A transferable active-learning strategy for reactive molecular force fields

Tom A. Young,[a] Tristan Johnston-Wood,[a] Volker L. Deringer,*[b] Fernanda Duarte*[a]

Predictive molecular simulations require fast, accurate and reactive interatomic potentials. Machine learning offers a promising approach to construct such potentials by fitting energies and forces to high-level quantum-mechanical data, but doing so typically requires considerable human intervention and data volume. Here we show that, by leveraging hierarchical and active learning, accurate Gaussian Approximation Potential (GAP) models can be developed for diverse chemical systems in an autonomous way, requiring only hundreds to a few thousand energy and gradient evaluations on a reference potential-energy surface. Our approach relies on decomposing the molecular system into intra- and inter-molecular components which are fit separatley. We also introduce a prospective error metric to quantify accuracy. We demonstrate applications to a range of molecular systems: from bulk water, organic solvents, a solvated metal ion and a metallocage onwards to chemical reactivity, including a bifurcating Diels–Alder reaction in the gas phase and non-equilibrium dynamics (S_N2 reaction) in explicit solvent. The method provides a route to routinely generating machine-learned force fields for reactive molecular systems.

Introduction

Molecular simulations are a cornerstone in computational chemistry, providing dynamical insights beyond experimental resolution.[1] Realistic simulation of (bio)chemical reactions require the inclusion of the chemical environment where they take place (e.g. solvent and/or enzyme) and often extended timescales. Therefore, the development of accurate and efficient approaches has been central to the development of this field.

Empirical interatomic potentials (force fields), in combination with molecular dynamics (MD) or Monte Carlo (MC) simulations, have been traditionally used to sample the potential-energy surface (PES). However, they are limited in accuracy and transferability.[2] Moreover, most of these potentials are parameterised for isolated entities with fixed connectivity and thus unable to describe bond breaking/forming processes. In contrast, *ab initio* methods provide an accurate description of the PES, which is particularly critical for reactions in solution. However, because of their high computational cost and unfavourable scaling behaviour, they are limited to a few hundred atoms and simulation times of picoseconds in *ab initio* molecular dynamics (AIMD) at the DFT level, and practically impossible at the computational 'gold-standard' [CCSD(T)].[3]

Machine learning (ML) approaches have the potential to revolutionise force-field based simulations, aiming to provide the best of both worlds,[4–6] and have indeed begun to provide new insights into a range of challenging research problems.[7–15] The development of an ML potential applicable to the whole periodic table mapping nuclear coordinates to total energies and forces is, however, precluded by the curse of dimensionality. Within small chemical subspaces, models can be achieved using neural networks (NNs),[6,16–20] kernel-based methods such as the Gaussian Approximation Potential (GAP) framework[21,22] or gradient-domain machine learning (GDML),[23] and linear fitting with properly chosen basis functions,[24,25] each with different data requirements and transferability.[26] GAPs have been used to study a range of elemental,[27–29] multicomponent inorganic,[30,31] gas-phase organic molecules,[14,32] and more recently condensed-phase molecular systems such as methane[33] and phosphorus.[34] These potentials, while accurate, have required considerable computational effort and human oversight. Indeed, condensed-phase NN[35,36] and GAP fitting approaches typically require several thousand reference ("ground truth") evaluations.

Active learning (AL), where new training data is added based on the current state of the potential, has been used for generating databases and accelerating the fitting process.[30,37–41] Notable examples in materials modelling include an early demonstration of a "query-by-committee" approach in fitting a high-dimensional NN potential for elemental copper,[38] the fitting of Moment Tensor Potential[25] models[42] to predict elemental crystal structures[37] and multicomponent alloys,[39] and the deep potential generator (DP-GEN)[43,44] that provides an interface to deep NN potential models for materials.[45] AL schemes have also been combined with GP based force fields including GAP,[46] and included within a first-principles MD implementation such that it allows the "on the fly" fitting of force fields for a specific simulation system.[47,48]

Efficient approaches to generate reactive ML potentials become even more important when exploring chemical reactions in molecular systems, which often require a description at a computational level beyond DFT, and therefore require reference data at the same level. Very recently, AL approaches have started to be adopted for fitting reactive potentials for organic molecules based on single point evaluations at quantum-chemical levels of theory. Notable examples include the modelling of gas-phase pericyclic reactions,[13] the exploration of reactivity during methane combustion,[49] and the decomposition of urea in water.[40]

In the present work – with a view to developing potentials to simulate solution phase reactions – we consider bulk water as a test case and develop a strategy capable of accurately describing the structure of liquid water. This strategy requires just hundreds of *total* ground truth evaluations and no *a priori* knowledge of the system, apart from the molecular composition. We then show how this methodology can be directly transferred to other chemical systems in the gas phase, and in implicit and explicit solvent, focusing on the applicability to a range of scenarios that are relevant in computational chemistry.

Results and Discussion

Despite GAP fitting being increasingly used for inorganic systems, we found that the same fitting strategies did not easily transfer to the description of complex molecular environments. Even with a high correlation and low error on energies in unseen test data, some potentials were not stable for more than a few femtoseconds. In the following section, we therefore outline a training strategy along with a prospective error metric to develop robust models for gas-phase and condensed-phase molecular systems.

A prospective error metric

The initial step in validating supervised machine learning (ML) tends to follow the splitting of a dataset into training and test sets, training the model, then evaluating its performance on the test set with a squared error (RMSE/MSE) or a correlation (R^2) metric. As with model overfitting, this 'retrospective' validation strategy ultimately limits the applicability of these models.[50–56] In an ML potential, the minimum required domain of applicability is the region of configuration space likely to be sampled during a simulation with the potential. However, this region is not known *a priori*, making the choice of test data problematic if not impossible for use in a standard train/test data split approach. In addition, one would also like to ensure high accuracy in regions sampled on the ground truth surface (especially for early versions of an evolving potential), but being able to quantify this accuracy requires dynamics at the ground truth method level in the first place, which is much more expensive than sampling with an efficient potential.

Using a train/test set split with high structural similarity between the two sets can lead to highly misleadingly accuracy whenever the potential is to be taken outside the training region in computational practice. For example, splitting an AIMD trajectory of water into a training and test set with an odd/even frame split (50:50) and training a simple GAP model yields an energy error on the order of 1 kcal mol^{-1} (**Figure S1a**). However, simulations with this potential in the same configuration space sample unphysical configurations within 10 fs (**Figure S1b**), making an RMSE over *a priori* test data an insufficient metric in quantifying the quality of a potential.

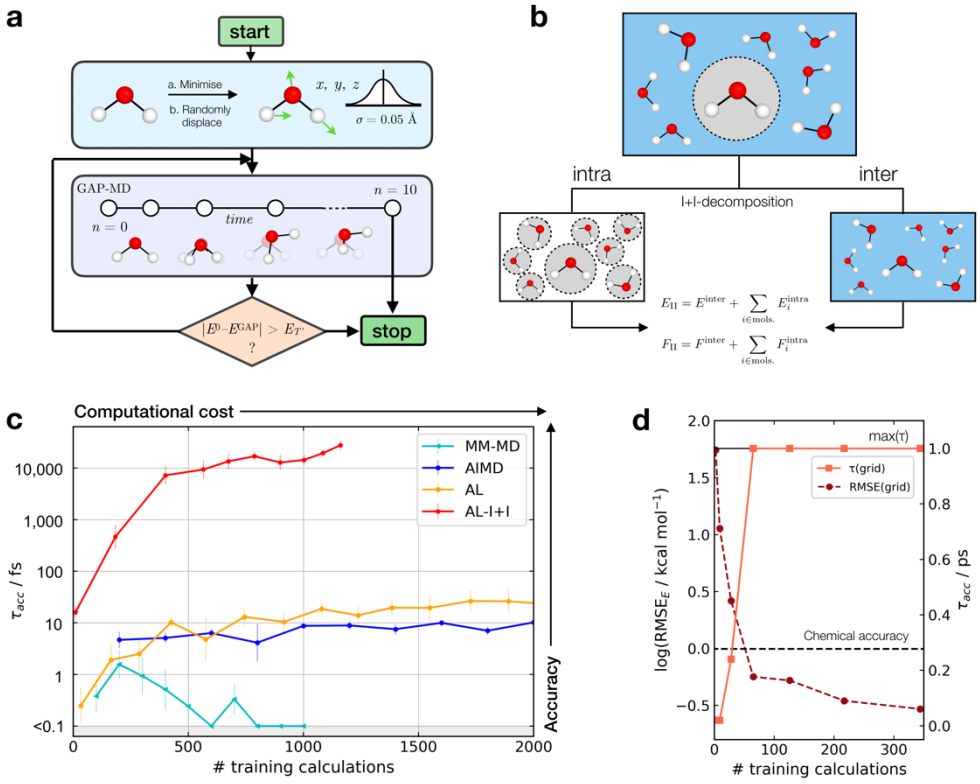

Figure 1: Active learning of machine-learning potentials for liquid water. (a) Schematic of the active learning loop implemented for fitting GAP models, where the GAP-MD exploration is run for n^3 +2 femtoseconds, where n (the number of evaluations) is incremented after each time the error is evaluated. (b) Schematic illustrating the separation into inter- and intra-molecular terms (I+I) for a bulk water system; these are described by separate GAP models and then added to give the combined prediction for energies, E, and forces, F. (c) Learning curves for a bulk water GAP model using different training strategies. τ_{acc} with E_I = 0.1 eV, E_T = 1 eV, 10 fs interval, 300 K, from the same random minimised configuration of 10 waters in a 7 Å cubic box. Error bars quoted as the standard errors in the mean from 5 independent repeats. The horizontal axis denotes the number of evaluations in training data generation. See Tables S1-2 for detailed methods. DFTB(3ob) ground truth. Minimum τ_{acc} is shown as 0.1 fs to enable plotting on a log scale. (d) Water monomer model training performance as characterised by τ_{acc} and RMSE over the full 3D PES; see **Figure S2** for details

Considering that single-point reference energy evaluations are reasonably cheap, a 'prospective' validation scheme is possible, where the error metric operates in the configuration space sampled in a simulation. With this in mind, we propose a temporal cumulative error metric (τ_{acc}, Eqn. 1), defined as the time required for the cumulative error (absolute difference between true (E^0) and predicted (E^{GAP})) to exceed a given threshold (E_T); the larger τ_{acc}, the more robust the potential. Note that a potential's stability can far exceed τ_{acc}, as shown in the following. Here only errors above a lower-bound threshold value (E_l) contribute to the cumulative error. The lower threshold is required to account for the residual error that is due to the finite radial cut-off of the model. In the following we take E_T to be 10 times E_l, but it may be adjusted depending on the simulation context.

$$\tau_{\mathrm{acc}} = \mathrm{time} \quad : \quad E_T < \sum_{i \in \mathrm{frames}} \max(|E_i^0 - E_i^{\mathrm{GAP}}| - E_l, 0) \qquad (1)$$

This metric has several advantages in that (a) it ensures that a potential with high accuracy will result in stable dynamics; (b) it allows the user to specify the level of accepted error according to the quality of the training method, thus not penalising where the error is within the difference between the ground truth and the true PES (i.e. a larger threshold may be suitable for a less accurate reference method); (c) it penalises large errors, even if they only occur for single configurations, which is important as such errors may lead to instabilities in the ML-driven MD trajectory and (d) it enables a quoted accuracy to include regions that may not be accessible to the direct evaluation at the ground-truth level (e.g. long-time behaviour). Overall, this metric depends on the lower bound and total error, interval between evaluations, and the simulation on which it is evaluated; so while not unique, it is – crucially – prospective. We found this metric to be essential in developing an efficient training strategy and accurate potentials for bulk water (**Figure 1**).

Water models

For bespoke ML potentials to be routinely developed for molecular systems, one would hope to complete the data generation, model training, and know the accuracy of the resulting potential within a matter of hours to days. With this in mind, here we train GAP models to simulate bulk water, aiming to minimise the number of required ground truth evaluations as well as the required human intervention, while maximising stability (measured by τ_{acc}). A selection of training strategies is discussed in the following paragraphs and their results are outlined in **Figure 1**.

We initially employed training strategies found to work well in elemental materials by, for example, fitting a combined potential with two- and three-body GAPs; this approach was, however, found to be detrimental to the potential's stability. This can be understood considering a water dimer (HO–H_c···OH_2); here, a two-body description that treats the two O–H_c interactions on the same footing is a poor approximation, in view of the different order of magnitude between the interactions at their respective minima (**Figure S3**). Therefore, the following GAPs only employ a smooth overlap of atomic positions[57] (SOAP) descriptor for an exclusively many–body description of atomic environments (except for AL-I+I where a different approach is used, see below).

Similarly, an emerging approach in generating training data for elemental GAPs is to initialise the database with randomised configurations (with reasonable constraints, as in *ab initio* random structure searching[58]), and to gradually explore configuration space with evolving versions of the potential (see, e.g., ref. 59). However, a similar approach of randomly placing water molecules does not in itself afford a stable potential. A similar result is observed when the most diverse configurations are selected using the CUR algorithm[60,59] (**Figure S4**) or when applying intramolecular displacements, following minimisation (**Figure S4**). Selecting frames from classical MD simulations at temperatures of 100–1000 K was also found to be an ineffective strategy (**Figure S5**), reaching τ_{acc} of only a few fs ("MM-MD", **Figure 1c**). This is in line with the results reported in ref. 14. Note that this is not because the GAP cannot fit reference energies and forces from MM configurations (**Figure S6**), but because of a poor configuration space overlap with the ground truth PES (**Figure S7**). Selecting configurations from an AIMD simulation at 300 K (AIMD, **Figure 1c**) was an improvement over training on random and MM-generated configurations, with τ_{acc} ~10 fs. Adding AIMD configurations, however, saturates in accuracy quickly even if those are obtained at higher temperatures (**Figure S8**). Using AIMD configurations can also involve a significant cost ($\gg 10^3$ evaluations). Finally, active learning from only a few randomly generated configurations provides a modest uplift in accuracy (AL, **Figure 1c**), with accuracy on-par with GAP trained on AIMD configurations at a third of the required reference data.

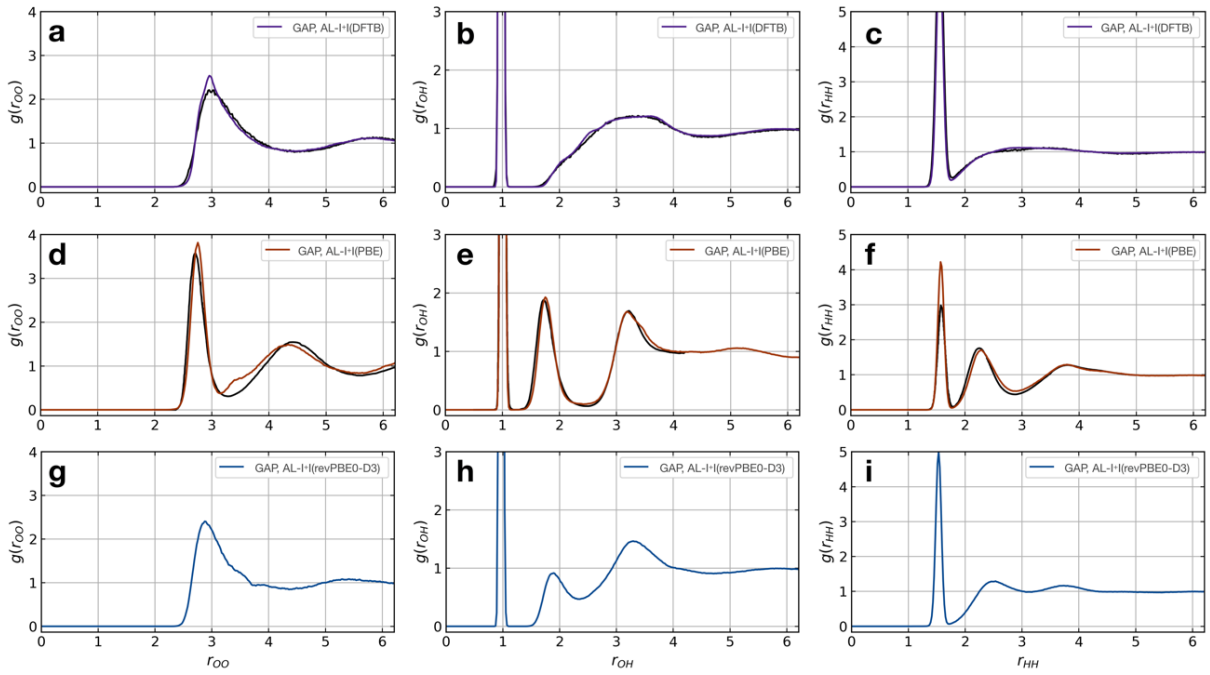

Figure 2: Liquid water simulations. Active learning of bulk water models at various levels of theory. Shown here are oxygen–oxygen, oxygen–hydrogen, and hydrogen–hydrogen radial distribution functions from NVT MD simulations of 64 water molecules in a 12.42 Å cubic box, with ground truth (black) and GAP (purple / red) simulations. (a–c) DFTB(3ob params) ground truth, 100 ps, 300 K, $r_c^{SOAP}(O)$ = 3 Å. (d–f) DFT(PBE) reference RDF data extracted from ref. [61], 30 ps, 330 K, $r_c^{SOAP}(O)$ = 3.5 Å. (g-i) DFT(revPBE0-D3) GAP, 30 ps, 330 K, $r_c^{SOAP}(O)$ = 4.0 Å.

Only when the relevant length and energy scales of the system are decomposed by treating intra- and inter-molecular components separately (**Figure 1b**) a potential that is stable for picoseconds is obtained (AL-I+I, **Figure 1c**). We note that this approach is related to but different from the hierarchical fitting of GAPs[33,62] and related ML models[32,63] using different levels of computational approaches, and the decomposition strategy by Wengert and co-workers.[64] Here, we employ the same level of input data throughout, but describe the stronger (e.g., covalent) and weaker intermolecular terms with separate fits that are afterwards combined to give the final model. The intramolecular GAP for water contains only 2- and 3-body terms and the training data are chosen using an evenly spaced grid over the full 3-atom PES (8×8×8 grid points in r_{OH} and r_{HH}, ~0.1 Å spacing, **Figure S9**). Energy and force evaluations of this potential are a simple sum of intra- and inter-molecular terms, but require the former to be evaluated in an expanded simulation box to ensure no non-bonded hydrogen atoms are present within the radius of the 2- and 3-body descriptors on oxygen (**Figure S10**). Here the intramolecular PES is fairly low-dimensional, so a full and reasonably dense grid is available, which in turn allows us to define an error measure over the whole PES, where we find the error to be inversely correlated with τ_{acc} (**Figure 1d**). Using an acceptable error of 0.2 kcal mol^{-1} per H_2O molecule for a description of bulk water, which is similar to that achieved in a recent NN fit of water,[35] we find that this potential affords τ_{acc} > 10 ps with just a few hundred ground truth evaluations (AL-I+I, **Figure 1c**). This accuracy is on-par with a bespoke NN[35] (τ_{acc} = 7.6±0.7 ps) when trained on DFTB energy and forces from a partially manually curated dataset of ~7000 training configurations. For comparison, an ANI[18] potential trained using ~6000 $(H_2O)_5$ clusters from the ANI-1x dataset gives τ_{acc} of 60±9 fs, implying that with these training data ANI does not afford an accurate bulk water model; somewhat expected considering the training data does not include any periodic configurations.

The model fitted using our approach (AL-I+I) yields a good description of the ground-truth radial distribution functions (RDFs), both considering the location and intensities of the peaks corresponding to the first and second coordination shells, with reference to the respective ground-truth method (**Figure 2a-c**). This is despite the relatively short-range atomic cut-offs (3 Å, O only) used. Interestingly, for a DFT-quality GAP simply re-evaluating energies and forces on DFTB-derived active-learnt configurations is insufficient, with the DFTB configurations being high in energy at the DFT level (~5 eV, **Figure S11**). However, applying an active learning strategy with a PBE reference method and a slightly larger 3.5 Å cut-off generates excellent agreement with the AIMD simulation from ref. [61], in only a few hours of total training time (**Figure 2d-f**). The real significance, of course, is in moving to more accurate ground-truth methods, for which a full MD would not be straightforward: indeed, using the same method, a hybrid DFT-quality water model can be generated within a few days, which would be inaccessible with other methods (the generation of the GAP model required ~5 days on 20 CPU cores, **Figure 2g-i**). These results suggest that the training strategy (and hyperparameter

selection) presented here is suitable independent of the reference method.

Other solvent systems

Organic reactions often take place in solvents other than water. Using an identical training strategy to the one described for water, we therefore trained GAPs for a selection of organic solvents with various types of intermolecular interactions. To quickly generate the reference simulation data for this proof-of-concept, a DFTB ground truth is employed; chlorinated solvents were not selected due to a large discrepancy between the DFTB-generated and experimental C–Cl bond dissociation energy (**Figure S12**). A uniform grid over the intramolecular PES is now no longer possible, thus AL is used to develop an initial intramolecular potential trained using GAP-MD at 1600 K (**Figure S13**). This temperature is used to sample higher-energy configurations more efficiently. In all cases, only hundreds of ground truth evaluations were necessary to generate GAPs affording stable dynamics, with τ_{acc} values on the order of picoseconds (**Figure S14**, **Table 1**, SI §**S1**). For a representative example, the computed RDFs for acetonitrile compare well with the ground truth (**Figure S15**). As with the water models above, to quantitatively evaluate bulk properties training an accurate reference method and the inclusion of nuclear quantum effects would be necessary.[33,35] Nevertheless, this example demonstrates that the training method is applicable to a range of chemical systems beyond water.

Table 1. Average number (N) of *total* ground truth evaluations (over 5 repeats quoted with a standard error in the mean) required to obtain a potential with τ_{acc} > 3 ps, where E_T = 1 eV, E_I = 0.1 eV, 300 K. All SOAP descriptors used 3.0 Å cut–offs; they are centred on the stated atomic species, and include all atoms within the neighbourhood of those atoms (including hydrogen). See **Table S3** for more detailed parameters.

Solvent	SOAP descriptors centred on	N_{intra}	N_{inter}
Acetonitrile	C, N	269±12	120±60
Methanol	C, O	221±13	292±49
Acetone	C, O	566±80	359±29
Pyridine	C, N	249±36	243±11
Ammonia	N	38±40	109±24

Aqueous Zn(II)

Modelling metal ions in solution remains one of the main challenges for general purpose force fields.[65] Historically, they have been developed by fitting van der Waals parameters to reproduce RDFs and hydration free energy of aquo complexes, which are expected to be transferrable to more chemically complex environments. However, while simple, these models have been often found to lead to unstable simulations or poorly describe structural properties.[65] Considering these challenges and their relevance in biomolecular modelling, we decided to use our strategy to generate a GAP for aqueous Zn(II) ion as a representative system. Here the system was decomposed into a $[Zn(H_2O)_6]^{2+}$ cluster and the remaining water molecules. A strategy identical to the one described for water was used, training the intermolecular interactions separately with a 4.0 Å intermolecular cut-off for the oxygen atoms. Using this potential, MD simulations were propagated at 300 K reproducing the experimental[66] coordination number (CN=6), and Zn–O distances of both the first (2.08 Å) and second hydration shells without further optimisation (**Figure 3**). From random points in the configuration space of $[Zn(H_2O)_6]^{2+}$ and 20 water molecules (intermolecular distances > 1.7 Å, 10 Å cubic box), τ_{acc} reached 0.5 ps (E_I = 0.8 kcal mol^{-1} per H_2O, 20 fs interval). Note this value is far short of the 100 ps simulations performed to generate the RDF and illustrates that a potential may be 'stable' and not sample any high energy regions for $t \gg \tau_{acc}$. Here, the force field for the Zn-water cluster was trained on almost 1000 configurations, suggesting that ~20 atom components may be the upper limit in component dimensionality for which a model can be trained within a day.

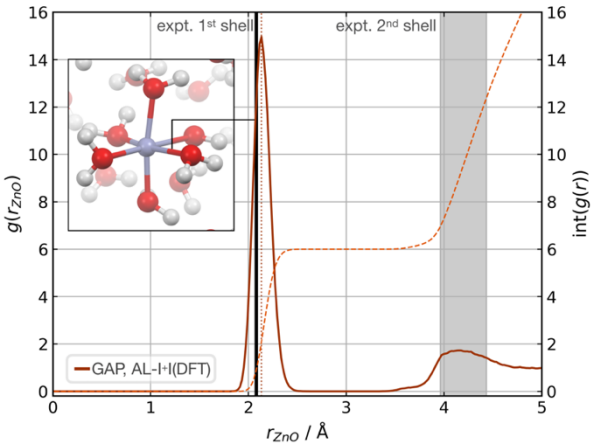

Figure 3: $Zn_{(aq)}$ simulation. Zn–O radial distribution function averaged from 1 ns of cumulative (10 × 100 ps) NVT MD simulations of Zn(II) in aqueous solution at 300 K, with the experimental modal Zn–O distance shown in black. Experimental (X-ray diffraction) Zn–O distances from ref. 66, octahedral first hydration shell. Shaded area denotes the range of experimental second hydration shell (ref. [66] and cited within). GAP trained as those in **Table 1** using a PBE/400 eV ground truth, intra-$Zn(H_2O)_6$ used a O SOAP r_c = 3.0 Å and inter r_c = 4.0 Å.

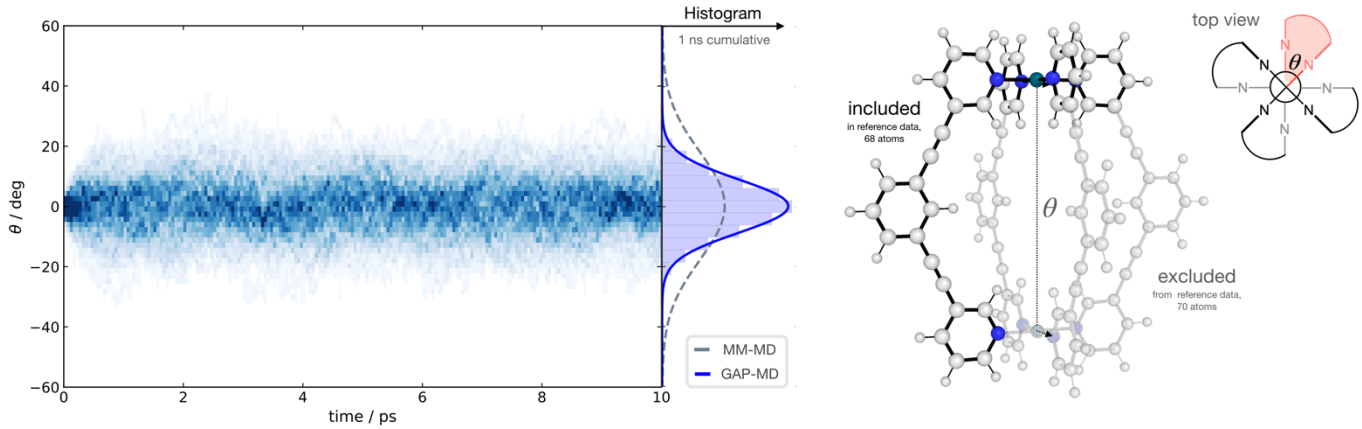

Figure 4: Metallocage dynamics. Temporal twist angle θ for an $[Pd_2L_4]^{2+}$ metallocage (right) obtained from 100 independent GAP-MD trajectories (each run for 10 ps at 300K), GAP trained on a 68-atom $[PdL(py)_3]^{2+}$ system (py = pyridine), representative of the whole metallocage as shown, at 500 K with a 0.2 eV error threshold for the active-learning protocol. Time-dependent histogram generated from 50 fs time chunks over the whole 10 ps time period. PBE0-D3BJ/def2-SV(P) ground truth surface.

Metallocage dynamics

With a method capable of generating high-quality potentials for modestly sized chemical systems, we next demonstrate the applicability of the strategy to investigate a supramolecular metallocage consisting of >100 atoms including metal ions. As a representative example, we selected the $[Pd_2L_4]^{4+}$ metallocage architecture (L = organic pyridine-based ligand), which occupies a prominent place in supramolecular chemistry. Previously, we have studied this system due to its catalytic proficiency in Diels-Alder reactions employing both classical and DFT modelling.[67] The different flexibility of two similar cage architectures was found to be key in explaining their contrasting catalytic activity. Taking advantage of the symmetry in the system, a representative fragment containing one full ligand and three pyridine molecules coordinated to a Pd^{2+} metal ion (68 atoms) was used to fit a GAP for the entire cage (138 atoms) in the gas phase. This potential was trained in a few days (~1400 CPUh). We used the resulting GAP to perform nanosecond MD simulations on the whole metallocage at 300 K in the gas phase. This simulation took 1 day and ~100 CPUh to complete. For comparison, an equivalent AIMD simulation would take around 50 years with the reference level of theory employed here. The flexibility of the system was monitored and compared to the one obtained using classical MD simulations in dichloromethane solvent. Compared to classical MD simulations, using helicity as a measurement of flexibility, our potential describes the cage as being more rigid; this suggests that the classical potential overestimates the dynamic flexibility (**Figure 4**). This difference is expected as the MM has no C-C≡C-C dihedral barrier, which is presumably correctly captured in the GAP. This example illustrates the general applicability of the approach to increasingly complex systems, where the training of a simpler but representative fragment is sufficient to capture the relevant features of the full system.

Reaction dynamics in gas and solvent phase

The high dimensionality and ensuing flexibility of machine-learned force fields makes them highly suitable in principle to study reaction dynamics – the latter usually require many costly electronic structure calculations to obtain atomic level descriptions of reaction mechanisms, solvent effects or post-transition state (TS) dynamics.[68,69] In the following section, we show that our data-efficient strategy enables accurate reactive potentials (τ_{acc} > 100 fs) with only a few hundreds of DFT evaluations for a set of prototypical organic reactions.

Gas phase bimolecular nucleophilic substitution

The S_N2 nucleophilic substitution reaction is fundamental in organic chemistry and has been extensively studied using AIMD and analytically fit PES.[68,70] However, even with efficient approaches to fitting PES, AIMD methods still require tens of thousands of energy evaluations.[71] Here, we generated a reactive GAP to study the reaction between chloride and methyl chloride. By initialising active learning from the transition state (TS), the true intrinsic reaction coordinate is reproduced to within 1 kcal mol^{-1} (**Figure S16**). Interestingly, and unlike our previous attempt to generate a DFT-quality GAP by evaluating energies and forces on DFTB active-learnt configurations, here uplifting a DFT-level GAP to an accurate wavefunction-level GAP is possible. This method allows coupled cluster-quality energy profile (**Figure 5a**) and dynamics to be propagated from the TS with just 55 energy and (numerical) force evaluations at the CCSD(T) level (**Figure 5b/S16**). The resultant GAP is considerably more accurate than the underlying DFT energy profile (dashed, **Figure 5a**).

Post-TS bifurcating pathway in a Diels-Alder reaction

GAPs for more complex reactions involving reactions that proceed on a bifurcating PES can also be trained. These reactions typically require AIMD simulations, where selectivity is determined from the average behaviour of many trajectories leading to either product. Other approaches have also been developed.[72] We here explored the dimerisation between cyclopentadiene, for which *endo* selectivity has been rationalised on the basis of bifurcating reaction pathways.[73] Once again, initiating active learning from the literature TS (TS_1, **Figure 5**) using a DFT method analogous to the one used in the original work by Caramella and co-workers affords a reactive potential from which 500 fs trajectories were propagated.

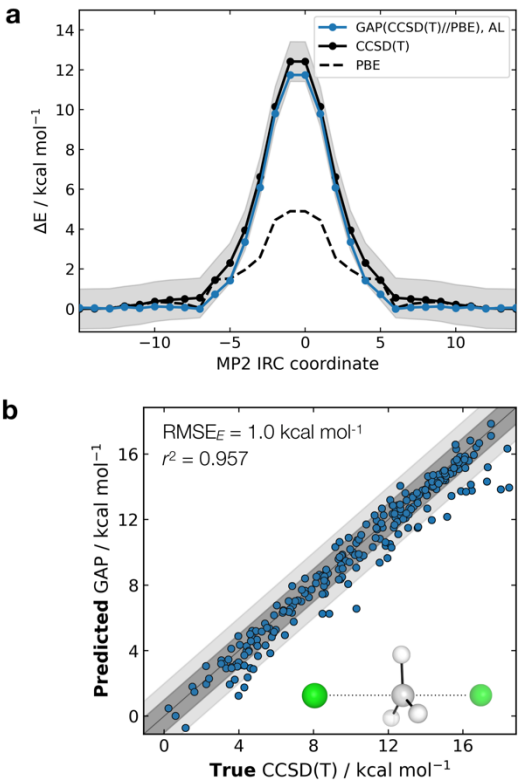

Figure 5: Gas phase S_N2 reactive dynamics. Energetics of a model S_N2 reaction in the gas phase. (a) Predictions (GAP) and ground-true (DLPNO-CCSD(T)/ma-def2-TZVPP) energy values on the MP2/ma-def2-TZVPP intrinsic reaction coordinate (IRC). Shaded region bounds the 'chemically accurate' (±1 kcal mol^{-1}) region. (b) Parity plots of between GAP predictions and true energies from ten 100 fs GAP-MD trajectories initialised from the TS (300 K). Dark and light shaded area bound the ±1 kcal mol^{-1} and ±2 kcal mol^{-1} error regions, respectively.

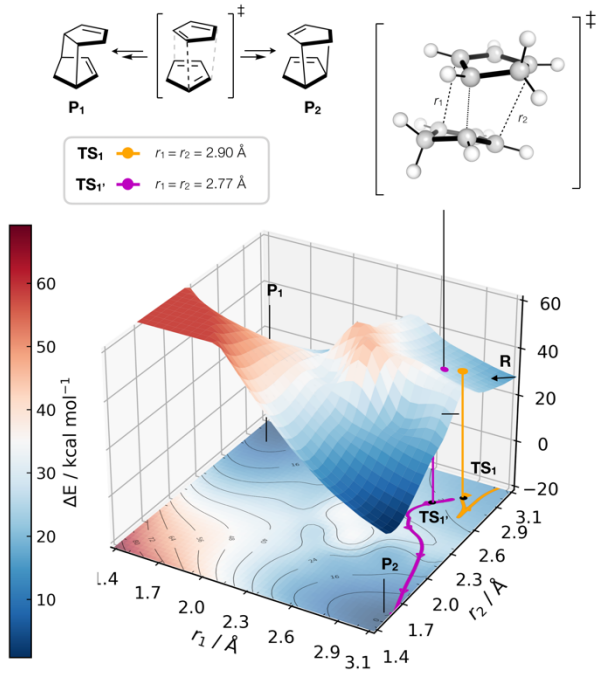

Figure 6: **GAP dynamics on a bifurcating surface**. 2D PES (B3LYP/def2-SVP) along the forming bond distances (r_1, r_2) in the dimerization of cyclopentadiene. An example of GAP-propagated reactive dynamics (300 K) is shown from TS$_1$ (7N in ref. [73]), which leads to reactants, (representative trajectory in orange), and from TS$_{1'}$ which leads to products (a representative trajectory is shown in purple). 3D projection is truncated at 2.5 eV above the minimum for plotting. Interpolated surface used a cubic spline using *scipy.interp2d* with the raw surfaces shown in **Figure S21**. All trajectories shown in **Figure S18** and **20**.

Interestingly, we found that propagating the system from this TS did not afford any products (P$_1$ or P$_2$, **Figure 6**), with all trajectories leading to the reactant state (**Figure S18**). Further investigation and generation of the relaxed 2D potential energy surface over the two possible forming C–C bonds (r_1, r_2) leading to products provided a rather different surface to the one suggested in ref. [73], with a flat portion then an incline as r_1, r_2 shorten below 2.9 Å, with a steeply exergonic reverse reaction (intrinsic reaction coordinate, IRC, shown in **Figure S19**). As noted by Caramella, following the IRC forwards from TS$_1$ the reaction proceeds to another TS$_{1'}$ which is similar in energy (ΔE = 2 kcal mol^{-1}). By training a GAP at 500 K and propagating GAP-MD from TS$_{1'}$, sampling the area of the PES around a valley-ridge inflection point (VRI), trajectories did lead to the expected two products (e.g., purple line, **Figure 6** and **Figure S20**). With no *a priori* knowledge, apart from the structure of TS$_1$, the topology of the bifurcating surface can therefore be revealed efficiently using GAP dynamics. Given the strategy is completely automated, and with training taking a few hours to days, this enables the routine examination of dynamical effects at the transition state following its location.

Solution phase bimolecular nucleophilic substitution

The ability to accurately describe bond-breaking/forming paths in the condensed-phase is crucial if this strategy is to be applied to increasingly complex processes, such as enzymatic reactions. Towards this goal, and having accurately generated potentials for condensed-phase molecular systems and gas-phase reactions, we decided to extend our active learning strategy to explicitly solvated reactions; once again, using the S_N2 reaction between chloride and methyl chloride as a test case. S_N2 reactions have been used as a test case for a recent ML potential, but those studies have been limited to the gas phase and used thousands of training points.[20] In terms of reactions in explicit solvent, a NN has been trained by Parrinello and co-workers to study urea decomposition in water.[40]

Adopting an identical strategy to the one employed for the systems described above, the intra- and inter-molecular PES dynamics can be propagated from the TS and the effect of explicit solvation interrogated (**Figure 7**). Using only the gas-phase TS as an *a priori* point on the PES the training is complete in a day (on a single compute node) if only short time (<1 ps) dynamics are of interest. Interestingly, the behaviour in explicit water (blue, **Figure 7**) differs from the implicit counterpart (red, **Figure 7**). Solvent reorganisation is instantaneous in the implicit case, which results in oscillations in the C-Cl bond characteristic

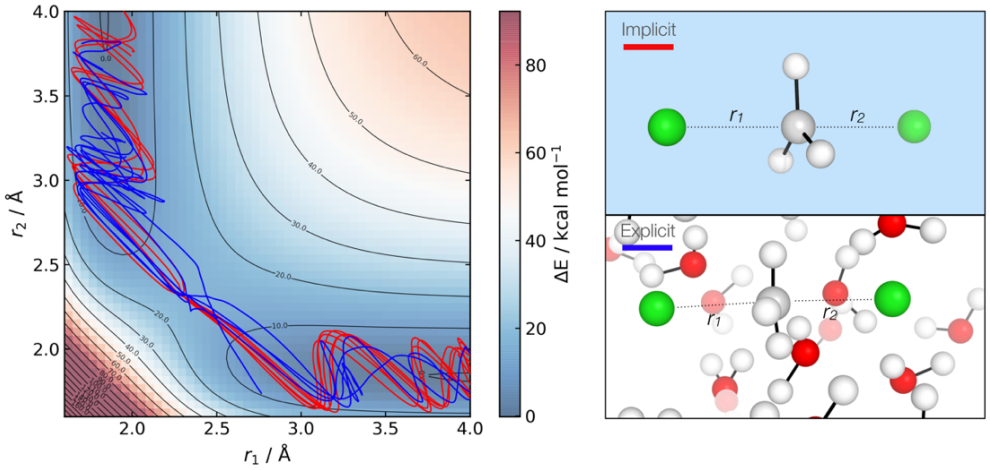

Figure 7: Solution-phase S_N2 reactive dynamics. Reactive GAP MD trajectories (lines) propagated from the TS, trained on implicitly solvated configurations (CPCM(Water)-PBE/def2-SVP, red) and explicitly solvated configurations (PBE/400 eV, blue). Ground truth implicitly solvated surface (cubic interpolated, 5:1) underneath with the error to the GAP prediction in **Figure S23**. Intramolecular GAP used C-only (r_c^{SOAP} = 6 Å) descriptor with the intermolecular explicit solvent using a SOAP on O (r_c^{SOAP} = 3.5 Å) and Cl (r_c^{SOAP} = 4.5 Å, Figure S24).

of a gas phase reaction. In contrast, in explicit solvent the dynamics are more complex, with a slower transition from the product channel. Additionally, one of the 10 trajectories re-crosses the barrier after 170 fs of simulation (**Figure S22**), where the solvent has not reorganised to accommodate the anionic chloride yet, making the path to products shallower in energy. The component-wise separation of the system also leads to the possibility of training to a more accurate *ab initio* surface for the gas phase reaction, in a similar way to QM/MM, but here a ML(A)/ML(B) partition is available where A and B are two different ground truth methods. Application of this kind of hierarchical ML potential fitting will be the subject of further work.

Conclusions

Studying dynamic processes and the effect of explicit solvation on chemical reactions demands a rapid method to develop bespoke force-field models with high accuracy. Here, we demonstrated that within the Gaussian Approximation Potential (GAP) machine learning framework, accurate and robust models can be developed efficiently for gas-phase and condensed-phase molecular reactions. Our strategy starts from a small number of randomly selected points in the configuration space, from which active learning training of intra- and inter-molecular components of the energy and forces is carried out. The developed method is completely automated and publicly available (https://github.com/duartegroup/gap-train). We also define a prospective error metric, which is found to be crucial in developing robust active-learning-based potentials, whereas correlation on a predefined test set is insufficient to assess the quality of such a potential. We illustrate the generality of this approach by modelling bulk water, Zn(II) in aqueous solution, and chemical reactions in the gas phase and explicit solvent, including post-TS cyclisation and S_N2 reactions. The diversity of the examples presented here demonstrates the general applicability of the strategy and encourages applying this approach in the modelling of more complex reactions in homogeneous and heterogenous environments.

Methods

All Gaussian Approximation Potentials (GAPs) were trained using the GAP and QUIP codes (Singularity distribution, commit #66c553f) and a Smooth Overlap of Atomic Positions (SOAP)[57] kernel with hyperparameters as defined in **Table S1**. GAP-MD simulations were performed with ASE[74] interfaced to QUIP with the quippy wrapper using the Langevin integrator with 0.5 fs timesteps at 300 K unless otherwise specified. Condensed phase MD simulations were performed with three-dimensional periodic boundary conditions following minimisation and equilibration for at least 20 ps. Initial configurations, CUR[60] selection and all learning curves were generated with the *gap-train* module, which was used to run the automated fitting (**Figure 8**).[75] All active learning was performed using a 'diff' strategy, where a configuration is added to the training set if $|E_0 - E_{GAP}|$ is larger than a threshold. With system-dependent hyperparameter optimisation, using a threshold on the maximum atomic variance predicted by the Gaussian process ('gp_var') can result in accelerated learning (**Figure S25**). CUR selection used SOAPs averaged over atoms in a configuration using the Dscribe[76] package ('inner' averaging, over entries of the expansion coefficient vector). Intra+Inter (I+I) energy and force evaluations used an expansion factor of 10 to ensure no intermolecular atoms were within the intra GAP cut-off. The NumPy[77] implementation introduces a negligible computational overhead for expanding the box (~0.1 ms step^{-1} real time) but requires two GAP calculations on the inter and intra components, currently carried out in serial. All generated potentials, with the exception of the revPBE0-D3 water potential and metallocage, were trained in less than a day on 10 CPU cores. The revPBE0-D3 water potential was constructed without any prior data in 5 days (1 intra + 4 inter) and used 20 CPU cores, while the metallocage fragment was trained for 3 days also on 20 CPU cores. Explicit S_N2 reaction dynamics was

performed using intra components for H_2O and $[Cl\cdots CH_3Cl]^-$, where the latter, due to the finite atomic cut-off employed, has the correct dissociation behaviour when Cl^- and CH_3Cl are distant.

Periodic DFTB calculations performed with DFTB+[78] using 3ob[79] parameters, and molecular equivalents using GFN2-XTB[80] in XTB v. 6.2.3. Periodic pure DFT calculations were performed with GPAW[81,82] v. 19.8.1 with the PBE[83] functional and a 400 eV plane-wave cut-off from a *dzp* LCAO initial guess at the gamma point. Hybrid periodic DFT calculations with the revPBE0[84,85] functional combined with the D3[86] dispersion correction were performed with CP2K.[87]

Molecular DFT, MP2 and coupled cluster [DLPNO-CCSD(T)] calculations performed with ORCA[88,89] v. 4.2.1 wrapped with autodE[90] using PBE[83] and PBE0[85] functionals, (ma)-def2-SVP, def2-TZVP, ma-def2-TZVPP basis sets.[91] AIMD calculations at the DFTB level performed with DFTB+ with 3ob parameters[79] and MM simulations performed with GROMACS[92,93] 2019.2 with TIP3P parameters.[94]

AN1-1x NN trained using on a subset (6276 configurations) of the full ANI-1x dataset,[95] containing only water clusters with energies and forces re-evaluated at DFTB(3ob) using the TorchANI[96] implementation with only O and H networks. All default parameters and training with forces using a batch size of 128, appropriate for fewer data. Best validation RMSE = 1.18 kcal mol^{-1} after 218 epochs. Behler NN[16] trained on 7258 RPBE configurations from ref. 97, re-evaluated at DFTB(3ob) with n2p2[98] using the default parameters and symmetry functions.

```python
import gaptrain as gt
gt.GTConfig.n_cores = 4

h2o = gt.System(box_size=[10, 10, 10])
h2o.add_solvent('h2o', n=20)

data, gap = gt.active.train_ii(h2o, method_name='dftb')
```

Figure 8: Example Python input script required to train a bulk water model from scratch at the DFTB level using four CPU cores.

Author Contributions

TAY developed and implemented the strategy, and carried out the calculations. TJW contributed to development and implementation of the strategy, and to MMMD calculations for training. All authors participated in data analyses. TAY and FD conceptualised the study. FD and VLD supervised the study and wrote the manuscript with TAY.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

TAY and TJW acknowledge the EPSRC Centre for Doctoral Theory and Modelling in Chemical Sciences (EP/L015722/1). TAY studentship is generously supported by AWE. Ministry of Defence ©British Crown Copyright 2020/AWE. The authors acknowledge the use of 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). We thank Prof. Gábor Csányi, Dr Grant Hill and Dr Robert Shaw for discussions and comments on the manuscript.

References

1 D. Frenkel and B. Smit, *Understanding Molecular Simulation : From Algorithms to Applications*, Academic Press, Cambridge, Massachusetts, 2nd edn., 2002.

2 K. Lindorff-Larsen, P. Maragakis, S. Piana, M. P. Eastwood, R. O. Dror and D. E. Shaw, *PLoS One*, 2012, **7**, e32131.

3 R. Iftimie, P. Minary and M. E. Tuckerman, *Proc. Natl. Acad. Sci.*, 2005, **102**, 6654–6659.

4 F. Noé, A. Tkatchenko, K.-R. Müller and C. Clementi, *Annu. Rev. Phys. Chem.*, 2020, **71**, 361–390.

5 T. Mueller, A. Hernandez and C. Wang, *J. Chem. Phys.*, 2020, **152**, 050902.

6 O. T. Unke, D. Koner, S. Patra, S. Käser and M. Meuwly, *Mach. Learn. Sci. Technol.*, 2020, **1**, 013001.

7 R. Z. Khaliullin, H. Eshet, T. D. Kühne, J. Behler and M. Parrinello, *Nat. Mater.*, 2011, **10**, 693–697.

8 G. C. Sosso, G. Miceli, S. Caravati, F. Giberti, J. Behler and M. Bernasconi, *J. Phys. Chem. Lett.*, 2013, **4**, 4241–4246.

9 M. A. Caro, V. L. Deringer, J. Koskinen, T. Laurila and G. Csányi, *Phys. Rev. Lett.*, 2018, **120**, 166101.

10 H. Niu, L. Bonati, P. M. Piaggi and M. Parrinello, *Nat. Commun.*, 2020, **11**, 2654.

11 B. Cheng, G. Mazzola, C. J. Pickard and M. Ceriotti, *Nature*, 2020, **585**, 217–220.

12 V. L. Deringer, N. Bernstein, G. Csányi, C. Ben Mahmoud, M. Ceriotti, M. Wilson, D. A. Drabold and S. R. Elliott, *Nature*, 2021, **589**, 59–64.

13 S. J. Ang, W. Wang, D. Schwalbe-Koda, S. Axelrod and R. Gómez-Bombarelli, *Chem*, 2021, **7**, 738–751.

14 D. J. Cole, L. Mones and G. Csányi, *Faraday Discuss.*, 2020, **224**, 247–264.

15 D. A. Rufa, H. Bruce Macdonald, J. Fass, M. Wieder, P. B. Grinaway, A. E. Roitberg, O. Isayev and J. D. Chodera, *bioRxiv*, , DOI:10.1101/2020.07.29.227959.

16 J. Behler and M. Parrinello, *Phys. Rev. Lett.*, 2007, **98**, 146401.

17 J. Behler, *Angew. Chem. Int. Ed.*, 2017, **56**, 12828–12840.

18 J. S. Smith, O. Isayev and A. E. Roitberg, *Chem. Sci.*, 2017, **8**, 3192–3203.

19 K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller and A. Tkatchenko, *Nat. Commun.*, 2017, **8**, 13890.

20 O. T. Unke and M. Meuwly, *J. Chem. Theory Comput.*, 2019, **15**, 3678–3693.

21 A. P. Bartók, M. C. Payne, R. Kondor and G. Csányi, *Phys. Rev. Lett.*, 2010, **104**, 136403.
22 A. P. Bartók and G. Csányi, *Int. J. Quantum Chem.*, 2015, **115**, 1051–1057.
23 S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt and K.-R. Müller, *Sci. Adv.*, 2017, **3**, e1603015.
24 A. P. Thompson, L. P. Swiler, C. R. Trott, S. M. Foiles and G. J. Tucker, *J. Comput. Phys.*, 2015, **285**, 316–330.
25 A. V. Shapeev, *Multiscale Model. Simul.*, 2016, **14**, 1153–1173.
26 Y. Zuo, C. Chen, X. Li, Z. Deng, Y. Chen, J. Behler, G. Csányi, A. V. Shapeev, A. P. Thompson, M. A. Wood and S. P. Ong, *J. Phys. Chem. A*, 2020, **124**, 731–745.
27 W. J. Szlachta, A. P. Bartók and G. Csányi, *Phys. Rev. B*, 2014, **90**, 104108.
28 V. L. Deringer and G. Csányi, *Phys. Rev. B*, 2017, **95**, 094203.
29 A. P. Bartók, J. Kermode, N. Bernstein and G. Csányi, *Phys. Rev. X*, 2018, **8**, 041048.
30 G. Sivaraman, A. N. Krishnamoorthy, M. Baur, C. Holm, M. Stan, G. Csányi, C. Benmore and Á. Vázquez-Mayagoitia, *npj Comput. Mater.*, 2020, **6**, 104.
31 F. C. Mocanu, K. Konstantinou, T. H. Lee, N. Bernstein, V. L. Deringer, G. Csányi and S. R. Elliott, *J. Phys. Chem. B*, 2018, **122**, 8998–9006.
32 P. O. Dral, A. Owens, A. Dral and G. Csányi, *J. Chem. Phys.*, 2020, **152**, 204110.
33 M. Veit, S. K. Jain, S. Bonakala, I. Rudra, D. Hohl and G. Csányi, *J. Chem. Theory Comput.*, 2019, **15**, 2574–2586.
34 V. L. Deringer, M. A. Caro and G. Csányi, *Nat. Commun.*, 2020, **11**, 5461.
35 B. Cheng, E. A. Engel, J. Behler, C. Dellago and M. Ceriotti, *Proc. Natl. Acad. Sci.*, 2019, **116**, 1110–1115.
36 C. Schran, K. Brezina and O. Marsalek, *J. Chem. Phys.*, 2020, **153**, 104105.
37 E. V. Podryabinkin, E. V. Tikhonov, A. V. Shapeev and A. R. Oganov, *Phys. Rev. B*, 2019, **99**, 064114.
38 N. Artrith and J. Behler, *Phys. Rev. B*, 2012, **85**, 045439.
39 K. Gubaev, E. V. Podryabinkin, G. L. W. Hart and A. V. Shapeev, *Comput. Mater. Sci.*, 2019, **156**, 148–156.
40 M. Yang, L. Bonati, D. Polino and M. Parrinello, *Catal. Today*, , DOI:10.1016/j.cattod.2021.03.018.
41 J. S. Smith, B. Nebgen, N. Lubbers, O. Isayev and A. E. Roitberg, *J. Chem. Phys.*, 2018, **148**, 241733.
42 E. V. Podryabinkin and A. V. Shapeev, *Comput. Mater. Sci.*, 2017, **140**, 171–180.
43 L. Zhang, D.-Y. Lin, H. Wang, R. Car and W. E, *Phys. Rev. Mater.*, 2019, **3**, 023804.
44 Y. Zhang, H. Wang, W. Chen, J. Zeng, L. Zhang, H. Wang and W. E, *Comput. Phys. Commun.*, 2020, **253**, 107206.
45 L. Zhang, J. Han, H. Wang, R. Car and W. E, *Phys. Rev. Lett.*, 2018, **120**, 143001.
46 J. Vandermause, S. B. Torrisi, S. Batzner, Y. Xie, L. Sun, A. M. Kolpak and B. Kozinsky, *npj Comput. Mater.*, 2020, **6**, 20.
47 R. Jinnouchi, J. Lahnsteiner, F. Karsai, G. Kresse and M. Bokdam, *Phys. Rev. Lett.*, 2019, **122**, 225701.
48 R. Jinnouchi, K. Miwa, F. Karsai, G. Kresse and R. Asahi, *J. Phys. Chem. Lett.*, 2020, **11**, 6946–6955.
49 J. Zeng, L. Cao, M. Xu, T. Zhu and J. Z. H. Zhang, *Nat. Commun.*, 2020, **11**, 5713.
50 S. Kearnes, *Trends Chem.*, 2021, **3**, 77–79.
51 G. C. Cawley and N. L. C. Talbot, *J. Mach. Learn. Res.*, 2010, **11**, 2079–2107.
52 C. Kramer and P. Gedeck, *J. Chem. Inf. Model.*, 2010, **50**, 1961–1969.
53 Y. Li and J. Yang, *J. Chem. Inf. Model.*, 2017, **57**, 1007–1012.
54 L. Chen, A. Cruz, S. Ramsey, C. J. Dickson, J. S. Duca, V. Hornak, D. R. Koes and T. Kurtzman, *PLoS One*, 2019, **14**, e0220113.
55 D. P. Kovács, W. McCorkindale and A. A. Lee, *Nat. Commun.*, 2021, **12**, 1695.
56 R. P. Sheridan, *J. Chem. Inf. Model.*, 2013, **53**, 783–790.
57 A. P. Bartók, R. Kondor and G. Csányi, *Phys. Rev. B*, 2013, **87**, 184115.
58 C. J. Pickard and R. J. Needs, *J. Phys. Condens. Matter*, 2011, **23**, 053201.
59 N. Bernstein, G. Csányi and V. L. Deringer, *npj Comput. Mater.*, 2019, **5**, 99.
60 M. W. Mahoney and P. Drineas, *Proc. Natl. Acad. Sci.*, 2009, **106**, 697–702.
61 L. Zheng, M. Chen, Z. Sun, H.-Y. Ko, B. Santra, P. Dhuvad and X. Wu, *J. Chem. Phys.*, 2018, **148**, 164505.
62 A. P. Bartók, M. J. Gillan, F. R. Manby and G. Csányi, *Phys. Rev. B*, 2013, **88**, 054104.
63 R. Ramakrishnan, P. O. Dral, M. Rupp and O. A. von Lilienfeld, *J. Chem. Theory Comput.*, 2015, **11**, 2087–2096.
64 S. Wengert, G. Csányi, K. Reuter and J. T. Margraf, *Chem. Sci.*, , DOI:10.1039/D0SC05765G.
65 P. Li and K. M. Merz, *Chem. Rev.*, 2017, **117**, 1564–1686.
66 H. Ohtaki and T. Radnai, *Chem. Rev.*, 1993, **93**, 1157–1204.
67 T. A. Young, V. Martí-Centelles, J. Wang, P. J. Lusby and F. Duarte, *J. Am. Chem. Soc.*, 2020, **142**, 1300–1310.
68 S. Pratihar, X. Ma, Z. Homayoon, G. L. Barnes and W. L. Hase, *J. Am. Chem. Soc.*, 2017, **139**, 3570–3590.
69 D. H. Ess, S. E. Wheeler, R. G. Iafe, L. Xu, N. Çelebi-Ölçüm and K. N. Houk, *Angew. Chem. Int. Ed.*, 2008, **47**, 7592–7601.
70 J. Xie and W. L. Hase, *Science*, 2016, **352**, 32–33.
71 I. Szabó and G. Czakó, *J. Phys. Chem. A*, 2017, **121**, 9005–9019.
72 S. Lee and J. M. Goodman, *J. Am. Chem. Soc.*, 2020, **142**, 9210–9219.
73 P. Caramella, P. Quadrelli and L. Toma, *J. Am. Chem. Soc.*, 2002, **124**, 1130–1131.
74 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. Peterson, C. Rostgaard, J. Schiøtz, O. Schütt, M. Strange, K. S. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng and K. W. Jacobsen, *J. Phys. Condens. Matter*, 2017, **29**, 273002.

75 T. Young and T. Johnston-Wood, *gap-train*, https://github.com/t-young31/gap-train, 2020.

76 L. Himanen, M. O. J. Jäger, E. V. Morooka, F. Federici Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke and A. S. Foster, *Comput. Phys. Commun.*, 2020, **247**, 106949.

77 C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke and T. E. Oliphant, *Nature*, 2020, **585**, 357–362.

78 B. Hourahine, B. Aradi, V. Blum, F. Bonafé, A. Buccheri, C. Camacho, C. Cevallos, M. Y. Deshaye, T. Dumitrică, A. Dominguez, S. Ehlert, M. Elstner, T. van der Heide, J. Hermann, S. Irle, J. J. Kranz, C. Köhler, T. Kowalczyk, T. Kubař, I. S. Lee, V. Lutsker, R. J. Maurer, S. K. Min, I. Mitchell, C. Negre, T. A. Niehaus, A. M. N. Niklasson, A. J. Page, A. Pecchia, G. Penazzi, M. P. Persson, J. Řezáč, C. G. Sánchez, M. Sternberg, M. Stöhr, F. Stuckenberg, A. Tkatchenko, V. W. -z. Yu and T. Frauenheim, *J. Chem. Phys.*, 2020, **152**, 124101.

79 M. Gaus, A. Goez and M. Elstner, *J. Chem. Theory Comput.*, 2013, **9**, 338–354.

80 C. Bannwarth, S. Ehlert and S. Grimme, *J. Chem. Theory Comput.*, 2019, **15**, 1652–1671.

81 J. J. Mortensen, L. B. Hansen and K. W. Jacobsen, *Phys. Rev. B*, 2005, **71**, 035109.

82 J. Enkovaara, C. Rostgaard, J. J. Mortensen, J. Chen, M. Dułak, L. Ferrighi, J. Gavnholt, C. Glinsvad, V. Haikola, H. A. Hansen, H. H. Kristoffersen, M. Kuisma, A. H. Larsen, L. Lehtovaara, M. Ljungberg, O. Lopez-Acevedo, P. G. Moses, J. Ojanen, T. Olsen, V. Petzold, N. A. Romero, J. Stausholm-Møller, M. Strange, G. A. Tritsaris, M. Vanin, M. Walter, B. Hammer, H. Häkkinen, G. K. H. Madsen, R. M. Nieminen, J. K. Nørskov, M. Puska, T. T. Rantala, J. Schiøtz, K. S. Thygesen and K. W. Jacobsen, *J. Phys. Condens. Matter*, 2010, **22**, 253202.

83 J. P. Perdew, K. Burke and M. Ernzerhof, *Phys. Rev. Lett.*, 1996, **77**, 3865–3868.

84 Y. Zhang and W. Yang, *Phys. Rev. Lett.*, 1998, **80**, 890–890.

85 C. Adamo and V. Barone, *J. Chem. Phys.*, 1999, **110**, 6158–6170.

86 S. Grimme, J. Antony, S. Ehrlich and H. Krieg, *J. Chem. Phys.*, 2010, **132**, 154104.

87 T. D. Kühne, M. Iannuzzi, M. Del Ben, V. V. Rybkin, P. Seewald, F. Stein, T. Laino, R. Z. Khaliullin, O. Schütt, F. Schiffmann, D. Golze, J. Wilhelm, S. Chulkov, M. H. Bani-Hashemian, V. Weber, U. Borštnik, M. Taillefumier, A. S. Jakobovits, A. Lazzaro, H. Pabst, T. Müller, R. Schade, M. Guidon, S. Andermatt, N. Holmberg, G. K. Schenter, A. Hehn, A. Bussy, F. Belleflamme, G. Tabacchi, A. Glöß, M. Lass, I. Bethune, C. J. Mundy, C. Plessl, M. Watkins, J. VandeVondele, M. Krack and J. Hutter, *J. Chem. Phys.*, 2020, **152**, 194103.

88 F. Neese, *WIREs Comput. Mol. Sci.*, 2012, **2**, 73–78.

89 F. Neese, *WIREs Comput. Mol. Sci.*, 2018, **8**, 1–6.

90 T. A. Young, J. J. Silcock, A. J. Sterling and F. Duarte, *Angew. Chemie Int. Ed.*, 2021, **60**, 4266–4274.

91 F. Weigend and R. Ahlrichs, *Phys. Chem. Chem. Phys.*, 2005, **7**, 3297.

92 H. J. C. Berendsen, D. van der Spoel and R. van Drunen, *Comput. Phys. Commun.*, 1995, **91**, 43–56.

93 M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess and E. Lindahl, *SoftwareX*, 2015, **1–2**, 19–25.

94 W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, *J. Chem. Phys.*, 1983, **79**, 926–935.

95 J. S. Smith, R. Zubatyuk, B. Nebgen, N. Lubbers, K. Barros, A. E. Roitberg, O. Isayev and S. Tretiak, *Sci. Data*, 2020, **7**, 134.

96 X. Gao, F. Ramezanghorbani, O. Isayev, J. S. Smith and A. E. Roitberg, *J. Chem. Inf. Model.*, 2020, **60**, 3408–3415.

97 T. Morawietz, A. Singraber, C. Dellago and J. Behler, *Proc. Natl. Acad. Sci.*, 2016, **113**, 8368–8373.

98 A. Singraber, T. Morawietz, J. Behler and C. Dellago, *J. Chem. Theory Comput.*, 2019, **15**, 3075–3092.