Modeling Chemical Processes in Explicit Solvents with Machine Learning Potentials

Hanwen Zhang, Veronika Juraskova, and Fernanda Duarte*

Chemistry Research Laboratory, 12 Mansfield Road, Oxford, OX1 3TA

E-mail: fernanda.duartegonzalez@chem.ox.ac.uk

Abstract

Solvent effects influence all stages of the chemical processes, modulating the stability of intermediates and transition states, as well as altering reaction rates and product ratios. However, accurately modelling these effects remains challenging. Here, we present a general strategy for generating reactive machine learning potentials (MLPs) to model chemical processes in solution. Our approach combines active learning with descriptor-based selectors and automation, enabling the construction of data-efficient training sets that span the relevant chemical and conformational space. We demonstrate the versatility of this strategy by applying it to investigate a Diels-Alder reaction in water and methanol. The generated MLPs exhibit excellent agreement with experimental data and provide insights into the differences in reaction rates observed between the two solvents. Our strategy offers an efficient approach to the routine modelling of chemical reactions in solution, opening up avenues for studying complex chemical processes in an efficient manner.

Introduction

(Bio)chemical and industrially relevant reactions occur predominantly in the liquid phase.[1,2] The influence of solvent on reaction rates was first documented by Berthelot and Pean de Saint-Gilles in 1862[3] and later formalised by Menschutkin in 1890.[4] Since then, experimental and computational techniques have been developed to elucidate and quantify the origin of solvent effects on reactivity and selectivity.[5] From an atomistic point of view, solvent effects arise from the interactions between solute and solvent molecules, which, although generally weak, have a significant impact on the overall reaction dynamics.

In computational chemistry, solvent effects are modelled using implicit or explicit solvent models. The former represents solvents as a polarizable continuum, offering computational simplicity and efficiency. However, it fails to capture the contributions arising from solute-solvent interactions, including entropy and pre-organisation effects.[6] On the other hand, explicit solvent models provide an atomistic representation of the solvent but at a much higher computational cost since they require the use of *ab initio* molecular dynamics (AIMD). This cost becomes significant when attempting to compute free energies, where extensive sampling is required to obtain statistically meaningful ensembles. Hybrid approaches, such as quantum mechanics/molecular mechanics (QM/MM), can alleviate this computational cost by describing only the reactive part at a QM level while the environment is described classically. However, the inclusion of mobile solvent molecules into the QM part creates additional technical difficulties, such as discontinuities at the boundary between the QM and MM regions.[7,8]

In recent years, machine learning-based potentials (MLPs) have emerged as powerful surrogates for existing modelling techniques, representing complex potential energy surfaces (PES) with an accuracy comparable to QM methods but at a significantly lower computational cost.[9,10] Since the pioneering work of Behler and Parinello (BP) on neural network-based potentials (NNP),[9] several MLP approaches have been developed, including models directly derived from BP, such as DeepMD[11] and ANI,[12] message-passing NN models,[13] such as

PhysNet,[14] SchNet,[15] and MACE,[16] kernel-based approaches, such as Gaussian Approximation Potential (GAP) and gradient-domain ML (GDML),[17–19] and linear regression-based models, such as linear Atomic Cluster Expansion (ACE).[20,21] These MLPs have been successfully applied to study organic molecules,[12,22] material properties[23,24] and simple chemical reactions on the gas phase and interfaces.[25–28] However, the application of MLPs to study chemical processes in solution remains underexplored. The few reported examples include urea decomposition in water,[29] 1,3-dipolar cycloadditions in water,[30] and reactions in alkali carbonate–hydroxide electrolytes,[31] with the NNPs being the state-of-the-art approach requiring thousands of AIMD configurations.

One of the main challenges limiting the application of MLPs to study chemical reactions in solution is the quality and size of the training data set, which needs to incorporate diverse configurations to effectively capture solute-solvent interaction at the minima and TS regions. Unlike solid-state systems, where particles are packed in highly ordered arrangements, molecules in solution are often flexible, resulting in a larger conformational space that needs to be sampled to accurately describe their PES. Moreover, the reliable description of reaction barriers and non-covalent interactions requires accurate reference electronic structure methods, further increasing the computational cost.[32] The complexity of the training data can be reduced in Δ-ML models,[33] which predict the differences between semi-empirical baseline and QM PES rather than the absolute energies. However, even in these approaches, the cost of energy and force evaluations remains a bottleneck, as the baseline computations need to be performed in each MD step.[34,35]

Recently, we demonstrated the promise of several MLP approaches, including ACE, GAP and NequIP, in modelling various molecular systems and chemical reactions.[36,37] Notably, when coupled with active learning (AL), where the MLP is retrained using new data collected based on preliminary versions of MLP, these methods can provide accurate potentials with much smaller data sets than traditional NN approaches. In this way, one can efficiently explore the chemical space required to describe reactions involving complex PESs.

The performance and efficiency of AL strategies heavily depend on the algorithm used to select suitable structures for retraining. Previous studies have utilised various metrics or selectors to evaluate the performance of MLPs within AL iterations. Probably the simplest and most straightforward metric is the direct comparison of energies or forces between the reference (ground-truth) method and the MLP.[36] However, this metric requires QM calculations at every evaluation step, which is only feasible for small systems in the gas phase. Gaussian Process-based methods, such as GAP, commonly use *energy uncertainty* as a metric, which refers to the variance in predicted energy.[28,38–40] However, this property is limited to Bayesian models, as they yield a predictive distribution characterised by its mean and variance. Other models only provide a point prediction for a given structure and are thus unable to utilise this metric.[41] Another approach widely used in NNPs,[30,42–44] although not limited to, is query-by-committee, where uncertainty is quantified through the variance of predicted energies or forces among an ensemble of MLPs, referred to as committees. In the case of NNs, multiple potentials are trained with the same architecture but with different initial parameters. This approach provides on-the-fly uncertainty estimation but incurs additional computational costs associated with the training of several NNPs. In all the uncertainty metrics mentioned above, a large variance indicates an under-representation of the corresponding data points and the need to include them in the training set.

In this work, we propose a general AL strategy for training MLPs to model chemical processes in explicit solvents and, more generally, large systems with complex PES. To achieve this, we employ an AL loop combined with descriptor-based selectors. In contrast to the metrics described above, these selectors use molecular descriptors, such as Smooth Overlap of Atomic Positions (SOAP)[45] or ACE,[20,21] to assess whether the training set accurately represents the chemical space of interest. By examining the descriptor space in the training set, we demonstrate that these selectors provide a general metric applicable across different MLP approaches at a low computational cost. The effectiveness and versatility of the selectors are showcased using the linear ACE model. Firstly, we investigate a simple water box as a test case, evaluating and comparing the performance of various selectors in terms of the quality of the potential and data efficiency. Building upon these findings, we further apply our strategy to study the Diels-Alder (DA) reaction between cyclopentadiene (CP) and methyl vinyl ketone (MVK) in both water and methanol. Through this example, we illustrate the transferability of our approach to accurately model chemical reactions in different solvent environments.

The generality and computational efficiency of our approach make it well-suited for a wide range of applications, contributing to the advancement of MLP in various fields. By harnessing its capabilities, accurate and efficient modelling of chemical processes in explicit solvents becomes attainable. This enables a deeper understanding of solvent effects and facilitates the investigation of more complex processes in solution.

Methods

Workflow

The AL strategy employed in this work to train MLPs is illustrated in Fig. 1(a). The first step involves generating a small set of configurations labelled with reference energies and forces. This information is used to train the initial version of the MLP. For a given reaction, we employ two different training sets, one containing the reacting substrates in the gas phase (or implicit solvent) and another including explicit solvent molecules. The latter is necessary to account for specific non-covalent interactions between the solute and solvent.

Different training strategies are used for each of the data sets mentioned above. For the gas phase or implicit solvated molecules, training configurations are generated by randomly displacing the atomic coordinates. In the case of a chemical reaction, the training typically starts from the corresponding TS. The dataset containing substrate and the explicit solvent is generated either from solvent molecules in a box under periodic boundary conditions (PBC) or cluster models containing only a handful of solvent molecules placed at relevant

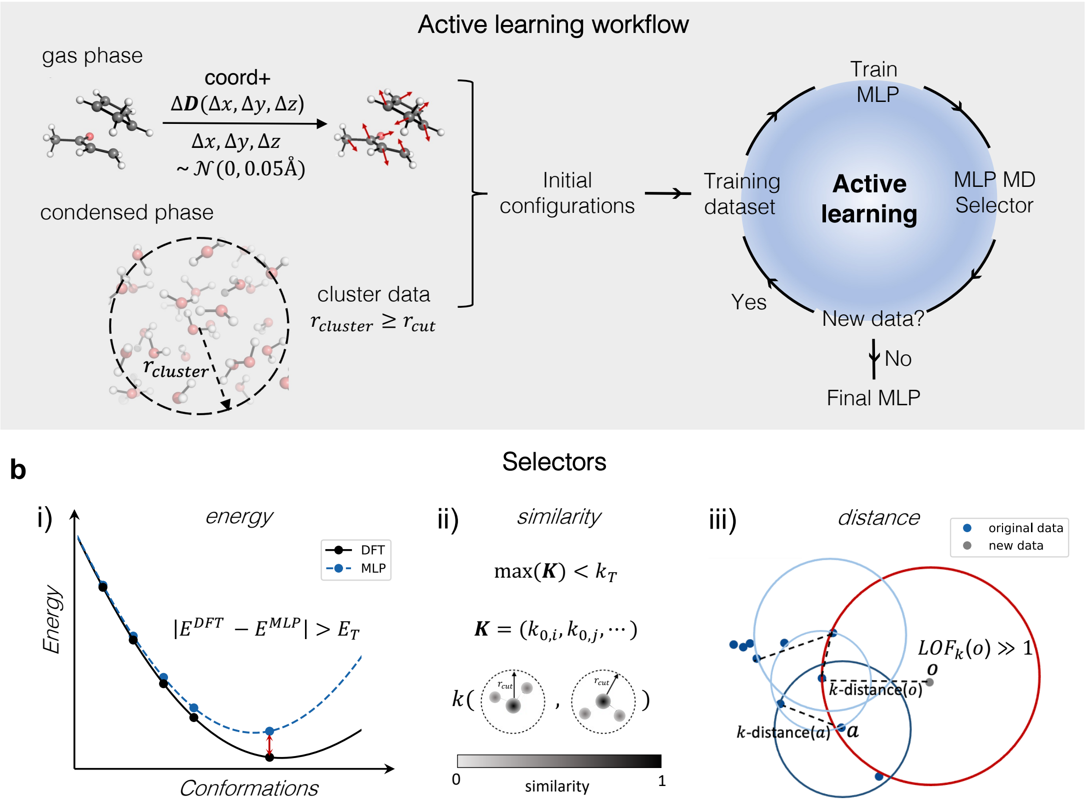

Figure 1: AL workflow and structure selection strategies. (a) Schematic representation of the initial configuration construction and AL process. (b) Illustration of the selection methods implemented in the AL process: i) *energy* selector selecting the data point whose energy difference (highlighted in red) between the ground truth (DFT) and MLP predicted values exceeds the threshold (E_T). ii) *similarity* selector collects configuration with dissimilar geometry to the original data set. iii) *distance* selector uses the local outlier factor method to identify outliers from the original data set. The circle centred at the outlier o with its k-distance as the radius is highlighted in red.

positions. While PBC reproduces well the structure of the bulk solvent and includes long-range interactions, generating such training data, in particular with AIMD, is computationally expensive as thousand of configurations are required to accurately describe all the interactions, making it unfeasible for most chemically relevant systems. An alternative approach for generating PBC data is to use classical MM force fields; however, they are often inaccurate and unable to describe bond-breaking/forming processes. Moreover, MM configurations exhibit a weak overlap with the true potential energy surfaces (PES), even for non-reactive systems making them unsuitable for the training of MLPs.[36] In this regard, cluster data labelled with molecular energy and forces provide all structural information for MLPs based on the local descriptors and offer access to a large spectrum of electronic structure methods, including long-range corrected and double-hybrid DFT functionals. The minimum radius of the solvent shell around the substrate should be at least two times the cut-off radius used for training the MLP to avoid artificial forces close to the solvent-vacuum interface in the cluster data.

As further discussed in the Results section, we observed good transferability of cluster-based MLPs to systems with PBC. Similar transferability has already been reported for an NNP applied to bulk water, where NNPs trained using both PBC and cluster training data demonstrate similar performance in the prediction of bulk properties, such as radial distribution functions (RDF), self-diffusion coefficients, and equilibrium densities.[46]

After the initial MLP training, one structure from the initial training set is selected as the starting point to propagate the molecular dynamics (MD) using the first version of the trained MLP. Several rounds of short MD simulations are then performed to assess the stability of the potential and generate new training structures. The simulation time is set to $(n^3 + 2)$ fs where n corresponds to the index of the MD run, starting from 0. More details can be found in Ref. 36. From each MD trajectory, the last frame is evaluated by the selector to determine whether to add or not this structure to the training set. If the structure is not selected, n is incremented, and MD runs are repeated until the maximum simulation time or a maximum number of MD iterations is achieved.

Descriptor-based selectors

In previous studies, we have utilised an energy-based selector (further referred to as *energy*) to determine whether a configuration should be added to the training set. This selector identifies structures that show an error in predicted energy higher than the threshold E_T, satisfying the condition $\left|E^{DFT} - E^{MLP}\right| > E_T$. Structures with prediction errors greater than $10E_T$ were excluded from the dataset as these were too distorted to provide meaningful information. Although reliable, this approach requires QM calculations for each selection step, which is computationally prohibitive for large systems.

In this work, we introduce selectors using descriptor-based selection criteria. Descriptors transfer molecular representation from Cartesian coordinates to a physical invariant description, encompassing both the geometrical and chemical information of molecular structures. Evaluating the SOAP descriptor over the training data thus provides information on how well the training set covers the relevant conformational and chemical space, enabling the identification of underrepresented data points.

During the AL loop, we apply so-called similarity- and distance-based selectors, referred here to as *similarity* and *distance* selectors, respectively. The former quantifies the similarity between a new data point p and existing configurations $p^{'}$ using the kernel function $k(p \cdot p^{'})$. The similarity matrix of the data point is defined as:

$$\boldsymbol{K} = \left(|k(p_0 \cdot p_i)|^{\zeta}, |k(p_0 \cdot p_j)|^{\zeta}, \cdots \right) \tag{1}$$

where p_0 is the SOAP vector of the new structure, p_i is the SOAP vector of the i-th configuration in the existing set, and ζ is a positive integer that increases the sensitivity of kernel to changes in atomic position.[45] The kernel is computed between the new configuration and all other configurations in the training data set. The selector adds structure to the training set if the maximum value of its similarity matrix, $\boldsymbol{K}$, is smaller than threshold k_T, i.e., $\max(\boldsymbol{K}) < k_T$. Selecting an appropriate threshold is key as too low values (e.g., similarity below 0.9) can result in the selection of non-physical structures that fail to converge in the self-consistent field (SCF) computations, while too high values (e.g., 1) do not provide any additional information.

For the *distance* selector, we use the local outlier factor (LOF) method[47] to determine whether the SOAP vector of the new configuration is an outlier compared to the SOAP vectors of the existing training data. LOF is based on the local density of each point, which is calculated by measuring the Euclidean distance between the target point and its k-nearest neighbours (Fig. 1 (b) iii)). The local reachability density of an object o, denoted as $lrd_k(o)$ is calculated as:

$$lrd_k(o) = \left(\frac{\sum_{i \in N_k(o)} rd_k(o,\, i)}{|N_k(o)|} \right)^{-1}. \tag{2}$$

$rd_k(o, a)$ in the equation corresponds to reachability distance defined as

$$rd_k(o, a) = \max(k\text{-distance}(a),\ d(o,\, a)). \tag{3}$$

Here, k-distance(a) represents the radius of the smallest circle with its origin in a, which includes the k-nearest neighbours of a (illustrated in dark blue in the Fig.1 b iii)), $d(o,\, a)$ is the Euclidean distance between points o and a, where point o is an outlier and point a is one of its neighbours. $N_k(o)$ is a set of k-nearest neighbours of o, which are illustrated by blue points. The local reachability density is thus expressed as the number of neighbours per distance unit. If the local reachability density of the target point is smaller than that of its neighbours, the point is considered an outlier and added to the training data set. The comparison among the local densities is achieved by computing the ratio of the average local density of neighbours and the local density of the point, as follows:

$$LOF_k(o) = \frac{\sum_{i \in N_k(o)} \frac{lrd_k(i)}{lrd_k(o)}}{|N_k(o)|} \tag{4}$$

A LOF value close to 1 indicates that a point is located in a similarly dense region as its neighbours. A LOF value less than 1 represents an inlier, meaning that the point is situated in a denser region, while a value greater than 1 indicates an outlier. Since there is no definitive rule for selecting a LOF threshold to identify outliers, in this study, we chose the threshold to encompass 20% of the outliers present in the training data.

In contrast to the variance metric, descriptor-based selectors can be applied to any regression method without incurring additional computational costs for training multiple models.

As discussed in the next section, both descriptor-based selectors accelerate the training of MLP models compared to the *energy* selector. This acceleration arises from the reduction in QM calculations as well as the ability of the selectors to explore the relevant chemical space more efficiently.

Results

Performance of selectors – water models

We assessed the performance of three different selectors (*energy*, *similarity*, and *distance*) during the AL training of MLPs for water. In each case, we evaluated the accuracy of the generated potential by measuring the mean absolute deviation (MAD) of total energy and atomic forces with respect to the ground truth method, PBE0-D3BJ. We also considered the efficiency of training by analysing the number of configurations required for training. The AL process was considered complete when no configurations were selected within the maximum simulation time of 5 ps.

Our results demonstrate that, in general, all descriptors provide accurate and stable potentials. This is evident from the direct comparison of predicted and ground-true energies for two systems: a small cluster system comprising 27 water molecules, from which we extracted 200 configurations from a 1 ps MD simulation, and a larger system consisting of 216 water molecules under PBC, where we performed 50 ps dynamics in the NVE ensemble (SI § S2). All MLPs achieved MAD errors in energy below 1 kcal mol^{-1}and errors in force of less than 2 kcal mol^{-1}Å^{-1}. Moreover, all MLPs remained stable during the NVE simulations, which is a significantly longer simulation time than the maximum simulation time in the AL process.

The main difference between the selectors compared here lies in the number of configurations selected during the training and computational efficiency. The descriptor-based selectors generally require a much smaller number of configurations, with 40 and 52 data points for *similarity* and *distance* selectors, respectively, compared to 281 data points for *energy*. To explore the geometrical similarity among the data selected by different selectors, we combined all the training sets (373 configurations in total, shown in grey in Fig. 2a) and analysed them with t-SNE maps based on global SOAP representation (cutoff 10 Å). Analysis of these results reveals three distinct clusters. The cluster at the bottom represents initial structures in training, where water molecules are randomly placed and have relatively high energy compared with the other clusters. The middle cluster contains configurations in which water molecules exhibit more structured arrangements and form hydrogen bonds (HB). Finally, the upper cluster comprises configurations where the system approaches equilibrium, as evidenced by much lower energies. Interestingly, the energy selector predominantly selects geometries near the equilibrium configurations (67%), with only a handful of configurations in the middle/bottom cluster. In contrast, the configurations selected by either the *similarity* or *distance* selectors are much more evenly distributed, as shown in Fig. 2b. The size of the training set does not influence the MLP performance, as all MLPs reproduced well the experimental radial distribution function (RDF) of water (Fig. 2 (c)). The small differences in the positions and amplitudes of peaks for the first and second solvation shells are likely due to the level of theory used and the lack of nuclear quantum effects.[48]

This demonstrates that both similarity and distance selectors explore the chemical space more efficiently, using only around 15% of the training data required by the *energy* selector. Furthermore, they reduce the structural correlation in the training sets, as illustrated in Fig. 2(a). Instead of containing numerous points with similar geometries, these training sets encompass data points distributed across the space. Descriptor-based selectors are also significantly faster than energy-based ones as they do not require QM computation at every selection step, making them suitable for larger and more complex systems.

Overall, the descriptor-based selectors demonstrate superior efficiency in terms of training speed and amount of data, outperforming the *energy* selector. All three selectors yield potentials with comparable accuracy and stability. However, the *distance* selector requires a

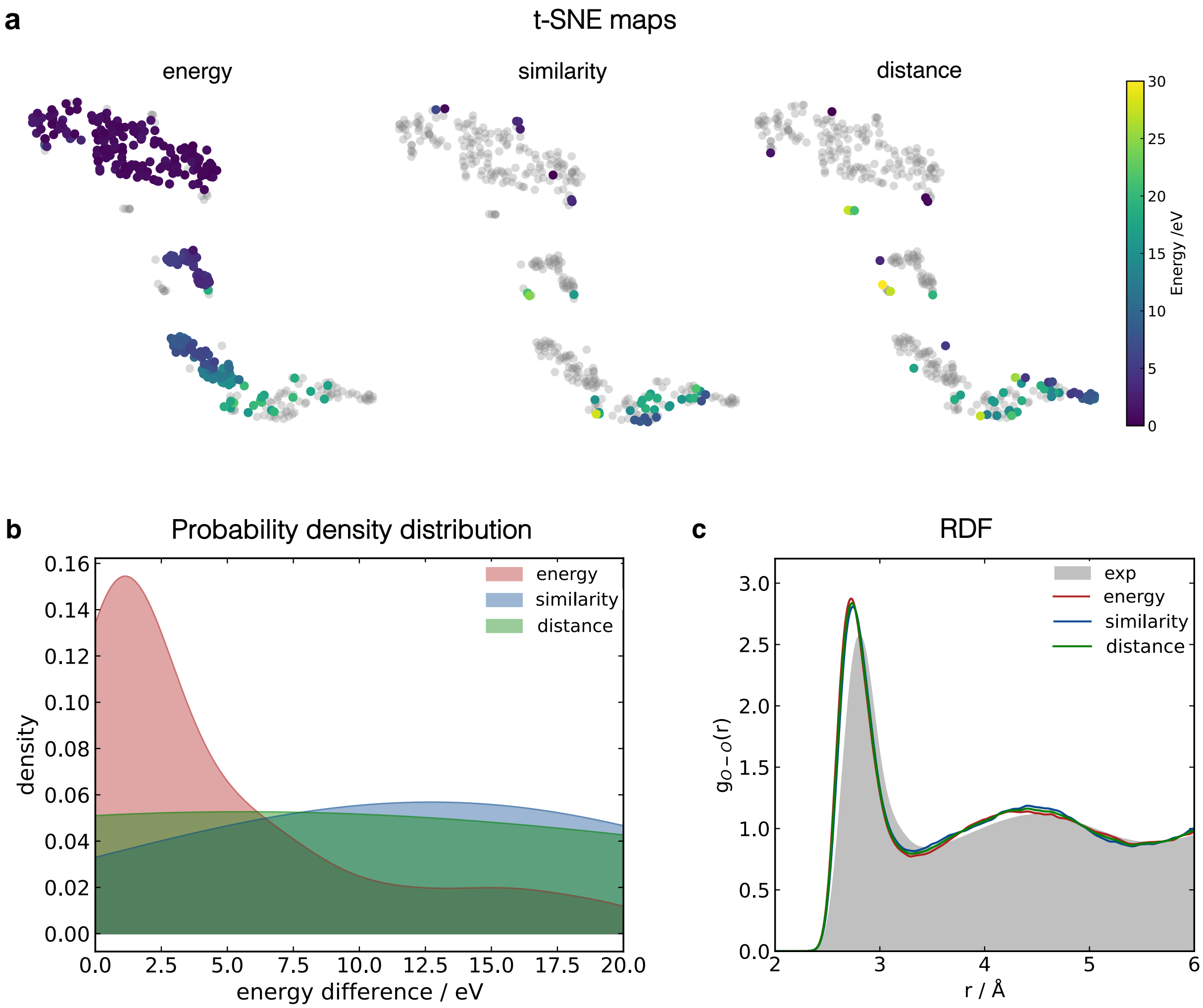

Figure 2: Comparative analysis of selectors. (a) t-SNE maps of configurations generated during AL using (i) *energy*, (ii) *similarity* and (iii) *distance* selectors. In each case, configurations are labelled by their energy relative to the lowest energy configuration within the corresponding dataset. (b) Distributions of energies across the configurations obtained using the different selectors for a 27-water cluster system. (c) Oxygen–oxygen radial distribution functions (RDF) from a 20 ps NVT ACE-MD simulation of 216 water molecules in an 18.65 Å cubic box under PBC. RDF obtained from ACE potential trained with *energy* (red), *similarity* (blue), or *distance* (green) selector. Experimental RDF in grey shading.[49]

more extensive initial data set to perform the selection criteria in the first iteration of the AL process as it relies on neighbouring information. Throughout this paper, we will use the similarity selector for MLP training.

Training Strategy – DA reaction of CP and MVK in explicit water

The ability of MLPs to accurately describe chemical processes in explicit solvents is critical for extending their application to more challenging systems. To this end, we investigated the solvent effects on the DA reaction between CP and MVK in explicitly modelled water and methanol. While this reaction exhibits only minor solvent effects compared to charged systems, several reports have shown rate acceleration and selectivity enhancement when water or aqueous solvent mixtures are employed.[50–52] The reaction is accelerated up to 58-fold in water compared to methanol,[51] and the *endo*/*exo* selectivity is enhanced by 8-fold in water over benzene.[52] This behaviour is widely explained by the formation of stronger HBs between solvent and substrate at the TS compared to the reactants state (RS) and product state (PS).[53–58] In addition, solvent polarity and hydrophobic effects have also been suggested to contribute to this enhancement.[50,56,59,60]

We trained four ACE MLPs for each *endo*/*exo* DA reaction between CP and MVK in either explicit water or methanol. For comparison, we trained the models also in implicit solvent, neglecting the presence of the explicit solvent molecules. Here, we describe the strategy employed to obtain the ACE MLP for the *endo* reaction in explicit water. A similar approach was used for the other systems (SI § S5 for further details). For the reaction in explicit solvents, the training set consisted of four subsets, each aimed at describing different types of interactions in the system. Subset 1 corresponds to the substrate complex (CP + MVK) and provides information about intramolecular interactions and the intrinsic reactivity of the system (Fig. 3(a)). Subsets 2 and 3 consist of the substrate with 2 and 33 water molecules, respectively, and aim at describing various solute-solvent interactions. Finally, subset 4 contains only water molecules, providing information about solvent-solvent interactions in bulk solvent. By combining these sub-training sets, we obtained 600 training points, which were used to train the final ACE MLP.

The accuracy of the resulting ACE MLP was assessed by conducting 500 fs ACE MLP-MD simulations starting from two configurations not included in the training set. The first

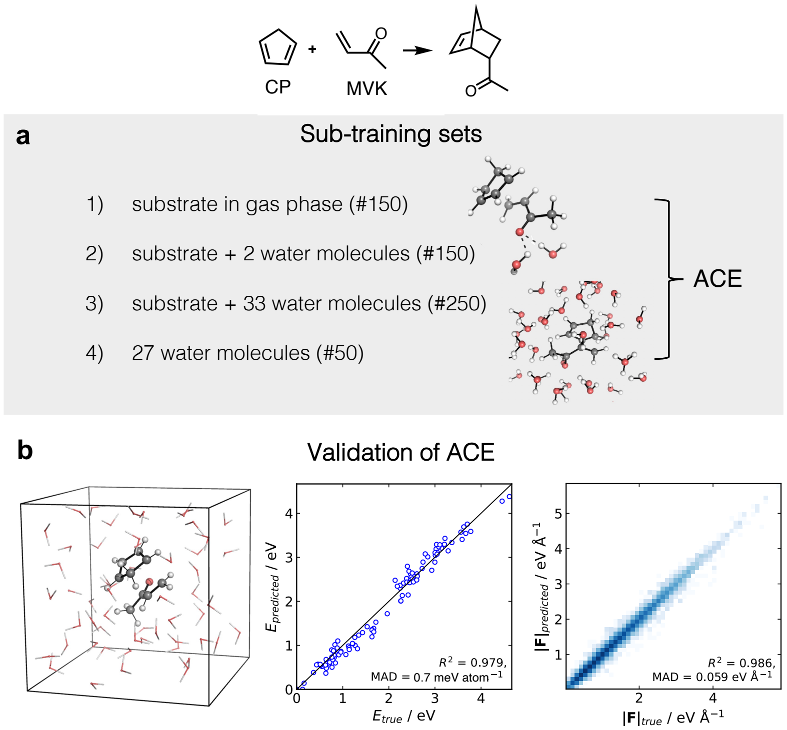

Figure 3: Training approach and accuracy of ACE MLP for *endo* DA reaction of CP and MVK in explicit water. (a) The training data consists of four subsets, each describing key interactions. (b) Comparison of ground-truth and ACE MLP energies and forces over a 500-fs ACE MLP-MD downhill dynamics for a system containing substrate and 55 water molecules with a timestep of 0.5 fs at 300 K.

configuration, which is similar to subset 2, consisted of the gas phase TS and three water molecules forming HBs with the carbonyl group of MVK. The resulting dynamics are stable with the energy error of 2 meV atom^{-1}, demonstrating the ability of the ACE MLP to represent reactions and specific HB interactions (Fig. S8). The second one corresponds to a TS immersed in a box containing 55 water molecules (box size: 12.42 Å, Fig. 3(b)). The resulting accuracy confirms that ACE MLP is reliable for investigating the reaction of CP and MVK in solution. Validation of the other systems is provided in the SI § S5.

Application of ACE MLPs – DA reaction of CP and MVK in solvents

After obtaining accurate and stable ACE MLPs for the reaction of CP and MVK in two solvent environments, we utilised these potentials to investigate the reaction pathway in more detail. This was done by conducting a relaxed 2D scan along the forming C-C bonds r_1 and r_2 in both implicit and explicit solvent (Fig. 4(a) and Fig. S14, respectively). Analysis of the *endo* 2D scan in both implicit and explicit water reveals the presence of a zwitterionic-like structure characterised by the formation of only one C-C bond in the region around $r_1 <$ 1.6 Å and $2.5 < r_2 < 3.0$ Å (an example marked by a cross in Fig. 4(a)). Unrestricted DFT calculations on these geometries confirmed that they do not exhibit any diradical character (SI § S7). The cross-labelled zwitterionic specie is slightly more stabilised in explicit solvent than in implicit solvent ($\Delta\Delta E = 2.2$ kcal mol^{-1}).

Using ACE MLP-MD in conjunction with umbrella sampling (US, ACE MLP-MD/US), we then computed the activation free energy, $\Delta G^{\ddagger}$, in implicit and explicit solvent (Fig. 4(b) and S17). $\Delta G^{\ddagger}$ in implicit solvent are 21.2 kcal mol^{-1}and 23.6 kcal mol^{-1}, for the *endo* and *exo* pathway, respectively ($\Delta\Delta G^{\ddagger} = 2.4$ kcal mol^{-1}). Incorporating explicit solvent reduces these values to 18.8 kcal/mol and 20.3 kcal mol^{-1}($\Delta\Delta G^{\ddagger} = 1.5$ kcal mol^{-1}), thereby improving the agreement with experimental data (19.2 kcal mol^{-1} and 21.1 kcal mol^{-1}, respectively,[51] Table S2). These results also illustrate the significance of solute-solvent interactions in the reaction.

Furthermore, the presence of solvent molecules influences the synchronicity of the reaction (Fig. 4(b)). In explicit solvent, the reaction exhibits an earlier and more asynchronous TS compared to the implicit solvent or gas phase. The difference in bond length between the forming C-C bonds $\Delta r = |r_2 - r_1|$ at the TS increases from 30 pm in the gas phase to 37 pm in implicit solvent and 46 pm in explicit water. Notably, explicit water molecules also altered the reaction mechanism from a concerted asynchronous to a "pseudo" stepwise mechanism, as evidenced by the presence of a shallow local minimum in the FESs immediately after the TS (Fig. 4(b)). This intermediate state, observed for both *endo* and *exo* reactions, corresponds to a zwitterionic state. Interestingly, such an intermediate is absent in the PES, where the structure corresponds to a high energy state (labelled as a cross in (Fig. 4(b)). This behaviour suggests that the intermediate arises from an entropic rather than an enthalpic contribution.

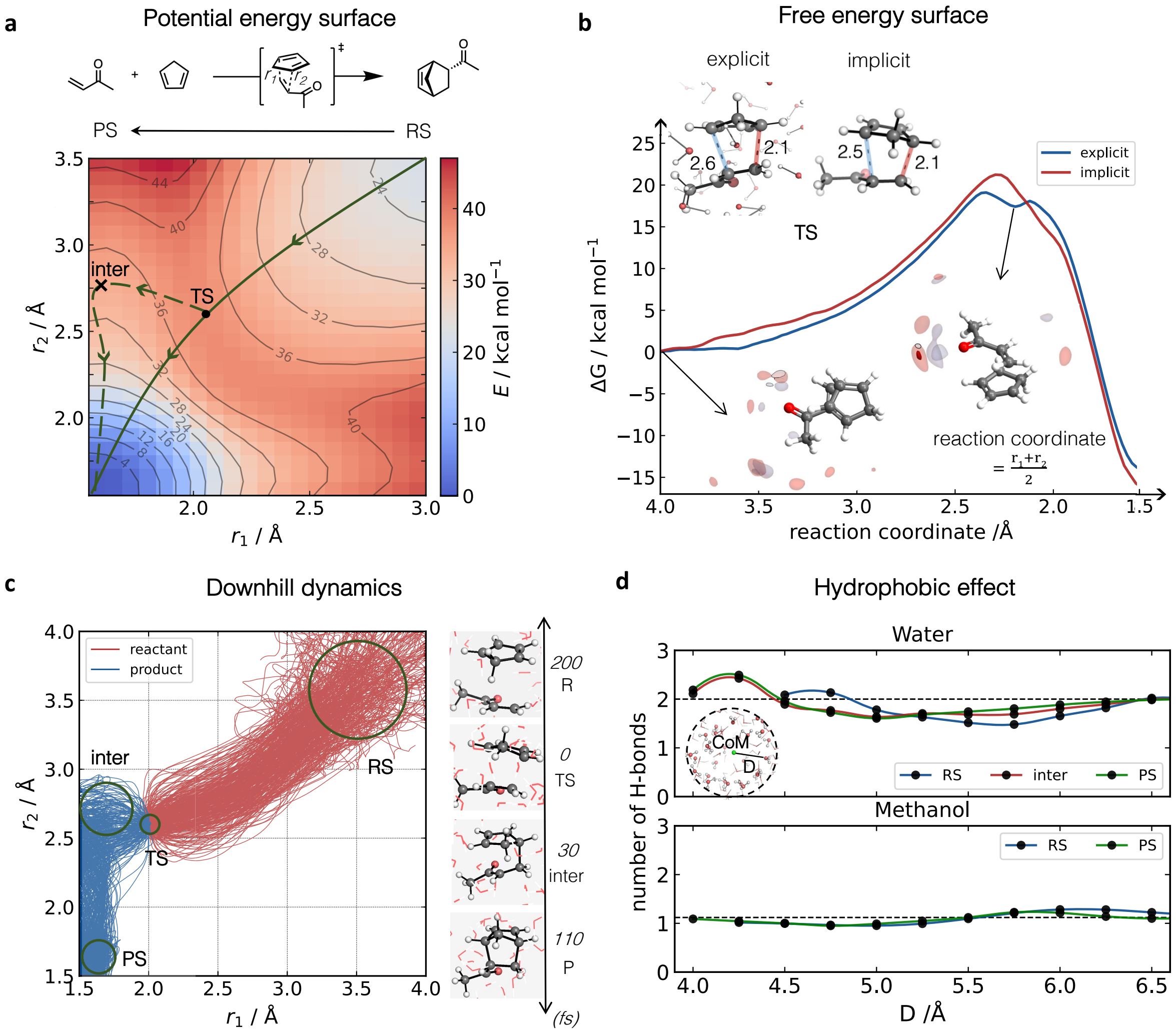

Figure 4: DA Reaction of CP+MVK. (a) 2D relaxed-PES scan along r_1 and r_2 generated by ACE MLP for the reaction of CP+MVK in explicit water (box size 18.5 Å). The solid and dashed lines indicate the reaction pathway obtained from the ACE MLP static scan and ACE MLP-MD, respectively. (b) FESs from ACE MLP-MD/US along the reaction coordinate. The density distribution of oxygen (red) and hydrogen (grey) atoms around the solute are shown for the reactants and intermediate. Representative TS geometries are also shown. (c) Downhill ACE MLP-MD dynamics in explicit water along r_1 and r_2 for 500 trajectories, including snapshots of forward/backward trajectories. (d) Number of HBs per solvent molecule from explicit solvent uphill trajectories at the reactant (RS), intermediate (if existent) and product (PS) states as a function of distance from the centre of mass of the reactive molecules.

The formation of an entropic intermediate has been previously reported by Singleton *et al.* for the reaction of cis-2-butene with dichloroketene, in which the free energy surface illustrates the entropic barrier and the mechanism change from concerted to stepwise.[61] In this study, we observed a similar phenomenon, highlighting the necessity of explicit solvent to capture the formation of an entropic intermediate.

The presence of this entropic intermediate is further confirmed by downhill dynamics initiated from the TS (Fig. 4(c) and Table S7). For *endo* reaction, the trajectories reveal a significantly more asynchronous reaction in explicit solvent compared to implicit solvent, in agreement with the ACE-MD/US data. The asynchronicity in the downhill dynamic is evident in the increased average time gap between the formation of the two C-C bonds, from 19.9 fs in an implicit water solvent to 84.3 fs in an explicit solvent. It is worth noting that although the average time gaps observed in water exceeded 60 fs (the time gap criterion proposed by Houk *et al.* to distinguish concerted and stepwise mechanisms[62]), some trajectories displayed time gaps below this threshold, indicating that not all trajectories passed through the intermediate region, and certain trajectories bypass the intermediate free energy well. The presence of this intermediate did not affect the product ratio, as the intermediate lifetime was shorter than the carbon-carbon bond rotation period, leaving no time to form alternative products by bond rotation.

A comparison of the average time gap for the reactions in explicit water and methanol reveals that the latter exhibits a more concerted mechanism, with a time gap of 24.8 fs. These distinct reaction mechanisms are in line with the differences in the synchronicity of this reaction in different solvents, where the TSs exhibit a Δr of 7 pm in methanol and 46 pm in water (listed in Table S6). Additionally, a shift in the stability of the intermediate species occurs in methanol (further discussed in SI § S9.1).

To assess the effect of solvent at the molecular level, we performed uphill trajectories, propagated from the optimised RS towards the TS to PS (further details in SI § S9.2). In contrast to downhill dynamics propagated from the TS, uphill dynamics allow the solvent sufficient time to reorganise before the trajectory passes the free energy barrier, providing a more realistic view of solvent behaviour during the reaction.

To determine the importance of HB stabilisation throughout the reaction, we analysed the number HBs as well as their bond lengths (O(carbonyl)-H(water)) and angles (O(carbonyl)-H(water)-O(water)) distributions at RS, TS and PS in explicit water and methanol (Tab. S9 and Fig. S22). We also analysed the density distribution of water molecules surrounding the reactive species at the RS and intermediate states obtained from the corresponding US windows (Fig. 4 (b)). The latter was chosen over the TS due to their similar geometry and the fact that only a few configurations representative of the TS were obtained. In both reactions, the number of HBs and the density of water's oxygen and hydrogen atoms around the substrate remained practically constant. However, the reaction in water exhibited stronger HB interactions compared to methanol during the reaction process, as evidenced by the existence of HBs with shorter bond lengths at TS in water and a larger decrease in the angle in methanol.

These observations challenge the hypothesis that the reaction is accelerated in water due to the stabilisation of the TS by enhanced HB interactions compared to the RS, as suggested in previous studies.[53,54,63] These differences can be attributed to the differing dynamics employed in the studies. For example, Houk *et al.* investigated the reaction in water using downhill dynamics from the TS and observed shorter HBs bond length and more linear bond angle at TS compared with RS and PS. In this approach, the TS with solvent molecules is fully optimised, and the solvent does not undergo complete reorganisation at RS or PS as the trajectory approaches the RS or PS at a faster rate than the reorganisation process, due to the absence of energy barriers to overcome.

The reactions in solution are further affected by hydrophobic effects,[59] which were investigated by analysing the change in cavity volume during uphill dynamics. The cavity volume was measured as empty space formed by the solvent when the solute was extracted from the system. Fig. S21 illustrates the reduction in cavity volume from the RS to the PS.

Reduction in the cavity is more pronounced in water with a change in cavity volume from RS to post-TS of -60 Å^3 compared to methanol (-40 Å^3). The decrease in cavity volume is entropically unfavourable as it implies the formation of a more ordered solvent structure. However, at the same time, it is enthalpically favourable due to the formation of HBs among solvent-solvent (rather than solvent-solute) molecules, resulting in a decrease in $\Delta G^{\ddagger}$. The importance of HB formation is supported by analysing the average number of HBs per solvent molecule at increasing distances from the solute. The distance was measured between the oxygen atom of the solvent and the centre of mass (CoM) of the RS, intermediate, and PS. Since the reaction passes the TS very quickly, the intermediate was chosen for analysis instead.

As the distance from the CoM of the substrate to solvent molecules increases, the number of HBs for each solvent converges to the average bulk value, with two HBs for water molecules and one for methanol. Fig. 4(d) displays a peak in the number of HB in the water close to the substrate, demonstrating the higher number of HBs in the first solvation shell for the RS (2.13), intermediate (2.43), and PS (2.49) compared to bulk water. Such an increase suggests the organisation of water molecules around the substrate during the reaction. In contrast, no such peak is observed in methanol, indicating that the solvent HB network is not influenced by the presence of substrate. The organisation of solvent molecules should lead to a reduction in free energy, as previously stated. Magsumov *et al.* demonstrated a linear relationship between the free energy of cavity formation and the volume of the cavity for various solvents through MD simulations utilising classical force fields.[64] Applying their parameters for water and methanol, the change in cavity contribution from RS to post-TS was approximately -2.6 kcal mol^{-1} for the *endo* reaction in water, compared to around -1.1 kcal mol^{-1} in methanol, further evidencing that the reactions benefit more from the hydrophobic effect in the water than in methanol. These observations thus further support the role of the hydrophobic effect in the acceleration of DA reaction in water.

Conclusion

In conclusion, we have presented a robust and efficient AL strategy that utilises descriptor-based selectors to train MLPs for condensed-phase reactions. We evaluated the impact of three different selectors: *energy*, *similarity*, and *distance* on data efficiency and MLP accuracy using bulk water as a model system. Our findings reveal that descriptor-based selectors outperform the classical *energy* selector in both efficiency and accuracy. This is evident from the smaller number of configurations required with comparable MAD in energies and forces compared with *energy* selector.

We applied this approach to investigate the influence of solvent molecules in the reaction between CP and MVK in water and methanol. The resulting MLPs enable us to efficiently obtain accurate PESs and FESs of these reactions. Consideration of explicit solvent significantly improves the agreement between the computed and experimental activation-free energies compared to implicit solvent models, providing detailed insights into the reaction mechanism.

Specifically, in the presence of explicit water molecules, we identify the formation of solvent-induced short-lived intermediates, revealing the intricate interplay between hydrogen bond interactions and hydrophobic effects in chemical reactions. The results emphasise the importance of incorporating explicit solvation in the modelling of chemical processes in order to provide a more realistic representation of experimental conditions.

By harnessing the speed and data efficiency of this framework, we demonstrate the feasibility of accurately modelling chemical processes in explicit solvents using MLPs. This approach paves the way for more thorough investigations of complex processes in solution, thereby contributing to a deeper understanding of the effects of solute-solvent interactions and entropy.

Computational Details

ACE Training ACE MLPs were trained using ACE.jl[65] wrapped by pyjulip in *mlp-train* package.[37] Unless specified otherwise, training for ACE MLPs uses hyperparameters listed in SI § S1. All ACE MLP-MD simulations were performed using the Atomic Simulation Environment (ASE) v.3.22.0.[66] package using a timestep of 0.5 fs. Initial velocities were sampled from a Maxwell-Boltzmann distribution. We used the following system-specific ground truth methods: PBE0-D3BJ/def2-SVP[67] for the water system, and ωB97M-D3BJ[68]/def2-TZVP level of theory for the Diels-Alder reaction. The PBE0-D3BJ method was selected as it provided good agreement with the bulk water structure.[69] The ground truth for the DA reaction was chosen based on a benchmark relative to SCS-MP2[70] (SI§S3). Implicit solvation was accounted for using the Conductor-like Polarizable Continuum Model (CPCM)[71] model. All QM computations were performed in ORCA v. 4.2.1[72] wrapped with autodE.[73]

Free energy computations The 1D free energy profiles in implicit and explicit solvents were computed by umbrella sampling (US). The reaction coordinate was defined as $(r_1 + r_2)/2$, where are the distance of C-C bonds formed during the reaction (Fig.4(a), upper part). The initial configurations were generated by the partial Nudged Elastic Band (pNEB)[74] method with 15 images. Solvent molecules were equilibrated before and during the NEB calculations. The ACE MLP-MD/US was performed with 30 windows for 10 ps per window at 300 K in the NVT ensemble. 15 windows were equally spaced in [1.55, 4.0] Å with $k = 10$ eV Å^{-2} and additional 15 equally spaced over [1.7, 2.5] Å with $k = 20$ eV Å^{-2}, except for the *exo* reaction in explicit water, for which $k = 13$ eV Å^{-2} and $k = 30$ eV Å^{-2} were applied for the wide- and narrow-range, respectively, to ensure overlap among the windows.

Trajectories with explicit solvents The dynamic properties of the DA reaction were investigated by downhill and uphill trajectories. The former refers to trajectories propagated from the TS, while the latter refers to ones generated by adding momentum to the reactants.

The initial configurations for the downhill trajectories (CP+MVK/ 200 water) were obtained by the approach inspired by the solvent-perturbed transition state (SPTS) sampling scheme proposed by Houk *et al.*[63] Solvent molecules surrounding the TS selected from umbrella sampling were equilibrated for 20 ps in the NVT ensemble with fixed TS. Initial configurations for the downhill trajectories were selected every 5 ps. The TSs were optimised using the Broyden–Fletcher–Goldfarb–Shanno (BFGS)[75–78] method, wrapped in ASE, with fixed solvent molecules and harmonic potentials to maintain the TS distance of the forming bonds. The trajectories were propagated until the products or reactants were formed within a maximum simulation time of 2 ps. Product formation was considered when both r_1 and r_2 distances were below 1.6 Å, whereas the reactant was reached when both forming bond distances exceeded 3.0 Å.

Uphill trajectory starting points were obtained by independent ACE-MD simulations of reactants at 300 K with r_1 and r_2 constrained by harmonic potentials at $3.0 \sim 5.0$ Å. The uphill trajectories were propagated with a harmonic potential (spring constant of 0.4 eV Å^{-2}) to overcome the energy barriers. Trajectories were propagated until products formed or reached the simulations time of 3 ps.

Conflicts of interest

There are no conflicts to declare.

Data availability

mlp-train package is available at https://github.com/duartegroup/mlp-train.

Author contributions

HZ and FD conceptualised the study. HZ carried out the calculations. All authors participated in data analyses and writing of the manuscript. HZ and VJ wrote the first draft. FD supervised the study.

Acknowledgements

The authors thank Dr T. Piskorz, Dr A. Sterling and Z. Zhu for insightful discussions. They also acknowledge H. Chan, Z. Bo, B. Lee, N. Frank, and V. Vitartas for their feedback on the manuscript. HZ thanks the EPSRC Centre for Doctoral Theory and Modelling in Chemical Sciences (EP/L015722/1). VJ acknowledges the funding from the Swiss National Science Foundation (SNSF, Postdoc.Mobility fellowship, grant no. 210737). This work used the University of Oxford Advanced Research Computing (ARC) facility and the Cirrus UK National Tier-2 HPC Service at EPCC (http://www.cirrus.ac.uk) funded by the University of Edinburgh and EPSRC (EP/P020267/1).

References

(1) Kitanosono, T.; Masuda, K.; Xu, P.; Kobayashi, S. Catalytic Organic Reactions in Water toward Sustainable Society. *Chem. Rev.* **2018**, *118*, 679–746.

(2) Mlynarski, J.; Baś, S. Catalytic asymmetric aldol reactions in aqueous media – a 5 year update. *Chem. Soc. Rev.* **2014**, *43*, 577–587.

(3) Berthelot, M.; de Saint-Gilles, P. Recherches Sur Les Affinites, Part 1. *Ann. Chim. Phys* **1862**, *65*, 385.

(4) Menschutkin, N. Über die Affinitätskoeffizienten der Alkylhaloide und der Amine. *Z. Phys. Chem.* **1890**, *6U*, 41–57.

(5) Orr-Ewing, A. J. Taking the plunge: chemical reaction dynamics in liquids. *Chem. Soc. Rev.* **2017**, *46*, 7597–7614.

(6) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions. *J. Phys. Chem. B* **2009**, *113*, 6378–6396.

(7) Duster, A. W.; Wang, C. H.; Garza, C. M.; Miller, D. E.; Lin, H. Adaptive quantum/molecular mechanics: what have we learned, where are we, and where do we go from here? *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2017**, *7*, 1–21.

(8) Watanabe, H. C.; Cui, Q. Quantitative Analysis of QM/MM Boundary Artifacts and Correction in Adaptive QM/MM Simulations. *J. Chem. Theory Comput.* **2019**, *15*, 3917–3928.

(9) Kocer, E.; Ko, T. W.; Behler, J. Neural Network Potentials: A Concise Overview of Methods. *Annu. Rev. Phys. Chem.* **2022**, *73*, 163–186.

(10) Deringer, V. L.; Caro, M. A.; Csányi, G. Machine Learning Interatomic Potentials as Emerging Tools for Materials Science. *Adv. Mater.* **2019**, *31*, 1902765.

(11) Wang, H.; Zhang, L.; Han, J.; E, W. DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics. *Comput. Phys. Commun.* **2018**, *228*, 178–184.

(12) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* **2017**, *8*, 3192–3203.

(13) Batzner, S.; Musaelian, A.; Sun, L.; Geiger, M.; Mailoa, J. P.; Kornbluth, M.; Molinari, N.; Smidt, T. E.; Kozinsky, B. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat. Commun.* **2022**, *13*, 2453.

(14) Unke, O. T.; Meuwly, M. PhysNet: A Neural Network for Predicting Energies, Forces, Dipole Moments, and Partial Charges. *J. Chem. Theory Comput.* **2019**, *15*, 3678–3693.

(15) Schütt, K. T.; Sauceda, H. E.; Kindermans, P. J.; Tkatchenko, A.; Müller, K. R. SchNet - A deep learning architecture for molecules and materials. *J. Chem. Phys.* **2018**, *148*, 241722.

(16) Batatia, I.; Kovács, D. P.; Simm, G.; Ortner, C.; Csányi, G. MACE: Higher Order Equivariant Message Passing Neural Networks for Fast and Accurate Force Fields. Advances in Neural Information Processing Systems. 2022; pp 11423–11436.

(17) Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons. *Phys. Rev. Lett.* **2010**, *104*, 136403.

(18) Chmiela, S.; Tkatchenko, A.; Sauceda, H. E.; Poltavsky, I.; Schütt, K. T.; Müller, K.-R. Machine learning of accurate energy-conserving molecular force fields. *Sci. Adv.* **2017**, *3*, e1603015.

(19) Chmiela, S.; Sauceda, H. E.; Müller, K. R.; Tkatchenko, A. Towards exact molecular dynamics simulations with machine-learned force fields. *Nat. Commun.* **2018**, *9*, 3887.

(20) Drautz, R. Atomic cluster expansion for accurate and transferable interatomic potentials. *Phys. Rev. B* **2019**, *99*, 014104.

(21) Kovács, D. P.; Oord, C. V. D.; Kucera, J.; Allen, A. E.; Cole, D. J.; Ortner, C.; Csányi, G. Linear Atomic Cluster Expansion Force Fields for Organic Molecules: Beyond RMSE. *J. Chem. Theory Comput.* **2021**, *17*, 7696–7711.

(22) Xue, L.-Y.; Guo, F.; Wen, Y.-S.; Feng, S.-Q.; Huang, X.-N.; Guo, L.; Li, H.-S.; Cui, S.-X.; Zhang, G.-Q.; Wang, Q.-L. ReaxFF-MPNN machine learning potential: a combination

of reactive force field and message passing neural networks. *Phys. Chem. Chem. Phys.* **2021**, *23*, 19457–19464.

(23) Deringer, V. L.; Bernstein, N.; Csányi, G.; Ben Mahmoud, C.; Ceriotti, M.; Wilson, M.; Drabold, D. A.; Elliott, S. R. Origins of structural and electronic transitions in disordered silicon. *Nature* **2021**, *589*, 59–64.

(24) Erhard, L. C.; Rohrer, J.; Albe, K.; Deringer, V. L. A machine-learned interatomic potential for silica and its relation to empirical models. *Npj Comput. Mater.* **2022**, *8*, 90.

(25) Brickel, S.; Das, A. K.; Unke, O. T.; Turan, H. T.; Meuwly, M. Reactive molecular dynamics for the [Cl-CH3-Br]a' reaction in the gas phase and in solution: A comparative study using empirical and neural network force fields. *Electron. Struct.* **2019**, *1*, 024002.

(26) Zeng, J.; Cao, L.; Xu, M.; Zhu, T.; Zhang, J. Z. Complex reaction processes in combustion unraveled by neural network-based molecular dynamics simulation. *Nat. Commun.* **2020**, *11*, 5713.

(27) Ang, S. J.; Wang, W.; Schwalbe-Koda, D.; Axelrod, S.; Gómez-Bombarelli, R. Active learning accelerates ab initio molecular dynamics on reactive energy surfaces. *Chem* **2021**, *7*, 738–751.

(28) Schaaf, L.; Fako, E.; De, S.; Schäfer, A.; Csányi, G. Accurate Reaction Barriers for Catalytic Pathways: An Automatic Training Protocol for Machine Learning Force Fields. *arXiv* **2023**, arXiv:2301.09931.

(29) Yang, M.; Bonati, L.; Polino, D.; Parrinello, M. Using metadynamics to build neural network potentials for reactive events: the case of urea decomposition in water. *Catal. Today* **2022**, *387*, 143–149.

(30) Yang, X.; Zou, J.; Wang, Y.; Xue, Y.; Yang, S. Role of Water in the Reaction Mechanism and endo/exo Selectivity of 1,3-Dipolar Cycloadditions Elucidated by Quantum Chemistry and Machine Learning. *Eur. J. Chem.* **2019**, *25*, 8289–8303.

(31) Mondal, A.; Kussainova, D.; Yue, S.; Panagiotopoulos, A. Z. Modeling Chemical Reactions in Alkali Carbonate-Hydroxide Electrolytes with Deep Learning Potentials. *J. Chem. Theory Comput.* **2022**, Article ASAP.

(32) Vassilev-Galindo, V.; Fonseca, G.; Poltavsky, I.; Tkatchenko, A. Challenges for machine learning force fields in reproducing potential energy surfaces of flexible molecules. *J. Chem. Phys.* **2021**, *154*, 094119.

(33) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Big Data Meets Quantum Chemistry Approximations: The Δ-Machine Learning Approach. *J. Chem. Theory Comput.* **2015**, *11*, 2087–2096.

(34) Rossi, K.; Jurásková, V.; Wischert, R.; Garel, L.; Corminbœuf, C.; Ceriotti, M. Simulating solvation and acidity in complex mixtures with first-principles accuracy: the case of CH3SO3H and H2O2 in phenol. *J. Chem. Theory. Comput.* **2020**, *16*, 5139–5149.

(35) Jurásková, V.; Célerse, F.; Laplaza, R.; Corminboeuf, C. Assessing the persistence of chalcogen bonds in solution with neural network potentials. *J. Chem. Phys.* **2022**, *156*, 154112.

(36) Young, T. A.; Johnston-Wood, T.; Deringer, V. L.; Duarte, F. A transferable active-learning strategy for reactive molecular force fields. *Chem. Sci.* **2021**, *12*, 10944–10955.

(37) Young, T. A.; Johnston-Wood, T.; Zhang, H.; Duarte, F. Reaction dynamics of Diels-Alder reactions from machine learned potentials. *Phys. Chem. Chem. Phys.* **2022**, *24*, 20820–20827.

(38) Sivaraman, G.; Krishnamoorthy, A. N.; Baur, M.; Holm, C.; Stan, M.; Csányi, G.; Benmore, C.; Álvaro Vázquez-Mayagoitia, Machine-learned interatomic potentials by active learning: amorphous and liquid hafnium dioxide. *Npj Comput. Mater.* **2020**, *6*, 104.

(39) Vandermause, J.; Xie, Y.; Lim, J. S.; Owen, C. J.; Kozinsky, B. Active learning of reactive Bayesian force fields applied to heterogeneous catalysis dynamics of H/Pt. *Nat. Commun.* **2022**, *13*, 5183.

(40) Xie, Y.; Vandermause, J.; Ramakers, S.; Protik, N. H.; Johansson, A.; Kozinsky, B. Uncertainty-aware molecular dynamics from Bayesian active learning for phase transformations and thermal transport in SiC. *Npj Comput. Mater.* **2023**, *9*, 36.

(41) Rasmussen, C.; Williams, C. *Gaussian Processes for Machine Learning*; Adaptive Computation and Machine Learning series; MIT Press, 2005.

(42) Schran, C.; Brezina, K.; Marsalek, O. Committee neural network potentials control generalization errors and enable active learning. *J. Chem. Phys.* **2020**, *153*, 104105.

(43) Zhang, L.; Lin, D. Y.; Wang, H.; Car, R.; Weinan, E. Active learning of uniformly accurate interatomic potentials for materials simulation. *Phys. Rev. Mater.* **2019**, *3*, 023804.

(44) Kulichenko, M.; Barros, K.; Lubbers, N.; Li, Y. W.; Messerly, R.; Tretiak, S.; Smith, J. S.; Nebgen, B. Uncertainty-driven dynamics for active learning of interatomic potentials. *Nat. Comput. Sci.* **2023**, *3*, 230–239.

(45) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Phys. Rev. B* **2013**, *87*, 184115.

(46) Zaverkin, V.; Holzmüller, D.; Schuldt, R.; Kästner, J. Predicting properties of periodic

systems from cluster data: A case study of liquid water. *J. Chem. Phys.* **2022**, *156*, 114103.

(47) Breunig, M. M.; Kriegel, H.-P.; Ng, R. T.; Sander, J. LOF: Identifying Density-Based Local Outliers. *SIGMOD Rec.* **2000**, *29*, 93–104.

(48) Marsalek, O.; Markland, T. E. Quantum Dynamics and Spectroscopy of Ab Initio Liquid Water: The Interplay of Nuclear and Electronic Quantum Effects. *J. Phys. Chem. Lett.* **2017**, *8*, 1545–1551.

(49) Skinner, L. B.; Benmore, C. J.; Neuefeind, J. C.; Parise, J. B. The structure of water around the compressibility minimum. *J. Chem. Phys.* **2014**, *141*, 214507.

(50) Rideout, D. C.; Breslow, R. Hydrophobic acceleration of Diels-Alder reactions. *J. Am. Chem. Soc.* **1980**, *102*, 7816–7817.

(51) Breslow, R.; Guo, T. Diels-Alder reactions in nonaqueous polar solvents. Kinetic effects of chaotropic and antichaotropic agents and of .beta.-cyclodextrin. *J. Am. Chem. Soc.* **1988**, *110*, 5613–5617.

(52) Lording, W. J.; Fallon, T.; Sherburn, M. S.; Paddon-Row, M. N. The simplest Diels–Alder reactions are not endo-selective. *Chem. Sci.* **2020**, *11*, 11915–11926.

(53) Liu, F.; Liang, Y.; Houk, K. N. Bioorthogonal Cycloadditions: Computational Analysis with the Distortion/Interaction Model and Predictions of Reactivities. *Acc. Chem. Res.* **2017**, *50*, 2297–2308.

(54) Acevedo, O.; Jorgensen, W. L. Understanding Rate Accelerations for Diels-Alder Reactions in Solution Using Enhanced QM/MM Methodology. *J. Chem. Theory Comput.* **2007**, *3*, 1412–1419.

(55) Jorgensen, W. L.; Lim, D.; Blake, J. F. Ab initio study of Diels-Alder reactions of

cyclopentadiene with ethylene, isoprene, cyclopentadiene, acrylonitrile, and methyl vinyl ketone. *J. Am. Chem. Soc.* **1993**, *115*, 2936–2942.

(56) Soto-Delgado, J.; Tapia, R. A.; Torras, J. Multiscale Treatment for the Molecular Mechanism of a Diels–Alder Reaction in Solution: A QM/MM-MD Study. *J. Chem. Theory Comput.* **2016**, *12*, 4735–4742.

(57) Li, P.; Liu, F.; Shao, Y.; Mei, Y. Computational Insights into Endo/Exo Selectivity of the Diels–Alder Reaction in Explicit Solvent at Ab Initio Quantum Mechanical/Molecular Mechanical Level. *J. Phys. Chem. B* **2019**, *123*, 5131–5138.

(58) Chandrasekhar, J.; Shariffskul, S.; Jorgensen, W. L. QM/MM Simulations for Diels-Alder Reactions in Water: Contribution of Enhanced Hydrogen Bonding at the Transition State to the Solvent Effect. *J. Phys. Chem. B* **2002**, *106*, 8078–8085.

(59) Meijer, A.; Otto, S.; Engberts, J. B. F. N. *J. Org. Chem.* **1998**, *63*, 8989–8994.

(60) Furlani, T. R.; Gao, J. Hydrophobic and Hydrogen-Bonding Effects on the Rate of Diels-Alder Reactions in Aqueous Solution. *J. Org. Chem.* **1996**, *61*, 5492–5497.

(61) Gonzalez-James, O. M.; Kwan, E. E.; Singleton, D. A. Entropic Intermediates and Hidden Rate-Limiting Steps in Seemingly Concerted Cycloadditions. Observation, Prediction, and Origin of an Isotope Effect on Recrossing. *J. Am. Chem. Soc.* **2012**, *134*, 1914–1917.

(62) Black, K.; Liu, P.; Xu, L.; Doubleday, C.; Houk, K. N. Dynamics, transition states, and timing of bond formation in Diels-Alder reactions. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 12860–12865.

(63) Yang, Z.; Doubleday, C.; Houk, K. N. QM/MM Protocol for Direct Molecular Dynamics of Chemical Reactions in Solution: The Water-Accelerated Diels–Alder Reaction. *J. Chem. Theory Comput.* **2015**, *11*, 5606–5612.

(64) Sedov, I.; Magsumov, T. The Gibbs free energy of cavity formation in a diverse set of solvents. *J. Chem. Phys.* **2020**, *153*, 134501.

(65) Ortner, C.; Zhang, L.; Ross, A.; Sachs, M.; van der Oord, C. ACE.jl. `https://github.com/ACEsuit/ACE.jl`.

(66) Larsen, A. H. et al. The atomic simulation environment—a Python library for working with atoms. *J. Phys. Condens. Matter* **2017**, *29*, 273002.

(67) Adamo, C.; Barone, V. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys.* **1999**, *110*, 6158–6170.

(68) Najibi, A.; Goerigk, L. The Nonlocal Kernel in van der Waals Density Functionals as an Additive Correction: An Extensive Analysis with Special Emphasis on the B97M-V and ωB97M-V Approaches. *J. Chem. Theory Comput.* **2018**, *14*, 5725–5738.

(69) Zhang, C.; Donadio, D.; Gygi, F.; Galli, G. First Principles Simulations of the Infrared Spectrum of Liquid Water Using Hybrid Density Functionals. *J. Chem. Theory Comput.* **2011**, *7*, 1443–1449.

(70) Fink, R. F. Spin-component-scaled Møller–Plesset (SCS-MP) perturbation theory: A generalization of the MP approach with improved properties. *J. Chem. Phys.* **2010**, *133*, 174113.

(71) Barone, V.; Cossi, M. Quantum Calculation of Molecular Energies and Energy Gradients in Solution by a Conductor Solvent Model. *J. Phys. Chem. A* **1998**, *102*, 1995–2001.

(72) Neese, F. Software update: the ORCA program system, version 4.0. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2018**, *8*, e1327.

(73) Young, T. A.; Silcock, J. J.; Sterling, A. J.; Duarte, F. autodE: Automated Calculation of Reaction Energy Profiles – Application to Organic and Organometallic Reactions. *Angew. Chem.* **2021**, *60*, 4266–4274.

(74) Bergonzo, C.; Campbell, A. J.; Walker, R. C.; Simmerling, C. A partial nudged elastic band implementation for use with large or explicitly solvated systems. *Int. J. Quantum Chem.* **2009**, *109*, 3781–3790.

(75) Broyden, C. G. The Convergence of a Class of Double-rank Minimization Algorithms 1. General Considerations. *IMA J. Appl. Math.* **1970**, *6*, 76–90.

(76) Fletcher, R. A new approach to variable metric algorithms. *Comput. J.* **1970**, *13*, 317–322.

(77) Goldfarb, D. A Family of Variable-Metric Methods Derived by Variational Means. *Math. Comp.* **1970**, *24*, 23–26.

(78) Shanno, D. F. Conditioning of Quasi-Newton Methods for Function Minimization. *Math. Comp.* **1970**, *24*, 647–656.