Supporting Information

Authors..

Contents

S1 GAPs for Reactions

Using our initial GAP training strategy,[1] found to be effective for developing an accurate reactive potential for the S_N2 reaction Cl^- + CH_3Cl in the higher-dimensional ethene+butadiene Diels-Alder case did not afford a chemically accurate GAP (Figure S1a). However, for a similarly complex H-abstraction reaction the obtained error is closer to the target 1 kcal mol^{-1} (0.04 eV) level of accuracy (Figure S1b). Note that these results are consistent within repeats of the partially random AL strategy and minor hyperparameter tweaks.

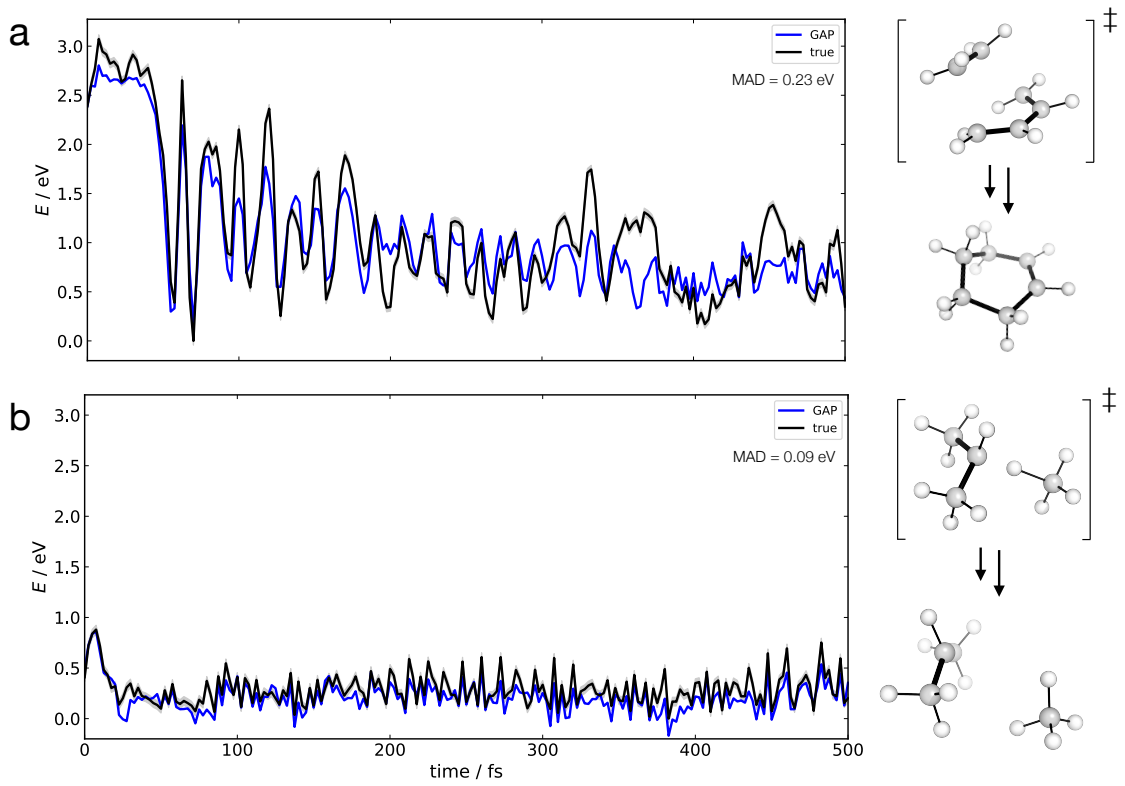

Figure S1: Comparison of predicted and true energies over a GAP-MD propagated trajectories using initial velocities suitable for 300 K and a Langevin thermostat (300 K) with a 0.5 fs time step. 'True' energies calculated at PBE0-D3BJ/def2-SVP in ORCA. GAPs trained using active learning (AL) and the same hyperparameters in ref. [1] at 500 K. AL used a $E_T = 2 \times 10^{-5}$ eV $atom^{-1}$ threshold on the maximum Gaussian process variance to select new configurations within a 1 ps window. Final datasets contained 217 and 223 configurations for ethene + butadiene (a) and methyl + propane (b) respectively.

S2 System Size Scaling

Extending our initial work on a selection of modestly-sized systems and configurational complexity,[1] the following section outlines the how the number of evaluations are required to train a chemically accurate GAP (over > 1 ps simulation time) scales with system size (hyperparameters shown in Table S1, standard active learning methodology described in ref. [1]).

Type	Parameter	Description	Value
GAP	σ_E	Expected error in energies	0.316 meV atom^{-1}
	σ_F	Expected error in force components	0.1 eV Å^{-1}
	ζ	Power the kernel is raised to, increasing the dissimilarity between environments ($\zeta > 1$)	4
	n_{sparse}	Number of atomic environments above which selection is performed	500
	sparse method	Method to select the maximumly diverse set of configurations	CUR points
SOAP descriptors	$\sigma_{\text{at}}^{\text{SOAP}}$	Spread of the Gaussian added to each nuclear coordinate	0.5 Å
	$n_{\text{max}}, l_{\text{max}}$	Expansion order in the radial (n) and angular (l) basis	6
	r_{cut}	Cut-off distance in the short range descriptor	4.0 Å

Table S1: Default parameter set for GAPs and SOAP descriptors. SOAP neighbour densities include all unique pairs.

S2.1 Alkanes

First we consider a semi-optimal case for training larger systems, that of gas phase linear alkanes. This is because as more CH_2 units are added, the 'bonded' component of the energy is already known to the potential, such that only the 'non-bonded' components must be learnt.

The number of evaluations required to reach a potential with $\tau_{\text{acc}} > 1$ ps increases roughly linearly up to pentane (Figure S2) and reaches a maximum at hexane (> 3000 evaluations).[a]

While 3000 evaluations is fast at the GFN2-XTB level, sampling using a more accurate DFT method would exceed the goal of building potentials within a day. Furthermore, alkanes with more than 20 atoms (e.g. 26 in octane) require more than 1000 training configurations (selected using AL). It is therefore necessary to perform some hyperparameter optimisation.

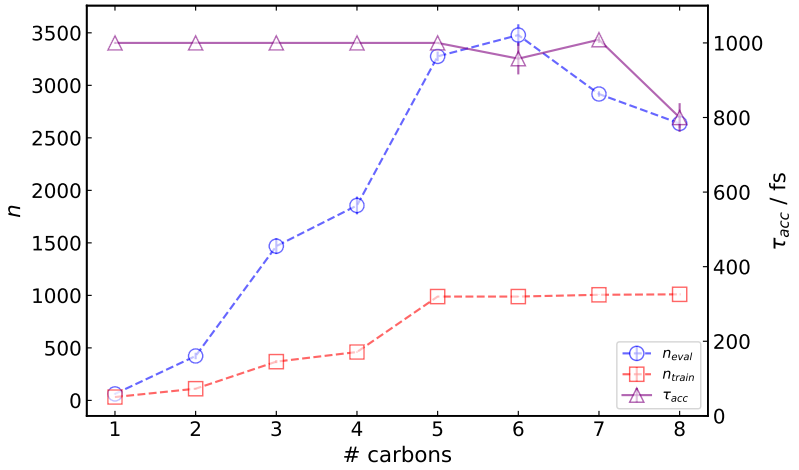

Figure S2: Number of reference evaluations (n_{eval}) and number of training configurations (n_{train}) generated in 100 active learning loops for a linear alkane chain (500 K, 1 kcal mol^{-1} AL threshold, GFN2-XTB reference method). τ_{acc} is an average from three simulations, $E_l = 1$ kcal mol^{-1}, $E_t = 10E_l$, $T = 300$ K, max(τ_{acc}) = 1 ps.

S2.2 Small Organic Molecules

A selection of solvents and small organic molecules (Figure S3, inc. > 2 elements) forms a diverse set over which system scaling can be more realistically determined.

Unlike the alkanes, there is no obvious trend in the number of evaluations required to train a chemically accurate potential. Even correlating against a general complexity metric[2] reveals little to no correlation aside from a general increase.

[a] As the system becomes larger, the less GAP-MD is run and fewer evaluations required to find a new configuration with a large enough error to be selected, hence the peak in n_{eval} for hexane.

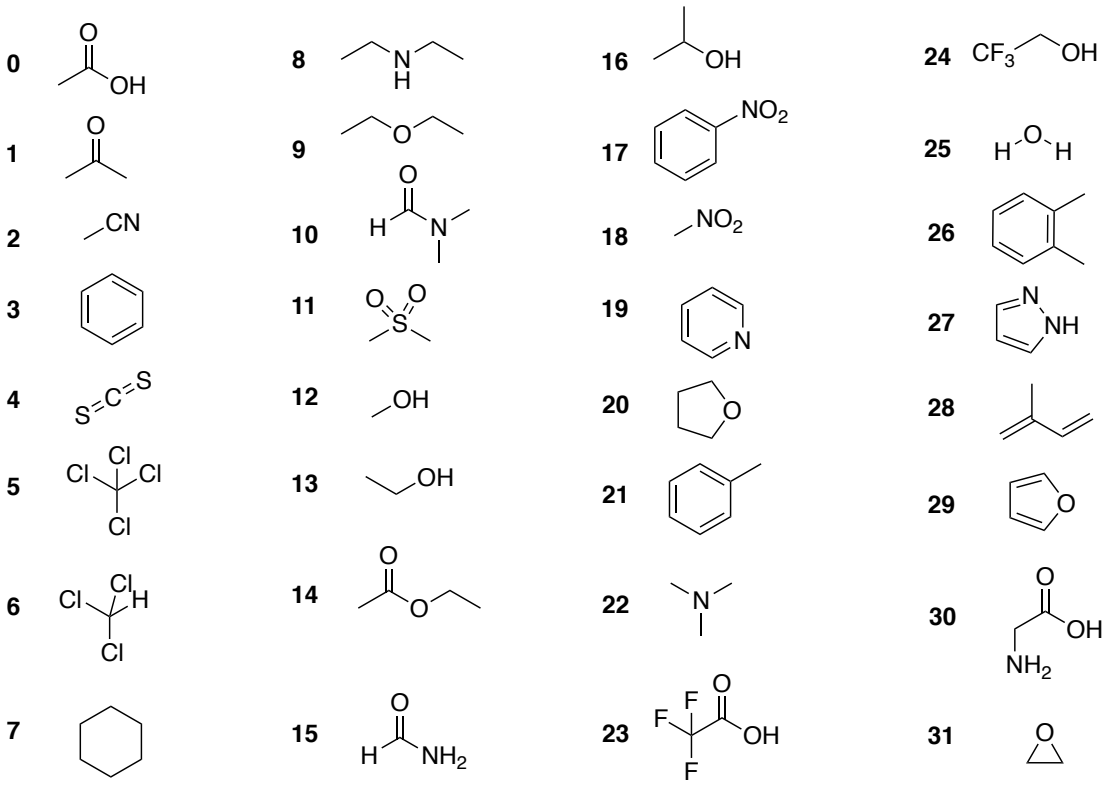

Figure S3: Small molecules used in Figure S4.

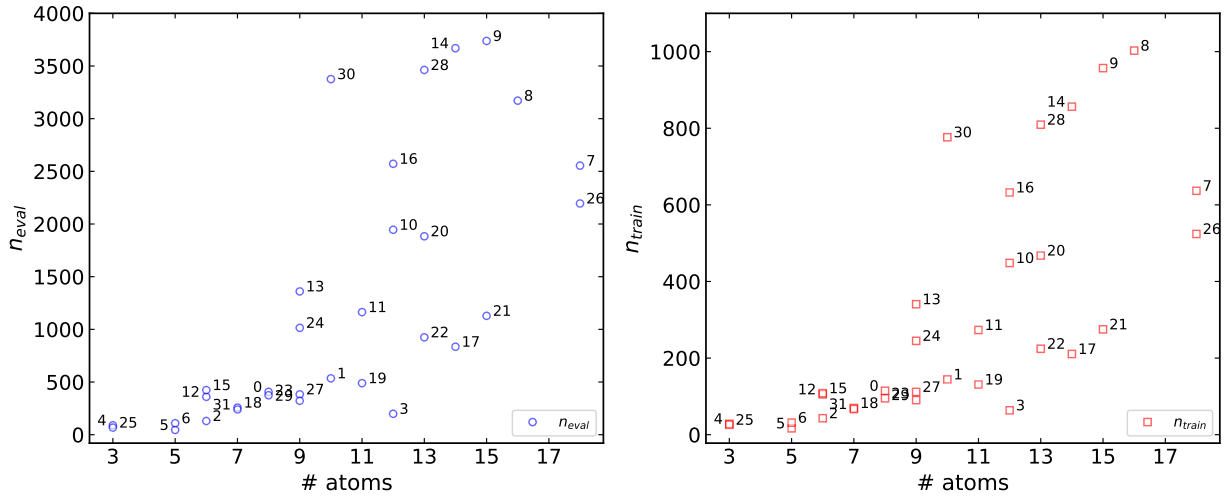

Figure S4: Number of reference evaluations (n_{eval}) and number of training configurations (n_{train}) required to generate a GAP with $\tau_{\text{acc}}>$ 1 ps (parameters as Figure S2) for a variety of small molecules. Values (averages) and standard errors of the mean are taken over three training repeats with different initial random displacements.

S3 Hyperparameter Optimisation

S3.1 QM Convergence

Using ORCA v. 4.2.1 the default SCF and DFT integration grids provide essentially converged forces (Figure S5) thus the defaults are used throughout. The error introduced is negligible compared to the 'expected error' in the forces used in the GAP fitting (σ_F = 0.1 eV Å by default), thus default ORCA parameters will be used herein.

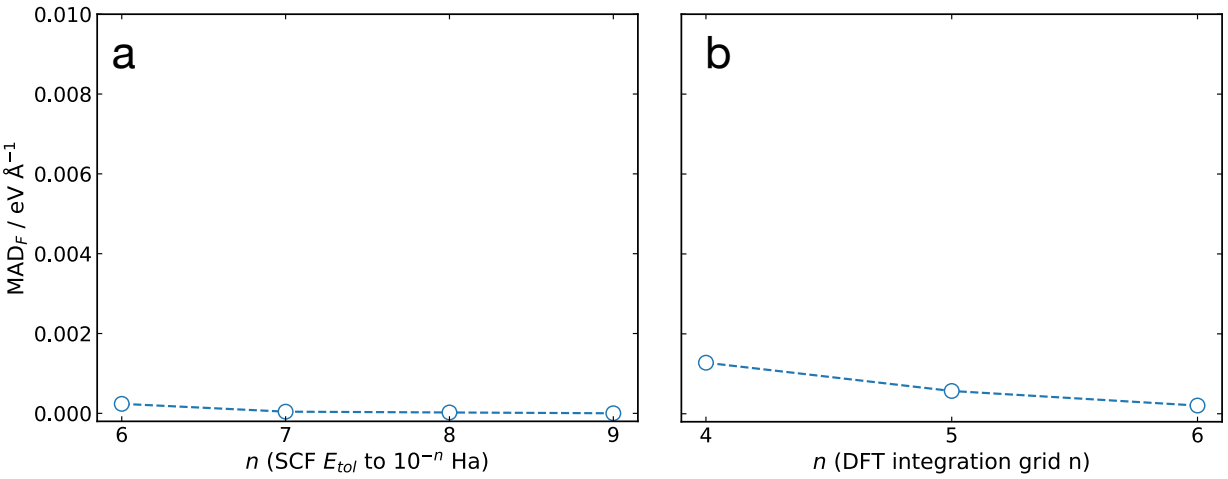

Figure S5: Convergence of the force components of 10 randomly selected frames from an active learning instance from the TS of ethene+butadiene at the PBE/def2-SVP level of theory in ORCA. Convergence of the mean absolute deviation in forcers for the SCF energy change (E_{tol}) convergence (a) and DFT integration grid (b) is with respect to their maximum values in ORCA (10^{-10} Ha and 7 respectively). The maximum MAD is 10× less than the default GAP σ_F.

S3.2 Expansion Order

Increasing the expansion order of the SOAP provides an improved approximation to the overlap between two atomic densities in the (dot product SOAP) kernel. Therefore, the energies and forces should be converged with respect to n_{max} and l_{max}, being aware of the increased computational cost of the SOAP computation. A slightly larger order of 8 in both l_{max} and n_{max} provides the optimum value, requiring a third of the training data to achieve a chemically accurate potential for propane (Figure S6). Similar results are obtained for pentane over a smaller n_{max}, l_{max} window, albeit without the exponential decay observed in Figure S6. As found in other works (e.g. ref. [3]), accuracy converges faster with respect to l_{max} than n_{max} (Figure S7b). Interestingly, decreasing the radial expansion order seems to reduce inhibit training slightly ($n_{\text{max}} = l_{\text{max}} = 8$ in Figure S7a cf.$n_{\text{max}} = 8, l_{\text{max}} = 4$ in Figure S7b).

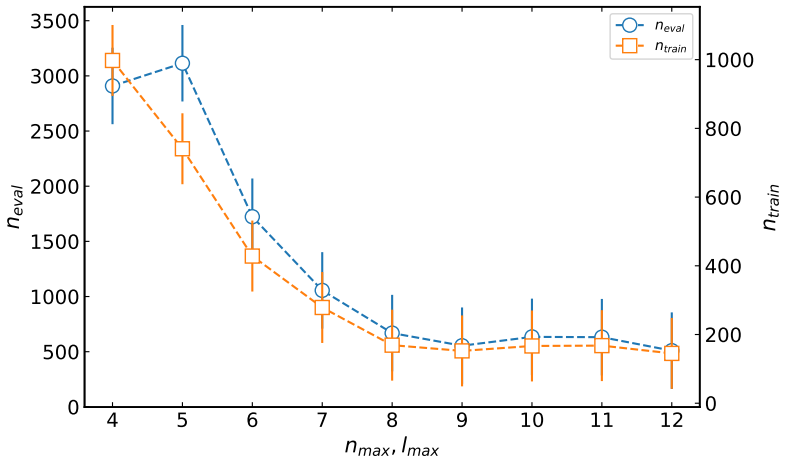

Figure S6: Number of reference evaluations (n_{eval}) and number of training configurations (n_{train}) required to propagate 10, 1 ps GAP-MD at 500 K without finding a prediction with an error > 1 kcal mol^{-1} from the true GFN2-XTB reference at a particular order of radial and angular SOAP expansion (n_{max}, l_{max}) for propane. Other hyperparameters as Table S1.

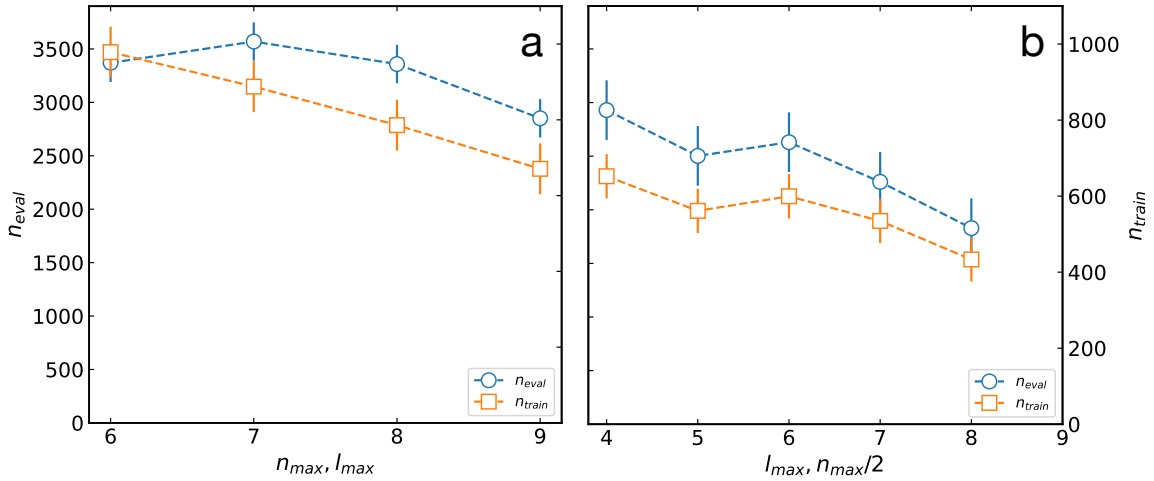

Figure S7: As Figure S6 for pentane with (a) $n_{\text{max}} = l_{\text{max}}$ and (b) $n_{\text{max}} = 2l_{\text{max}}$.

S3.3 Expected Errors

Using $n_{\text{max}} = l_{\text{max}} = 8$ and optimising the 'expected errors' in the energy and forces in training a gas-phase GAP for propane, increasing the 'smoothness' (larger σ_E) is not beneficial to the active learning rate (Figure S8). Instead, fitting energies and forces more *strongly* (top right, Figure S8) enables only 75 training configurations required for an accurate propane potential.[b] Repeating the same along a slice in the (σ_E, σ_F) space suggests that this result it not limited to just small systems (Figure S9), no the choice of expansion order.

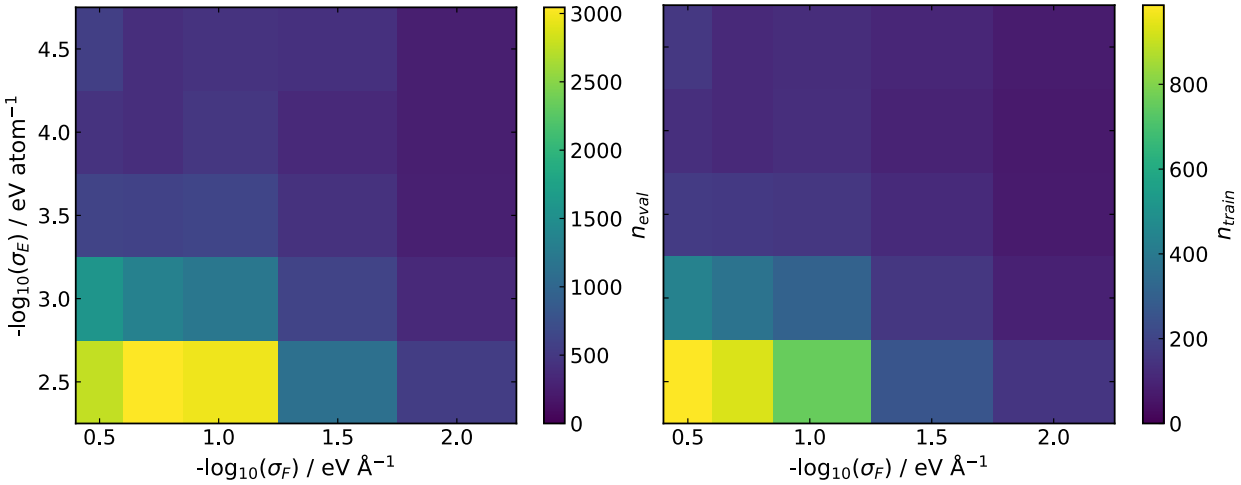

Figure S8: Number of evaluations required to generate a chemically accurate GAP (as Figure S6) for propane as a function of σ_E and σ_F. Values are averages over three independent repeats.

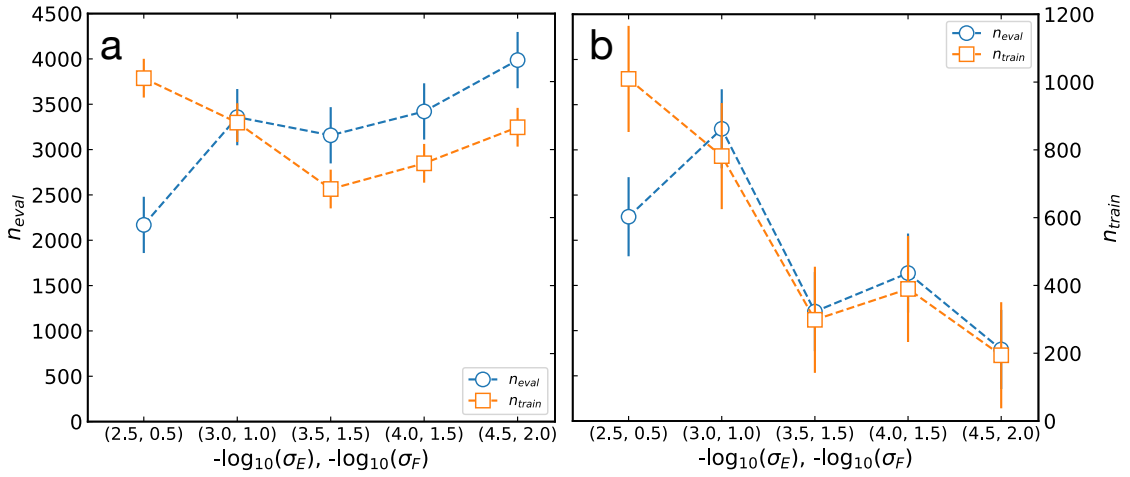

Figure S9: As Figure S8 for pentane along a slice of the 2D space with (a) $n_{\text{max}} = l_{\text{max}} = 8$, (b) $n_{\text{max}} = 12, l_{\text{max}} = 6$.

[b]Note that visualisation of a short 1 ps MD trajectory at 500 K suggests the potential is reasonable and is not circumventing the τ_{acc}error metric.

Here, it is important to emphasize that the pentane potential trained using 500 K active learning at the GFN2-XTB level with the *strongest* fitting ($\sigma_E = 10^{-4}$ eV atom^{-1}, $\sigma_F = 10^{-2}$ eV Å^{-1}, bottom right Figure S9) is highly accurate over 'long-time' dynamics (Figure S10).

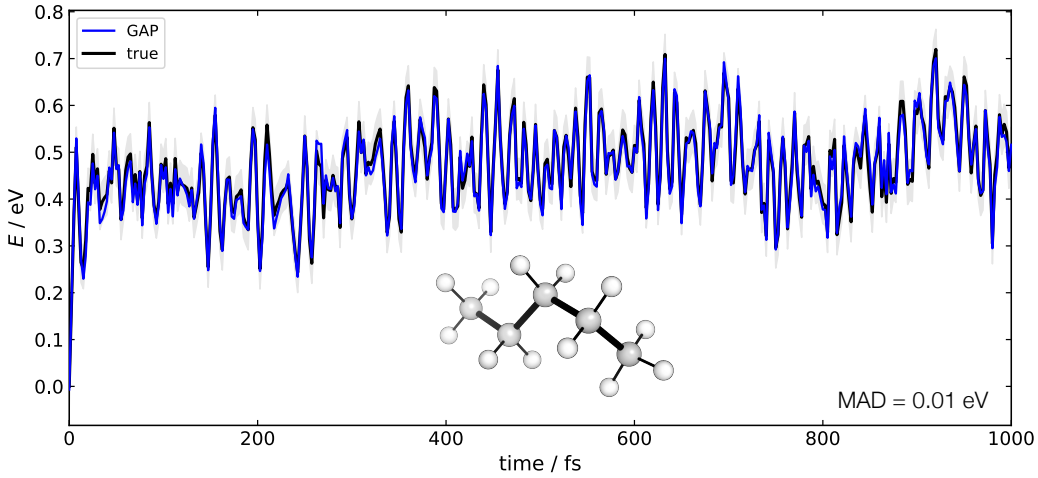

Figure S10: GAP-MD dynamics at 300 K (dt = 0.5 fs, 50 Å cubic box equivalent to a vacuum) compared to GFN2-XTB ground truth. $\sigma_E = 10^{-4}$ eV atom^{-1}, $\sigma_F = 10^{-2}$ eV Å^{-1}, 500 K active learning.

S3.4 Sparse Points

Previous studies have shown that GAP accuracy can converge exponentially with the number of atomic environments ('sparse points').[3]. Of course this will be system dependent, thus the scaling is evaluated for alkanes with 3, 4 and 5 carbons (Figure S11). For these systems all GAPs are converged with respect to n_{sparse} at 800 points. For pentane a substantial drop in the number of required evaluations is observed after 400 configurations, suggesting for 'large' systems (> 15 atoms) n_{sparse} closer to 1000 is more suitable.

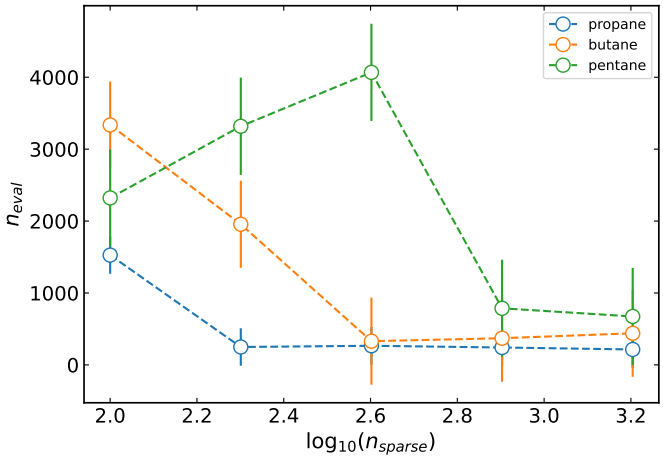

Figure S11: Convergence of the number of evaluations required to generate a potential as Figure S6 for linear alkanes as a function of the sparse points. Error bars are standard errors of the mean over three independent repeats.

S3.5 Summary

Based on these data, we suggest using the updated hyperparameters shown in Table S2 for small molecules in the gas phase.

Type	Parameter	Value
GAP	σ_E	0.1 meV atom^{-1}
	σ_F	0.01 eV Å^{-1}
SOAP descriptors	$n_{\text{max}}/2$, l_{max}	6
	n_{sparse}	1000

Table S2: Updated hyperparameters parameter set for GAPs and SOAP descriptors, all other hyperparameters as Table S1.

S4 Mixing Energy and Force Methods

Prior CCSD(T)-quality GAPs[1] have used computationally demanding numerical gradients.[c] It would therefore be beneficial to use a cheaper QM method to evaluate the gradients in combination with CCSD(T) energies. The following examples are chosed to be of modest computational cost, and are assumed to be generalisable.

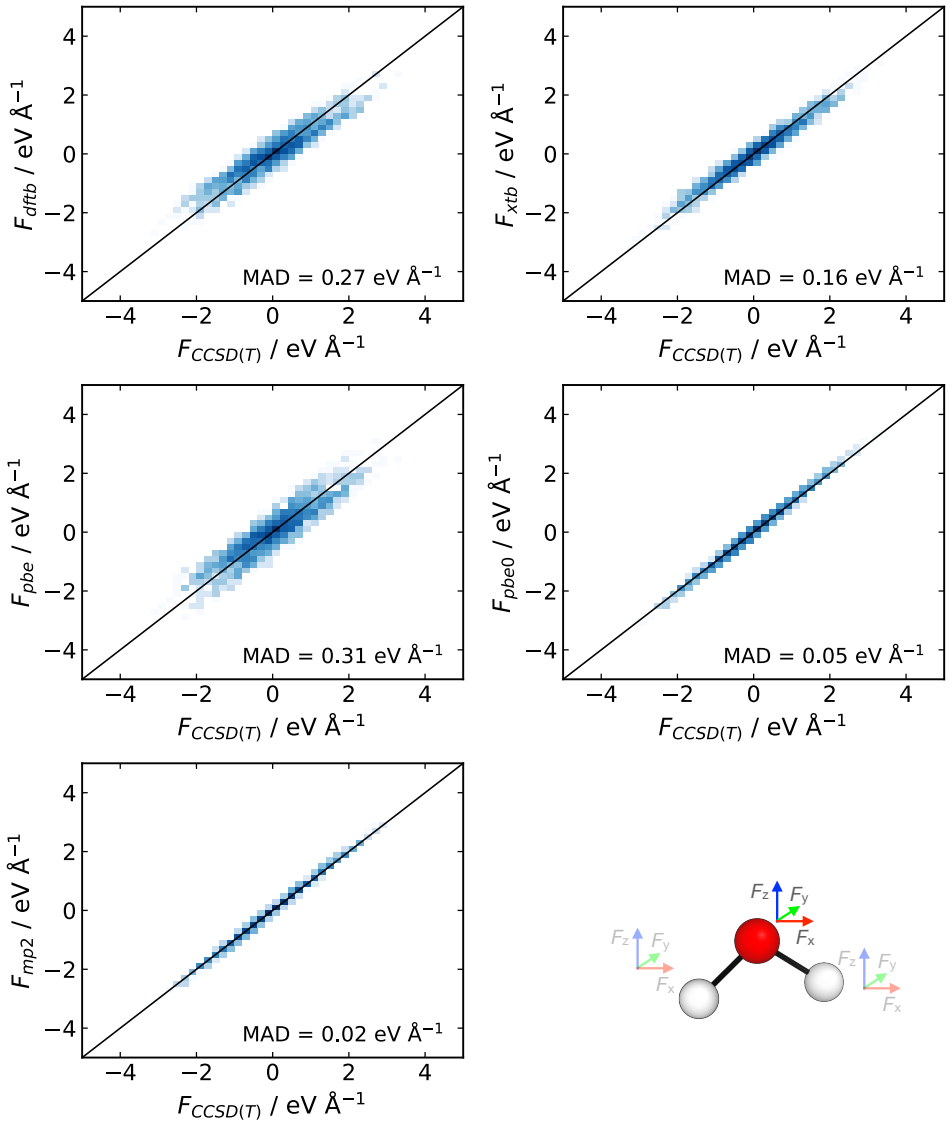

Figure S12: Parity plots of the force components (log density colouring) encountered in a DFTB(3ob) MD simulation (500 K, δt = 0.5 ps, 1 ps, 10 step print interval) of a gas phase water molecule at a QM method compared to CCSD(T)/de2-TZVP reference values. Methods: dftb ≡ DFTB(3ob), xtb ≡ GFN2-XTB, pbe ≡ PBE/def2-SVP, PBE0 ≡ PBE0/def2-SVP, mp2 ≡ MP2/def2-TZVP. Dispersion is not expected to alter the forces.

[c] $3N_{\text{atoms}}$ single points at the minimum and $6N_{\text{atoms}}$ by default in ORCA v. 4.2.1, which uses the central difference approximation to the numerical gradient.

Generating frames using a DFTB(3ob) MD simulation and evaluating the force components at several levels of theory suggests that using hybrid DFT or MP2 forces in combination with CCSD(T) energies should be sufficient to train a CCSD(T)-quality GAP. The average error is less than or similar to an 'expected error' in the forces that optimises the active learning rate ($\sigma_F =< 0.1$ eV Å^{-1}, Figure S12).

S4.1 Methane

Training a GAP on CCSD(T) energies and MP2 forces for a gas-phase methane molecule is sufficient to generate a GAP within 1 kcal mol^{-1} of the true CC surface. Using active learning at XTB to sample the configuration space, MP2 forces and CC energies, highly accurate methane dynamics (Figure S13) can be propagated in just 5 minutes of training time (10 cores).

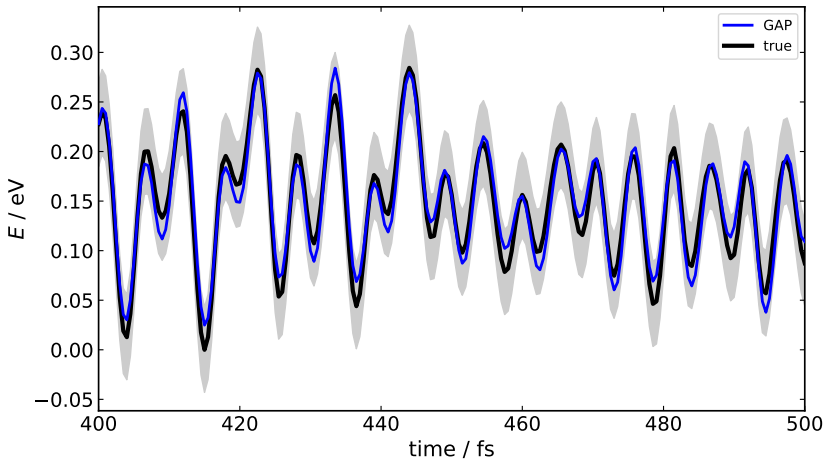

Figure S13: GAP predicted energies compared to true CCSD(T)/def2-TZVP values for a GAP trained using GFN2-XTB active learning, MP2/def2-TZVP energy and force evaluations then CCSD(T)/def2-TZVP single point energies on those configurations. Methane AL at 1000 K, GAP-MD at 500 K with a single frame printing interval. The shaded region bounds the 1 kcal mol^{-1} area of accuracy. Hyperparameters as Table S1.

S4.2 Methanol

Generating a GAP using active learning (AL) in parallel can lead to an overcomplete set of configurations. This is because in the initial stages the independent MD can sample similar regions of configuration space, which in turn do not provide much information for the fit. Selecting the most diverse set of configurations is therefore advantageous when the energy and forces are going to be reevaluated. Training a GAP for methanol using XTB AL then evaluating MP2/def2-TZVP energies and gradients using different number of CUR[4] selected configurations on the SOAP kernel matrix leads a reasonably stable τ_{acc} with the number selected (Table S3).

N_{configs}	τ_{acc} / fs
190	1000 ± 0
160	1000 ± 0
130	1000 ± 0
100	1000 ± 0
70	1000 ± 0
40	100 ± 30

Table S3: GAP accuracy on CUR selected configurations generated as Figure S13 for methanol values are averages over three τ_{acc} evaluations and errors in standard error of the mean.

S4.3 Acetic acid

Employing both mixed energy and forces and CUR selection of the configurations allows dynamics broadly within (MAD = 0.7 kcal mol^{-1}) chemical accuracy to the ground truth CC using only 167 configurations. For comparison this is $\sim$ 100 times fewer CC calculations than would have been required using numerical gradients on the whole set of configurations.[d]

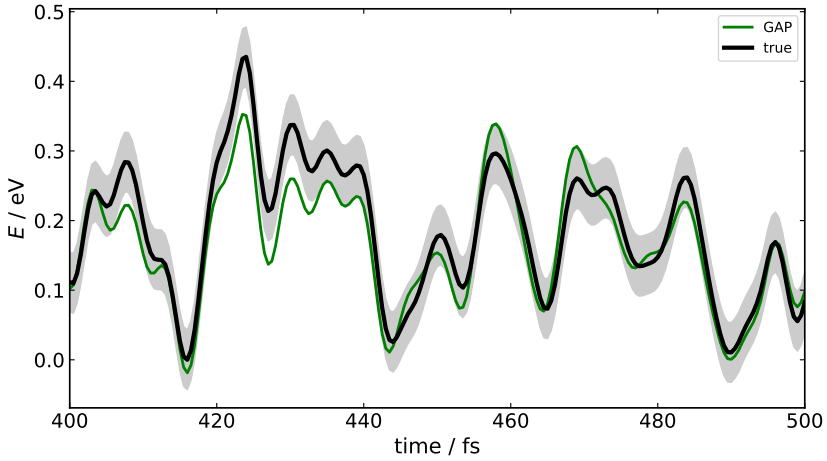

Figure S14: As Figure S13 for acetic acid using a 50% CUR selection of configurations. Hyperparameters as Table S1.

[d]334 configurations × 3 Cartesian coordinates × 2 for central differences × 8 atoms = 16,032 total calculations.

S5 Ethene + Butadiene

S5.1 Sampling Methodology

We have shown previously that the AL sampling method (e.g. DFT functional) is important in obtaining accurate uplifted GAPs,[1] where the energy/forces of generated configurations are recalculated at a higher level. For the [4+2] Diels-Alder reaction between ethene and butadiene, sampling using a cheaper method than PBE does not seem to be suitable over the intrinsic reaction coordinate (IRC, Figure S15a). Although the thermodynamics at PBE are closer to the reference MP2 values, kinetics at PBE0 are slightly closer (Figure S15b). Sampling is therefore performed with the more transferable PBE0 functional.

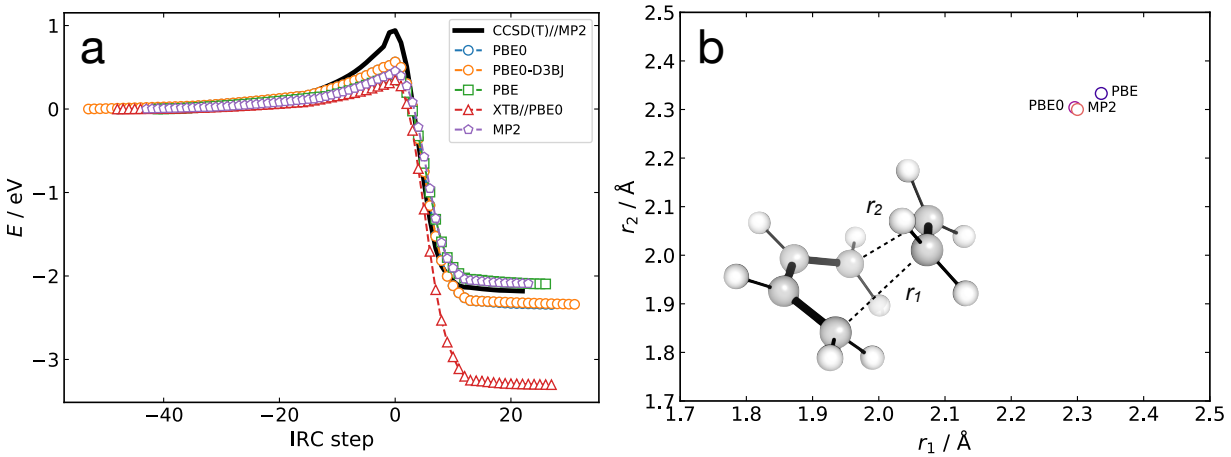

Figure S15: Intrinsic reaction coordinates for ethene+butadiene [4+2] cyclisation at different levels of theory. DFT use def2-SVP basis sets, MP2 a def2-TZVP basis and XTB//PBE0 corresponds to GNF2-XTB energies on PBE0 IRC geometries. All TSs are confirmed as such by the presence of a single imaginary mode.

Instead of using the difference between reference and predicted energies as the criteria for adding new configurations in the AL loop the GP variance may also be used.[1] For this system, using a threshold around 2×10^{-5} eV atom^{-1} samples in a similar region of the energy space as using a 0.09 eV (2 kcal mol^{-1}) difference threshold (Figure S16), but at a much lower computational cost.[e] Therefore, a 'gp_var' sampling strategy will be used for maximum training efficiency.

[e] Using a 'diff' strategy requires evaluating $|E_0 - E_{\mathrm{GAP}}|$ at intermediate points in the GAP-MD trajectory, which are avoided using 'gp_var'.

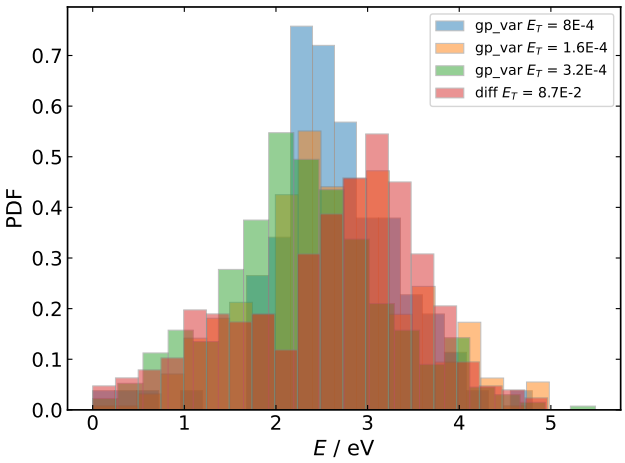

Figure S16: Normalised probability density function over the energies sampled using AL from the ethene+butadiene [4+2] TS. gp_var = x corresponds to running GAP-MD at 500 K until the predicted variance is above x eV for the whole system. diff= y eV corresponds to running GAP-MD at 500 K until $|E_0 = E_{\mathrm{GAP}}| > y$.

S5.2 Conformational Rearrangement

Upon cyclisation, the formed cyclohexene molecule undergoes conformational rearrangement to the more stable (and $> 99\%$ populated) chair conformer. While often assumed to be fast, in cyclohexane solution the interconversion at room temperature only occurs at a rate of 55 s^{-1}.[5] An MM forcefield (GAFF) affords rapid interconversion between the conformers, while the GAP (trained at 1200 K on cyclohexene) interconverts more slowly (Figure S17). The MM however does afford the qualitatively correct distribution (Figure S18), with the half-chair favoured over the half-boat. Developing potentials which allow for conformational changes without reparametrisation is necessary if accurate long-time simulations are to be possible.

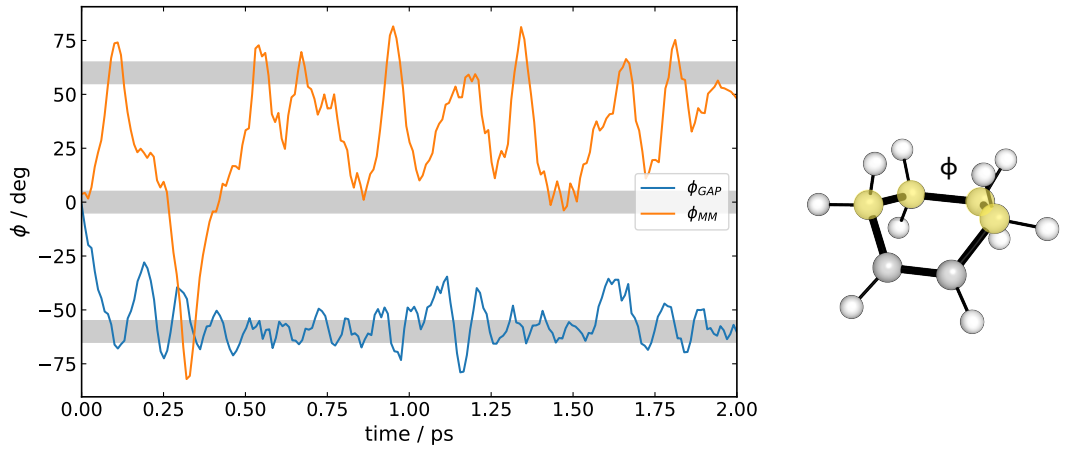

Figure S17: Sample trajectories of cyclohexene at 900 K (0.5 fs timestep) using GAFF-parametrised (RESP charges) and GAP-AL generated potentials (latter uplifted from $\sim$ 500 GFN2-XTB AL configurations to PBE0 CUR selected to 50%). Shaded areas highlight the approximate regions of the half-chair, boat and chair conformations of cyclohexene. Plotted using a 5 fs block average.

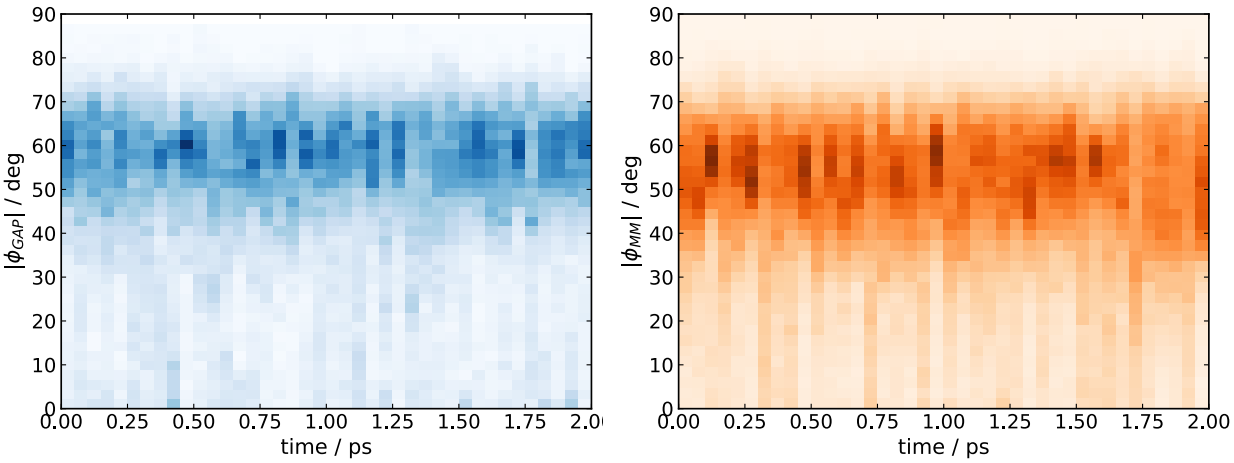

Figure S18: Histogram of absolute dihedral angles over 100 trajectories of cyclohexene, as Figure S17.

S5.3 Fragmentation

With a view to reducing the number of evaluations required to generate a chemically accurate potential for ethene+butadiene cyclisation we explored an additive approach to obtaining a potential. Training the intermolecular interaction between components is an effective fragmentation of a condensed phase molecular[1] or solid state system.[6] Extending this approach to covalent bonds however did not lead to a gain in accuracy for a specific number of configurations (i.e. $\tau_{\text{acc}}\sim$ 100 fs cf. $\tau_{\text{acc}}\sim$ 200 fs using the same 'gp_var' selection strategy).

Fragmentation over a covalent bond in this manner produces two limitations (1) the intra GAP used to train the fragments may never have encountered the configuration adopted in the full system and (2) the energy scale over which the inter GAP must be accurate is much larger than without an I+I decomposition. Specifically for this system, even if the ethene and butadiene potentials are high quality ($\tau_{\text{acc}} > 1$ ps, Figure S20) they will not have well sampled the pyramidal sp^3 geometrics present in the cyclohexene product. This increases the amount of data required for the inter GAP to learn. Furthermore, the intra component distortion is only possible in the product making the energy scale larger that otherwise required (Figure S16). The one advantage is that the reactant state (reactants separated by $>$ 3Å) is well approximated by the intra+inter decomposition, as observed for e.g. bulk water.

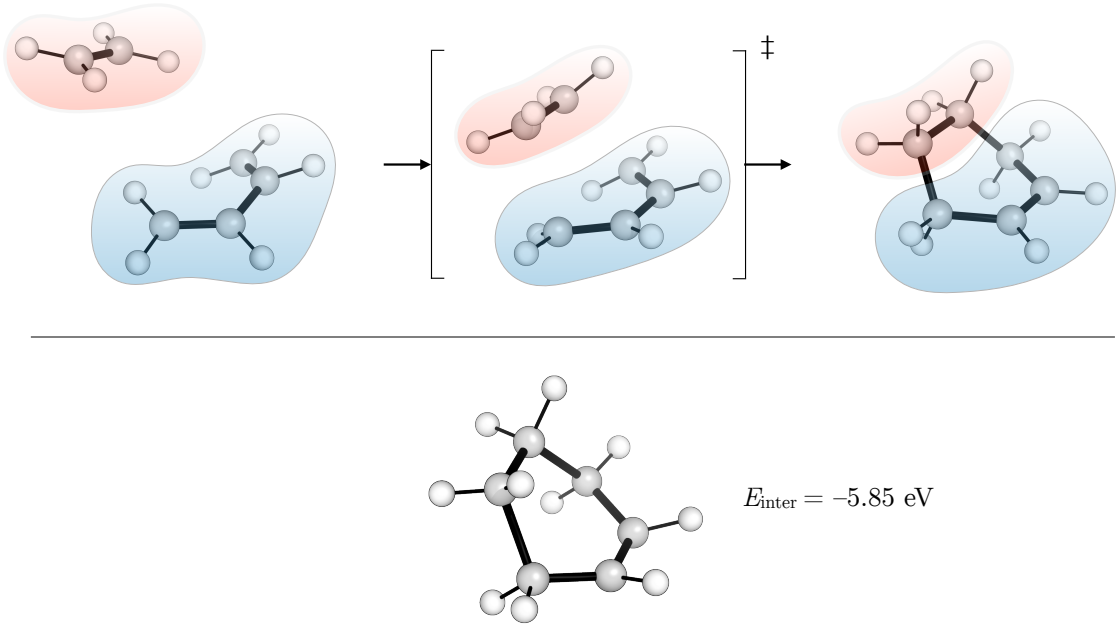

Figure S19: Sample intra+inter fragmentation strategy for ethene+butadiene. An example configuration is highlighted in the bottom panel along with the residual interaction energy between the two fragments (i.e. true total energy minus $E_{\text{intra}}^{\text{GAP}}$) for each component.

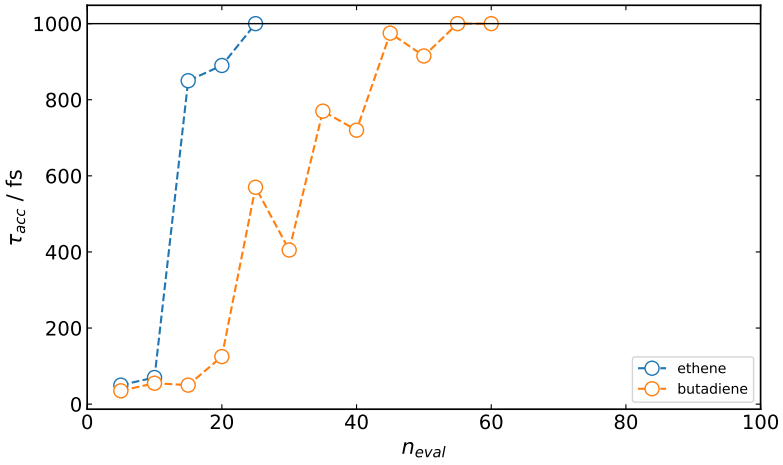

Figure S20: Learning curves for ethene and butadiene using a 'gp_var' selection strategy with $E_t = 1 \times 10^{-5}$ eV atom^{-1}. τ_{acc} uses $E_l = 1$ kcal mol^{-1}, $E_T = 10E_l$, 25 fs interval and a maximum time of 1 ps. n_{eval} is the total number of reference PBE0/def2-SVP calculations performed.

S5.4 Reaction Training

Training a GAP for the whole ethene+butadiene reaction with the optimised hyperparameters (Table S2) from the TS using a 'gp_var' selection strategy generated a MLP capable of a $\tau_{\text{acc}} \sim 500$ fs at 1 kcal mol^{-1}. A representative trajectory back and forwards from the TS illustrates the achieved accuracy (Figure S21) available in just 30 CPUh.

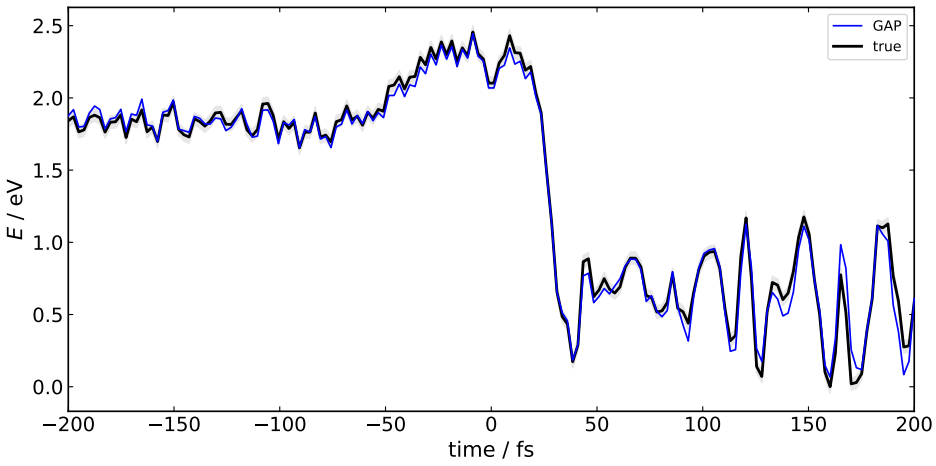

Figure S21: Comparison of true (PBE0/def2-SVP) and GAP predicted energies over one trajectory propagated from the TS to reactants ($t < 0$) and products ($t > 0$). GAP trained using 'gp_var' ($E_t = 2 \times 10^{-5}$ eV Å^{-1}, 300 K) and contained 181 configurations.

S6 Atomic Energy Errors

In addition to the complexity increasing with system size, the accuracy *per atom* increases when the goal is to generate a potential that is accurate to 1 kcal mol^{-1} in total energy. To evaluate if the total (relative) energy is an important quantity for our target properties (free energies or reaction dynamics) we use the set of S_N2 reactions: Cl^- + {MeCl, EtCl, ^{n}PrCl}.

Generating highly accurate (error $\ll$ 1 kcal mol^{-1}) GAPs for the reaction is possible (Figure S22). Using these potentials and increasing the regularisation ('expected error') on energies *per atom* leads to larger errors on the potential energy barriers (Figure S23). Note that these GAPs are trained purely on energies, to isolate the effect of adding larger atomic errors. This scenario simulates training different potentials to the same *per atom* accuracy, which – as expected – leads to larger errors on barriers for these small systems. Based on these data the total energy is an important quantity in predicting free energy barriers.

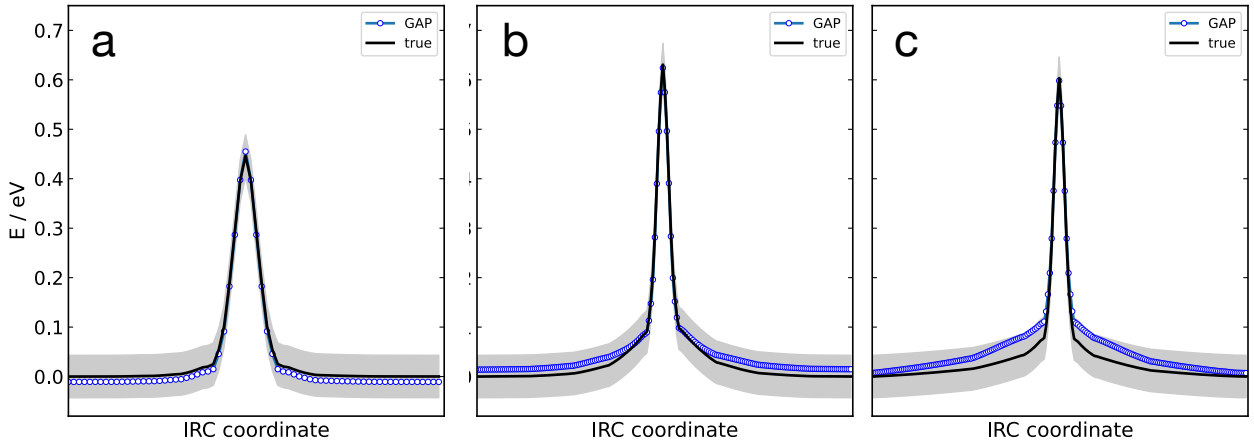

Figure S22: True intrinsic reaction coordinates (IRCs) for Cl^- + MeCl, EtCl, ^{n}PrCl (a–c respectively), calculated at PBE0-D3BJ/ma-def2-SVP with GAP predicted values overlaid. GAPs trained at 500 K from their respective TSs up to a maximum time of 0.5 ps using a 'gp_var' strategy ($E_t = 1 \times 10^{-5}$ eV atom^{-1}. $\sigma_E = 10^{-3.5}$ eV atom^{-1}, $\sigma_F = 10^{-1.5}$ eV Å^{-1}). The shaded area bounds the 1 kcal mol^{-1} region of accuracy.

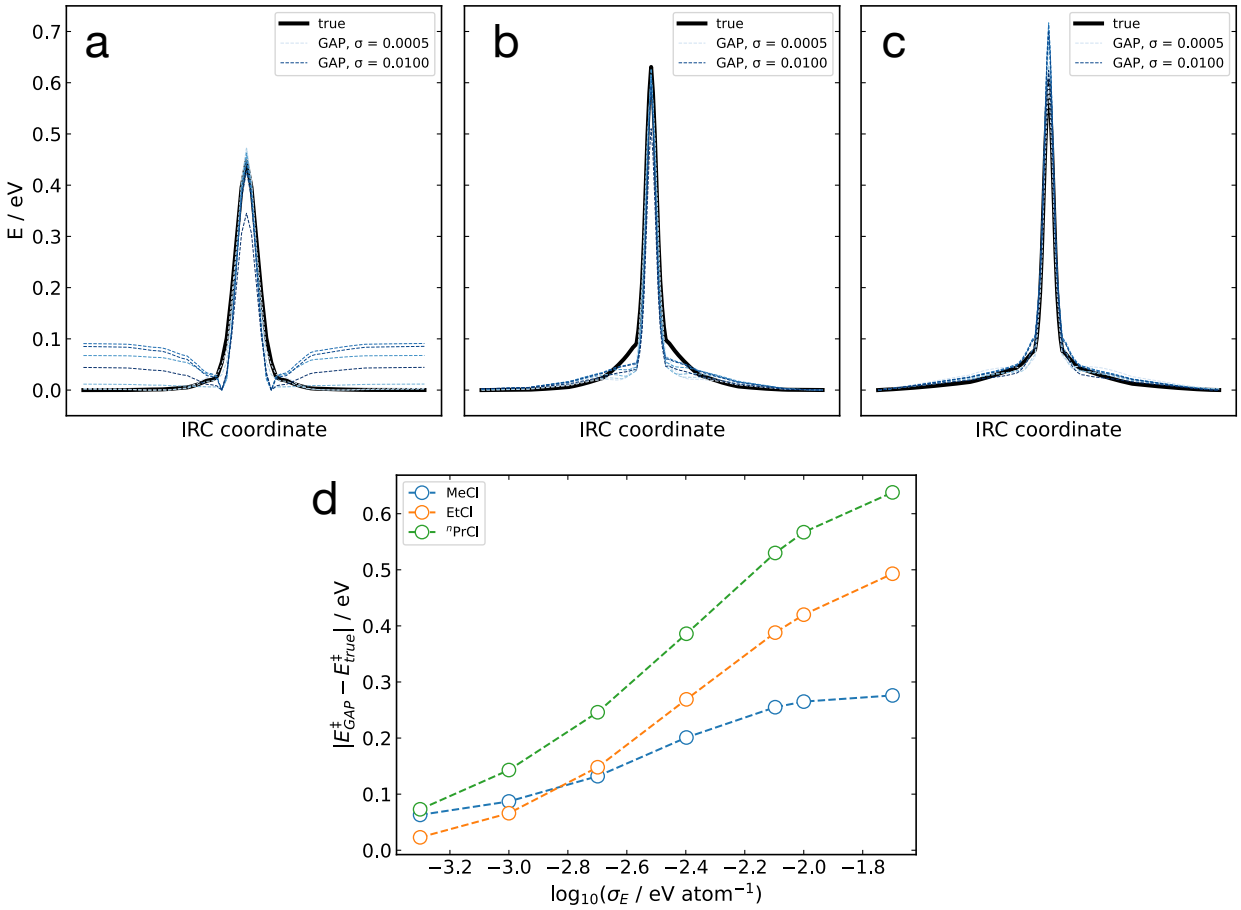

Figure S23: True and predicted IRCs (training data identical Figure S22) with GAPs trained on energies only with different σ_E values for MeCl, EtCl, ^{n}PrCl (a–c respectively). Absolute error on the TS energy for each GAP is plotted in (d).

S7 Method Comparison

Training different MLP methods on the ethene+butadiene reaction using active learning with a selection criteria of 0.1 eV[f] affords highly accurate potentials in all cases (Figure S24, MAD $\sim$ 0.04 eV, 1 kcal mol^{-1}). The data requirement of the GAP (406 configurations) was significantly higher than both ACE (114) and NeQUIP (126) potentials and required significant hyper-parameter tuning (S3). The total training time on 10 CPU cores was 7, 4 and 14 hours for GAP, ACE and NeQUIP potentials respectively, the latter also utilised an Nvidia RTX 2080 GPU.

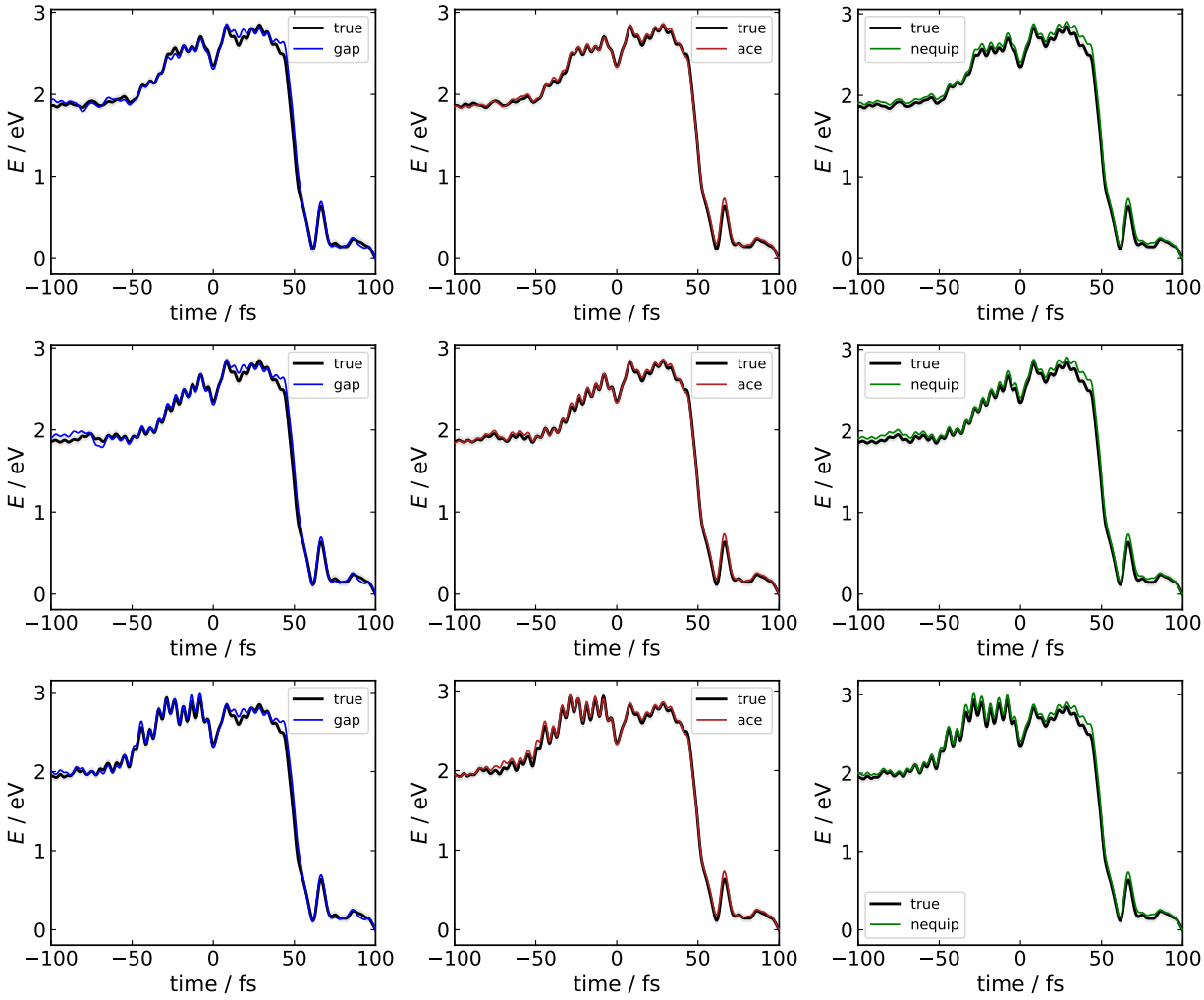

Figure S24: Comparison of MLP methods trained using an identical AL strategy from the TS of ethene+butadiene to 'ground truth' AIMD data. Both the training and AIMD used the PBE0/def2-SVP level of DFT theory. Dynamics propagated from the TS at 300 K using a Berendsen thermostat, as implemented in ORCA v. 4.2.1. Trajectories are stitched from two that proceeded forwards and backwards.

[f]If $|E_{\mathrm{MLP}} - E_{\mathrm{true}}| > 0.1$ eV then the configuration is selected in MLP-driven MD, propagated at 500 K with a 0.5 fs time step. AL is halted if 10 trajectories reach the maximum MD time of 500 fs.

Evaluating the potentials on the intrinsic reaction coordinate again each perform comparably (Figure S25), and all provide smooth 2D surfaces (Figure S26). The latter is in spite of the extrapolation within high-energy regions. Interpolating the surface

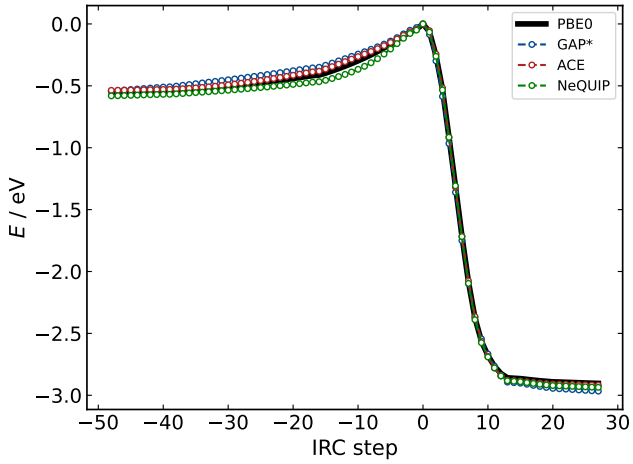

Figure S25: Comparison of MLP methods (as Figure S24) on the PBE0/def2-SVP intrinsic reaction coordinate (IRC).

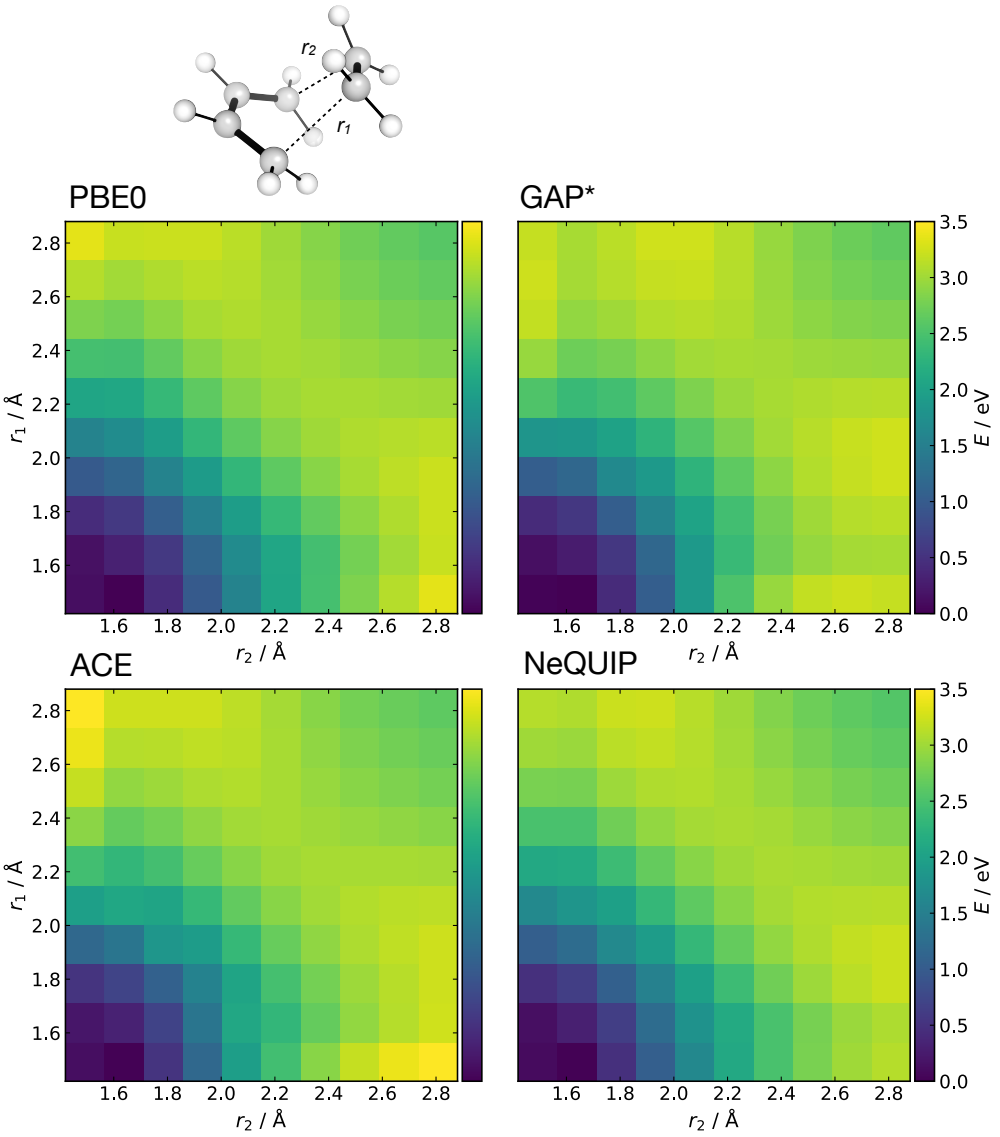

Figure S26: Comparison of MLP methods (as Figure S24) on the PBE0/def2-SVP 2D relaxed PES.

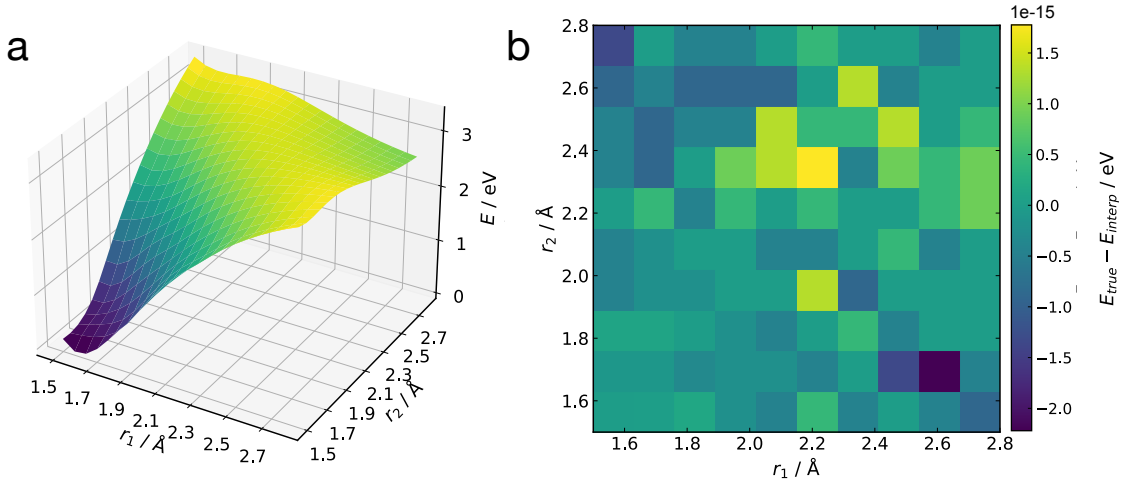

Figure S27: (a) Interpolated 2D surface (PBE0/def2-SVP 2D relaxed PES, as Figure S26) and (b) residuals between the interpolated and true values. Interpolation performed with *RectBivariateSpline* from *SciPy*.

S8 Active Learning Selection Strategies

Using the predicted GP variance on a new configuration can be a highly effective selection strategy for sampling new configurations (see e.g. Figure S20). However, when training other kinds of MLP there may be no analogue to accelerate the 'diff' selection strategy.[g] Using a threshold on the maximum distance ('max_dist') to any of the training set can afford a 10× speed-up in training for non-GAP MLPs where the reference evaluations dominate the execution time. Specifically, using a selection criteria defined by,

$$\max(\boldsymbol{k}^*) < k_T \quad : \quad \boldsymbol{k}^* = ((\boldsymbol{p}_0 \cdot \boldsymbol{p}^*)^\zeta, \cdots) \tag{S8.0.1}$$

where $\boldsymbol{p}_i$ is the normalised SOAP vector for the i-th configuration in the training data and ζ is an positive power to sharpen differences. This is exactly the form of the kernel used in our GAPs and can provide a quantification of the similarity between one molecular configuration and another. With an appropriately chosen k_T, potentials can be trained efficiently (Figure S28) despite not being correlated with the absolute energy difference (Figure S29).

Using this strategy it is essential to backtrack until $\max(\boldsymbol{k}^*)$ is not *too* small, to prevent high-energy structures (or SCF convergence failures) making their way into the training data. We found an upper threshold of $(k_T)^2$ to be sufficient without much tuning and $\zeta = 8$ to be optimal.

Direct use of this strategy to train a ACE potential for the DA/cope reactions between tropone and cycloheptatriene did afford a reasonable potential, but during AL training there was no sampling of one of the DA products.

[g] Using $|E_{\text{true}} - E_{\text{MLP}}|$ and evaluating potentially 8 DFT evaluations with no selected configuration, for a 1 ps max time approaching the end of AL cycle.

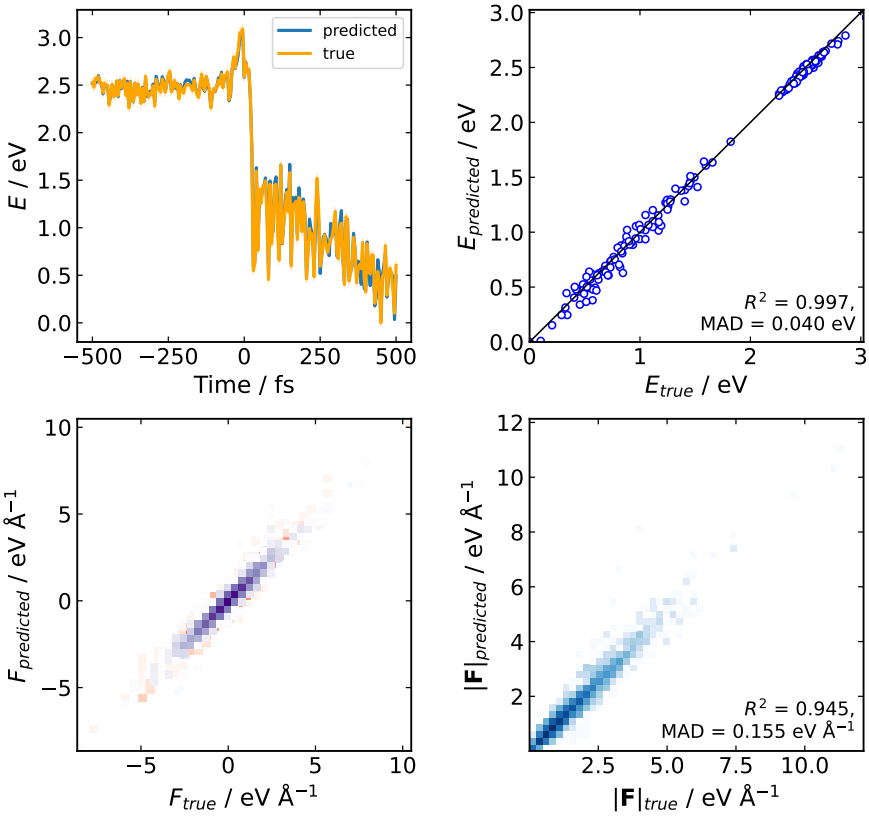

Figure S28: Comparison of true (PBE0/def2-SVP) and predicted (ACE) energies and forces over a 500 fs ACE-propagated trajectory from the TS (300 K, $\delta t = 0.5$ fs). ACE potential trained using a 'max_dist' AL strategy ($k_T = 0.999$), which generated 104 configurations.

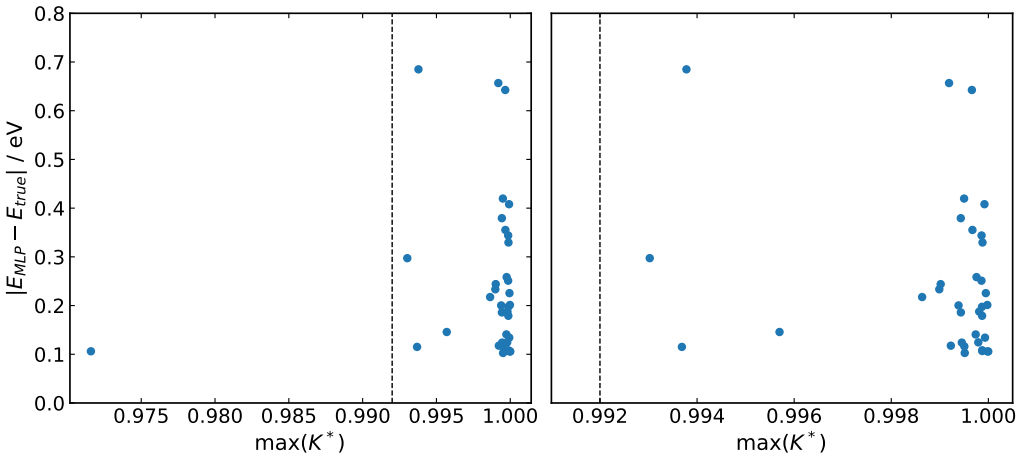

Figure S29: Correlation of maximum SOAP kernel vector values with true differences, showing a no expected negative correlation. Frames selected over a 5000 K active learning trajectory of methane, using GAP MLP.

S9 Method Effects on Product Distributions

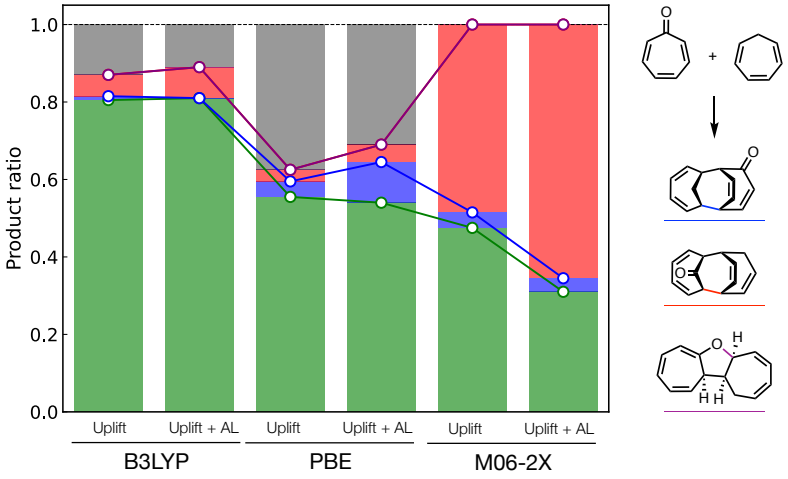

Figure S30: Product distribution generated from 200 classical NVE trajectories propagated from 'the' TS at B3LYP, PBE and M06-2X levels of theory (def2-SVP) using initial velocities for 300 K for 1 ps using a 0.5 fs time step. Colours correspond to the products, with green the reactant state and grey no defined state being formed in 1 ps. Uplifted corresponds to single point energy and force evaluations on ACE AL configurations (PBE0/def2-SVP) then retraining the ACE potential. Uplift+AL uses PBE0 configurations reevaluated plus 5 iterations of active learning at 500 K from the appropriate TS.

References

(1) Young, T. A.; Johnston-Wood, T.; Deringer, V. L.; Duarte, F. *Chemical Science* **2021**, *12*, 10944–10955.

(2) Bertz, S. H. *Journal of the American Chemical Society* **1981**, *103*, 3599–3601.

(3) Rowe, P.; Deringer, V. L.; Gasparotto, P.; Csányi, G.; Michaelides, A. **2020**, *153*, 034702.

(4) Mahoney, M. W.; Drineas, P. *Proceedings of the National Academy of Sciences* **2009**, *106*, 697–702.

(5) Jensen, F. R.; Noyce, D. S.; Sederholm, C. H.; Berlin, A. J. *Journal of the American Chemical Society* **1962**, *84*, 386–389.

(6) Wengert, S.; Csányi, G.; Reuter, K.; Margraf, J. T. *Chemical Science* **2021**, *12*, 4536–4546.