



# *Ab initio* molecular dynamics benchmarking study of machine-learned potential energy surfaces for the $\text{HBr}^+ + \text{HCl}$ reaction

Kazuumi Fujioka, Eric Lam, Brandon Loi, Rui Sun\*

<sup>a</sup> University of Hawaii at Manoa, Department of Chemistry, 2545 McCarthy Mall, Honolulu, HI 96822-2275, United States



## ARTICLE INFO

### Keywords:

Molecular dynamics  
 Ab initio molecular dynamics  
 Machine learning  
 Neural network  
 Kernel regression  
 Bimolecular reaction  
 Potential energy surface  
 Cross section  
 Scattering angle  
 Collision energy  
 Rotational excitation

## ABSTRACT

Machine learning has grown in use for constructing potential energy surfaces for their ability to theoretically recreate any function given enough training as well as their fast predictive powers after being trained. When trained on *ab initio* data, this enables simulation of a large number of *ab-initio*-quality trajectories. Here, rigorous benchmarking of these machine-learned potential energy surfaces—both in terms of their static errors and dynamics errors—is carried out for the  $\text{HBr}^+ + \text{HCl}$  system. In a novel comparison, both neural networks and a kernel regression method are compared for a global potential energy surface, covering multiple dissociation channels. Further, comparison with *ab initio* molecular dynamics simulations enables one of the first direct comparisons of dynamic, ensemble-average properties of the system. Finally, comparison with experimental results reveals remarkable agreement for the sGDML method for training sets of thousands to tens of thousands of molecular configurations.

## 1. Introduction

Potential energy surfaces (PESs) can be accurately modelled with *ab initio* methods from low levels of theory (e.g., Hartree-Fock[1,2], DFT with LDA functionals[3,4]) to high levels of theory (e.g., coupled cluster[5,6], configuration interaction[7,8]). These PESs may be used to propagate molecular dynamics (MD) simulations, thereby providing unparalleled detail into the reaction dynamics of chemical systems by tracking exactly how the atoms move over time. While more accurate, the high levels of theory often are more computationally expensive and make extensive MD study of even small molecules intractable. And even if a low level of theory can be used, to make statistically meaningful results, hundreds or thousands of *ab initio* molecular dynamics (AIMD) trajectories must be gathered and averaged over for all of the different possible initial configurations of the system. In total, AIMD studies of bimolecular reactions take millions of energy gradients calculations and as a result take months to carry out for just a single set of initial conditions: that is, one particular collision energy and rotational/vibrational excitation [9–14]. Thus, often only a few initial conditions are selected and the larger cross sections and other rotational energies are unexplored.

To tackle the large computational expense, machine learning (ML) methods may be used to fit these PESs from training data: a large, diverse selection of geometries on the PES evaluated with the *ab initio* methods. These resulting ML-PESs have been shown to have little error

or deviation from the underlying *ab initio* PESs they were trained on while computing energies much faster than their *ab initio* counterparts [15–22]. Thus, reactions under study with AIMD can easily instead be studied with a ML-PES. These systems often have few atoms and thus large training sets can easily be constructed for a high-accuracy ML-PES.

Molecular dynamics studies of complex molecular systems (e.g., those involving molecular dissociation to reactants/products) have begun making more use of these ML-PESs. For example, extensive sampling of bimolecular collisions of small molecules have been carried out with permutationally-invariant polynomial neural networks (PIP-NNs)[20–22] and to a lesser extent with Behler-style atom-centered neural networks (e.g., HD-NN)[15,16,23–25]. Constructing and applying *global* PESs (global meaning that they cover at least the reactant and one product channel) for MD studies of bimolecular reactions has so far been confined to neural networks[22,26–37], as opposed to other ML methods like kernel-based regressions (e.g., gradient-domain machine learning and gaussian approximation potentials) [18,38–42]. However, a number of benchmarking studies suggest these latter methods are just as capable of modeling global PESs with the level of chemical accuracy (1 kcal/mol)[43] sufficient for MD studies [17,18,44]. After constructing the PES, the efficiency of these ML-PESs enables bimolecular reaction studies to simulate very many trajectories (millions or more) or very long trajectories (microseconds or more). Sampling many trajectories for bimolecular reactions is necessary for accurate rate constants

\* Corresponding author.

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

and cross sections. [18,38–42] Sampling long trajectories for gas mixtures is necessary for accurate time-average properties (e.g., diffusion) [45,46].

*Ab initio* molecular dynamics (AIMD) has been shown to accurately model the dynamics of the  $\text{HBr}^+ + \text{HCl}$  reaction [47] and for other similar ion-molecule reactions like  $\text{HCl}^+ + \text{HCl}$  [9,12], e.g., replicating the change in reaction cross sections vs collision and rotation excitations. Of fundamental interest is the role excitations, particularly rotational excitations, have on different types of reactions: either by mechanism (e.g., charge transfer, proton transfer) or thermochemistry (i.e., exothermic, endothermic, or thermoneutral). Following studies on a number of simpler reactions, the  $\text{HBr}^+ + \text{HCl}$  reaction, with multiple product channels of varying thermochemical properties, is a well-rounded example, exhibiting a variety of behaviors with respect to rotational excitation [47].

Here, we report (1) a benchmarking study of the accuracy of the ML-PESs of the  $\text{HBr}^+ + \text{HCl}$  system created by three ML methods: Schnet (<https://github.com/atomistic-machine-learning/schnetpack>) [17], sGDML (<https://github.com/stefanch/sGDML>) [38], and NequIP (<https://github.com/mir-group/nequip>) [48], (2) dynamics results for each method for the reactive and nonreactive channels, and (3) comparison of extended analysis with the best ML-PESs with experimental results. The first two points demonstrate that while all methods meet chemical accuracy (1 kcal/mol) on the PES, disagreement can still exist in dynamics results. Of particular note is the nearly quantitative agreement for the kernel-based ML method (sGDML) with AIMD. The last point demonstrates that the ML-PES which showed agreement with AIMD is able to extend the AIMD-quality simulations to the other unexplored collision energies and rotational excitations. A few remarks are made about the novelty of this study, in particular for comparing multiple, diverse energy-gradient-trained ML methods for a global PES, as well as in benchmarking MLMD with AIMD results.

## 2. Methods

Following the previous AIMD study of  $\text{HBr}^+ + \text{HCl}$ , the inexpensive *ab initio* method, frozen core MP2 [49] (FCmp2) with def2-SVP [50], is employed as the ML training and benchmarking reference here. This method was found to (1) accurately calculate critical points on the PES benchmarked by energies calculated with coupled cluster theory with single, double, and perturbative triple excitations [51] (CCSD(T)) extrapolated to the complete basis set limit [52] (CBS) and experimental heats of reaction, when available, and (2) capture the dynamics of the reaction to agree well by trend and fairly by absolute value of experimental cross sections [47]. In the AIMD study, the orientations of reactants were uniformly sampled at various impact parameters and bond lengths are sampled at their ground vibrational states.

Training sets were constructed based on molecular geometries by choosing only a set of molecules with unique distance matrices (DMs). Given the coordinates  $q_a$  for molecule  $a$ , the distance matrix  $\text{DM}_a$  stores exact information on the molecule's geometry by storing all inverse pairwise distances between atoms, as shown below.

$$\text{DM}_a(i, j) = |\overline{q_{a,i}} - \overline{q_{a,j}}|^{-1}$$

$$\text{DMD}_{a,b} = \min_P \left| \text{vec}(\text{DM}_a) - \text{vec}(P \text{DM}_b) \right|$$

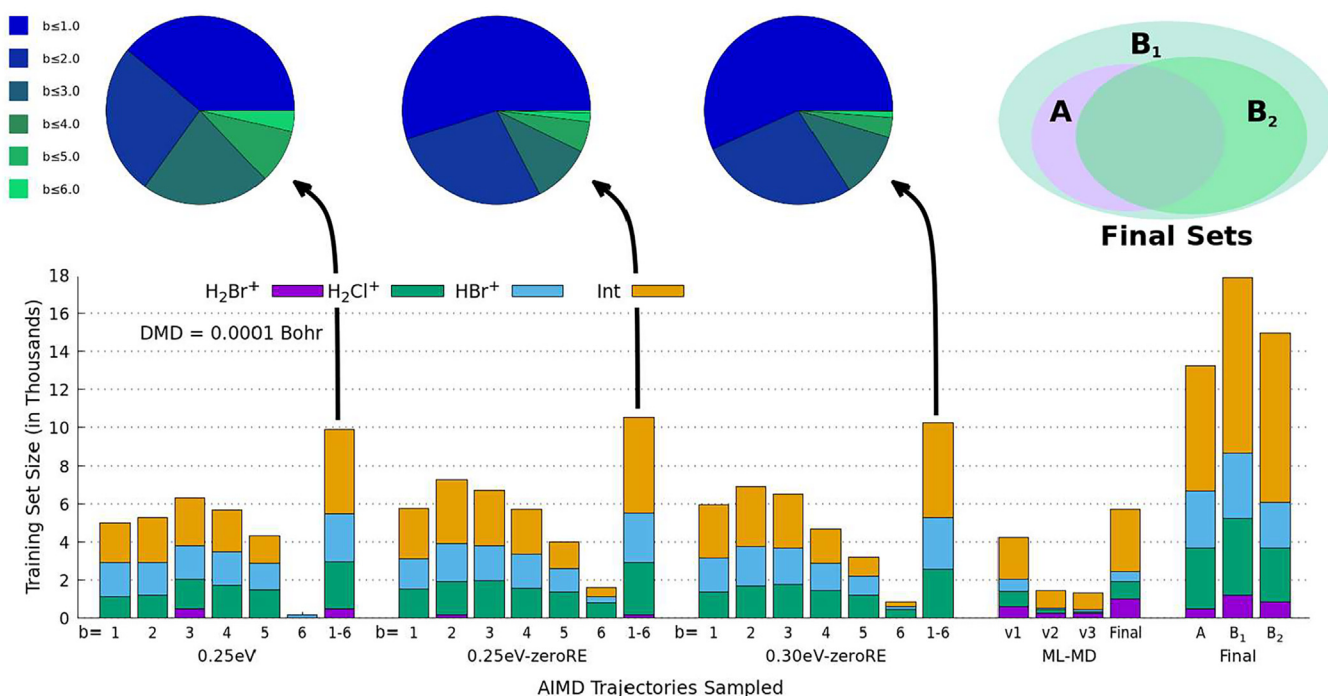
Pairs of molecular configurations have a positive, nonzero distance matrix deviation (DMD) and the uniqueness of any two DMs is determined by a DMD threshold. In computing the DMD of a pair of molecules, the deviation between DMs is minimized over all possible permutations  $P$  of indistinguishable atoms (i.e., the hydrogens), resulting in the DMD being symmetry-, rotation-, and translation-invariant. As a result, when using these training sets, ML models which are able to take advantage of these symmetries cover a larger portion of the PES without redundancy.

In addition to filtering the training set by geometry, molecular configurations corresponding to a large change (e.g., greater than 1 kcal/mol) in total energy (i.e., an “energy jump” in AIMD trajectories) are discarded. In this way, training set  $\mathbf{A}$ , selected purely from AIMD trajectories, is made. A larger training set sampling configurations not seen in AIMD is then constructed, labelled  $\mathbf{B}_1$ . These additional configurations are generated through machine-learned molecular dynamics (MLMD) trajectories, which are propagated with ML forces trained on  $\mathbf{A}$ , and added with the same DMD thresholds. Molecular configurations are also discarded if the Mulliken population calculation assigns a formal positive charge on the dissociated HCl when it is separated from HBr, as the charge transfer pathway makes up less than 0.1% of the total reaction and is not observed in the AIMD trajectories. Additionally, molecular configurations are excluded if their *ab initio* energy is too high (i.e., more than 100 kcal/mol relative to the separated reactants). While discarding configurations with a DMD threshold aims to remove *redundancy* in the training set, discarding configurations of questionable charge and energy aims to remove *discontinuities* in the training set. After three iterations of MLMD, the number of novel configurations found converges (see more in Section 3.1). Training set  $\mathbf{B}_1$  is further refined into  $\mathbf{B}_2$  with a stricter DMD threshold in selecting geometries, making it smaller.

Three machine learning methods are tested, two neural networks: Schnet [17] and NequIP [48], and one kernel regression: sGDML [38]. These three capture a good cross-section of the ML methods used for PES modeling today, with Schnet and sGDML in particular being prominently used ML methods [47]. All three are trained on both the energies and energy gradients of the exact same training sets, namely: the original only-AIMD training set  $\mathbf{A}$  and the further-sampled MLMD training set  $\mathbf{B}_s$ . All training sets can be found on Github at <https://github.com/kakazuumi/HBrHCl-ML-PES>. Of the training sets, for the neural networks, only 75% is used to train while the other 25% is used to validate. With default settings of sGDML, given 480 GB of memory, the kernel matrix construction and inverting caps out at about 8000 geometries for a system of four atoms like the current one; thus, of the training set, 8000 geometries are used to train while the remaining set are halved for validation and testing. It is important that training sets of comparable sizes are provided to the neural networks and sGDML for the interest of a fair comparison. In the training-validation splitting, subsets are chosen so their energy distributions mimic the energy distribution of the entire set.

For each ML-method/training-set combination (herein called a *ML model*), the default training parameters supplied by the program are used unless stated otherwise. For sGDML, this involves scanning over a number of values for the hyperparameter  $\sigma$  and choosing the one with the lowest error over the validation set,  $\sigma = 60$ . For the two neural networks, which depend on local atomic environments, a number of distance cutoffs are tested before settling on 6 Å and for NequIP, using  $l_{\max}$  of 2 or 3 (see Table S5 for the full hyperparameter scan). Schnet models have 3 interaction layers with 128 features each: 228,865 parameters altogether. NequIP models have a 4 interaction layers with 32 features: 392,24 ( $l_{\max}=2$ ) or 764,536 ( $l_{\max}=3$ ) parameters altogether. The two neural networks use early-stopping to decide on the best model after 50 epochs of no improvement in the loss function, and the Adam optimizer with batch sizes of 100 and 5 for Schnet and NequIP, respectively [17,48].

MLMD is done with the python Atomic Simulation Environment package (ASE) for all three ML methods [53]. Initial coordinates and momenta are used from the original AIMD trajectories generated in VENUS/NWChem [54–56], thus three sets of initial conditions are simulated: (1) 0.25 eV collision energy with the  $\text{HBr}^+$  given zero rotational energy, (2) 0.25 eV collision energy with the  $\text{HBr}^+$  given 0.05 eV rotational energy, and (3) 0.30 eV collision energy with the  $\text{HBr}^+$  given zero rotational energy. 840 trajectories are simulated for each set of initial conditions with 0.25 eV collision energy while 600 trajectories are simulated for the set with 0.30 eV collision energy. For the range of col-



**Fig. 1.** Sizes of sets of geometries are described in five clusters from left to right. The first three clusters sample AIMD trajectories at different impact parameters (i.e.  $b = 1, 2, 3, \dots$  in Angstroms) and different collision and rotational energies with the specified DMD threshold. The fourth cluster samples MLMD trajectories for three iterations (v1, v2, v3). Of the final training sets on the far right, set A comes from just the  $b = 1, 2, 3 \text{ \AA}$  AIMD trajectories while the two B sets incorporate the final set of geometries from MLMD. A breakdown of which reactant/product channel each geometry belongs to ( $H_2Br^+ + Cl$ ,  $H_2Cl^+ + Br$ , or  $HBr^+ + HCl$ ) from each set is given as the color guide; intermediate-like geometries defined as a center-of-mass distance less than 2.5 Å are marked as “Int”. For the example training sets gathered from all impact parameters 1–6 Å, the blue-green pie charts above signify how many novel configurations come from adding each successive impact parameter from only 1 Å (blue) to those from 1 to 6 Å (green). Finally, the overlap in geometries between sets A and B are displayed as a Venn diagram in the top right.

lision energies of 0.25eV–0.30 eV, a timestep of 0.2 fs is used with the Verlet algorithm. For larger collision energies (CE), timesteps ( $\Delta t$ ) are decreased inversely proportionally to the square of the collision energy, i.e.  $CE \sim \Delta t^{-2}$ . All trajectories have less than 1 kcal/mol sudden or overall changes in total energy. Trajectories are stopped when any two atoms are separated by a distance greater than 12 Å and reactant/product labels are assigned with distance-based clustering.

### 3. Results

#### 3.1. Creating training sets with molecular dynamics

The original AIMD study of the  $HBr^+ + HCl$  reaction has over 2000 trajectories across three different initial conditions, resulting in a little over ten million molecular configurations spanning energies from –24 to 20 kcal/mol, with respect to the reactants [10]. This large set of geometries, with energies and energy gradients already computed, is a natural start point for training set creation. In theory, with more geometries, the wider the selection and the more diverse the training set will be. However, in practice, there is a diminishing return in searching through more trajectories, as the likelihood of finding unique molecular configurations decreases.

AIMD trajectories are sampled by impact parameter,  $b$ : trajectories may collide head-on ( $b = 0$ ) or more indirectly ( $b > 0$ ). While often trajectories “colliding” very indirectly ( $b > 5 \text{ \AA}$ ) react with a low probability, these glancing collisions are more frequent in nature and are thus sampled proportionally more heavily in the AIMD study ( $N(b) \sim b$ ) [9,12,14]. As seen in Fig. 1, for the six impact parameters studied, although many more trajectories are simulated at  $b = 6 \text{ \AA}$ , very few unique configurations are seen: with a DMD threshold of 0.0001 Bohr, less than 2000 unique configurations are detected. This is com-

pared to the set of trajectories at  $b = 1 \text{ \AA}$  (sampled six times less than  $b = 6 \text{ \AA}$ ), which sees many more unique configurations: with a DMD threshold of 0.0001 Bohr<sup>-1</sup>, more than 5000 unique configurations are detected. The large impact parameter restricts the types of approaches leading to complex-formation: consequently, with a less diverse set of approaches, fewer unique configurations are visited. Overall, while a large number of trajectories are reactive, all AIMD trajectories start as reactants and a majority are non-reactive (i.e. forming reactants at the end). This results in a large number of redundant reactant-like geometries visited over the course of the study: with a DMD threshold of 0.0001 Bohr<sup>-1</sup>, each set of trajectories at one set of initial conditions (e.g. 0.25 eV with zero rotational energy) has only about 10,000 unique geometries, as seen in Fig. 1. Meanwhile, the reactant-like configurations (light-blue) make up less than a third of the unique geometries.

The training set A, constructed only from AIMD trajectories, needs to visit only a fraction of the entire set to get most of the unique molecular configurations. As seen in Fig. 1, for each set of initial conditions, the first three impact parameters (1–3 Å) comprise more than 75% of the unique geometries in each set. The trajectories at these three low impact parameters were ultimately used to construct training set A. The set, comprised of 13,225 unique configurations, has geometries from the two lowest energy product channels ( $H_2Cl^+ + Br$ ,  $H_2Br^+ + Cl$ ), from the reactant channel ( $HBr^+ + HCl$ ), and a majority from the intermediate region. While sampling from a single set of initial conditions may produce about 10,000 unique geometries, as seen in Fig. 1, extending this sampling to all three sets of initial conditions (as in training set A) only increases the set by about 30%. This suggests that sampling more trajectories at different initial conditions is not necessarily more efficient than sampling more trajectories at low impact parameters, at least in visiting unique configurations.

**Table 1**

The total number of geometries (Ntotal) and the minimum and maximum energies of the training sets are shown for the three sets described in the main text.

Training	A	B <sub>1</sub>	B <sub>2</sub>
Ntotal	13,225	17,909	14,982
Min Energy (kcal/mol)	-24	-24	-24
Max Energy (kcal/mol)	20	100	100

Although training set **A** was generated from a large number of AIMD trajectories in an unbiased manner, MLMD (with training set **A**) is employed to explore parts of the PES that the AIMD trajectories failed to explore. This can be seen in Fig. 1, where the first set of new configurations not seen in **A**, called **v**<sub>1</sub>, has about 4000 unique configurations. Another iteration of MLMD (with the updated training set, **A** + **v**<sub>1</sub>) was carried out to explore more novel molecular configurations (**v**<sub>2</sub>). Training set **B**<sub>1</sub> was generated after three such iterations (**B**<sub>1</sub> = **A** + **v**<sub>1</sub> + **v**<sub>2</sub> + **v**<sub>3</sub>), having 17,909 molecular configurations. As shown in Fig. 1, training set **B**<sub>1</sub> is ~35% larger than training set **A**, and most of these novel molecular configurations correspond to intermediates and product channels. The energies (with respect to reactants) of some of these molecular configurations far exceed the maximum energy found in AIMD trajectories, highlighting the risk embedded in insufficiently-trained MLMD trajectories. Finally, a slightly smaller training set **B**<sub>2</sub> is generated from **B**<sub>1</sub> using a tighter DMD threshold. This creates a training set comparable in size to the original set **A** for a fair comparison, while covering a space of geometries of roughly the same size as **B**<sub>2</sub>. The three training sets are summarized in Table 1.

### 3.2. Reproduce the potential energy surface

The performance of nine ML models (three ML methods: Schnet, NequIP, and sGDML; three training sets: **A**, **B**<sub>1</sub>, and **B**<sub>2</sub>) in reproducing the potential energy surface are compared in Table 2. As the models are all trained on both energies and forces, both energy and force errors are reported. The training error analysis provides a general idea of how close the ML-PES resembles the true PES, and is commonly used to measure the accuracy of a ML model. [17,21,22,26-37,44,48,57-59] As the table shows, the energy and force errors are on par with errors found on systems of similar size with similar ML models. All ML models have errors of chemical accuracy (i.e., less than 1 kcal/mol). Among the three ML methods, sGDML and Schnet have comparable energy errors, while the energy errors of NequIP are roughly twice as large. Interestingly, expanding the AIMD-only training set (**A**) to include molecular configurations from MLMD (**B**<sub>1</sub> and **B**<sub>2</sub>) does not necessarily improve the performance of any ML methods. However, it is important to note that the validation sets for **B**<sub>1</sub> and **B**<sub>2</sub> include high energy configurations found in MLMD. The energy errors associated with these high energy molecular configurations are expected to be large due to a lack of sampling, thus skewing the average.

All ML models are tested on their ability to replicate critical points on the *ab initio* PES: in particular, three intermediates (two hydrogen bond complexes and one van der Waals complex), separated reactants (HBr<sup>+</sup> + HCl), and separated products (H<sub>2</sub>Br<sup>+</sup> + Cl and H<sub>2</sub>Cl<sup>+</sup> + Br) [10]. Geometry optimizations are performed on the ML-learned PESs from hundreds of initial geometries spanning the PES. Similar to PES benchmarking done for AIMD [9,11-14] and other ML-PESs [22,26-37] in the literature, the overall error of each model is summarized in the RMSE between the true *ab initio* and ML-predicted *ab initio* energies of the critical points and shown in Table 2 under "Benchmarking". As a general observation, all ML models are able to locate all known critical points found by *ab initio* method (i.e. Nmissing=0 in Table 2) with energies close to their *ab initio* counterparts (RMSE < 1 kcal/mol for most). Further, none of the initial geometries are optimized to structures that are not critical points on the *ab initio* PES. Overall, the NequIP models

fare poorly for this benchmarking while most of the sGDML and Schnet models fare acceptably (< 1 kcal/mol energy error, bolded in the table). The poorer results of the NequIP models prompted an extended hyperparameter scan over two important variables: *l*<sub>max</sub>, the maximum order of the equivariant tensor and *r*<sub>cut</sub>, the local atomic environment cutoff distance. This scan, summarized in Table S5, reveals errors do not significantly vary across hyperparameters.

### 3.3. Reproduce the reaction dynamics

If the ML-PES fits the original *ab initio* PES well, the dynamics of MLMD simulation should ideally reproduce the dynamics of AIMD simulation within statistical errors. Therefore, a more rigorous performance assessment of the quality of ML models is to analyze how closely the MLMD dynamics resemble the AIMD dynamics. Due to the chaotic nature of a multi-body system, any single MLMD trajectory should not be expected to reproduce an AIMD trajectory of the same initial conditions. Therefore, ensemble-averaged properties of MLMD trajectories are compared to their AIMD counterparts. The same number of MLMD trajectories and sampling of initial conditions are employed as in the original AIMD study for the three sets of initial conditions [47].

Both the AIMD and MLMD simulations found H<sub>2</sub>Cl<sup>+</sup> as the major product of the bimolecular collision (i.e., not H<sub>2</sub>Br<sup>+</sup>). The cross section ( $\sigma_r$ ) of the reaction is computed as:

$$\sigma_r = 2\pi \sum_b b \frac{N_r(b)}{N(b)}$$

where *N*<sub>r</sub>(*b*) is the number of reactive trajectories at impact parameter *b*, while *N*(*b*) is the total number of trajectories simulated at impact parameter *b*. The cross sections of the three collision/rotational energy combinations are summarized in Table 3. AIMD simulations demonstrate that  $\sigma_r$  decreases as either the collision or rotational energy increases, but is more sensitive to the rotational energy. This trend is accurately represented by all three sGDML models, while for the other two ML methods only with the Schnet/**B**<sub>1</sub> and NequIP/**B**<sub>1</sub> models. The absolute magnitudes of the cross sections are also comparable: Table 3 bolds ML model cross sections which agree with the AIMD cross sections within a 95% confidence interval (*p*>0.05) [60]. sGDML performs well across the board in this regard, although the Schnet and NequIP also both have a few models that agree. In particular, for these neural networks, the AIMD-only training set **A** performs worse than the iteratively-constructed training sets **B**<sub>1</sub> and **B**<sub>2</sub>. Of these, the sGDML/**B**<sub>2</sub> model performs the best, being the only model with *t*-values in [-1,1] for all three initial conditions. Similar reactive cross sections suggest similar reactivities (e.g., with respect to impact parameters), and this is confirmed for the sGDML/**B**<sub>2</sub> model in Figure S1.

Reactive cross sections for the trace product channels may also be compared, although with less statistical certainty. The original AIMD results noted no formation of H<sub>2</sub>Br<sup>+</sup> + Cl and cross sections of 2-3 Å<sup>2</sup> for hydrogen exchange (reforming reactants). Similarly small cross sections of these two product channels are noted for the ML models, summarized in tables S2 and S3. In addition, zero cross section (i.e., no reactive trajectories) are reported by AIMD and MLMD for all other possible products (i.e., HBrCl<sup>+</sup> + H or BrCl<sup>+</sup> + H<sub>2</sub>).

AIMD simulations, with a large enough ensemble of trajectories, also provide scattering angle and time of flight (i.e., relative translational energy) distributions for select products. These two properties, measuring the direction and speed at which the product scatters, can be measured in experiments and thus are often used to validate AIMD simulations [11,13,14]. Given the chaotic nature of molecular dynamics, these distributions are nearly impossible to predict *a priori*, thus are employed as another metric of the performance of MLMD in reproducing the AIMD results. For the title reaction, only the major product channel (i.e., H<sub>2</sub>Cl<sup>+</sup> + Br) has enough reactive trajectories to generate scattering angle and relative translational energy distributions (e.g., as seen in

**Table 2**

At top, training for each model is summarized as: the number of geometries in the training set,  $N_{\text{train}}$ , the number of geometries in the validation and/or test set,  $N_{\text{validate}}$ , and the range of energies ( $\Delta$  Energy) spanned by those geometries. The mean absolute errors (MAE) and root-mean-squared errors (RMSE) for the energy and force are given of various ML-method/training-set combinations. At bottom, the critical point benchmarking errors are given as well as the number of critical points unsuccessfully optimized,  $N_{\text{missing}}$ .

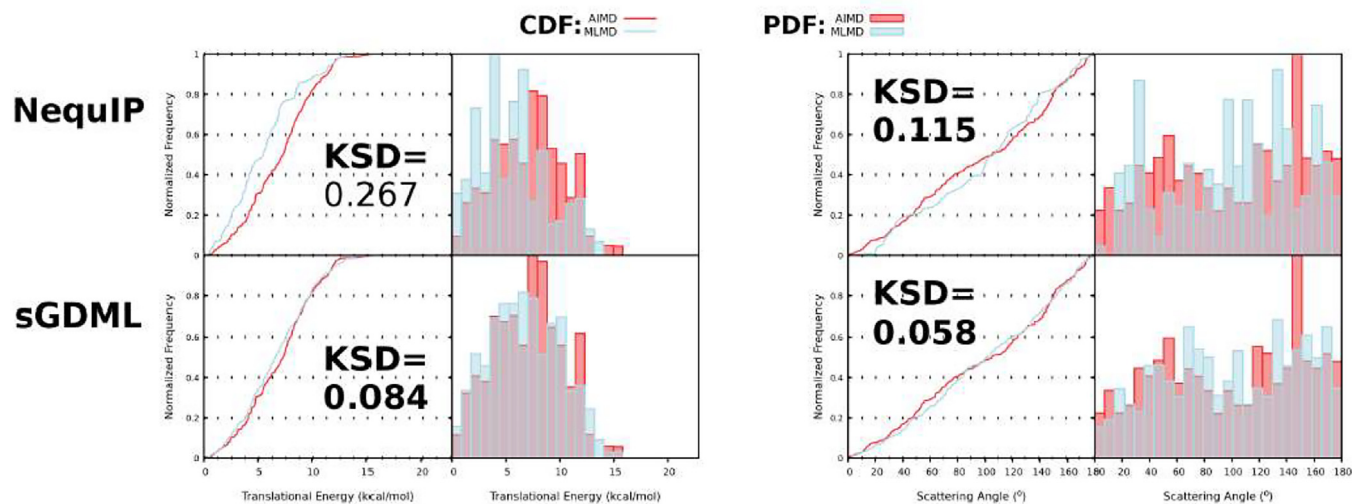
Training	sGDML			Schnet			NequIP		
	A	B <sub>1</sub>	B <sub>2</sub>	A	B <sub>1</sub>	B <sub>2</sub>	A	B <sub>1</sub>	B <sub>2</sub>
$N_{\text{train}}$	8000	8000	8000	9919	13,432	11,237	9919	13,432	11,237
$N_{\text{validate}}$	5225*	9909*	6982*	3306	4477	3745	3306	4477	3745
$\Delta$ Energy (kcal/mol)	44	124	124	44	124	124	44	124	124
Energy Error (kcal/mol)									
MAE	<b>0.15</b>	<b>0.27</b>	<b>0.28</b>	<b>0.19</b>	<b>0.24</b>	<b>0.28</b>	<b>0.47</b>	<b>0.64</b>	<b>0.62</b>
Force Error (kcal/mol/Å)									
MAE	0.46	1.23	1.27	0.47	0.63	0.74	0.38	0.72	0.81
Benchmarking	0	0	0	0	0	0	0	0	0
$N_{\text{missing}}$									
RMSE (kcal/mol)	<b>0.33</b>	<b>0.12</b>	<b>0.92</b>	<b>0.25</b>	<b>0.51</b>	1.83	1.41	1.53	1.64

\* For sGDML, the validation set is further split into halves: one tested per candidate  $\sigma$  to choose the final  $\sigma$  and one held out to test the final  $\sigma$ .

**Table 3**

Reactive cross sections in  $\text{Å}^2$  for the  $\text{H}_2\text{Cl}^+$  product channel from AIMD and MLMD simulations for each model for each set of initial conditions: some collision energy (CE) and rotational energy (RE) of the  $\text{HBr}^+$ . AIMD cross sections are listed with their  $\pm$  standard error. MLMD cross sections which are not rejected as being different from the AIMD cross sections within  $p < 0.05$  are bolded. MLMD standard errors are reported in Table S1.

Initial Conditions	AIMD	sGDML			Schnet			NequIP		
	FCmp2/def2-SVP	A	B <sub>1</sub>	B <sub>2</sub>	A	B <sub>1</sub>	B <sub>2</sub>	A	B <sub>1</sub>	B <sub>2</sub>
0.25 eV CE + 0.05 eV RE	20.26 $\pm$ 1.53	<b>23.59</b>	<b>23.93</b>	<b>21.71</b>	27.22	<b>23.78</b>	30.56	<b>19.52</b>	<b>22.76</b>	33.40
0.25 eV CE + 0.00 eV RE	43.98 $\pm$ 2.13	<b>47.28</b>	<b>43.07</b>	<b>41.68</b>	21.98	35.00	<b>43.93</b>	16.85	<b>36.32</b>	22.87
0.30 eV CE + 0.00 eV RE	38.33 $\pm$ 1.96	<b>36.33</b>	<b>35.04</b>	<b>36.71</b>	21.53	23.93	27.08	14.92	31.07	18.25



**Fig. 2.** The  $\text{H}_2\text{Cl}^+$  product translational energy and scattering angle distributions are compared for AIMD (red) and two ML models (blue) at the 0.25eV-zeroRE set of initial conditions. Both models use training set B<sub>1</sub> as described in the main text. KSDs are inset on graphs of the cumulative distribution function (drawn as lines) next to the original probability density function (drawn as a histogram).

**Fig. 2).** These are compared to their counterparts generated from the nine ML models.

To assess how closely the MLMD distribution resembles the AIMD distribution, the Kolmogorov-Smirnov deviation (KSD)[61–63] between them is measured. We note that KSD does not assume any specific parametric form of the underlying distribution (e.g., Gaussian, Poisson, etc.), but instead, simply calculates the maximum difference between the cumulative distribution functions (CDFs) of the two, as shown below:

$$\text{CDF}_A(\theta_0) = \sum_{\theta \in A, \theta < \theta_0} \frac{1}{N_A}$$

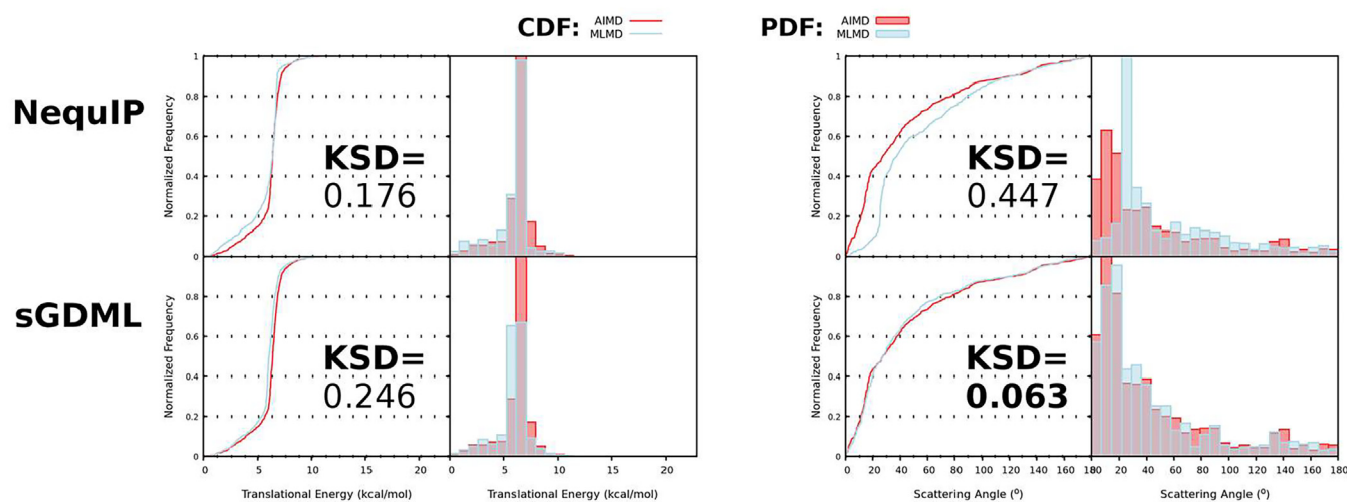
$$\text{KSD}(A, B) = \max_{\theta} |\text{CDF}_A(\theta) - \text{CDF}_B(\theta)|$$

where A and B are two discrete distributions of  $\theta$  (i.e., the scattering angle distribution in this example). Here, A and B would be the AIMD and MLMD product distributions and so  $N_A$  and  $N_B$  are the total numbers of reactive trajectories in the AIMD and MLMD simulations, respectively.  $N_{\text{AIMD}}$  and  $N_{\text{MLMD}}$  are approximately 250 in this case. Two identical distributions will have a KSD of zero, and the larger the KSD, the more different the two distributions are (the maximal KSD value is 1). For example, for the distributions in Fig. 2, the difference in  $\theta$ -value between the AIMD and MLMD CDFs, shown on left of each pair

**Table 4**

Kolmogorov-Smirnov deviations between discrete samples of the  $\text{H}_2\text{Cl}^+$  scattering angle(top) and relative translational energy (bottom) between the AIMD distributions and the MLMD distributions. Each set of initial conditions and ML model correspond to a row and column, respectively, above. Those below the critical value 0.1216 defined in the main text are bolded.

Initial Conditions	sGDML			Schnet			NequIP		
	A	B <sub>1</sub>	B <sub>2</sub>	A	B <sub>1</sub>	B <sub>2</sub>	A	B <sub>1</sub>	B <sub>2</sub>
<b>Scattering Angle</b>									
0.25 eV CE + 0.05 eV RE	<b>0.098</b>	<b>0.067</b>	<b>0.089</b>	<b>0.082</b>	<b>0.111</b>	<b>0.093</b>	0.164	0.142	0.173
0.25 eV CE + 0.00 eV RE	<b>0.085</b>	<b>0.058</b>	<b>0.069</b>	0.144	<b>0.077</b>	0.143	<b>0.070</b>	<b>0.115</b>	0.121
0.30 eV CE + 0.00 eV RE	0.120	<b>0.110</b>	<b>0.077</b>	<b>0.110</b>	<b>0.098</b>	<b>0.091</b>	0.127	<b>0.108</b>	0.130
<b>Translational Energy</b>									
0.25 eV CE + 0.05 eV RE	<b>0.091</b>	<b>0.106</b>	<b>0.090</b>	<b>0.089</b>	<b>0.089</b>	0.129	<b>0.092</b>	<b>0.076</b>	0.225
0.25 eV CE + 0.00 eV RE	<b>0.065</b>	<b>0.084</b>	<b>0.087</b>	0.215	0.207	0.275	0.280	0.267	0.155
0.30 eV CE + 0.00 eV RE	<b>0.115</b>	<b>0.069</b>	<b>0.075</b>	0.179	<b>0.097</b>	0.166	0.212	<b>0.086</b>	0.145



**Fig. 3.** The non-reactive translational energy and scattering angle distributions are compared for AIMD (red) and two ML models (blue) at the 0.30eV-zeroRE set of initial conditions. Both models use training set B<sub>2</sub> as described in the main text. KSDs are inset on graphs of the cumulative distribution function (drawn as lines) next to the original probability density function (drawn as a histogram).

of graphs, has a larger maximum for the NequIP/B<sub>1</sub> model compared to the sGDML/B<sub>1</sub> model, for both the scattering angle and relative translational energy, leading to larger KSDs. Within a 95% confidence interval, the MLMD and AIMD samples cannot be considered as indistinguishable if their KSD is larger than 0.1216 ( $\alpha = 0.05$ ;  $N_{\text{AIMD}}, N_{\text{MLMD}} \sim 250$ ) [63]. The KSDs of these distributions are summarized in Table 4, comparing between AIMD and each of the MLMD. Those below the critical value (KSD = 0.1216) defined above are bolded. Each distribution is also visualized in Figs. S2 and S3. Again, sGDML performs the best among three ML methods – compared to AIMD, it is able to provide statistically indistinguishable scattering angle and translational energy distributions for the  $\text{H}_2\text{Cl}^+$  product in all but one case. Both Schnet and NequIP fare reasonably well, particularly with the largest, MLMD-enhanced training set B<sub>1</sub>.

A similar sort of analysis can be done for the non-reactive trajectories (i.e., those reforming the reactants). A large majority of the trajectories are non-reactive (e.g., ~80% for the AIMD) and so their scattering angle and relative translational energy distributions are well-defined, with results summarized in Table S4, and Figs. S4 and S5 in the Supporting Information. As the figures show, the KSDs associated with the non-reactive trajectories are overall larger than the reactive trajectories. This is due to the KSD being sensitive to the overall distribution changing; with non-reactive, large-impact-parameter trajectories forming sharp, low-variance, well-defined peaks (behaving similarly due to the lack of collision), small changes in mean manifest themselves in small x-shifts in the CDF but large y-changes in the CDF. For example, the relative translational energy distributions for AIMD and sGDML/B<sub>2</sub>

(Fig. 3, bottom left) have a small shift in their CDF but the large y-value change in the CDF at the peak results in a large KSD. Similar to the reactive trajectories, sGDML performs the best among three ML methods, matching the AIMD scattering angle distribution almost exactly while having some of the smallest errors in the relative translational energy distribution. Schnet and NequIP have comparable performance and both show strong training set dependence.

## 4. Discussion

### 4.1. Extending MLMD comparison with experiments

MLMD has the unique opportunity to extend analysis with experiments much further than AIMD. For the title reaction, while the AIMD results cover three sets of initial conditions, experimental results cover a much larger range of collision and rotational energies: 0–6 eV and 0–0.05 eV, respectively. Given more energy in a system, higher-energy molecular configurations are available in a trajectory and thus other mechanisms and products may be visited. Generally, trends are known about the effects of increasing collision or rotational energy on various reactions (e.g. proton transfer (PT), charge transfer (CT), hydrogen abstraction (HA), etc.), but even for simple systems, nonintuitive reactions may take place. These experimental results provide an excellent point of comparison for the MLMD, and in particular the fidelity of the best ML PES.

Additional MLMD trajectories are carried out with sGDML/B<sub>2</sub> at higher collision energies and sampled in a similar way as the original

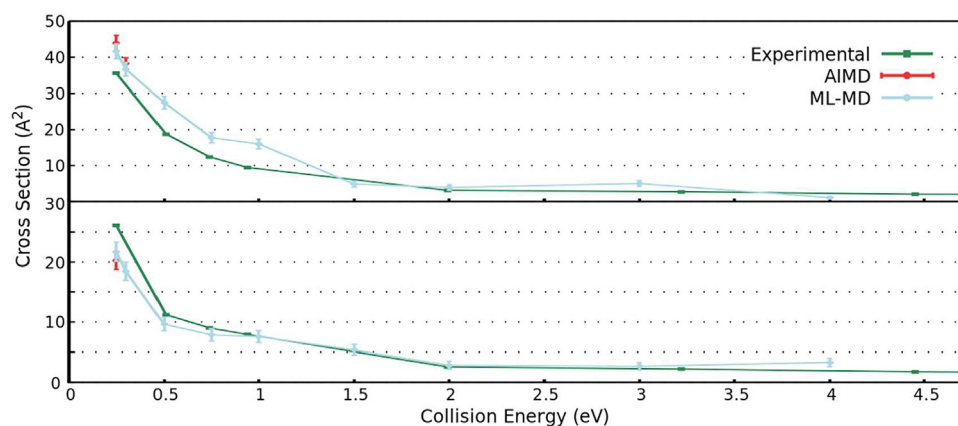


Fig. 4. The reactive cross section of the  $\text{H}_2\text{Cl}^+ + \text{Br}$  product channel is compared across collision energies at the lowest rotational excitation (at top, 0 meV for MD and 3.4 meV for experiments) and highest rotational excitation (at bottom, 50 meV for MD and 46.8 meV for experiments) between the experiment (green), AIMD (light-red), and MLMD (light-blue).

set. No new products are observed. Reactive cross sections are again calculated for the major product,  $\text{H}_2\text{Cl}^+$ . For both rotational energies simulated, the cross sections are compared as a function of collision energy in Fig. 4. Just like for the AIMD results, the cross sections are within an order of magnitude of experiments and follow a very similar trend. The only significant difference between the MLMD and experiments occurs for the zero rotational energy set: there is a deviation of about  $7 \text{ \AA}^2$  in the 1.00 eV region before agreeing again for the larger collision energies.

Overall, the agreement of the  $\text{H}_2\text{Cl}^+ + \text{Br}$  product channel with experiments is remarkable, and the other product channels are reasonable. The charge transfer product, which is not in the training set, is never observed and similarly so in experiments, as the CT cross sections make up only 0.008% (lowest rotational energy) to 0.054% (highest rotational energy) of the total observed reactive cross section. Even in AIMD, this would only translate to at most 0.135 trajectories for one set of initial conditions, and none were observed. For the only other product channel,  $\text{H}_2\text{Br}^+ + \text{Cl}$ , a trace number of trajectories are observed for each ML model. This product could not be measured directly by the experiment, however, deuteration experiments ( $\text{HBr}^+ + \text{DCl}$ ) detected a small amount of  $\text{HDBr}^+$  (between 0.2 to 1.0  $\text{ \AA}^2$ ). If we expect similar relative amounts between the isotopic analogs, then the  $\text{H}_2\text{Br}^+ + \text{Cl}$  product possesses a cross section of 1.02  $\text{ \AA}^2$  (lowest rotational energy) to 0.92  $\text{ \AA}^2$  (highest rotational energy) compared to the primary product at the same CE, and for the same reason, this would predict about 7–9 trajectories for one set of initial conditions. Therefore, the MLMD results is consistent with the experiment that a small, if not non-zero, reactive cross section is associated with the  $\text{H}_2\text{Br}^+ + \text{Cl}$ .

#### 4.2. Accuracy of machine-learned PESs

Previous studies using ML-PESs, specifically for bimolecular reaction dynamics, go ahead with MLMD after analyzing training errors, skipping AIMD altogether. [22,26–37] This is efficient, as *ab initio* calculations are often expensive, and the ML-PES does not necessitate finding an affordable *ab initio* method for MD. The current study, by having an established AIMD and experimental benchmark, provides deep comparative value and is the only study to our knowledge to: (1) compare multiple energy-gradient-trained ML methods, including a kernel-based regression method, for a global PES in use for a bimolecular reaction (2) compare MLMD to AIMD results for a variety of observations, including non-reactive trajectories. Here, we highlight two points from the comparison.

First, the kernel-based regression, sGDML, performs better given the same training set compared to its neural-network peers, SchNet and NequIP. Here, a training set like  $\mathbf{B}_2$  with 14,982 geometries constructed ML-PESs with MAEs of 0.28 kcal/mol and 1.27 kcal/mol/ $\text{ \AA}$  with sGDML (60% training, 40% validation) and MAEs of 0.34 kcal/mol

and 0.72 kcal/mol/ $\text{ \AA}$  with NequIP (75% training, 25% validation). As seen in the Results, Table 2, while the training errors make them appear comparable, the critical point analysis and MLMD show otherwise. In general, neural networks require more data to reach the same level of accuracy as kernel-based regression methods [58], so their underperformance may indicate that the  $\sim 10,000$  geometries used to train the neural networks is not sufficient for this four-atom system. This is the only study to our knowledge to use sGDML for a global PES like this, where previous studies focus on MLMD reproducing either short isomerizations (e.g., hydrogen migration)[39] or vibrations around a local minima [18,38,39]. Other kernel-regression methods have similarly not focused on global PESs [57,64,65], with the reason most likely being the difficulty in scaling to large amounts of training: memory costs involved in solving the linear equation  $Ax = b$  scales quadratically with the training set size. For this four-atoms system, given 480 GB with default settings, sGDML runs into memory issues when asked to train for 9000 geometries (thus, 8000 geometries were used).

The neural networks have the ability to scale up in training-set-size much better than sGDML. In the literature, a comparable study constructed a global multi-channel ML-PES of  $\text{H}_2 + \text{CO}^+$  with a local-gradient-based neural network and found energy and force RMSEs of 0.48 kcal/mol and 0.28 kcal/mol/ $\text{ \AA}$ , respectively, for training on 77,755 geometries (95% training, 5% validation) [35]. Another study of a global ML-PES for  $\text{CO} + \text{CH}_4$  with PhysNet, found energy RMSEs of 1.2 kcal/mol for training on 432,399 geometries ( $\sim 95\%$  training,  $\sim 5\%$  validation) [28]. Various factors like the different training set sizes, the number of energy gradients compared, the range of energies studied, and the percentage of validation make direct comparison difficult between the current study and those in the literature. In general, many studies report energy RMSEs less than 0.5 kcal/mol and some less than 0.1 kcal/mol [22,26–37], with MAEs likely to be at least a factor of two lower—errors lower than seen here. But of key insight, all studies with these levels of training errors report smooth dynamics and some agreement with experiment. Being able to reach a level of accuracy to recreate dynamics with as few expensive *ab initio* energy (and energy gradient) calculations is of large interest.

Second, for the training sets and PESs of interest here, training errors may not suffice to predict accurate dynamics, in which case some AIMD comparison is necessary. For the dynamics, the ML methods may be tested qualitatively as in Fig. S7, where the ML predicted energies along a trajectory are shown to vary smoothly and the dynamics of an *individual* trajectory seems acceptable. As seen in the Results, Table 2, the training errors suggest that all models would be acceptable for MLMD, save for perhaps sGDML/ $\mathbf{B}_1$  or sGDML/ $\mathbf{B}_2$ , as they have the largest force errors ( $> 1$  kcal/mol/ $\text{ \AA}$ ). However, as Tables 3 and 4 show, not all ML-PESs generate comparable MLMD dynamics, and sGDML/ $\mathbf{B}_2$  in particular outperforms. In other words, the training errors, which are supposed to capture the overall agreement between the ML-PES and the

*ab initio* PES, do not differentiate the performance of the ML models well. Therefore, the most rigorous benchmarking would be to compare MLMD dynamics with AIMD dynamics (e.g., scattering angle distribution of the products). This simple, diatom-diatom bimolecular reaction may be completely characterized and benchmarked across several ML method, but in general this may not be possible for large systems with more product channels and reaction mechanism.

The need for AIMD benchmarking complicates the actual cost for future MLMD studies. AIMD studies are very expensive as they require thousands of trajectories which entails millions of *ab initio* energy gradient calculations. For the AIMD study of the title reaction with FCmp2/def2-SVP, each step costs on average 6.615 s when distributed across 9 CPUs. All of the MLMD calculations are at least an order of magnitude faster, with sGDML/B<sub>2</sub> in particular costing only 0.0074 s per step on 1 CPU, costing only 0.1% that of the AIMD. Over 10 CPUs, the fastest simulation (a non-reactive trajectory of about 5000 steps) costs only 0.5 min for sGDML/B<sub>1</sub>, while costing 1.5 min for Schnet/B<sub>1</sub> and 5.5 min for NequIP/B<sub>1</sub>. Although training is also required for MLMD, the sGDML method requires only hours while the neural networks, for training sets of this size, require less than a week (~200 s/epoch for Schnet/B<sub>1</sub> (228,865 wt) and ~1500 s/epoch for NequIP/B<sub>1</sub> (764,536 wt)). Now, on one extreme, there is no need for MLMD for anyone who can afford AIMD. But usually, some compromise is necessary, reducing the number of trajectories to make the study either faster or more cost-efficient. Accurate MLMD is invaluable to chemical dynamics simulations for its sampling ability—with essentially the same cost as AIMD (which is necessary for benchmarking), MLMD is able to sample orders of magnitudes more trajectories. Additional sampling would be essential to characterize the mechanism of chemical reaction more accurately (with better statistics on ensemble average properties), to investigate how the dynamics of the reaction depends on temperature, pressure, etc., (with trajectories at various conditions), and to better shed light on the experiments with isotopic labeling (with trajectories using the same PES but the masses are changed [66]).

### Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests.

### Data Availability

Data will be made available on request.

### Acknowledgments

The authors appreciate the information technology service (ITS) from the University of Hawai'i at Manoa for the computational resources. This manuscript is based upon research supported by the National Science Foundation under Grant no. 2144031.

### Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cartre.2023.100257.

### References

- V. Fock, Naherungsmethode Zur Lösung Des Quantenmechanischen Mehrkörperproblems, Z. Angew. Phys. 61 (1–2) (1930) 126–148, doi:10.1007/BF01340294.
- P.A.M. Dirac, Note on exchange phenomena in the Thomas atom, Math. Proc. Cambridge Philos. Soc. 26 (3) (1930) 376–385, doi:10.1017/S0305004100016108.
- J.P. Perdew, Density-functional approximation for the correlation energy of the inhomogeneous electron gas, Phys. Rev. B 33 (12) (1986) 8822–8824, doi:10.1103/PhysRevB.33.8822.
- W. Kohn, L.J. Sham, Self-consistent equations including exchange and correlation effects, Phys. Rev. 140 (4A) (1965) A1133–A1138, doi:10.1103/PhysRev.140.A1133.
- J. Čížek, On the correlation problem in atomic and molecular systems. Calculation of wavefunction components in ursor-type expansion using quantum-field theoretical methods, J. Chem. Phys. 45 (11) (1966) 4256–4266, doi:10.1063/1.1727484.
- R. Krishnan, J.A. Pople, Approximate fourth-order perturbation theory of the electron correlation energy, Int. J. Quantum Chem. 14 (1) (1978) 91–100, doi:10.1002/qua.560140109.
- C. David Sherrill, H.F. Schaefer, The configuration interaction method: advances in highly correlated approaches, in: Advances in Quantum Chemistry, 34, Academic Press, 1999, pp. 143–269, doi:10.1016/S0065-3276(08)60532-8.
- R.J. Buenker, S.D. Peyerimhoff, Individualized configuration selection in CI calculations with subsequent energy extrapolation, Theor. Chim. Acta 35 (1974) 33–58.
- Y. Luo, T. Kreuzscher, C. Kang, W.L. Hase, K.-M. Weitzel, R. Sun, A chemical dynamics study of the HCl + HCl+ reaction, Int. J. Mass spectrom. 462 (2021) 116515, doi:10.1016/j.ijms.2020.116515.
- K. Fujioka, K.-M. Weitzel, R. Sun, The potential energy profile of the HBr+ + HCl bimolecular collision, J. Phys. Chem. A 126 (9) (2022) 1465–1474, doi:10.1021/acs.jpca.1c08300.
- 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>2</sub>H<sub>2</sub> carbenes, Proc. Natl Acad. Sci. 117 (48) (2020) 30142–30150, doi:10.1073/pnas.2019257117.
- Y. Luo, K. Fujioka, A. Shoji, W.L. Hase, K.-M. Weitzel, R. Sun, Theoretical study of the dynamics of the HBr+ + CO<sub>2</sub> → HOCO+ + Br reaction, J. Phys. Chem. A 124 (44) (2020) 9119–9127, doi:10.1021/acs.jpca.0c05323.
- 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 (HC<sub>2</sub>SH), Proc. Natl Acad. Sci. 117 (37) (2020) 22712–22719, doi:10.1073/pnas.2004881117.
- C. He, K. Fujioka, A.A. Nikolayev, L. Zhao, S. Doddipatla, V.N. Azyazov, A.M. Mebel, R. Sun, R.I. Kaiser, A chemical dynamics study of the reaction of the methylidyne radical (CH, X<sup>2</sup> Π) with dimethylacetylene (CH<sub>3</sub>CCCH<sub>3</sub>, X 1 A 1 g), Phys. Chem. Chem. Phys. 24 (1) (2022) 578–593, doi:10.1039/D1CP04443E.
- J. Behler, M. Parrinello, Generalized neural-network representation of high-dimensional potential-energy surfaces, Phys. Rev. Lett. 98 (14) (2007) 146401.
- J. Behler, Neural network potential-energy surfaces in chemistry: a tool for large-scale simulations, Phys. Chem. Chem. Phys. 13 (40) (2011) 17930–17955, doi:10.1039/c1cp21668f.
- K.T. Schütt, H.E. Sauceda, P.J. Kindermans, A. Tkatchenko, K.R. Müller, S.chNet - a deep learning architecture for molecules and materials, J. Chem. Phys. 148 (24) (2018) 1–11, doi:10.1063/1.5019779.
- 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), doi:10.1126/sciadv.1603015.
- B.J. Braams, J.M. Bowman, Permutationally invariant potential energy surfaces in high dimensionality, Int. Rev. Phys. Chem. 28 (4) (2009) 577–606.
- B. Jiang, H. Guo, Permutation invariant polynomial neural network approach to fitting potential energy surfaces, J. Chem. Phys. 139 (5) (2013) 054112, doi:10.1063/1.4817187.
- B. Jiang, J. Li, H. Guo, Potential energy surfaces from high fidelity fitting of *Ab initio* points: the permutation invariant polynomial - neural network approach, Int. Rev. Phys. Chem. 35 (3) (2016) 479–506, doi:10.1080/0144235X.2016.1200347.
- J. Li, B. Jiang, H. Song, J. Ma, B. Zhao, R. Dawes, H. Guo, From *Ab initio* potential energy surfaces to state-resolved reactivities: X + H<sub>2</sub>O ↔ HX + OH [X = F, Cl, and O(<sup>3</sup>P)] Reactions, J. Phys. Chem. A 119 (20) (2015) 4667–4687, doi:10.1021/acs.jpca.5b02510.
- K.V.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.
- J. Behler, First principles neural network potentials for reactive simulations of large molecular and condensed systems, Angewandte Chemie - International Edition 56 (42) (2017) 12828–12840, doi:10.1002/anie.201703114.
- M. Gastegger, P. Marquetand, High-dimensional neural network potentials for organic reactions and an improved training algorithm, J. Chem. Theory Comput. 11 (5) (2015) 2187–2198, doi:10.1021/acs.jctc.5b00211.
- J. Li, K. Song, J. Behler, A critical comparison of neural network potentials for molecular reaction dynamics with exact permutation symmetry, Phys. Chem. Chem. Phys. 21 (19) (2019) 9672–9682, doi:10.1039/C8CP06919K.
- Z. Yang, S. Wang, J. Yuan, M. Chen, Neural network potential energy surface and dynamical isotope effects for the N<sup>+</sup>(<sup>3</sup>P) + H<sub>2</sub> → NH<sup>+</sup> + H reaction, Phys. Chem. Chem. Phys. 21 (40) (2019) 22203–22214, doi:10.1039/C9CP02798J.
- S. Käser, O.T. Unke, M. Meuwly, Isomerization and decomposition reactions of acetaldehyde relevant to atmospheric processes from dynamics simulations on neural network-based potential energy surfaces, J. Chem. Phys. 152 (21) (2020) 214304, doi:10.1063/5.0008223.
- Y. Liu, J. Li, An accurate potential energy surface and ring polymer molecular dynamics study of the Cl + CH<sub>4</sub> → HCl + CH<sub>3</sub> reaction, Phys. Chem. Chem. Phys. 22 (1) (2020) 344–353, doi:10.1039/C9CP05693A.
- J. Chen, X. Xu, S. Liu, D.H. Zhang, A neural network potential energy surface for the F + CH<sub>4</sub> reaction including multiple channels based on coupled cluster theory, Phys. Chem. Chem. Phys. 20 (14) (2018) 9090–9100, doi:10.1039/C7CP08365C.
- J. Qin, J. Li, An accurate full-dimensional potential energy surface for the reaction OH + SO → H + SO<sub>2</sub>, Phys. Chem. Chem. Phys. 23 (1) (2021) 487–497, doi:10.1039/D0CP05206J.
- D. Lu, J. Behler, J. Li, Accurate global potential energy surfaces for the H + CH<sub>3</sub>OH

- reaction by neural network fitting with permutation invariance, *J. Phys. Chem. A* 124 (28) (2020) 5737–5745, doi:10.1021/acs.jpca.0c04182.
- [33] A.T.H. Le, N.H. Vu, T.S. Dinh, T.M. Cao, H.M. Le, Molecular dynamics investigations of chlorine peroxide dissociation on a neural network Ab initio potential energy surface, *Theor. Chem. Acc.* 131 (3) (2012) 1158, doi:10.1007/s00214-012-1158-2.
- [34] X. Zhang, J. Chen, X. Xu, S. Liu, D.H. Zhang, A neural network potential energy surface for the  $F + H_2O \leftrightarrow HF + OH$  reaction and quantum dynamics study of the isotopic effect, *Phys. Chem. Chem. Phys.* 23 (14) (2021) 8809–8816, doi:10.1039/D1CP00641J.
- [35] H. Xiang, L. Tian, Y. Li, H. Song, Energy- and local-gradient-based neural network method for accurately describing long-range interaction: application to the  $H_2 + CO^+$  reaction, *J. Phys. Chem. A* 126 (2) (2022) 352–363, doi:10.1021/acs.jpca.1c09719.
- [36] Y. Liu, J. Li, Permutation-invariant-polynomial neural-network-based  $\Delta$ -machine learning approach: a case for the  $HO_2$  self-reaction and its dynamics study, *J. Phys. Chem. Lett.* 13 (21) (2022) 4729–4738, doi:10.1021/acs.jpclett.2c01064.
- [37] M. Pan, H. Xiang, Y. Li, H. Song, Study on the kinetics and dynamics of the  $H_2 + NH_2^-$  reaction on a high-level Ab initio potential energy surface, *Phys. Chem. Chem. Phys.* 23 (33) (2021) 17848–17855, doi:10.1039/D1CP02423J.
- [38] Chmiela, S.; Sauceda, H.E.; Poltavsky, I. SGDMML : constructing Accurate and Data Efficient Molecular Force Fields Using. 2018, No. February 2019, 1–6.
- [39] 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) 68–70, doi:10.1063/1.5078687.
- [40] 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.
- [41] A.P. Bartók, R. Kondor, G. Csányi, On representing chemical environments, *Phys. Rev. B Condens. Matter Mater. Phys.* 87 (18) (2013) 1–19, doi:10.1103/PhysRevB.87.184115.
- [42] T.S. Ho, H. Rabitz, Reproducing Kernel Hilbert Space interpolation methods as a paradigm of high dimensional model representations: application to multidimensional potential energy surface construction, *J. Chem. Phys.* 119 (13) (2003) 6433–6442, doi:10.1063/1.1603219.
- [43] J.A. Pople, Nobel lecture: quantum chemical models\*, *Rev. Mod. Phys.* 71 (5) (1998) 1267–1274.
- [44] K.R. Brorsen, Reproducing global potential energy surfaces with continuous-filter convolutional neural networks, *J. Chem. Phys.* 150 (20) (2019) 204104, doi:10.1063/1.5093908.
- [45] J. Zeng, L. Cao, M. Xu, T. Zhu, J.Z.H. Zhang, Complex reaction processes in combustion unraveled by neural network-based molecular dynamics simulation, *Nat. Commun.* 11 (1) (2020) 5713, doi:10.1038/s41467-020-19497-z.
- [46] J.P. Mailoa, M. Kornbluth, S. Batzner, G. Samsonidze, S.T. Lam, J. Vandermause, C. Ablitt, N. Molinari, B. Kozinsky, A fast neural network approach for direct covariant forces prediction in complex multi-element extended systems, *Nat Mach Intell* 1 (10) (2019) 471–479, doi:10.1038/s42256-019-0098-0.
- [47] D. Plamper, S. Schmidt, K.-M. Weitzel, Kazuomi Fujioka; Rui Sun, *Private Commun.* (2022).
- [48] S. Batzner, A. Musaelian, L. Sun, M. Geiger, J.P. Mailoa, M. Kornbluth, N. Molinari, T.E. Smidt, B. Kozinsky, E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials, *Nat. Commun.* 13 (1) (2022) 2453, doi:10.1038/s41467-022-29939-5.
- [49] Chr. Møller, M.S. Plesset, Note on an approximation treatment for many-electron systems, *Phys. Rev.* 46 (7) (1934) 618–622, doi:10.1103/PhysRev.46.618.
- [50] F. Weigend, R. Ahlrichs, Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: design and assessment of accuracy, *Phys. Chem. Chem. Phys.* 7 (18) (2005) 3297, doi:10.1039/b508541a.
- [51] 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, doi:10.1016/S0009-2614(89)87395-6.
- [52] D.G. Truhlar, Basis-set extrapolation, *Chem. Phys. Lett.* 294 (1–3) (1998) 45–48, doi:10.1016/S0009-2614(98)00866-5.
- [53] A. Hjorth Larsen, J. Jørgen Mortensen, J. Blomqvist, I.E. Castelli, R. Christensen, M. Dułak, J. Friis, M.N. Groves, B. Hammer, C. Hargus, E.D. Hermes, P.C. Jennings, P. Bjerre Jensen, J. Kermode, J.R. Kitchin, E. Leonhard Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. Bergmann Maronsson, T. Maxson, T. Olsen, L. Pastewka, A. Petersson, C. Rostgaard, J. Schiøtz, O. Schütt, M. Strange, K.S. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng, K.W. Jacobsen, The atomic simulation environment—a python library for working with atoms, *J. Phys. Condens. Matter* 29 (27) (2017) 273002, doi:10.1088/1361-648X/aa680e.
- [54] X. Hu, W.L. Hase, T. Pirraglia, Vectorization of the general Monte Carlo classical trajectory program VENUS, *J. Comput. Chem.* 12 (8) (1991) 1014–1024, doi:10.1002/jcc.540120814.
- [55] G.H. Peslherbe, H. Wang, W.L. Hase, Monte Carlo sampling for classical trajectory simulations, *Adv. Chem. Phys.* 105 (1999) 171–201.
- [56] U. Lourderaj, R. Sun, S.C. Kohale, G.L. Barnes, W.A. de Jong, T.L. Windus, W.L. Hase, The VENUS/NWChem software package. Tight coupling between chemical dynamics simulations and electronic structure theory, *Comput. Phys. Commun.* 185 (3) (2014) 1074–1080, doi:10.1016/j.cpc.2013.11.011.
- [57] S. Käser, D. Koner, A.S. Christensen, O.A. von Lilienfeld, M. Meuwly, ML models of vibrating H<sub>2</sub>S<sub>2</sub>CO: comparing reproducing kernels, FCHL and PhysNet, *J. Phys. Chem. A* 124 (42) (2020) 8853–8865, doi:10.1021/acs.jpca.0c05979.
- [58] O.T. Unke, S. Chmiela, H.E. Sauceda, M. Gastegger, I. Poltavsky, K.T. Schütt, A. Tkatchenko, K.R. Müller, Machine learning force fields, *Chem. Rev.* (2021) 10142–10186 American Chemical Society August 25, doi:10.1021/acs.chemrev.0c01111.
- [59] 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.
- [60] G. Cumming, Inference by eye: reading the overlap of independent confidence intervals, *Stat. Med.* 28 (2) (2009) 205–220, doi:10.1002/sim.3471.
- [61] A. Kolmogorov, Sulla determinazione empirica Di Una Legge Di Distribuzione, *Inst. Ital. Attuari, Giorn.* 4 (1933) 83–91.
- [62] N. Smirnov, Table for estimating the goodness of fit of empirical distributions, *Ann. Statist. Math.* 19 (2) (1948) 279–281.
- [63] F.J. Massey, The Kolmogorov-Smirnov test for goodness of fit, *J. Am. Stat. Assoc.* 46 (253) (1951) 68–78, doi:10.1080/01621459.1951.10500769.
- [64] A.S. Christensen, L.A. Bratholm, F.A. Faber, O. Anatole Von Lilienfeld, FCHL revisited: faster and more accurate quantum machine learning, *J. Chem. Phys.* 152 (4) (2020), doi:10.1063/1.5126701.
- [65] J. Westermayr, F.A. Faber, A.S. Christensen, O.A. von Lilienfeld, P. Marquetand, Neural networks and kernel ridge regression for excited states dynamics of CH<sub>2</sub>NH<sub>2</sub><sup>+</sup>: from single-state to multi-state representations and multi-property machine learning models, *Mach. Learn. Sci. Technol.* (2) (2020) 1, doi:10.1088/2632-2153/ab88d0.
- [66] M. Wolfsberg, E.B. Wilson, J.C. Decius, P.C. Cross, Theoretical evaluation of experimentally observed isotope effects, *Acc. Chem. Res.* 5 (7) (1972) 225–233.