gradient for the target.

The efficacy of IMRR at small  $t_{cut}$  indicates that it can be closely coupled with AIMD as a time interval ( $\delta t$ , time between updating configurations of the system) multiplier in addition to the active learning

discussed earlier. In AIMD,  $\delta t$  should be chosen as large as possible while conserving the physical properties of the system (e.g., total energy, momentum, etc.) and an *ab initio* energy gradient is calculated every  $\delta t$  to propagate the trajectory.  $\delta t$  is usually sub-femtosecond and thus the configurations of consecutive steps are geometrically very close [66]. Therefore, if IMRR could estimate an energy gradient better representing the target than the energy gradient from the previous step does, one can propose an integer multiple  $n$  (e.g.,  $n = 2, 3, 4, \dots$ ) 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 various values of  $h$  are provided in Fig. 8. The targets of these IMRR are from an AIMD trajectory of the  $\text{HBr}^+ + \text{CO}_2$  reaction ( $\sim 4000$  targets), and  $(15-h)$  inputs of  $t_{cut} = 0.15$  Å were selected from the library of 4000 trajectories.

The results show that including just one or two “history” can dramatically decrease the error of IMRR—for example, the error of IMRR with 2 “history” + 13 “library” (of a  $t_{cut}$  of 0.15 Å) is on the same level as the error of IMRR with 15 “library” that are geometrically much closer to the target (i.e.,  $t_{cut} = 0.01$  Å, see Fig. 3). It is also important to note that the error of IMRR does not further decrease monotonically with more “history” which could be allotted to several factors. First, as more “history” is included, inputs that are geometrically further and further away from the target (i.e., larger  $t_{cut}$ ) are included at the cost of excluding “library” of smaller  $t_{cut}$ . Previous results (Fig. 3) 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” aligns well with the nature of the local interpolation of IMRR.

### 4.2. The risk of IMRR

The risk of IMRR 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. 10) and the terms related to the relative position of the inputs with respect to the target ( $R_1$  and  $R_2$  terms in Eq. 10)—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. Consider two different  $q_0$ : all else (the geometrical closeness of the inputs, number of inputs, and hyperparameter) being equal, the one associated with

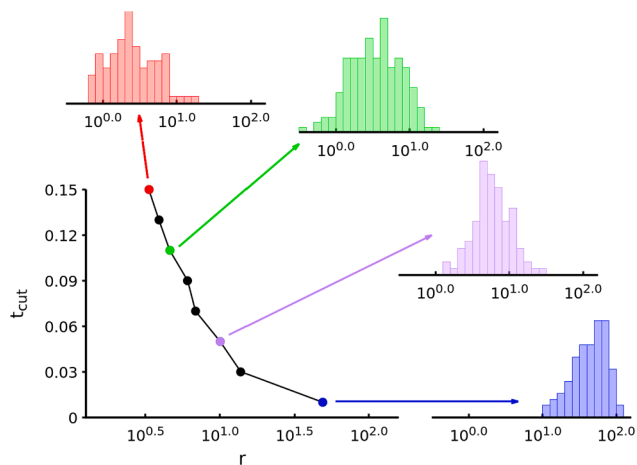


Fig. 7. The average efficacy of IMRR ( $r$ ) is plotted across the range of  $t_{cut}$  used in the dataset for Fig. 3.

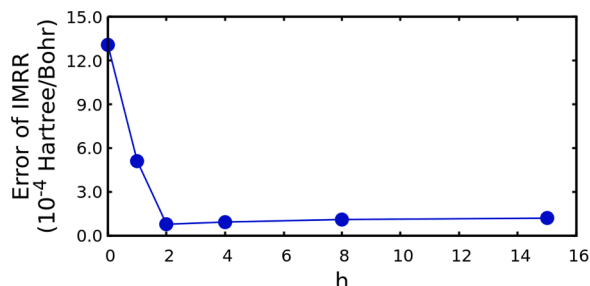


Fig. 8. The error of IMRR vs the number of “history”,  $h$ .



smaller  $R_1$  and  $R_2$  values is expected to have a smaller error of IMRR. To verify, a series of IMRR with 23,000 targets uniformly sampling the phase space and their corresponding  $R_1 + \alpha R_2$  values are summarized in Fig. 9. The inputs of the IMRR are obtained from a library containing 4,000 AIMD trajectories and no “history” is included. This figure demonstrates that  $R_1 + \alpha R_2$  is a reasonable measurement of the error of IMRR, as its average value (bold green line) decrease as  $R_1 + \alpha R_2$  gets smaller. Therefore, enforcing a small  $R_1 + \alpha R_2$  value on average decreases the risk of IMRR. It is important to note that since AIMD trajectory involves many steps of propagation of a chaotic system, it can be thrown off by any single step with large energy gradient error [67]. Practically, this nature unfortunately means that the minimal IMRR error, instead of its average, should be used to select a threshold value of  $R_1 + \alpha R_2$  for AIMD simulations. It is also of interest to note that given a finite-size library and history, not every target can be guaranteed with an IMRR that yields small enough  $R_1 + \alpha R_2$ . As a result, enforcing a small  $R_1 + \alpha R_2$  could potentially result in many IMRRs being too risky to use. Nonetheless, being able to assess the risk of the predicted energy gradient of a ML method locally is valuable in precisely controlling its usage -only low-risk results from the ML method should be adapted for the simulation. This feature is critical for the application of IMRR in AIMD: With access to the risk 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 and propagate the trajectory. 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.

#### 4.3. The versatility of IMRR

By examining its performance on the  $\text{HBr}^+ + \text{CO}_2$  system, IMRR has been established so far to be a viable option to predict energy gradients for AIMD simulations with small and tunable errors. The underlying theory of IMRR does not directly speak to how different chemical systems of different size could impact its performance, thus we apply IMRR to two different reaction systems,  $\text{H}_2\text{S} + \text{CH}$  (same size, different chemistry) and  $\text{C}_4\text{H}_2 + \text{CH}$  (different size and chemistry), to probe IMRR’s versatility. Both of the reactions have been studied by our group with AIMD and are of great importance to astrochemistry: the former reaction is believed to form thioformaldehyde ( $\text{H}_2\text{CS}$ ) and its thiohydroxycarbene isomer ( $\text{HCSH}$ ) in star-forming regions such as Sagittarius

B2 [64], while the latter forms triplet pentadiynylidene ( $\text{HCCCCCH}$ ) and singlet ethynylcyclopropenylidene ( $\text{c-C}_5\text{H}_2$ ) carbene and is a prototype reaction to study the chemistry in extreme, hydrocarbon-rich outer space [65].

To assess the performance of IMRR, the targets ( $q_0$ ) of the  $\text{H}_2\text{S} + \text{CH}$  and  $\text{C}_4\text{H}_2 + \text{CH}$  systems are selected unbiasedly in their respective phase space, which are characterized by two CVs to distinguish key configurations involved in the reaction (e.g., reactant, intermediates, transition states, products). For the  $\text{H}_2\text{S} + \text{CH}$  system, the two CVs are the S-C distance and the root mean square of the three C-H distances. The configurations of 1080 AIMD trajectories are binned into the CV space, and  $78\ 1\ \text{\AA} \times 1\ \text{\AA}$  cells are populated (Fig. S3 in the Supporting Information). These cells are determined to be relevant to the reaction and one target is picked from each cell. Similar to the  $\text{HBr}^+ + \text{CO}_2$  system, configurations are generated by randomly displacing atom(s) of  $q_0$  with respect to different  $t_{\text{cut}}$ , whose energy gradients are computed by NWChem with B3LYP/aug-cc-pvdz level of theory. The configurations and their energy gradients are the inputs of IMRR to estimate the energy gradient of the targets. The procedure of the  $\text{C}_4\text{H}_2 + \text{CH}$  systems is similar, except the two CVs are the largest C-C distance and the largest H-H distance, rendering  $73\ 1\ \text{\AA} \times 1\ \text{\AA}$  cells (Fig. S4 in the Supporting Information), and the energy gradients are computed with B3LYP/cc-pvdz level of theory.

The performance of IMRR on the  $\text{H}_2\text{S} + \text{CH}$  and  $\text{C}_4\text{H}_2 + \text{CH}$  systems is depicted in Fig. 10 and they closely resemble its performance on the  $\text{HBr}^+ + \text{CO}_2$  system: as more inputs becomes available and as those inputs becomes geometrically closer to the target, the error of IMRR decreases to a similar level. It is also of interest to note that the instability of IMRR displayed with a limited number of inputs that are geometrically far from the target (i.e., large  $t_{\text{cut}}$ ) is consistent in all three tested systems. Surprisingly, the error of IMRR for both  $\text{HBr}^+ + \text{CO}_2$  and  $\text{H}_2\text{S} + \text{CH}$  system shows a local maximum with nine inputs ( $K = 9$ ). This value is exactly the number of degrees of freedom (DOF) of the system (3 N-6 or 9). For the  $\text{C}_4\text{H}_2 + \text{CH}$  systems, this local maximum appears at 19 inputs ( $K = 19$ ), close to (or, if taking into the consideration that the system is largely linear, exactly is) the number of DOF of the system. The reason behind why the local maximum of the error of IMRR occurs at the number of DOF of the system, and why it exists at all, remains unclear. However, empirically, it provides a reliable guide for the minimal number of inputs that IMRR should have access to when applied to AIMD simulations. Overall, the consistent performance of IMRR on  $\text{HBr}^+ + \text{CO}_2$ ,  $\text{H}_2\text{S} + \text{CH}$ , and  $\text{C}_4\text{H}_2 + \text{CH}$  systems demonstrate its ability as a viable option to dramatically accelerate AIMD simulations of chemical

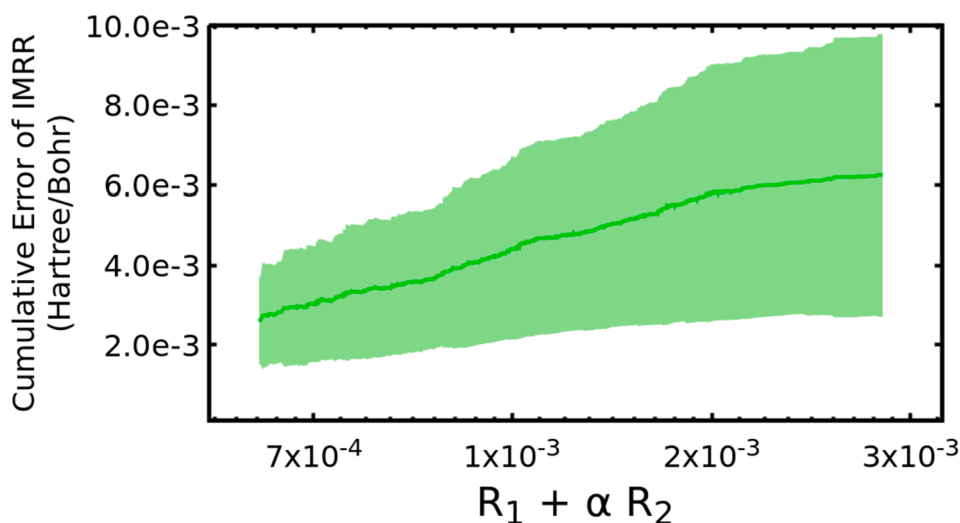
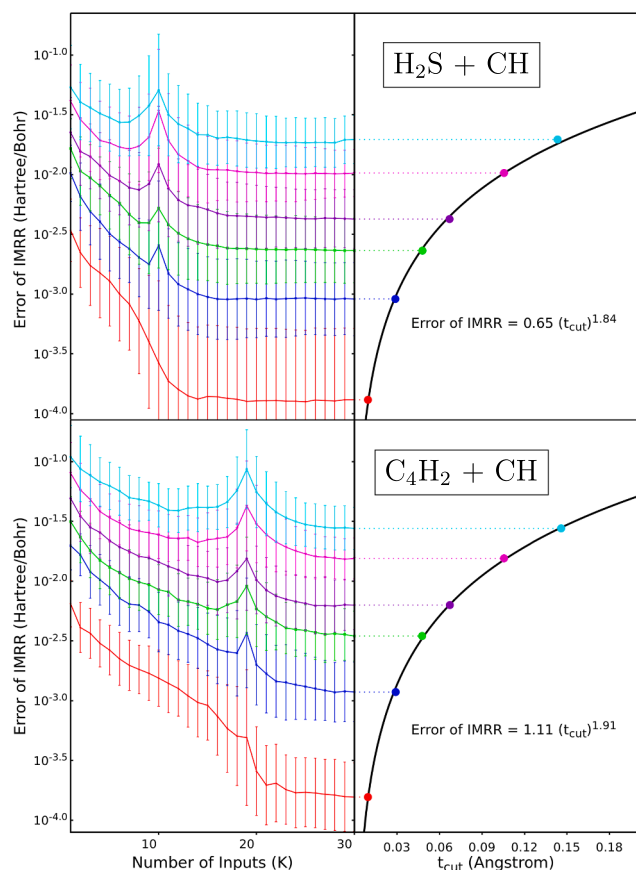


Fig. 9. The sum  $R_1 + \alpha R_2$  may be used to estimate risk; the cumulative error of all IMRR predictions with  $R_1 + \alpha R_2$  less than some threshold are reported above. Data is gathered as described in the text with  $\alpha = 10^{-1}$ . The average error is shown as a bright green line with the standard deviation shown as a light green region.



**Fig. 10.** Left panels: the error of IMRR with vs. the number of inputs at various  $t_{\text{cut}}$  values. Right panels: the error of IMRR vs.  $t_{\text{cut}}$  corresponding to  $K = 40$  inputs. The targets of IMRR uniformly sample the phase space of each system.

reaction.

### Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Rui Sun reports was provided by University of Hawai'i at Manoa. Rui Sun reports a relationship with University of Hawai'i at Manoa that includes: employment.

### Acknowledgements

This manuscript is based upon research supported by the National Science Foundation under Grant No. 2144031. The authors are in debt of insights from Dr. William L. Hase at Texas Tech University. The authors appreciate the information technology service (ITS) from the University of Hawai'i, Mānoa for the computational resources. The authors are grateful to the financial support from the University of Hawai'i, Mānoa

### Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at <https://doi.org/10.1016/j.chemphys.2022.111482>.

### References

- [1] S. Pratihar, X. Ma, Z. Homayoon, G.L. Barnes, W.L. Hase, Direct chemical dynamics simulations, *J. Am. Chem. Soc.* 139 (10) (2017) 3570–3590.
- [2] J.M. Bowman, P.L. Houston, Theories and simulations of roaming, *Chem. Soc. Rev.* 46 (24) (2017) 7615–7624.
- [3] J. Xie, W.L. Hase, Rethinking the  $\text{sn}2$  reaction, *Science* 352 (6281) (2016) 32–33.

- [4] J. Mikosch, S. Trippel, C. Eichhorn, R. Otto, U. Lourderaj, J. Zhang, W. Hase, M. Weidemüller, R. Wester, Imaging nucleophilic substitution dynamics, *Science* 319 (5860) (2008) 183–186.
- [5] J. Mikosch, J. Zhang, S. Trippel, C. Eichhorn, R. Otto, R. Sun, W.A. De Jong, M. Weidemüller, W.L. Hase, R. Wester, Indirect dynamics in a highly exoergic substitution reaction, *J. Am. Chem. Soc.* 135 (11) (2013) 4250–4259.
- [6] R. Sun, C.J. Davda, J. Zhang, W.L. Hase, Comparison of direct dynamics simulations with different electronic structure methods.  $\text{f} + \text{ch}_3\text{i}$  with mp2 and dft/b97-1, *PCCP* 17 (4) (2015) 2589–2597.
- [7] J. Zhang, U. Lourderaj, R. Sun, J. Mikosch, R. Wester, W.L. Hase, Simulation studies of the  $\text{cl}^- + \text{ch}_3\text{i}$   $\text{sn}2$  nucleophilic substitution reaction: comparison with ion imaging experiments, *J. Chem. Phys.* 138 (11) (2013) 114309.
- [8] Y.T. Lee, Y.R. Shen, Molecular beam studies of ir laser induced multiphoton dissociation and vibrational predissociation, Tech. rep., Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States), 1980.
- [9] M.E. Tuckerman, Ab initio molecular dynamics: basic concepts, current trends and novel applications, *J. Phys.: Condens. Matter* 14 (50) (2002) R1297.
- [10] R. Iftimie, P. Minari, M.E. Tuckerman, Ab initio molecular dynamics: Concepts, recent developments, and future trends, *Proc. Nat. Acad. Sci.* 102 (19) (2005) 6654–6659.
- [11] M. Paranjthy, R. Sun, Y. Zhuang, W.L. Hase, Direct chemical dynamics simulations: coupling of classical and quasiclassical trajectories with electronic structure theory, *Wiley Interdiscipl. Rev. Comput. Mol. Sci.* 3 (3) (2013) 296–316.
- [12] U. Lourderaj, K. Park, W.L. Hase, Classical trajectory simulations of post-transition state dynamics, *Int. Rev. Phys. Chem.* 27 (3) (2008) 361–403.
- [13] K. Raghavachari, G.W. Trucks, J.A. Pople, M. Head-Gordon, A fifth-order perturbation comparison of electron correlation theories, *Chem. Phys. Lett.* 157 (6) (1989) 479–483.
- [14] R.A. Kendall, T.H. Dunning Jr, R.J. Harrison, Electron affinities of the first-row atoms revisited. systematic basis sets and wave functions, *J. Chem. Phys.* 96 (9) (1992) 6796–6806.
- [15] A.D. Beck, Density-functional thermochemistry. III. The role of exact exchange, *J. Chem. Phys.* 98 (7) (1993) 5648–6.
- [16] C.M. Aikens, S.P. Webb, R.L. Bell, G.D. Fletcher, M.W. Schmidt, M.S. Gordon, A derivation of the frozen-orbital unrestricted open-shell and restricted closed-shell second-order perturbation theory analytic gradient expressions, *Theoret. Chem. Acc.* 110 (4) (2003) 233–253.
- [17] T.H. Dunning Jr, Gaussian basis sets for use in correlated molecular calculations. i. the atoms boron through neon and hydrogen, *J. Chem. Phys.* 90 (2) (1989) 1007–1023.
- [18] R. Ditchfield, W.J. Hehre, J.A. Pople, Self-consistent molecular-orbital methods. ix. an extended gaussian-type basis for molecular-orbital studies of organic molecules, *J. Chem. Phys.* 54 (2) (1971) 724–728.
- [19] W.J. Hehre, R. Ditchfield, J.A. Pople, Self-consistent molecular orbital methods. xii. further extensions of gaussian-type basis sets for use in molecular orbital studies of organic molecules, *J. Chem. Phys.* 56 (5) (1972) 2257–2261.
- [20] K.A. Peterson, D.E. Woon, T.H. Dunning Jr, Benchmark calculations with correlated molecular wave functions. iv. the classical barrier height of the  $\text{h} + \text{h}_2 \rightarrow \text{h}_2 + \text{h}$  reaction, *J. Chem. Phys.* 100 (10) (1994) 7410–7415.
- [21] K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O.A. Von Lilienfeld, A. Tkatchenko, K.-R. Müller, Assessment and validation of machine learning methods for predicting molecular atomization energies, *J. Chem. Theory Comput.* 9 (8) (2013) 3404–3419.
- [22] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O.A. Von Lilienfeld, K.-R. Müller, A. Tkatchenko, Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space, *J. Phys. Chem. Lett.* 6 (12) (2015) 2326–2331.
- [23] F.A. Faber, L. Hutchison, B. Huang, J. Gilmer, S.S. Schoenholz, G.E. Dahl, O. Vinyals, S. Kearnes, P.F. Riley, O.A. Von Lilienfeld, Prediction errors of molecular machine learning models lower than hybrid dft error, *J. Chem. Theory Comput.* 13 (11) (2017) 5255–5264.
- [24] F. Noe, A. Tkatchenko, K.-R. Müller, C. Clementi, Machine learning for molecular simulation, *Annu. Rev. Phys. Chem.* 71 (2020) 361–390. URL <https://www.annualreviews.org/doi/full/10.1146/annurev-physchem-042018-052331>.
- [25] J. Behler, First principles neural network potentials for reactive simulations of large molecular and condensed systems, *Angew. Chem. Int. Ed.* 56 (42) (2017) 12828–12840.
- [26] J. Behler, M. Parrinello, Generalized neural-network representation of high-dimensional potential-energy surfaces, *Phys. Rev. Lett.* 98 (14) (2007) 146401.
- [27] K.J. Jose, N. Artrith, J. Behler, Construction of high-dimensional neural network potentials using environment-dependent atom pairs, *J. Chem. Phys.* 136 (19) (2012) 194111.
- [28] M. Gastegger, J. Behler, P. Marquetand, Machine learning molecular dynamics for the simulation of infrared spectra, *Chem. Sci.* 8 (10) (2017) 6924–6935.
- [29] J.S. Smith, O. Isayev, A.E. Roitberg, Ani-1: an extensible neural network potential with dft accuracy at force field computational cost, *Chem. Sci.* 8 (4) (2017) 3192–3203.
- [30] V. Botu, R. Ramprasad, Learning scheme to predict atomic forces and accelerate materials simulations, *Phys. Rev. B* 92 (9) (2015) 094306.
- [31] F. Brockherde, L. Vogt, L. Li, M.E. Tuckerman, K. Burke, K.-R. Müller, Bypassing the kohn-sham equations with machine learning, *Nat. Commun.* 8 (1) (2017) 1–10. URL <https://www.nature.com/articles/s41467-017-00839-3>.
- [32] V. Botu, R. Batra, J. Chapman, R. Ramprasad, Machine learning force fields: construction, validation, and outlook, *J. Phys. Chem. C* 121 (1) (2017) 511–522.

- [33] Y. Wu, N. Prezhdo, W. Chu, Increasing efficiency of nonadiabatic molecular dynamics by hamiltonian interpolation with kernel ridge regression, *J. Phys. Chem. A* 125 (41) (2021) 9191–9200.
- [34] A.P. Bartók, M.C. Payne, R. Kondor, G. Csányi, Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons, *Phys. Rev. Lett.* 104 (13) (2010) 136403.
- [35] A.P. Bartók, R. Kondor, G. Csányi, On representing chemical environments, *Phys. Rev. B* 87 (18) (2013) 184115.
- [36] J. Ischtwan, M.A. Collins, Molecular potential energy surfaces by interpolation, *J. Chem. Phys.* 100 (11) (1994) 8080–8088.
- [37] M.A. Collins, Molecular potential energy surfaces by interpolation, in: *International Conference on Computational Science*, Springer, 2003, pp. 159–167.
- [38] G.G. Maisuradze, A. Kawano, D.L. Thompson, A.F. Wagner, M. Minkoff, Interpolating moving least-squares methods for fitting potential energy surfaces: Analysis of an application to a six-dimensional system, *J. Chem. Phys.* 121 (21) (2004) 10329–10338.
- [39] Y. Guo, A. Kawano, D.L. Thompson, A.F. Wagner, M. Minkoff, Interpolating moving least-squares methods for fitting potential energy surfaces: Applications to classical dynamics calculations, *J. Chem. Phys.* 121 (11) (2004) 5091–5097.
- [40] B.J. Braams, J.M. Bowman, Permutationally invariant potential energy surfaces in high dimensionality, *Int. Rev. Phys. Chem.* 28 (4) (2009) 577–606.
- [41] T.-S. Ho, H. Rabitz, A general method for constructing multidimensional molecular potential energy surfaces from ab initio calculations, *J. Chem. Phys.* 104 (7) (1996) 2584–2597.
- [42] T. Hollebeek, H. Rabitz, L.B. Harding, G.C. Schatz, A global h<sub>2</sub>o potential energy surface for the reaction o (1 d)+ h<sub>2</sub> → oh+ h, *J. Chem. Phys.* 105 (23) (1996) 10472–10486.
- [43] T. Hollebeek, T.-S. Ho, H. Rabitz, A fast algorithm for evaluating multidimensional potential energy surfaces, *J. Chem. Phys.* 106 (17) (1997) 7223–7227.
- [44] K.C. Thompson, M.J. Jordan, M.A. Collins, Polyatomic molecular potential energy surfaces by interpolation in local internal coordinates, *J. Chem. Phys.* 108 (20) (1998) 8302–8316.
- [45] T. Hollebeek, T.-S. Ho, H. Rabitz, Constructing multidimensional molecular potential energy surfaces from ab initio data, *Annu. Rev. Phys. Chem.* 50 (1) (1999) 537–570.
- [46] T.-S. Ho, H. Rabitz, F.J. Aoiz, L. Banares, S.A. Vázquez, L.B. Harding, Implementation of a fast analytic ground state potential energy surface for the n (2 d)+ h<sub>2</sub> reaction, *J. Chem. Phys.* 119 (6) (2003) 3063–3070.
- [47] S. Chmiela, A. Tkatchenko, H.E. Sauceda, I. Poltavsky, K.T. Schütt, K.-R. Müller, Machine learning of accurate energy-conserving molecular force fields, *Sci. Adv.* 3 (5) (2017) e1603015.
- [48] H.E. Sauceda, S. Chmiela, I. Poltavsky, K.-R. Müller, A. Tkatchenko, Molecular force fields with gradient-domain machine learning: Construction and application to dynamics of small molecules with coupled cluster forces, *J. Chem. Phys.* 150 (11) (2019) 114102.
- [49] M. Gastegger, P. Marquetand, Molecular dynamics with neural network potentials, in: *Machine Learning Meets Quantum Physics*, Springer, 2020, pp. 233–252.
- [50] J.P. Janet, S. Ramesh, C. Duan, H.J. Kulik, Accurate multiobjective design in a space of millions of transition metal complexes with neural-network-driven efficient global optimization, *ACS Central Sci.* 6 (4) (2020) 513–524.
- [51] I.S. Novikov, Y.V. Suleimanov, A.V. Shapeev, Automated calculation of thermal rate coefficients using ring polymer molecular dynamics and machine-learning interatomic potentials with active learning, *PCCP* 20 (46) (2018) 29503–29512.
- [52] J.S. Smith, B. Nebgen, N. Lubbers, O. Isayev, A.E. Roitberg, Less is more: Sampling chemical space with active learning, *J. Chem. Phys.* 148 (24) (2018) 241733.
- [53] G. Imbalzano, Y. Zhuang, V. Kapil, K. Rossi, E.A. Engel, F. Grasselli, M. Ceriotti, Uncertainty estimation by committee models for molecular dynamics and thermodynamic averages, *arXiv preprint arXiv:2011.08828* (2020) 30–35.
- [54] C. Schran, K. Brezina, O. Marsalek, Committee neural network potentials control generalization errors and enable active learning, *J. Chem. Phys.* 153 (10) (2020) 104105.
- [55] A.P. Bartók, J. Kermode, N. Bernstein, G. Csányi, Machine learning a general-purpose interatomic potential for silicon, *Phys. Rev. X* 8 (4) (2018) 041048.
- [56] E. Uteva, R.S. Graham, R.D. Wilkinson, R.J. Wheatley, Active learning in gaussian process interpolation of potential energy surfaces, *J. Chem. Phys.* 149 (17) (2018) 174114.
- [57] N. Bernstein, G. Csányi, V.L. Deringer, De novo exploration and self-guided learning of potential energy surfaces, *npj Comput. Mater.* 5 (1) (2019) 1–9.
- [58] E.V. Podryabinkin, E.V. Tikhonov, A.V. Shapeev, A.R. Oganov, Accelerating crystal structure prediction by machine-learning interatomic potentials with active learning, *Phys. Rev. B* 99 (6) (2019) 064114.
- [59] E.A. Coutsiyas, C. Seok, K.A. Dill, Using quaternions to calculate rmsd, *J. Comput. Chem.* 25 (15) (2004) 1849–1857.
- [60] L. Paetow, F. Unger, W. Beichel, G. Frenking, K.-M. Weitzel, Rotational dependence of the proton-transfer reaction hbr<sup>+</sup> + co<sub>2</sub> → hoco<sup>+</sup> + br. i. energy versus angular momentum effects, *J. Chem. Phys.* 132 (17) (2010) 174305.
- [61] L. Paetow, F. Unger, B. Beutel, K.-M. Weitzel, Rotational dependence of the proton-transfer reaction hbr<sup>+</sup> + co<sub>2</sub> → hoco<sup>+</sup> + br. ii. comparison of hbr<sup>+</sup> (2π/2) and hbr<sup>+</sup> (2π1/2), *J. Chem. Phys.* 133 (23) (2010) 234301.
- [62] A. Shoji, D. Schanzenbach, R. Merrill, J. Zhang, L. Yang, R. Sun, Theoretical study of the potential energy profile of the hbr<sup>+</sup> + co<sub>2</sub> → hoco<sup>+</sup> br. reaction, *J. Phys. Chem. A* 123 (45) (2019) 9791–9799.
- [63] Y. Luo, T. Kreuscher, C. Kang, W.L. Hase, K.-M. Weitzel, R. Sun, A chemical dynamics study of the hcl+ hcl<sup>+</sup> reaction, *Int. J. Mass Spectrom.* 462 (2021) 116515.
- [64] S. Doddipatla, C. He, R.I. Kaiser, Y. Luo, R. Sun, G.R. Galimova, A.M. Mebel, T. J. Millar, A chemical dynamics study on the gas phase formation of thioformaldehyde (h<sub>2</sub>cs) and its thiohydroxycarbene isomer (hchsh), *Proc. Nat. Acad. Sci.* 117 (37) (2020) 22712–22719.
- [65] C. He, G.R. Galimova, Y. Luo, L. Zhao, A.K. Eckhardt, R. Sun, A.M. Mebel, R. I. Kaiser, A chemical dynamics study on the gas-phase formation of triplet and singlet c<sub>5</sub>h<sub>2</sub> carbenes, *Proc. Nat. Acad. Sci.* 117 (48) (2020) 30142–30150.
- [66] M.P. Allen, D.J. Tildesley, *Computer simulation of liquids*, Oxford University Press, 2017.
- [67] T. Thornton, J.B. Marion, *Classical dynamics of particles and systems*, Brooks/Cole, New York, 2004.