LEARNING TRANSITION-METAL COMPLEXES REACTION BARRIERS WITH ACTIVE LEARNING AND MACHINE LEARNING INTERATOMIC POTENTIAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Fine-tuning a machine learning interatomic potential (MLIP) to accurately model a chemical reaction needs training data along the reaction path. However, that path is exactly what is unknown beforehand. Here we demonstrate a method to efficiently build such a dataset from the only knowledge of the reactant and product. We show that fine-tuning only works when this method is robust. We use a combination of active learning and a custom path generator. This generator is named *curved reactive scan*, in which each reactive bond follows its own power law, feeding a certified saddle search whose by-products constitute the training set. As a result these data points sit on the reaction path. We show that the fine-tuning is data efficient: for 25 transition-metal complex reactions from the benchmark MOBH35 (15–67 atoms), we generate 86 to 284 density functional theory (ωB97M-V/def2-TZVPD) labels per reaction, and the fine-tuned versions of UMA S 1.2 improve the retrieving of their transition states, forward and reverse energy barriers, at chemical accuracy on average. Those results outperform the pretrained model, even though it was trained on OMol25, a data set that contains reactive configurations generated from these very MOBH35 systems. Our work shows that blind generation of molecular geometries along the reaction path is a crucial step towards efficient MLIP fine-tuning, to potentially unlock applications in metalloprotein studies, and transition metal catalyst discovery.

1 INTRODUCTION

Within the Born–Oppenheimer approximation (Born & Oppenheimer, 1927), the energy of a molecule is obtained by solving the electronic Schrödinger equation at fixed nuclear positions. Repeating this for every geometry $\mathbf{x} \in \mathbb{R}^{3N}$, with N the number of atoms, defines the potential energy surface (PES) $E(\mathbf{x})$, whose negative gradient gives the forces acting on the nuclei. The PES underlies most of computational chemistry: its minima and first-order saddle points give reaction energies and barriers, and sampling it with molecular dynamics (MD) gives access to free energies and rate constants. The bottleneck is the electronic structure calculation itself. The most widely used electronic structure method is density functional theory (DFT) (Hohenberg & Kohn, 1964; Kohn & Sham, 1965), which replaces the many-electron wavefunction by the electron density and maps the problem onto a system of non-interacting electrons, folding all many-body effects into an approximate exchange–correlation functional. Modern functionals typically reach a few kcal/mol on reaction energies and barriers (Mardirossian & Head-Gordon, 2017) at a cost scaling as $\mathcal{O}(\mathcal{N}^3)$ to $\mathcal{O}(\mathcal{N}^4)$ with system size. This compromise has made DFT the default workhorse for decades, while more accurate wavefunction methods remain limited to small systems or single-point energies. Even so, DFT costs far too much for the 10^5–10^7 evaluations required by MD or systematic reaction-path searches. Machine learning interatomic potentials (MLIPs) (Behler & Parrinello, 2007) remove this bottleneck. Trained on DFT energies and forces, they reproduce the reference PES at a small fraction of its cost, so that the price of accuracy is paid once, when generating the training data, rather than at every evaluation.

However, even with the current best foundational models (Wehrhan et al., 2025) such as UMA (Wood et al., 2025) or MACE (Batatia et al., 2022), some corners of the chemical space are not captured well enough. This is not surprising given the size of the space to cover. Counting molecules alone gives an

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

estimated $\sim 10^{60}$ drug-like compounds (Bohacek et al., 1996), and $\sim 10^{11}$ organic molecules with at most 17 heavy atoms (Ruddigkeit et al., 2012). Each molecule then comes with a $3N$-dimensional configuration space and a choice of charge and spin state. No dataset can cover this uniformly, and the largest current effort, OMol25 (Levine et al., 2025), makes deliberate choices. It gathers about 10^8 single-point calculations at the ωB97M-V/def2-TZVPD level on molecules of up to 350 atoms and 83 elements. Geometries come mostly from conformer search, docking, and classical or MLIP-driven MD, which concentrates the data near equilibrium. Reactive configurations come from recomputed reaction datasets, dominated by main-group organic chemistry (e.g. Transition1x (Schreiner et al., 2022)), and, for transition metal complexes (TMCs), from AFIR paths (Maeda et al., 2016) built on the reactions of MOR41, ROST61 and MOBH35 after swapping metals and ligands. AFIR pushes reactive atoms together or apart with an artificial force, so the resulting path is only a guess of the minimum energy path, and nothing guarantees that structures close to the true saddle point are sampled. TMCs add their own difficulties: several accessible spin and oxidation states, near-degenerate electronic configurations, and a combinatorial number of metal–ligand combinations, each sparsely represented. Saddle points of TMCs thus remain among the least well described regions of the PES (Marks et al., 2026).

In this work, we ask how to build a small dataset that teaches a foundation MLIP the reactive region of a TMC reaction, starting only from its reactant (R) and product (P). Our test bed is the MOBH35 benchmark (Iron & Janes, 2019; Semidalas & Martin, 2022), which provides R, P and transition-state (TS) geometries for 35 reactions of TMCs, of which we use 25 (§3). The pretrained model is UMA-S-1.2 (Wood et al., 2025). As explained above, its training set, OMol25, contains reactive structures derived from these very reactions, so the chemistry is familiar to the model. Yet, evaluated on the published geometries, UMA misses chemical accuracy (1 kcal/mol) on the forward or backward barrier of 13 of the 25 reactions (§5.1). Improving it raises a chicken-and-egg problem: the useful training structures lie along a reaction path that is not known in advance, and the only tool available to find this path is the pretrained MLIP itself. Labelling structures that lie too high in energy or far from the path wastes the DFT budget and can degrade the model where it matters. We therefore build the dataset by active learning: each round generates candidate structures with the current MLIP, labels a small selected batch with DFT, and fine-tunes. As the model improves, so does the path it generates.

Contributions. (i) A path generator, the *curved reactive scan*, which needs only R and P: each reactive bond advances along its own power law, and a certified saddle search followed by a pseudo-intrinsic reaction coordinate (IRC) turns the scan into training data (§4.1). (ii) A fully automated active-learning pipeline combining this generator with energy filters, batch selection in the MLIP's own feature space, DFT labelling and fine-tuning (§4.2). (iii) Two metrics that separate the accuracy of the energy (barriers on fixed, published geometries) from the accuracy of the TS search (barriers between the stationary points the model finds by itself). The first does not guarantee the second: for several reactions, the fine-tuned model reproduces the barriers at published geometries, but its search converges to a different, higher saddle point (§5.1). (iv) Results on 25 reactions (15–67 atoms, 13 metals): with 86 to 284 DFT labels per reaction, the mean barrier error on the model's own stationary points drops from 1.52/1.66 to 0.87/0.89 kcal/mol (forward/backward), and on the published geometries from 1.10/1.19 to 0.39/0.37 kcal/mol. All datasets and analysis scripts are released. (v) A controlled comparison of the curved reactive scan with climbing-image nudged elastic band (NEB), the standard generator: both are run inside the same loop, on five reactions at a fixed budget of 160 labels (§5.2). The scan reaches chemical accuracy on all five reactions with fewer labels, NEB on four only due to a weaker harvesting of the important chemical structures.

2 Related work

Transition-state search. Locating the saddle point that connects two known minima is a classic problem, reviewed by Dewyer et al. (2018) and, for machine learning approaches, by Wang et al. (2025). Double-ended methods relax a discretised path between R and P: the NEB and its climbing-image variant (Jónsson et al., 1998; Henkelman et al., 2000), the string method (E et al., 2002; 2007) and the growing string method (Peters et al., 2004), all sensitive to the initial interpolation (Smidstrup et al., 2014; Zhu et al., 2019). Local saddle-point optimisers instead refine a single starting structure by climbing along the lowest-curvature mode while minimising along all others, as in the dimer

method (Henkelman & Jónsson, 1999) and Sella (Hermes et al., 2022), which we use. They converge only from a good starting TS guess. Among them, we compare our path generator against NEB: it is the most widely used TS search method, the one used to build reactive MLIP datasets such as Transition1x (Schreiner et al., 2022), and it takes exactly the same input as our scan, the reactant and product alone, so the two can be swapped inside the same active learning loop with everything else held fixed.

Driven and biased path generation. A second family obtains TS guesses by driving the reaction explicitly: rather than relaxing a path or a guess on the unbiased PES, it adds a bias on a few chosen reactive coordinates, hypothesised from the reactant or read from a known product. The bias takes one of two forms. Coordinate-driving methods impose hard constraints: the driven distances are fixed at prescribed values while all other degrees of freedom are minimised, and the energy maximum along the constrained path, or the saddle of a constrained grid, serves as the guess. In MD/CD (Yang et al., 2017), the driven atom pairs are selected in reactant conformers sampled by molecular dynamics; in autodE (Young et al., 2021), the forming and breaking bonds are read from the atom-mapped reactant and product, and one or two of them are scanned on a full grid; and pyMCD (Lee et al., 2023) drives several internal coordinates simultaneously towards target values. The artificial force induced reaction (AFIR) method (Maeda & Morokuma, 2010; Maeda et al., 2016) instead applies a constant force, a bias linear in a weighted inter-fragment distance; it needs only the reactant and leaves the order of bond events to the PES. Our curved reactive scan differs from both on two points. First, like autodE, it reads all M reactive bonds from R and P, but it replaces the grid, whose size grows exponentially with M, by a few one-dimensional paths: forming and breaking bonds advance on different power-law schedules λ^{p_m}, scanned in both directions, which sample asynchronous mechanisms at a cost independent of M. Second, it uses harmonic restraints, which interpolate between the two forms above: they reduce to hard constraints for $K \to \infty$ and to a constant force for $K \to 0$ at fixed Kt_m. Like constraints, and unlike a constant force, they control positions rather than forces. This matters at the barrier top: a constant force adds no curvature along the stretched bonds, so no image can rest in the negative-curvature region around the saddle, which the optimiser crosses in a few steps, whereas a restraint stiffer than this negative curvature keeps every image a stable minimum (Appendix A). In all these methods, the guesses are then refined by a saddle optimiser and checked for connectivity to R and P.

Reactive data and MLIPs for barriers. Reactive datasets for MLIPs remain dominated by main-group organic chemistry: Transition1x (Schreiner et al., 2022) collects NEB paths for about 10^4 reactions of C, H, N and O, and RGD1 (Zhao et al., 2023) enumerates about 1.8×10^5 reactions of the same elements. OMol25 (Levine et al., 2025) covers transition-metal complexes, but its reactive structures come from AFIR paths on metal- and ligand-swapped variants of benchmark reactions, which sample the saddle region only sparsely. The closest works to ours build reactive datasets from GFN2-xTB transition states and IRC paths generated from reaction templates, add normal-mode sampling around a few points of each path, label with DFT and train MACE models, for Rh-catalysed hydrogenation (Tang et al., 2025) and for organic reactions of ten main-group elements with a Δ-learning correction to GFN2-xTB (Ren et al., 2026). Both show that IRC points alone, where forces nearly vanish, are insufficient for MLIP-driven transition-state optimisation, which our thermal perturbations also address. Their pipelines, however, require the semi-empirical method to locate the correct saddle for every reaction before any DFT label is computed, the step least reliable for transition metals (Ren et al., 2026). We instead bootstrap the path from the foundation model and correct it with DFT.

3 Problem setup

Benchmark. We use 25 of the 35 MOBH35 reactions, containing molecules of 15 to 67 atoms, spanning metals atoms Sc, Ti, Fe, Nb, Mo, Ru, Rh, Pd, Ta, W, Os, Ir and Pt, in a single automated campaign. All labels are ωB97M-V/def2-TZVPD (Mardirossian & Head-Gordon, 2016; Weigend & Ahlrichs, 2005) from `gpu4pyscf` (Wu et al., 2025), energies and forces. For each reaction the P, R and TS geometries were optimised using a TPSS-D3(BJ)/def2-SVP (Iron & Janes, 2019). We also re-optimized all geometries with ωB97M-V/def2-TZVPD for consistency.

The MLIP An MLIP takes as input the cartesian positions of N atoms and predicts the energy of the system. It can be trained or fine-tuned using a dataset $\{x_i, [E(x_i), F(x_i)]\}$, where $x_i \in \mathcal{R}^{3N}$ is the *geometry* of the molecule, and $E(\cdot)$ and $F(\cdot)$ are energy and forces computed with an algorithm (eg. DFT). We use UMA-S-1.2 (task `omol`) (Wood et al., 2025). It is an SO(3)-equivariant message-passing network with a mixture-of-linear experts conditioned on task, charge and spin. It sums atomic contributions read from the final node embedding, whose invariant (L=0) channels form the 128-dimensional vector our acquisition rule reuses (§4.2), and takes forces as exact gradients $F_i = -\partial E_{\mathrm{MLIP}}/\partial r_i$, so it is *conservative*, without which the curvature the saddle certification reads would be the curvature of no energy. We fine-tune end to end, no frozen layer, pretrained multi-task head kept, AdamW (10^{-4}, weight decay 10^{-3}), cosine schedule with warm-up, gradient clipping at 100.

Reactive bonds For a system of N atoms, we compare R and P and focus on reactive bonds only. We automatically detect a set of M atom pairs, close enough to form a covalent bond on R or P, and whose separation changes between R and P. Across the reactions we studied, M ranged from 1 to 8. For each configuration $x \in \mathbb{R}^{3N}$ of the same molecule, each reactive bond m has a length $d_m(x)$. We measure similarity between two geometries x, y by the RMSD over those bonds only, $d_{\mathrm{react}}(x, y) = [\frac{1}{M}\sum_m (d_m(x) - d_m(y))^2]^{1/2}$ [Å]. In the following, we use d_{react} throughout because the reactive bonds are what the reaction is about and because a all-atoms RMSD would pick the wrong chemical state (a non-reactive functional group rotating would highly change the RMSD).

Error metrics We compare the pretrained and the fine-tuned MLIP on activation barriers. A barrier is a difference of two energies evaluated on two geometries. The energy depends on the method that computes it, $E_{\mathrm{DFT}}(\cdot)$ with our DFT, $E_{\mathrm{pre}}(\cdot)$ with UMA pretrained or $E_{\mathrm{ft}}(\cdot)$ with UMA after final fine-tuning. We write $E_{\mathrm{MLIP}}(\cdot)$ when the statement holds for any MLIP. Each geometry is subscripted by where it comes from:

- $\mathrm{R}_{\mathrm{MOBH35}}$, $\mathrm{TS}_{\mathrm{MOBH35}}$, $\mathrm{P}_{\mathrm{MOBH35}}$, the stationary points published with the benchmark.
- $\mathrm{R}_{\mathrm{DFT}}$, $\mathrm{TS}_{\mathrm{DFT}}$, $\mathrm{P}_{\mathrm{DFT}}$, our own reoptimisation of those at the labelling level, minima (R, P) by L-BFGS and the transition state by Sella (§4.1), so that the gradient actually vanishes on the surface we grade against.
- $\mathrm{R}_{\mathrm{MLIP}}$, $\mathrm{TS}_{\mathrm{MLIP}}$, $\mathrm{P}_{\mathrm{MLIP}}$ for MLIP $\in \{\mathrm{pre}, \mathrm{ft}\}$, what a MLIP finds by itself, the minima (R, P) again by relaxation and the TS by the blind search of §4.

We now define the errors, here written for the forward barrier ($\rightarrow$) and the fine-tuned MLIP (ft), the backward barrier ($\leftarrow$) and the pretrained MLIP (pre) being identical up to substitution:

$$\varepsilon_{\rightarrow,1} = \left[E_{\mathrm{ft}}(\mathrm{TS}_{\mathrm{ft}}) - E_{\mathrm{ft}}(\mathrm{R}_{\mathrm{ft}})\right] - \left[E_{\mathrm{DFT}}(\mathrm{TS}_{\mathrm{DFT}}) - E_{\mathrm{DFT}}(\mathrm{R}_{\mathrm{DFT}})\right] \quad (1)$$

$$\varepsilon_{\rightarrow,2} = \left[E_{\mathrm{ft}}(\mathrm{TS}_{\mathrm{MOBH35}}) - E_{\mathrm{ft}}(\mathrm{R}_{\mathrm{MOBH35}})\right] - \left[E_{\mathrm{DFT}}(\mathrm{TS}_{\mathrm{MOBH35}}) - E_{\mathrm{DFT}}(\mathrm{R}_{\mathrm{MOBH35}})\right] \quad (2)$$

The first metric is quantifying the error of the MLIP energy surface on the geometries found by our method. The second is just comparing the energy prediction on independent geometries. Therefore, if ε_2 is good but ε_1 is bad, it can mean that the MLIP correctly predicts the energy, but that our search found another saddle point than the MOBH35-DFT one.

4 METHOD

Our active learning pipeline repeats rounds of data acquisition and fine-tuning. Each round generates geometries with the current fine-tuned MLIP, filters out high energy ones, selects a few points, labels them in DFT, filters again and fine-tune.

4.1 THE PATH GENERATOR: A CURVED REACTIVE SCAN

A restrained, curved scan. We identify the length of every reactive bond (see definition above) m at R and at P (d_m^{R} and d_m^{P}). We introduce a progress variable λ, worth 0 at the reactant and 1 at the product. We discretize λ into 24 values, λ_k, leading 24 target lengths $t_m(\lambda_k)$ per bond. Each geometry along the path is minimized according to $E(x) = E_{\mathrm{MLIP}}(x) + \frac{K}{2}\sum_m (d_m(x) - t_m(\lambda_k))^2$, with $K = 20\,\mathrm{eV/Å}^2$. The resulting geometry is x_k. Each optimization is started from x_{k-1}. This

elastically pulls the geometry towards the targeted reactive bond length ($t_m(\lambda)$), while minimizing its global energy on all atoms. The harmonic restraint is what lets images sit near the saddle: with K larger than the negative curvature of the barrier top, every image is a stable minimum of the restrained energy (see Appendix A).

Choices of $t_m(\lambda)$: a first choice would be a simple linear interpolation $t_m(\lambda) = d_m^{\mathrm{R}} + \lambda\big(d_m^{\mathrm{P}} - d_m^{\mathrm{R}}\big)$. This places evenly spaced targets between R and P in the same way way for all bonds. However, physical TS do not lye on this linear path (see Fig. 1).

To improve the interpolation path, we introduce a variable p used to raise λ^p. This allows us to move away from the linear division of the path. Moreover we pick a different value p_m for each bond. We simply differentiate two types of bonds to obtain smooth curvatures:

$$t_m(\lambda) = d_m^{\mathrm{R}} + \lambda^{\,p_m}\big(d_m^{\mathrm{P}} - d_m^{\mathrm{R}}\big), \qquad p_m = \begin{cases} p & \text{if bond breaks } (d_m^{\mathrm{P}} > d_m^{\mathrm{R}}) \\ 1/p & \text{if bond forms } (d_m^{\mathrm{P}} < d_m^{\mathrm{R}}), \end{cases} \tag{3}$$

so $p > 1$ means forming bonds are explored sooner than breaking ones. We took $p \in \{0.5, 1, 2\}$ both ways ($\mathrm{R} \to \mathrm{P}$ and $\mathrm{P} \to \mathrm{R}$), for a total of 6 paths, as shown on Fig.1 a).

Our scan is different from other initialization methods such as IDPP (Smidstrup et al., 2014) or geodesic interpolation (Zhu et al., 2019). Note that the NEB algorithm (Jónsson et al., 1998; Henkelman et al., 2000) could be used after our scan, but we used Sella (Hermes et al., 2022) instead (see below). Our harmonic bias is inspired by umbrella sampling (Torrie & Valleau, 1977), but we use it to relax single geometries on the potential energy surface rather than to sample a free energy.

Saddle search The 6 resulting paths sit above the true reaction path. We select 8 geometries[1] out of 24 to serve as starting point for *Sella* (Hermes et al., 2022), a saddle algorithm . It finds the softest direction of the energy surface, given by the Hessian matrix, H, of the energy. Computing H outright is costly, Sella instead starts from a guess and improves it, using simple gradients of the MLIP. It then walks uphill along the softest direction while relaxing all the others, which makes it a first-order saddle (transition state) attractor. Each converged point is then certified from the low end of its Hessian spectrum, obtained by a Lanczos scheme. Indeed, at a stationary point, each eigenvalue of H is the curvature of the energy along its eigenvector: all positive is a minimum, and exactly one negative is a first-order saddle. Each eigenvalue μ of the mass-weighted H is reported as a frequency $\nu = \sqrt{\mu}/2\pi$, which is imaginary when $\mu < 0$ and quoted as a negative number in cm^{-1}. We accept a point only if it has exactly one such frequency, below $-50\,\mathrm{cm}^{-1}$ so that numerical noise does not count. Among the certified saddles, we keep the most populated cluster within $k_B T = 0.59$ kcal/mol of the best energy (see Fig. 1 b).

Pseudo-IRC towards the minima. To check that the saddle point found by Sella connects the right pair of minima, we displace it along its imaginary mode in both directions and relax, verifying that the two ends reach R and P. We also use this pseudo-IRC as a low-cost source of structures for our dataset as we keep all the geometries along the relaxation path. In addition, we heat up 8 structures along the path with Langevin runs (300 K, 100 fs) to increase diversity.

4.2 From a pool of candidates to a training set

A round ends with a large pool of candidate structures (a median of 1373) produced by the generator. We label only 22 of them with DFT and add them to the fine-tuning set.

(i) Filtering. Much of the harvest sits far too high in energy, which is a waste of DFT budget. We discard it with a threshold, anchored on the two minima. A geometry x is discarded when $E_{\mathrm{MLIP}}(x) > \max\big(E_{\mathrm{MLIP}}(\mathrm{R}_{\mathrm{MLIP}}), E_{\mathrm{MLIP}}(\mathrm{P}_{\mathrm{MLIP}})\big) + 60$ kcal/mol. A second filter acts after labeling, on the DFT energies with the same threshold as before. However, to avoid wasting DFT data, if some labels fall above the threshold we also include up to 10 of them in the data set. A few high structures teach the model how the energy rises.

(ii) Selecting. We select only 22 geomtries to be labeled at each round. To select diverse but useful ones, we use a kernel method inspired from (Varga-Umbrich et al., 2026) in an online version. We

[1]For each path we keep the highest energy point (total of 6 points). For the two paths with lowest peak, we keep one extra point each, raising the total number of point to 8.

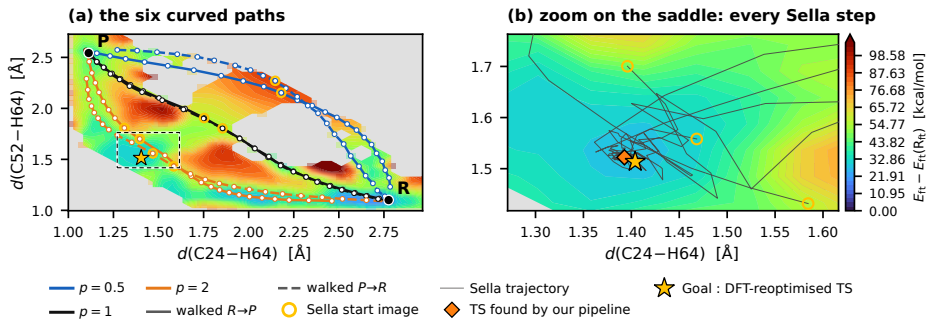


Figure 1: Path generator on reaction R1 (65 atoms), projected on two of its 6 reactive bonds. (a) the six paths along the surface, with exponents p of Eq. 3. $p = 0.5$ and 2 take different paths than the linear $p = 1$. (b) zoom in (a) showing every Sella step of every start, descending onto one certified saddle 0.023 Å from $\mathrm{TS}_{\mathrm{DFT}}$. The background energy was obtained through extensive sampling.

compare points in the MLIP's own feature space: a structure is represented by $\phi_{\mathrm{MLIP}}(x) \in \mathbb{R}^{128}$, the mean over atoms of the backbone's final invariant channels, ℓ_2-normalised, and the batch is picked by largest-cluster maximum-distance selection (LCMD).

(iii) Active learning. The labelled set grows by 22 points per round, and round restarts from the previous checkpoint on everything collected so far. Hyperparameters : batch 4, 50 epochs per fine-tuning, a per-atom energy MAE (weight 20) plus an L_2 force term (weight 2), and refitted atomic reference energies. The loop stops when all four errors stay below 1 kcal/mol for two consecutive rounds. The final model is then retrained from the best checkpoint on all the points for 100 epochs.

5 Results

5.1 Main results

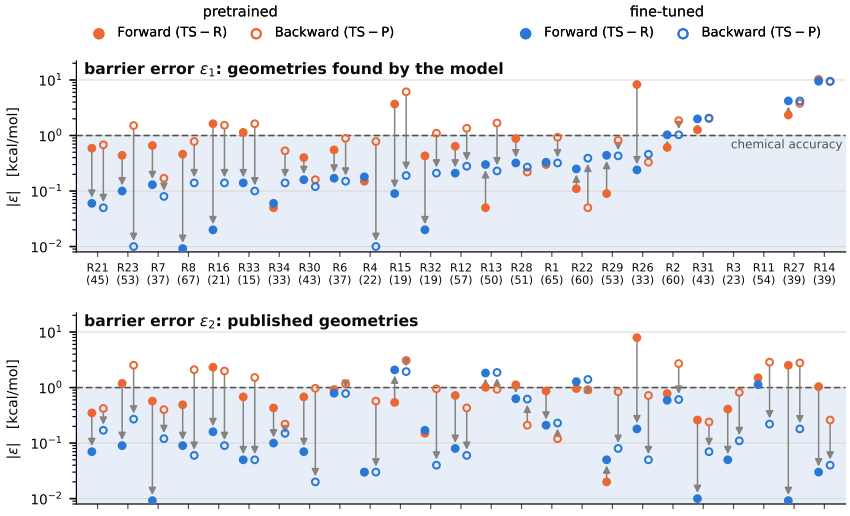


Figure 2: Results on 25 TMC reactions before and after fine-tuning. For each reaction (with #atoms below), we compare the two metrics (see 3) between the pretrained and the fine-tuned MLIP. An arrow pointing down is an improvement due to the fine-tuning. Upper panel: the energy barrier error of the R, P, TS found by the MLIP. Lower panel: the energy barrier error on three fixed geometries. R3 and R11 are absent from the upper panel as they don't have a TS reference.

Our active learning loop is able to find and label structures next to the real TS: for each reaction, between 9 to 123 of them fall within 0.3 Å of the TS, even though dataset are small : between 86 and 284 labels per reaction. Fine-tuned MLIPs found better TS starting only from R and P. This improves energy barriers with average errors $\epsilon_{\rightarrow,1}/\epsilon_{\leftarrow,1}$ going from 1.52 / 1.66 kcal/mol (pre) to 0.87 / 0.89 (ft), crossing chemical accuracy of 1 kcal/mol on both forward and backward. Among the 23 reactions [2], we went from 11 (pre) to 19 (ft) solved in both directions. On the second barrier errors $\epsilon_{\rightarrow,2}/\epsilon_{\leftarrow,2}$, we improve from 1.10/1.19 kcal/mol (pre) to 0.39/0.37 (ft), below chemical accuracy on both directions. This solves 21 (ft) of 25 reactions in both directions against 12 (pre). This shows that the potential energy surface was learned more precisely. Datasets are provided at this link.

A few failures to report (see Fig.2): for two reactions, the geometry optimization of the transition state with ωB97M-V/def2-TZVPD did not converge and were left out: R3 had a local minmum, and R11 an 2nd order saddle. The TS-search driven by our MLIP for R14 found another certified and connected saddle, 9.5 kcal/mol too high. This is a failure of search but the model reproduces the published barriers exactly. A similar behavior is observed for R27. R31 is not improving, and we suspect that the single reactive bond detected is not enough for studying the reaction.

Implementation details: We fine-tuned starting from the checkpoint of the best round of the active learning. Only one seed per reaction. We used A10G and L40S GPUs (3.5 hours per 100 DFT labels).

5.2 CURVED REACTIVE SCAN VERSUS NEB

We compare our curved reactive scan against NEB, a standard path generator, on five different reactions (15–37 atoms). We use its climbing-image variant (Henkelman et al., 2000), initialised by geodesic interpolation (Zhu et al., 2019) between R and P, and swap it into the loop with everything else frozen: a fixed budget of 8 rounds $\times$ 20 labels $=$ 160 per reaction (no stopping rule), the same filters, selector, fine-tuning recipe and final retraining, one replicate per reaction and three for R33. The NEB generator relaxes R and P, seeds a climbing-image band of 15 images by geodesic interpolation, and harvests the band, its optimisation trajectory and thermal perturbations of it (see Appendix B). Every final model, pretrained included, is then graded with one common TS searcher, the scan reactive curve of §4.1, so that ε_1 compares the models and not the search; ε_2 needs no search.

Both models improve on the pretrained one, the scan-built on all five reactions and the NEB-built on four (Fig. 3a). The scan model is ahead everywhere, but on R33, R16, R4 and R34 by 0.08–0.16 kcal/mol only, as much as the spread of the three R33 replicates (0.05–0.08 with the scan, 0.04–0.18 with NEB, Fig. 5a). On R7 however the NEB model ends at $\varepsilon_1 = 1.67$ kcal/mol, outside chemical accuracy and above the 0.68 kcal/mol of the pretrained model it started from. Its ε_2 is off by 2.2 kcal/mol as well, so the surface itself is wrong at the saddle, not only the search.

R4 has a single reactive bond, N7–C8 (2.72 Å in R, 2.13 at the TS, 1.50 in P), so its paths can be drawn in full with one bond length and one energy, and we use it to look at the two generators in detail (Fig. 3b). The position along a path is the image index divided by the number of images minus one, 0 at R and 1 at P. The paths scanned from P are flipped, $s \rightarrow 1 - s$, and the 13 interior images of the NEB band are placed between the same endpoints. Energies are relative to each model's own relaxed R. The six scan paths carry the bond monotonically from 2.7 to 1.5 Å while staying within 2 kcal/mol of R, and all eight Sella starts, launched from six different paths, converge close to $\mathrm{TS}_{\mathrm{DFT}}$. The NEB band from the same R and P first opens the bond to 4.0 Å, wider than in the reactant, reaches 67.7 kcal/mol there, and closes it in a single image (3.63 $\rightarrow$ 1.52 Å). Tighter NEB settings, i.e. a tighter clash gate, 1000 optimiser steps instead of 250 and 31 images instead of 15, all return the same band within 1 kcal/mol, and the MLIP-relaxed endpoints lie within 0.04 Å of the DFT minima. The NEB band is wrong and far from the saddle, yet the NEB-trained model still scores a good ε_1 in Fig. 3a.

The training set explains this (Fig. 3c). Of the 160 NEB harvested points on R4 80% of the set lies where the scan never goes. The remaining 20% is what matters. The pretrained model already had R4's saddle right (1.79/11.84 against 1.94/12.62 kcal/mol), and fine-tuning did not damage that region because it was not empty of labels: the 35 product- and TS-like points sit within 0.5 Å of a saddle that is itself 0.23 Å from P. On the scan-generated structures, the NEB-trained model's energy

[2] Two reactions are not taken into account since their geometry re-optimization did not converge.

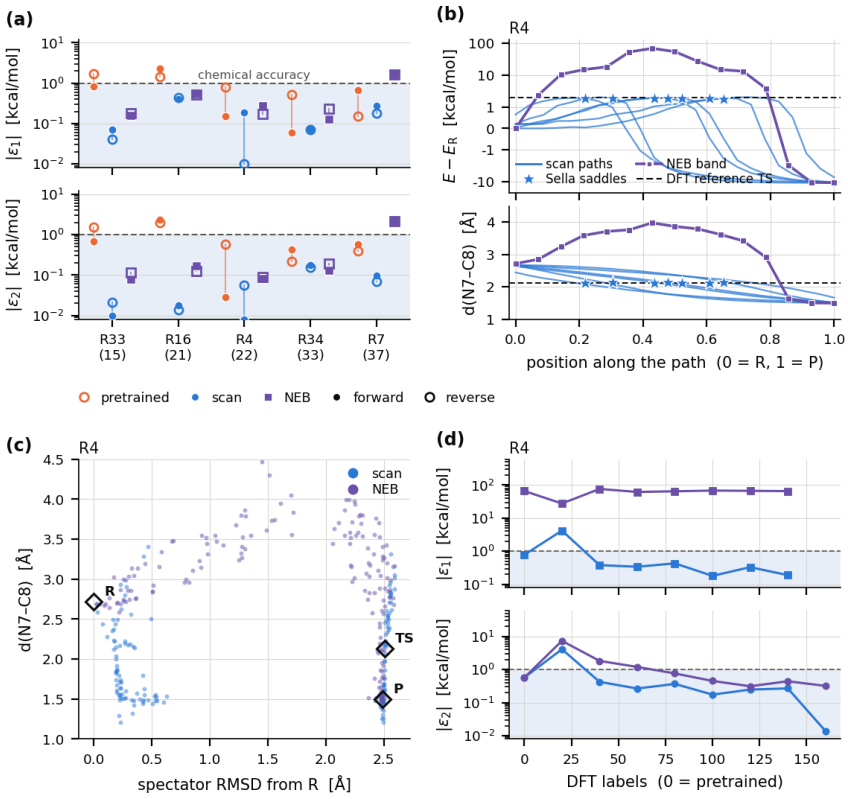


Figure 3: Replacing the curved scan by a climbing-image NEB. Colours throughout: pretrained model (orange), model fine-tuned on the 160 labels generated by the curved scan (blue) and by NEB (purple). Shaded: chemical accuracy, 1 kcal/mol. (a) Errors of the final models on five reactions (number of atoms in parentheses), forward barrier filled and reverse open. ϵ_1 is measured at the stationary points each model finds when the curved scan search, as described in section 4.1, is applied to it, so it grades each model with one common TS searcher; ϵ_2 is measured at the published geometries (Eq. (2)). (b) The two generators on R4, driven by the finetuned models (scan curves are obtained with the scan model and the NEB curve with the NEB model), against position along the path (0 at R, 1 at P, see main text): energy relative to R (top) and the length of the single reactive bond in R4 (bottom). (c) Training data of R4: the 160 DFT-labelled points of each generator. The data points are plotted as the length of the single reactive bond versus the RMSD of the non-reactive atoms from the reactant. Diamonds: reference R, TS and P. (d) Both metrics, Eqs. (1) and (2), on R4 against the number of DFT labels, i.e. the number of points in the finetuning data set. ϵ_1 is not defined as in (a). Here it is evaluated during the finetuning rounds where each generator uses its own search so the purple curve is the NEB search failing on the surface it is given.

MAE is $0.363\,\mathrm{kcal/mol}$ against 0.370 for the pretrained one, neither better nor worse (Appendix B.2). A local saddle search only ever interrogates that neighbourhood, so the 125 far-away labels are invisible to it.

However this makes the NEB finetuning less efficient in the number of data points as shown in (Fig. 3d). Inside the active learning loop, the NEB search on R4 reports a barrier error ϵ_1 of $\sim 60\,\mathrm{kcal/mol}$ each round: the improved surface does not move a band that is routed wrongly. R34 and R7 behave similarly (Fig. 6). NEB thus spends most of its DFT budget far from the saddle, harmlessly on R4 but harmfully on R7, where no label came within $0.5\,\text{Å}$ of the transition state and the fine-tuned surface is lifted by $+1.7$ to $+2.1\,\mathrm{kcal/mol}$ there (Appendix B.2). The curved reactive scan is a more data-efficient generator: on the learning curves at the published geometries (Fig. 3d, top, and Fig. 6), the scan-trained model crosses chemical accuracy with fewer labels on every reaction.

Finally, we note that the generator is a minor part of the full active-learning cost: generating candidates takes 4–13 min per round against 2.5–19 h per campaign, of which 90–95 % is DFT labelling and fine-tuning. Since the curved scan reaches chemical accuracy with fewer labels, the saving it brings is in rounds of labelling and fine-tuning, not in the generator itself.

6 CONCLUSION

Building a small but efficient dataset for fine-tuning is the hard part of specialising an MLIP to a reaction, and it should be done from the reactant and the product alone. We have developed an automated, active learning loop to obtain those datasets and sucessfully fine-tune MLIPs on 25 transition-metal complexes reactions from MOBH35. Thanks to our custom path generator and selection rules, we obtained rather small datasets. The resulting MLIP found transition states blindly and reached chemical accuracy on average for both forward and backward energy barriers improving on their pretrained versions. Compare to NEB, our curved reactive scan generator proved more robust and data efficient.

Four directions follow from this work.: i) UMA was pretrained on configurations derived from MOBH35. We will test our pipeline on datasets unseen by Omol25 (with known TS) to see how far this transfers. ii) To capture real chemistry, one should simulate strongly-correlated transition-metal complexes with post-DFT method such as CCSD(T). This turns the problem into multi-level fine-tuning. iii) For families of systems, for instance TMCs with a common metal atom center, fine-tune one single MLIP instead of several. iv) Transferability to molecular dynamics and free energy surface sampling.

AI USE STATEMENT

In this work, we have not used generative AI tools for synthetic data generation, theoretical modelling, proof formulation, methodology design or result interpretation, which are the tasks with required disclosure; none of them applies to this work beyond what is stated here. Additionally, we used generative AI tools as a coding and writing aid (code scaffolding, editing and formatting of the manuscript). We have reviewed all AI-assisted work: every reported number is generated by scripts from the raw campaign data, and the text was checked by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses public benchmark geometries and quantum-chemistry calculations only. It involves no human subjects, no personal data and no deployment. Cheaper barrier prediction lowers the cost of designing catalysts; a surrogate accurate on fixed geometries but wrong on the saddle it finds itself, the failure mode documented on R27, could mislead an automated design loop, and reporting ε_1 alongside ε_2 is the safeguard we propose.

REPRODUCIBILITY STATEMENT

Section 3 gives the benchmark subset, the level of theory (ωB97M-V/def2-TZVPD, `gpu4pyscf`) and the two metrics; Section 4 gives the path generator, the filters, the selector, the loss, the optimiser, the budget per round and the stopping rule (learning rate 10^{-4}, batch 4, 50 epochs per round and 100 for the final model, loss weights 20 and 2, refitted atomic references). The three published stationary points are never trained on. The benchmark geometries are public (Iron & Janes, 2019). The 3954 DFT-labelled structures forming each final training set are released in extended XYZ with energies and analytic forces, together with the CSV files and the script from which every reported number is derived. One active-learning run per reaction was performed, so no error bars are reported.

REFERENCES

Ilyes Batatia, Dávid P. Kovács, Gregor N. C. Simm, Christoph Ortner, and Gábor Csányi. MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Jörg Behler and Michele Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.*, 98:146401, 2007.

Regine S. Bohacek, Colin McMartin, and Wayne C. Guida. The art and practice of structure-based drug design: A molecular modeling perspective. *Medicinal Research Reviews*, 16(1):3–50, 1996. doi: 10.1002/(SICI)1098-1128(199601)16:1<3::AID-MED1>3.0.CO;2-6.

Max Born and Robert Oppenheimer. Zur quantentheorie der molekeln. *Ann. Phys.*, 389:457–484, 1927.

Amanda L. Dewyer, Alonso J. Argüelles, and Paul M. Zimmerman. Methods for exploring reaction space in molecular systems. *WIREs Computational Molecular Science*, 8(2):e1354, 2018. doi: 10.1002/wcms.1354.

Weinan E, Weiqing Ren, and Eric Vanden-Eijnden. String method for the study of rare events. *Physical Review B*, 66:052301, 2002. doi: 10.1103/PhysRevB.66.052301.

Weinan E, Weiqing Ren, and Eric Vanden-Eijnden. Simplified and improved string method for computing the minimum energy paths in barrier-crossing events. *The Journal of Chemical Physics*, 126:164103, 2007. doi: 10.1063/1.2720838.

Graeme Henkelman and Hannes Jónsson. A dimer method for finding saddle points on high dimensional potential surfaces using only first derivatives. *The Journal of Chemical Physics*, 111 (15):7010–7022, 1999. doi: 10.1063/1.480097.

Graeme Henkelman, Blas P. Uberuaga, and Hannes Jónsson. A climbing image nudged elastic band method for finding saddle points and minimum energy paths. *J. Chem. Phys.*, 113:9901–9904, 2000. doi: 10.1063/1.1329672.

Eric D. Hermes, Khachik Sargsyan, Habib N. Najm, and Judit Zádor. Sella, an open-source automation-friendly molecular saddle point optimizer. *J. Chem. Theory Comput.*, 18:6974–6988, 2022. doi: 10.1021/acs.jctc.2c00395.

Pierre Hohenberg and Walter Kohn. Inhomogeneous electron gas. *Phys. Rev.*, 136:B864–B871, 1964.

Mark A. Iron and Trevor Janes. Evaluating transition metal barrier heights with the latest density functional theory exchange–correlation functionals: The MOBH35 benchmark database. *J. Phys. Chem. A*, 123:3761–3781, 2019. doi: 10.1021/acs.jpca.9b01546.

Hannes Jónsson, Greg Mills, and Karsten W. Jacobsen. Nudged elastic band method for finding minimum energy paths of transitions. In Bruce J. Berne, Giovanni Ciccotti, and David F. Coker (eds.), *Classical and Quantum Dynamics in Condensed Phase Simulations*, pp. 385–404. World Scientific, 1998. doi: 10.1142/9789812839664_0016.

Walter Kohn and Lu Jeu Sham. Self-consistent equations including exchange and correlation effects. *Phys. Rev.*, 140:A1133–A1138, 1965.

Kyunghoon Lee, Jinwon H. Kim, and Woo Youn Kim. pyMCD: Python package for searching transition states via the multicoordinate driven method. *Computer Physics Communications*, 291: 108831, 2023. doi: 10.1016/j.cpc.2023.108831.

Daniel S. Levine et al. The open molecules 2025 (OMol25) dataset, evaluations, and models. *arXiv:2505.08762*, 2025.

Satoshi Maeda and Keiji Morokuma. Communications: A systematic method for locating transition structures of A+B→X type reactions. *The Journal of Chemical Physics*, 132:241102, 2010. doi: 10.1063/1.3457903.

Satoshi Maeda, Yu Harabuchi, Makito Takagi, Tetsuya Taketsugu, and Keiji Morokuma. Artificial force induced reaction (AFIR) method for exploring quantum chemical potential energy surfaces. *The Chemical Record*, 16(5):2232–2248, 2016. doi: 10.1002/tcr.201600043.

Narbe Mardirossian and Martin Head-Gordon. ωB97M-V: A combinatorially optimized, range-separated hybrid, meta-GGA density functional with VV10 nonlocal correlation. *J. Chem. Phys.*, 144:214110, 2016. doi: 10.1063/1.4952647.

Narbe Mardirossian and Martin Head-Gordon. Thirty years of density functional theory in computational chemistry: an overview and extensive assessment of 200 density functionals. *Mol. Phys.*, 115:2315–2372, 2017.

Jonah Marks, Jonathon Vandezande, and Joseph Gomes. Reliable and efficient automated transition-state searches with machine-learned interatomic potentials. *arXiv:2604.00405*, 2026.

Baron Peters, Andreas Heyden, Alexis T. Bell, and Arup Chakraborty. A growing string method for determining transition states: Comparison to the nudged elastic band and string methods. *J. Chem. Phys.*, 120:7877–7886, 2004. doi: 10.1063/1.1691018.

Kaipai Ren, Kun Tang, Yujing Zhao, Lei Zhang, Jian Du, Qingwei Meng, and Qilei Liu. Reactive machine learning potential for accelerating transition state search in organic synthesis. *Nature Communications*, 17:6253, 2026. doi: 10.1038/s41467-026-72945-0.

Lars Ruddigkeit, Ruud van Deursen, Lorenz C. Blum, and Jean-Louis Reymond. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *Journal of Chemical Information and Modeling*, 52(11):2864–2875, 2012. doi: 10.1021/ci300415d.

Mathias Schreiner, Arghya Bhowmik, Tejs Vegge, Jonas Busk, and Ole Winther. Transition1x – a dataset for building generalizable reactive machine learning potentials. *Sci. Data*, 9:779, 2022. doi: 10.1038/s41597-022-01870-w.

Emmanouil Semidalas and Jan M. L. Martin. The MOBH35 metal–organic barrier heights reconsidered: Performance of local-orbital coupled cluster approaches in different static correlation regimes. *J. Chem. Theory Comput.*, 18:883–898, 2022. doi: 10.1021/acs.jctc.1c01126.

Søren Smidstrup, Andreas Pedersen, Kurt Stokbro, and Hannes Jónsson. Improved initial guess for minimum energy path calculations. *J. Chem. Phys.*, 140:214106, 2014. doi: 10.1063/1.4878664.

Kun Tang, Yujing Zhao, Lei Zhang, Jian Du, Qingwei Meng, and Qilei Liu. Accelerating transition state search and ligand screening for organometallic catalysis with reactive machine learning potential. *Journal of Chemical Theory and Computation*, 21:8933–8949, 2025.

Glenn M. Torrie and John P. Valleau. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.*, 23:187–199, 1977. doi: 10.1016/0021-9991(77)90121-8.

Emese Varga-Umbrich, Shubhendu Surana, Paul Duckworth, Jules Tilly, Olivier Peltre, and Zachary Weller-Davies. Pretrained model representations as acquisition signals for active learning of machine-learned interatomic potentials. *arXiv:2605.03964*, 2026.

Xingyu Wang, Yu Mao, and Ziyun Wang. Machine learning approaches for transition state prediction. *Chem Catalysis*, 5:101458, 2025. doi: 10.1016/j.checat.2025.101458.

Leon Wehrhan, Lucien Walewski, Marie Bluntzer, Heloise Chomet, Jules Tilly, Christoph Brunken, and Silvia Acosta-Gutiérrez. Mlipaudit: A benchmarking tool for machine learned interatomic potentials. *arXiv preprint arXiv:2511.20487*, 2025.

Florian Weigend and Reinhart Ahlrichs. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn. *Phys. Chem. Chem. Phys.*, 7:3297–3305, 2005. doi: 10.1039/B508541A.

Brandon M. Wood et al. UMA: A family of universal models for atoms. *arXiv:2506.23971*, 2025.

Xiaojie Wu, Rui Qi, et al. Enhancing GPU-acceleration in the python-based simulations of chemistry framework. *J. Chem. Theory Comput.*, 21:3452–3463, 2025. doi: 10.1021/acs.jctc.4c01527.

Manyi Yang, Jingxiang Zou, Guoqiang Wang, and Shuhua Li. Automatic reaction pathway search via combined molecular dynamics and coordinate driving method. *The Journal of Physical Chemistry A*, 121(6):1351–1361, 2017. doi: 10.1021/acs.jpca.6b12195.

Tom A. Young, Joseph J. Silcock, Alistair J. Sterling, and Fernanda Duarte. autodE: Automated calculation of reaction energy profiles – application to organic and organometallic reactions. *Angewandte Chemie International Edition*, 60(8):4266–4274, 2021. doi: 10.1002/anie.202011941.

Qiyuan Zhao, Sai Mahit Vaddadi, Michael Woulfe, Lawal A Ogunfowora, Sanjay S Garimella, Olexandr Isayev, and Brett M Savoie. Comprehensive exploration of graphically defined reaction spaces. *Scientific data*, 10(1):145, 2023.

Xiaolei Zhu, Keiran C. Thompson, and Todd J. Martínez. Geodesic interpolation for reaction pathways. *J. Chem. Phys.*, 150:164103, 2019. doi: 10.1063/1.5090303.

A STABILITY OF BIASED IMAGES NEAR THE SADDLE POINT

Driven path generators obtain each image by minimising, from the previous image, a biased energy

$$F(\mathbf{x}) = E(\mathbf{x}) + V(\mathbf{x}), \tag{4}$$

where the bias V acts on the reactive coordinates. Since an optimiser converges only to minima, an image can lie near the saddle point only if F has a local minimum there; otherwise the optimiser falls into a neighbouring basin and the barrier region is skipped. Such an image is a minimum of F but not of E, on which it lies close to a saddle point.

Let $\mathbf{v}$ be the reaction mode, i.e. the unit eigenvector of $\nabla^2 E$ with negative eigenvalue at the transition state, and $\mu = \mathbf{v}^\top \nabla^2 E\, \mathbf{v} < 0$ its (Cartesian, non-mass-weighted) curvature. A necessary condition for a minimum of F near the saddle is a positive curvature along $\mathbf{v}$,

$$\mathbf{v}^\top \nabla^2 F\, \mathbf{v} = \mu + \mathbf{v}^\top \nabla^2 V\, \mathbf{v} > 0, \tag{5}$$

that is, the bias must add more curvature than the barrier removes. We evaluate this condition for three choices of V: a constant force (AFIR), a harmonic restraint (our scan), and hard constraints on the reactive distances (coordinate driving). The harmonic restraint interpolates between the other two, reducing to a constant force for $K \to 0$ at fixed Kt_m and to hard constraints for $K \to \infty$.

Constant force (AFIR). AFIR uses $V = \alpha\, s(\mathbf{x})$, where

$$s = \frac{\sum_{i\in A}\sum_{j\in B} \omega_{ij}\, r_{ij}}{\sum_{i\in A}\sum_{j\in B} \omega_{ij}}, \qquad \omega_{ij} = \left(\frac{R_i + R_j}{r_{ij}}\right)^6, \tag{6}$$

is a weighted mean of the distances r_{ij} between the atoms of two fragments A and B, with R_i the covalent radii. The sixth-power weights concentrate s on the closest atom pairs. Near a reaction, the atoms about to bond dominate, and $s \approx r_{ij}$ for that single pair. In this limit,

$$\mathbf{v}^\top \nabla^2 F\, \mathbf{v} = \mu + \alpha\, \mathbf{v}^\top \nabla^2 r_{ij}\, \mathbf{v}, \qquad \nabla^2 r_{ij} \propto \frac{1}{r_{ij}}\, \mathbf{P}_\perp, \tag{7}$$

where $\mathbf{P}_\perp$ projects out the stretching direction of the pair: a distance grows linearly when its two atoms move apart along their axis, and curves only when the bond rotates. Along the stretch, which dominates the reaction mode, the bias adds no curvature, and Eq. (5) fails in the whole region where E curves downward, for any α. When several pairs contribute comparably, the derivatives of the weights add terms of order α/r; these are fixed by the force needed to cross the barrier and cannot be tuned independently to exceed $|\mu|$. As α grows, the reactant-side minimum disappears at an inflection point of E and the optimiser falls to the product side. Geometries near the saddle appear only as intermediate optimisation steps.

Harmonic restraint (curved reactive scan). With $F_k(\mathbf{x}) = E(\mathbf{x}) + \frac{K}{2}\sum_m \big(d_m(\mathbf{x}) - t_m(\lambda_k)\big)^2$,

$$\nabla^2 F_k = \nabla^2 E + K \sum_m \nabla d_m \nabla d_m^\top + K \sum_m \big(d_m - t_m\big)\, \nabla^2 d_m, \tag{8}$$

so that along $\mathbf{v}$

$$\mathbf{v}^\top \nabla^2 F_k\, \mathbf{v} \;\approx\; \mu + K \sum_m \big(\mathbf{v}\cdot\nabla d_m\big)^2. \tag{9}$$

The last term of the Hessian is negligible near the saddle, since stationarity gives $|d_m - t_m| = |\partial E/\partial d_m|/K$, which vanishes along the reaction mode at the transition state. Every image is therefore a stable minimum provided

$$K \sum_m \big(\mathbf{v}\cdot\nabla d_m\big)^2 > |\mu|. \tag{10}$$

With $K = 20$ eV Å^{-2}, comparable to a stiff bond-stretching force constant, this condition holds for all reactions studied here.

Hard constraints (coordinate driving). In the limit $K \to \infty$ the restraints become the constraints $d_m(\mathbf{x}) = t_m$ of coordinate-driving methods (Yang et al., 2017; Young et al., 2021). The driven directions are removed from the optimisation, which amounts to an infinite curvature along them, so the stability condition holds for any μ; each image minimises E on the constraint surface, and the Lagrange multipliers take the place of the restraint forces $K(d_m - t_m)$. The image is then stable if E has positive curvature within that surface, which holds when the reaction mode lies in the span of the driven coordinates. When it does not, images can still jump between basins, the known failure of relaxed scans along a poorly chosen coordinate, and the same caveat applies at finite K. Compared with hard constraints, finite restraints trade a small lag, $|d_m - t_m| = |\partial E/\partial d_m|/K$, for two practical advantages: the bias is an ordinary energy term usable with any optimiser, and possibly incompatible targets on up to eight coupled bonds are resolved by compromise rather than by an infeasible or ill-conditioned constrained problem.

B THE NEB COMPARISON

B.1 PROTOCOL

One axis moves, the generator that proposes candidate structures at each round. Everything downstream is shared: reactions R33 (15 atoms, $M = 1$), R16 (21, 6), R4 (22, 1), R34 (33, 4) and R7 (37, 2); a fixed recipe of 8 rounds $\times$ 20 DFT labels with no stopping rule; at round 1 the MLIP-relaxed R and P forced into the batch and 18 points chosen by LCMD, then 20 by LCMD on the current checkpoint's 128-dimensional features; the energy filters of §4.2; DFT labels (ωB97M-V/def2-TZVPD, `gpu4pyscf`, energies and forces); the fine-tuning of §4.2 (learning rate 10^{-4}, batch 4, 50 epochs, cosine schedule with warm-up, pretrained heads kept, atomic references refitted), round 1 from the pretrained weights and rounds 2–8 warm-started from the previous round on the cumulative set; and a final model retrained for 100 epochs on all 160 points from the last checkpoint. Harvest seeds are $1000\,r + 10^5 k$ for round r and replicate k, the fine-tuning seed is 0 throughout. This differs from the main campaign of §5.1 in four places, each of which applies identically to both arms: 20 labels per round instead of 22, no stopping rule, the final retraining from the last rather than the best checkpoint, and the filter anchored on the DFT energies of the published endpoints.

The scan generator is that of §4.1: six restrained curved scans ($p \in \{0.5, 1, 2\}$, each R $\to$ P and P $\to$ R, 24 images, $K = 20\,\mathrm{eV\,Å^{-2}}$), Sella starts from the highest images, Lanczos certification (exactly one imaginary frequency below $-50\,\mathrm{cm^{-1}}$, reactive mode, pseudo-IRC ends at R and P), the pseudo-IRC descents, and a thermal necklace of 8 seeds along the retained path (Langevin, 300 K, 100 fs, 11 snapshots each). The pool it hands to the filter and selector holds scan images, saddles, Sella and IRC descents, thermal snapshots and the two endpoints.

The NEB generator relaxes R and P with the same free relaxation as the scan. The band is seeded by geodesic interpolation (Zhu et al., 2019) with a fixed seed, so a round is reproducible, and optimised as a climbing-image NEB (Henkelman et al., 2000) with 15 images, spring constant $0.1\,\mathrm{eV\,Å^{-2}}$, improved tangent, a first phase to 0.30 eV/Å (80 steps) and a climbing phase to 0.05 eV/Å (250 steps); convergence is judged on the residual at the last accepted step, and the whole optimiser trajectory is archived. Eight images of the converged band, the climbing image and seven evenly spaced interior ones, are then perturbed by short Langevin runs (300 K, 100 fs), as for the scan.

B.2 RESULTS

For every labelled point we compute a reaction progress, the mean over the reactive bonds of $(d - d^{\mathrm{R}})/(d^{\mathrm{P}} - d^{\mathrm{R}})$, signed and unbounded, and a spectator RMSD, the Kabsch RMSD over the atoms of no reactive bond between the point and $\mathrm{R}_{\mathrm{DFT}}$. We plot those in Fig. 4.

To test the robustness of the active learning, R33 was run three times per generator, replicates differing by the harvest seed only. For each of the six final models we take the larger of its two barrier errors, $\max(|\varepsilon_{\rightarrow,1}|, |\varepsilon_{\leftarrow,1}|)$, with the transition state found by a final scan search driven by the finetuned model. The three scan-trained models give 0.05 to $0.08\,\mathrm{kcal/mol}$, the three NEB-trained ones 0.04 to 0.18, see Fig. 5a. This spread is as large as the final accuracy gaps between generators reported in Fig. 3. The narrower spread with the scan is expected, since its paths are fixed by R, P and p and a

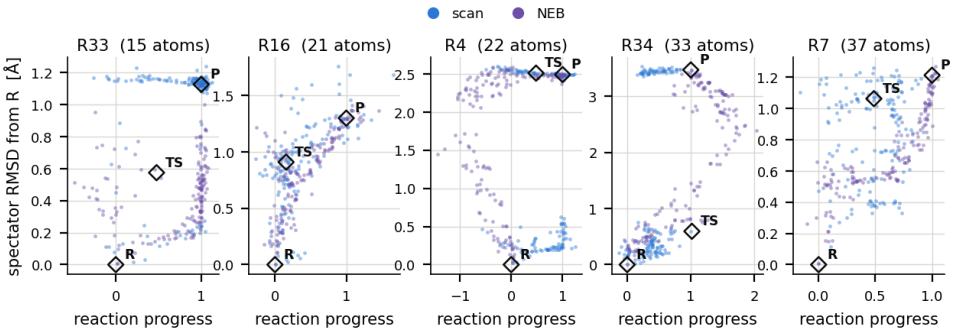


Figure 4: Where the 160 DFT-labelled points of each generator sit, on the five reactions. Horizontal: reaction progress, the mean over the reactive bonds of $(d - d^R)/(d^P - d^R)$, 0 at R and 1 at P. Vertical: RMSD of the non-reactive atoms from the relaxed reactant, i.e. how far the rest of the molecule has moved. Diamonds: our DFT reoptimised R, TS and P.

replicate changes only the perturbations and the selection, whereas a NEB replicate changes the seed of the interpolation and hence possibly the band.

For each reaction, we also evaluate the scan and NEB models on the data set harvested by the other generator. Since the NEB-generated data set contains points of much higher energy than the scan-generated one, raw MAEs cannot be compared directly. Therefore, we plot each fine-tuned model's MAE as a ratio to the pretrained model's on the same structures in Fig. 5b. The forgetting is confined to the three reactions on which the two generators' structures differ most: the fine-tuning generalises poorly to a different path through the same molecule.

For each reaction, we evaluate both models along a common path, a geodesic interpolation of 11 images from $\mathrm{R_{DFT}}$ to $\mathrm{TS_{DFT}}$ and 11 from $\mathrm{TS_{DFT}}$ to $\mathrm{P_{DFT}}$. Only images 0, 10 and 20 are stationary points; the other 18 are unrelaxed interpolation and strained. On four reactions the models agree at the saddle to 0.06–0.12 kcal/mol and diverge only at strained images, by up to 130 kcal/mol. On R7 the gap peaks at the saddle itself, at 2.26 kcal/mol.

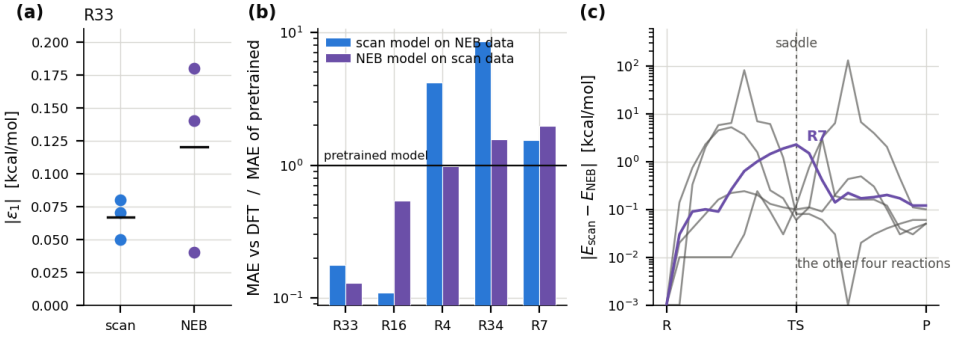


Figure 5: Replicate spread, transfer between the two training sets, and disagreement between the two models. (a) R33, the one reaction run three times per generator: $|\epsilon_1|$ of each final model under the scan search. (b) Each fine-tuned model evaluated on the 160 DFT-labelled structures produced by the other generator: the MAE of its energies against the DFT labels, divided by the MAE of the pretrained model against the same labels. (c) The gap between the two fine-tuned models at identical geometries, along a common 21-image geodesic path through the reference R, TS and P, each model's energy taken relative to its own value at R.

Finally, in Fig. 6 we show the evolution of the two metrics, ϵ_1 and ϵ_2, as defined in Eqs. (1) and (2). The first fine-tune damages every model, in both arms, on every reaction. The pretrained model starts at 0.43–2.32 kcal/mol and is already inside chemical accuracy on three of the five reactions; after the first 20 labels it is 4 to 40 times worse, and it takes 60–120 more labels to return. The training logs show why: a training loss of 0.05–0.28 against 0.9–2.2 on held-out structures from the first round,

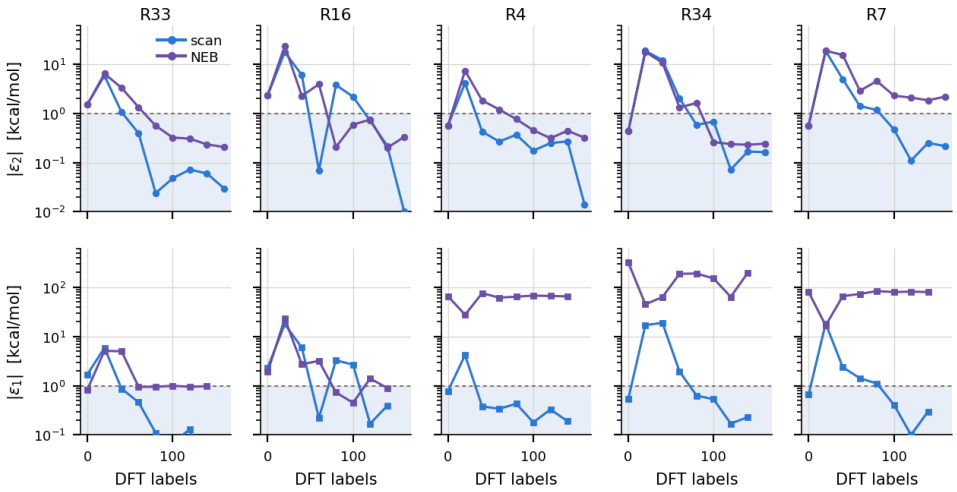


Figure 6: Learning curves of the two generators on the five reactions. Both error metrics, defined in Eqs. (1) and (2), against the number of DFT labels, blue for the model fine-tuned on scan-generated labels and purple for NEB-generated ones; x = 0 is the pretrained model, and every point is the worse of the forward and reverse barrier errors. Shaded: chemical accuracy.

a gap of 10–30$\times$ that closes only around 160 labels. This is overfitting to a small and chemically narrow batch, 250 gradient steps at learning rate 10^{-4} on twenty geometries of one molecule, with no anchor to the pretrained weights, not underfitting. It is a property of our finetuning recipe. This forgetting is acceptable here, since each model is meant to describe one reaction and is graded on that reaction alone; it would not be for a model expected to keep the pretrained reactivity elsewhere. The usual remedies, a lower learning rate, fewer epochs, restarting from the pretrained weights at each round or replaying a slice of OMol25, all change the recipe for both arms and would break comparability with the campaign of §5.1, so we leave them to future work.

The bottom plots (showing ϵ_1) shows the limit of the NEB search and not of the corresponding model. On R4, R34 and R7 the NEB curve is flat near 10^2 kcal/mol at every round while the scan's descends normally. On the other hand, the NEB-trained models of those three reactions score 0.27, 0.23 and 1.67 under the common scan search. Note that the lines ends at 140 DFT labels because the round-8 checkpoint is never harvested again.