

Global properties of the energy landscape: a testing and training arena for machine learned potentials

Received: 22 August 2025

Accepted: 17 November 2025

Cite this article as: Cărare, V., Thiemann, F.L., Morrow, J.D. *et al.* Global properties of the energy landscape: a testing and training arena for machine learned potentials. *npj Comput Mater* (2025). <https://doi.org/10.1038/s41524-025-01878-x>

Vlad Cărare, Fabian L. Thiemann, Joe D. Morrow, David J. Wales, Edward O. Pyzer-Knapp & Luke Dicks

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

Global properties of the energy landscape: a testing and training arena for machine learned potentials

Vlad Cărare,^{1,2,3,*} Fabian L. Thiemann,^{3,†} Joe D. Morrow,^{3,‡} David

J. Wales,^{1,2,§} Edward O. Pyzer-Knapp,^{1,¶} and Luke Dicks^{1,3,**}

¹*Xyme, Manchester, UK*

²*Yusuf Hamied Department of Chemistry,
University of Cambridge, Cambridge, UK*

³*IBM Research Europe, Daresbury, UK*

(Dated: November 4, 2025)

Machine learning interatomic potentials (MLIPs) have achieved remarkable accuracy on standard benchmarks, yet their ability to reproduce molecular kinetics, critical for reaction rate calculations, remains largely unexplored. We introduce Landscape17, a dataset of complete kinetic transition networks (KTNs) for the six molecules of the rMD17 dataset, computed using hybrid-level density functional theory. Each KTN contains minima, transition states, and approximate steepest-descent paths, along with energies, forces, and Hessian eigenspectra at stationary points. We develop a comprehensive test suite to evaluate the MLIPs’ ability to reproduce these reference landscapes and apply it to state-of-the-art architectures. Our results reveal limitations in current MLIPs: all models considered miss over half of the DFT transition state paths and generate stable unphysical structures throughout the potential energy surface. Data augmentation with pathway configurations improves reproduction of DFT potential energy surfaces, resulting in significant improvement in global kinetics. However, these models still produce many spurious stable structures, indicating that current MLIP architectures face underlying challenges in capturing the topology of molecular potential energy surfaces. The Landscape17 benchmark provides a straightforward, lightweight, but demanding test of MLIPs for kinetic applications, and we propose this test for validation of next-generation MLIPs targeting reaction discovery and rate prediction.

* vcarare@xyme.ai

† fabian.thiemann@ibm.com

‡ joe@culp.ai

§ dw34@cam.ac.uk

¶ ed@xyme.ai

** Corresponding author: ldicks@xyme.ai

I. INTRODUCTION

Physics-based approaches to molecular simulation have long served as the foundation for *in silico* molecular investigation and discovery. Expanded computational power has enabled these tools to answer sets of increasingly complex questions, with notable success.¹

Over the past decades, machine learning interatomic potentials (MLIPs) have emerged as an alternative component for molecular simulations. These models, often neural networks, are trained on high-quality reference data and hold the promise of providing *ab initio* accuracy at a fraction of the computational cost.^{2–5} This efficiency provides access to longer simulation timescales, and the computation of a wide range of derived properties.

Various architectures^{6–15} and molecular representations^{16–20} have been developed for the construction of MLIPs. These models are trained on energies, forces and, for periodic materials, stresses, which are most commonly computed using density functional theory (DFT). Foundational models have been proposed for materials^{21–23} and molecules^{24–27} based on large-scale datasets,^{28–31} which aim to improve out-of-distribution generalizability, potentially offering improved representations for system-specific applications.

Underlying such advances are both comprehensive benchmarks³² and lightweight tests that allow for on-the-fly tracking of architectural or training dataset changes. One example is the popular rMD17 benchmark,^{33,34} which provides energy and force labels for samples obtained from *ab initio* molecular dynamics (MD) trajectories of small molecules. Given its simplicity and the desire for models to faithfully sample from the Boltzmann distribution, rMD17 has been used as a first benchmark for model accuracy.^{11–13,35}

While valuable, standard MD simulations do not address broken ergodicity problems. At low temperature they tend to sample low-diversity, low-energy states around local energy minima, trapped between energy barriers that restrict exploration of the complete potential energy surface (PES) on the accessible MD time scale. Increasing temperature can help overcome such barriers, but may then miss the configurations and pathways essential for accurate molecular kinetics calculations at the temperatures of interest (e.g. for reaction rates).^{36–40} These bottlenecks can also compromise thermodynamic predictions when competing low-energy minima remain unobserved due to insufficient sampling. Consequently, due to its relatively low diversity, models released over the past years have saturated the energy and force errors on the rMD17 benchmark.^{11–13,35}

While enhanced sampling methods, such as parallel tempering^{41–43} MD replica exchange,⁴⁴ multicanonical,⁴⁵ Wang-Landau,⁴⁶ and biasing schemes,^{47,48} can help to overcome broken ergodicity

problems, the computational expense limits routine application for generating large databases. Methods to explicitly treat rare event dynamics^{49–57} are also computationally demanding.

An alternative approach is to construct a kinetic transition network^{58–60} using discrete path sampling,^{61,62} employing methods based on geometry optimization to characterize local minima and the pathways that connect them via transition states in the potential energy surface. Here, a transition state is defined as a stationary point with a single negative Hessian eigenvalue.⁶³ Global kinetic properties are then extracted using unimolecular rate theory for the individual minimum-to-minimum rates.^{64,65} This formalism involves solution of a master equation⁶⁶ for the global dynamics, as for dynamics-based schemes that produce a Markov state representation.^{53,56,67,68}

To address the need for kinetically-relevant structures in developing MLIPs, specialized datasets have been developed that capture transition paths. Examples include Transition1x,⁶⁹ which features samples from nudged-elastic-band (NEB) constructions across various reactions, and more comprehensive, data-rich collections such as OMol25³⁰ and SPICE.²⁹ The latter examples, created as training sets for MLIPs, include specialized subsets that capture configurations along transition state pathways. The incorporation of more extensive sampling of molecular conformations has been highlighted as of key importance for MLIP performance.^{70–72} Automated transition state finding is an important objective of many studies, and recent works explore the applicability of generative models: SDiff,⁷³ TSGen,⁷⁴ and React-OT,⁷⁵ as well as workflows to improve MLIPs performance in NEBs.⁷⁶

Given this context, there is a need for lightweight benchmarks that assess both the accuracy and transferability of MLIPs across the kinetically-relevant paths in the energy landscape. Such a benchmark should allow the validation of MLIPs beyond energy and force errors to include the organization of the potential energy landscape. In particular, we wish to reproduce stationary points faithfully, without additional spurious local minima (or transition states). We present two major contributions in this direction.

Firstly, we introduce the Landscape17 dataset, which systematically expands upon rMD17 by providing complete kinetic transition networks (KTNs)^{58–60} for the six molecules within rMD17 that have more than one distinct local minimum structure. This dataset features global potential energy surface representations generated using the energy landscape framework,^{77,78} and includes regions crucial for accurately reproducing both thermodynamic and kinetic properties. For each of the selected six molecules (ethanol, malonaldehyde, paracetamol, salicylic acid, azobenzene, and aspirin) we provide all the minima and transition states, along with configurations from the two approximate steepest-descent paths connecting each transition state to the corresponding minima.

These paths underpin the most probable routes between minima at finite temperature, offering essential configurations for understanding system kinetics.

Secondly, we establish a comprehensive, yet lightweight, testing suite for evaluating energy landscape fidelity, and apply this to assess current MLIP architectures. The ability of MLIPs to capture entire reference DFT KTNs, with the corresponding energies and atomic structures of both minima and transition states, extends molecular validation metrics beyond conventional force and energy errors. Since PES organization strongly determines physical properties, reproducing these properties provides an appropriate and generalized test of overall physical property reproduction.

Our analysis suggests that current MLIP models struggle to accurately predict KTNs even for the small, semi-rigid organic molecules present in rMD17, despite their apparent simplicity. These small molecules span many ubiquitous functional groups: alkane, alcohol, ketone, carboxylic acid, azo, amine and ester, so we expect the results to be relevant for many small molecules. Moreover, we find that MLIPs often exhibit unphysical minima, an issue that is not usually a problem for semi-empirical methods such as GFN2-xTB.⁷⁹ We therefore propose this dataset and benchmark suite as initial tests of the models’ physicality and ability to recreate kinetics, both of which prove difficult, even for small systems. We demonstrate systematic improvement strategies using configurations sampled along the pathways.

II. RESULTS

In this contribution, we extend the commonly used rMD17 benchmark dataset to produce Landscape17, which provides KTNs for the six molecules within rMD17 that exhibit multiple distinct minima: ethanol, malonaldehyde, salicylic acid, azobenzene, paracetamol, and aspirin. These networks were computed using hybrid-functional DFT, at an estimated computational cost exceeding 10^5 CPU hours, and capture the pathways that are essential for a proper description of global kinetics. The dataset includes configurations from the pathways to supplement the stationary points and is publicly available.⁸⁰

We benchmark several MLIP architectures for the selected molecules to evaluate how well they reproduce key features of the reference DFT potential energy surfaces, and investigate whether incorporating pathway data improves MLIP performance. We begin with a brief overview of the DFT potential energy surfaces and data acquisition regimes, followed by an evaluation of MLIPs trained using the corresponding datasets. We then examine the resulting MLIP surfaces, and compare with the DFT reference.

A. DFT landscapes

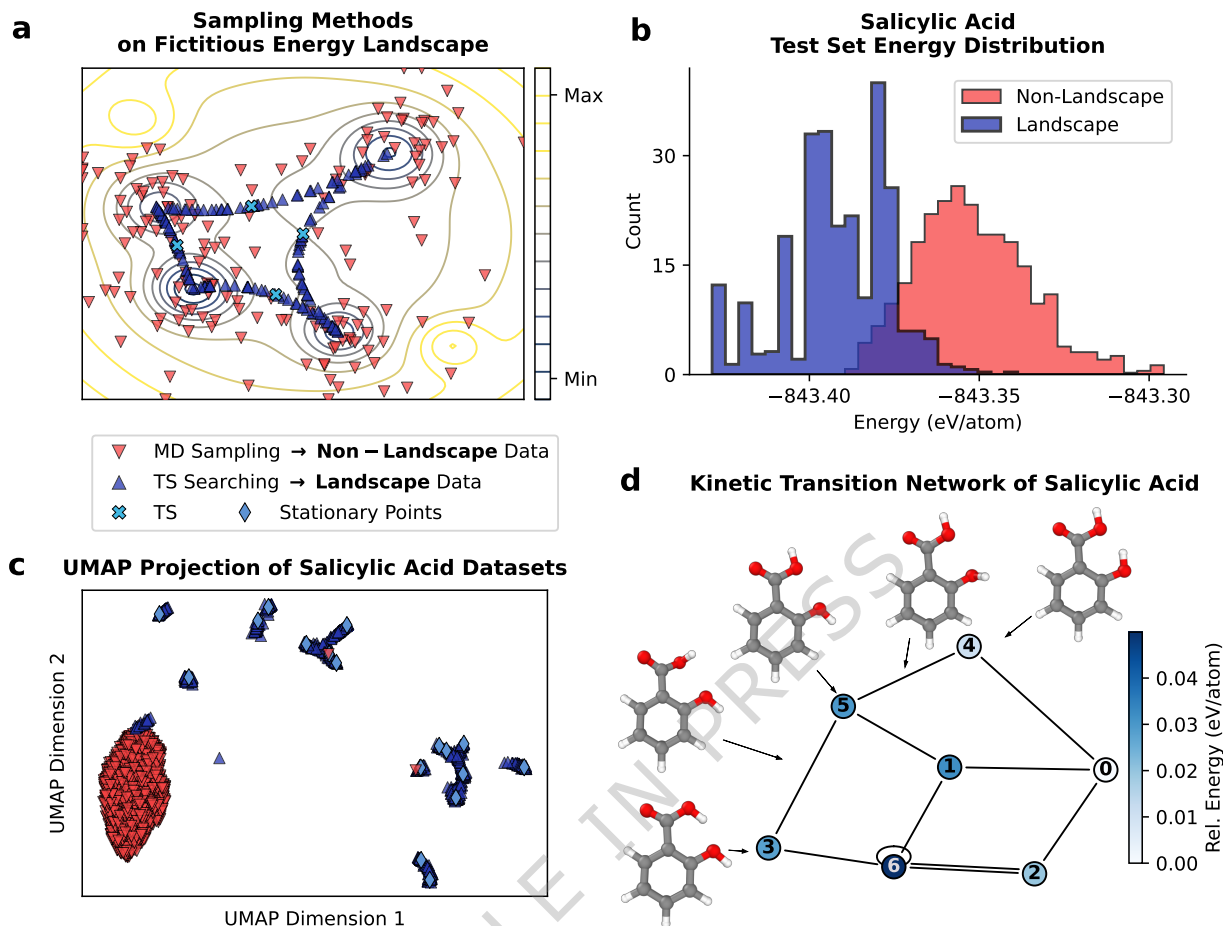
To generate DFT KTNs, we follow an established procedure that begins with basin-hopping global optimization^{81–83} to identify low-energy minima, followed by combined single- and double-ended searches to locate transition states; these methods have been explained in detail elsewhere.^{77,78,84,85} Here, transition states are defined geometrically as points of (approximate) zero gradient and exactly one negative Hessian eigenvalue.⁶³ We utilized TopSearch,⁸⁶ an open-source Python package developed by some of the authors, to perform landscape exploration, complemented by the OPTIM program⁸⁷ for some difficult cases. For each KTN, we excluded repeated permutational isomers and structures related by the inversion operation, as these can be reconstructed through symmetry operations. Further details of the landscape generation process are available in the Methods section and in the Supplementary Information section I.

Table I presents the number of minima and transition states for the six rMD17 molecules in our study. We also collated data along approximate steepest-descent paths from transition states to their connected minima, following small displacements parallel and antiparallel to the eigenvector associated with the unique negative Hessian eigenvalue. The complete dataset – Landscape17 – includes atomic coordinates, energies, forces, and Hessian eigenspectra for all the stationary points, along with positions, energies, and forces for configurations along the pathways.

In this work, we distinguish between two data categories: **Landscape** (L) data from approximate steepest-descent pathways, and **Non-Landscape** (N-L) data from molecular dynamics simulations (e.g., rMD17). Landscape data captures minimum energy paths between stationary points,

TABLE I: Statistics of the Landscape17 dataset, reflecting the counts of minima and transitions states in the DFT landscape of each molecule. The number of stationary points includes only distinct structures, lumping together permutation-inversion isomers.

Molecule	# Min.	# TS
Ethanol	2	2
Malonaldehyde	2	4
Salicylic acid	7	11
Azobenzene	2	4
Paracetamol	4	9
Aspirin	11	37
Total	28	67



representing the essential topology of the potential energy surface, while Non-Landscape data samples beyond this minimal representation of the connectivity. This distinction is illustrated in Fig. 1. The key difference is that the geometry optimization procedures employed in discrete path sampling^{61,62} can treat both high and low barriers, and the associated long and short time scales, directly addressing rare event dynamics and broken ergodicity.

Panel **a** provides a schematic of a hypothetical energy landscape that contrasts the two sampling methods. Panel **b** shows energy distributions for one of the rMD17 molecules: salicylic acid. The distributions reveal that 500 K MD (N-L data) explores higher energy regions than transition state searching (L data) but produces less diverse distributions, as shown in panel **c**. These UMAP⁸⁸ projections of structurally-averaged MACE¹² descriptors further demonstrate a clear separation between regions sampled by the two methods. Landscape data explores diverse environments around stationary points, while, in contrast, Non-Landscape data concentrates in a single region, despite the higher energies. Panel **d** presents the DFT KTN for salicylic acid. KTNs can be visualized as graphs, where nodes represent minima and edges represent transition states. Self-

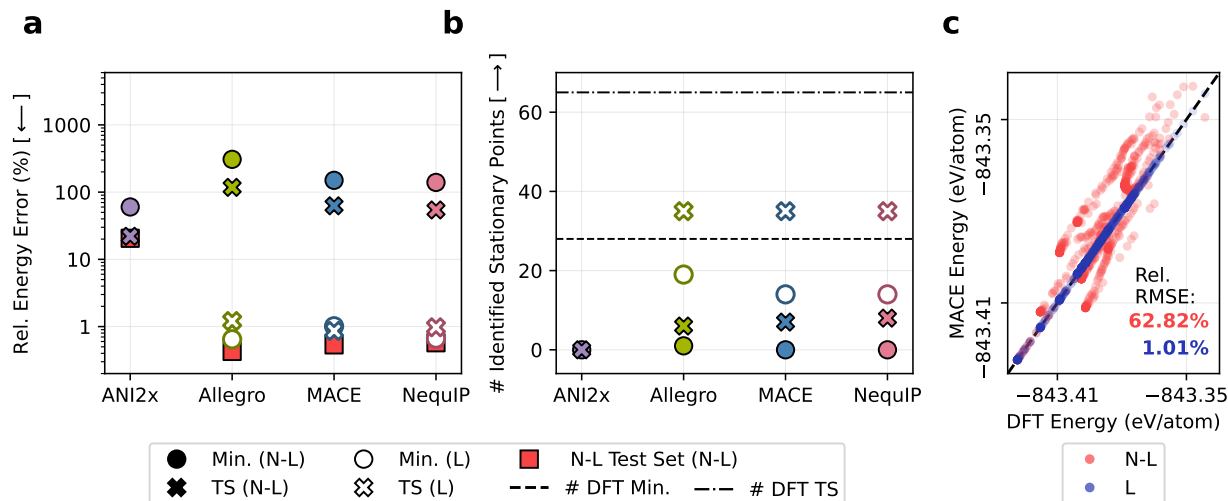
edges indicate transition states between minima related by permutation-inversion symmetry; these are termed degenerate rearrangements.⁸⁹ Similar figures for the other molecules are provided in the Supplementary Information section II.

The targeted, discrete path sampling^{61,62} approach employed in constructing the Landscape17 dataset, with explicit exploration of transition states, enables both rigorous benchmarking of these critical areas of the potential energy surface and systematic analysis of training data effects on model performance. This sampling strategy provides clear insight into how different machine learning potential architectures handle the challenging task of reproducing saddle point regions, which are harder to represent accurately due to the inherently unstable eigendirection with negative curvature.

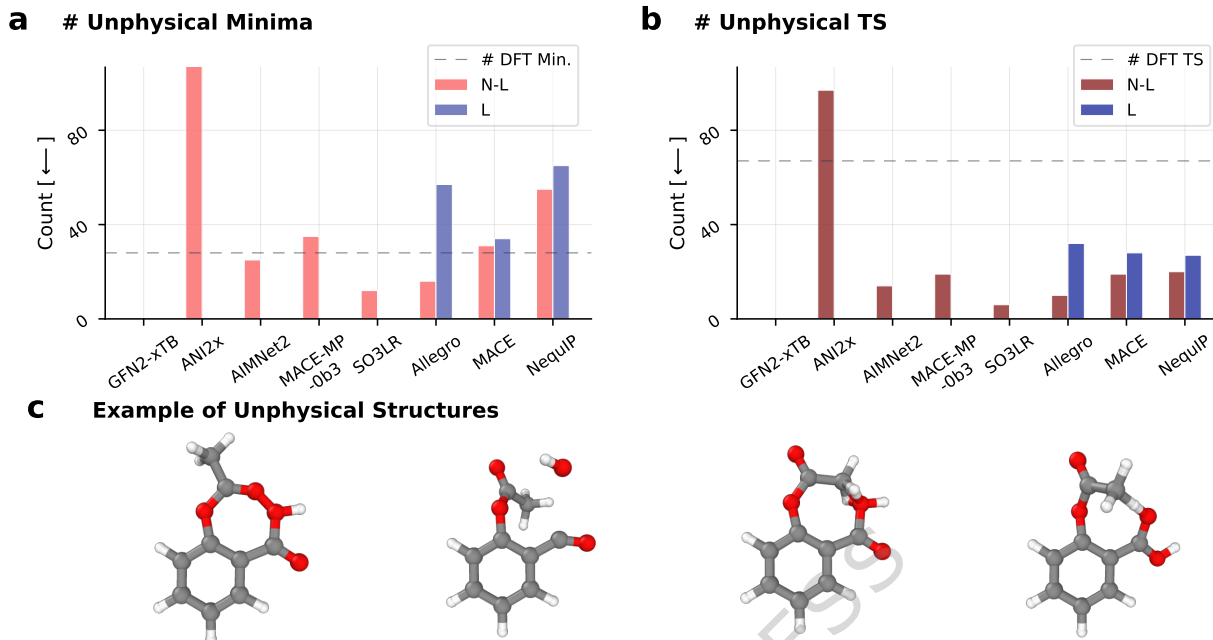
B. Performance Impact of Landscape Data Inclusion

With the Landscape17 dataset we can systematically test whether the inclusion of landscape-specific data into training sets based on non-landscape-specific data (such as molecular dynamics trajectories) can help achieve accurate molecular KTNs. We trained individual models, for each of the six selected rMD17 molecules, following the protocols established in the original publications, using the same data across models. The system-specific architectures employed include NequIP,¹¹ MACE¹² and Allegro,¹³ supplemented by pretrained foundational models: ANI2x,¹⁴ AIMNet2,¹⁵ MACE-MP-0b3²¹ and SO3LR,²⁵ and the semi-empirical method GFN2-xTB.⁷⁹ These models constitute our baseline **Non-Landscape** (N-L) set. Subsequently, we augmented the Non-Landscape training datasets with portions of the newly generated Landscape17 data (see Methods) to create corresponding **Landscape** (L) Allegro, MACE and NequIP models.

The Landscape models demonstrate superior performance in predicting energies and forces at critical points on DFT energy surfaces. Fig. 2 reveals a consistent improvement across all architectures, with systematic reductions in both energy and force errors following landscape data inclusion (with results for ANI2x displayed for reference). Panel **a** shows the average relative root-mean-squared errors (RMSE) for energy predictions on minima and transition states, calculated by normalizing RMSE values against the standard deviations of target distributions. Panel **b** quantifies instances where MLIPs correctly predict maximum force components below convergence thresholds of 10^{-3} eV/Å for minima and 10^{-2} eV/Å for transition states - criteria matching those used for DFT stationary point optimization. Panel **c** presents an example energy parity plot comparing MACE N-L and L model predictions on the aspirin landscape test set. Consistent trends across all



models and molecules are documented through comprehensive MAE and RMSE statistics provided in the Supplementary Information, in Tables S2–S5. It is apparent that N-L models fail, by two orders of magnitude, to accurately predict the energies and forces of stationary points, despite achieving low errors for MD samples, while L models attain low errors on both test sets. This result underlines the difference between models fitted with Non-Landscape and Landscape data, even for the small molecules considered here. Further experiments presented in the Supplementary Information section IV explore the errors in predictions of the lowest energy Hessian eigenmodes and the sum of log (positive) Hessian eigenvalues to corroborate the accuracy of L models.



C. Unphysical structures in MLIP landscapes

To evaluate the performance of machine learning potentials beyond traditional energy and force metrics, we constructed energy landscapes for each MLIP using the same algorithms employed for DFT landscape generation. We report KTNs for each MLIP and molecule after combining results from 20 independent landscape exploration runs, each initiated from different conformations. Permutation-inversion isomers are lumped together (see Figs. S10 and S14 in the Supplementary Information). This multi-run workflow was not feasible for DFT calculations due to the prohibitive computational cost; instead, we employed single extended runs followed by thorough manual enumeration of candidate stationary points (see Methods).

The KTNs for each model were initially assessed through self-consistent analysis without reference to DFT benchmarks, focusing on the physicality of discovered stationary points. We define a structure to be unphysical if its SMILES representation⁹⁰ differs from that of the global minimum, or if it contains hydrogen atoms with multiple bonds.

Our analysis highlights the prevalence of unphysical minima for all MLIPs, Fig. 3. This phenomenon affects both system-specific models trained on N-L/L data (Allegro, MACE, NequIP) and foundational models pretrained on diverse datasets (ANI2x, AIMNet2, MACE-MP-0b3, SO3LR), indicating a systematic limitation across different MLIP architectures and training paradigms. It is very interesting to note that the effects are magnified for the system-specific models trained on Landscape data, suggesting that adding more training data results in models with rougher energy

landscapes. Landscape roughness, and its relation to training data, has been explored in detail elsewhere.^{91–93} Extended plots for each molecule are shown in Fig. S15.

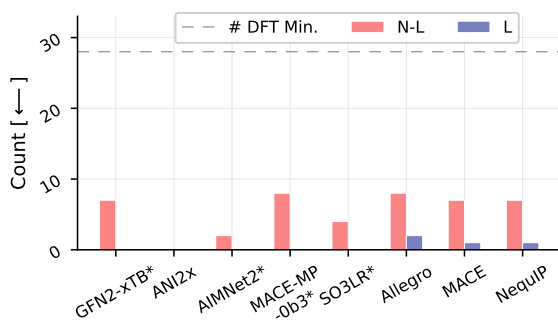
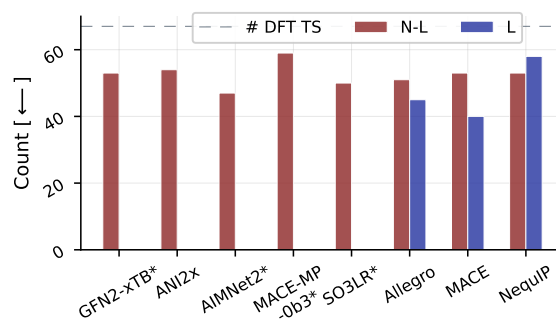
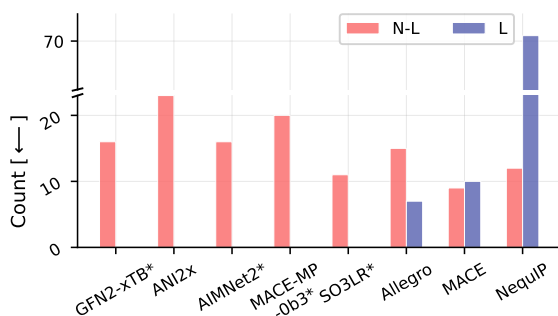
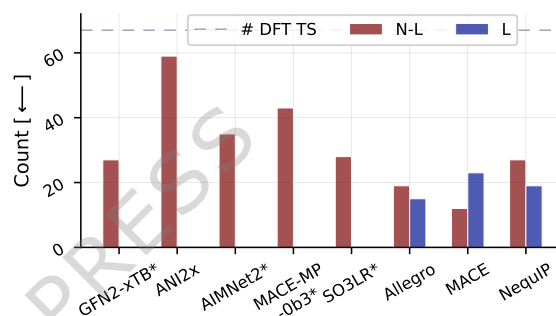
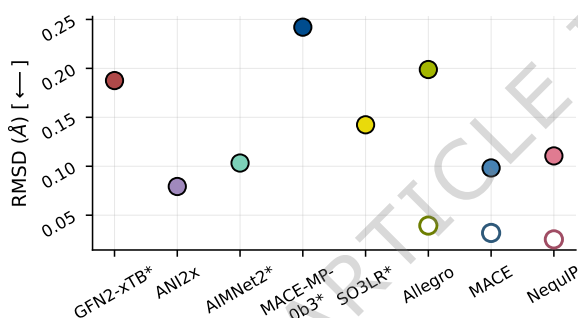
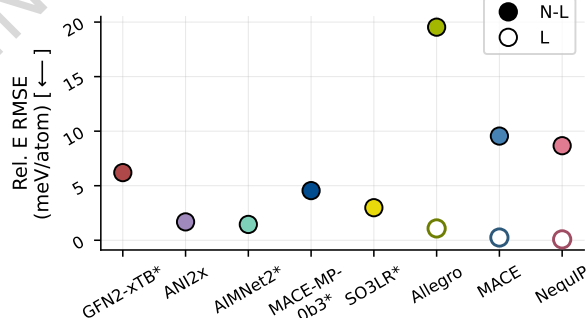
Importantly, our landscape exploration methodology exclusively employs dihedral angle rotations (see Methods), and application to GFN2-xTB reveals no unphysical stationary points. Therefore, the prevalence of such structures serves as a quantitative metric to assess the global MLIP energy landscape. The emergence of unphysical local minima will hinder enhanced sampling methods, since these artificial configurations may have significant statistical weights.

D. Reproduction of reference DFT landscapes by MLIPs

Faithfully reproducing the KTNs for the underlying quantum chemical level of theory is a demanding test for machine learning potentials. To evaluate this capability, we removed the unphysical stationary points from each MLIP landscape and mapped the remaining structures onto the reference DFT landscape using the algorithms detailed in the Methods section, and in the Supplementary Information, Figs. S11–S13. These mapping procedures rely on a structure comparison module that calculates the minimal root-mean-squared-deviation (RMSD) between configurations to establish correspondence between DFT and MLIP stationary points.

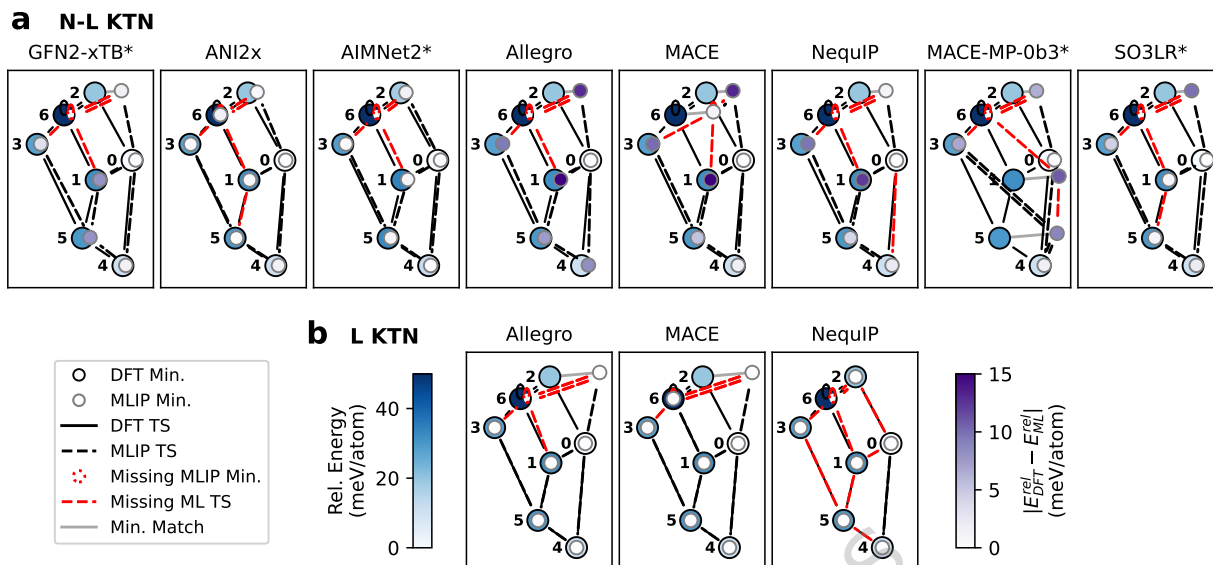
Our comparative analysis addresses two complementary issues: the extent to which DFT KTN stationary points can be *exactly matched* to MLIP analogues, and the structural and relative energy accuracy of the *closest* DFT-MLIP pairs. We define exact matching by an RMSD threshold of 0.3 Å for structural equivalence between DFT and MLIP configurations. For transition state exact matching, we imposed the additional requirement that the connected minima in the MLIP network must themselves be exactly matched to corresponding DFT minima. This stricter criterion for TS is needed as rates and other observables depend strongly on the KTN connectivity graph (see the Supplementary Information section VI). To assess structural and energy fidelity for the closest MLIP analogues of each DFT minimum, we computed average RMSD and relative (to the global minimum) energy RMSE values, intentionally excluding (physical) transition states from this analysis due to their systematic underrepresentation in MLIP landscapes compared to DFT references (see the Supplementary Information Fig. S15). The results for exact matching are shown in Fig. 4 **a–d** (and Fig. S16), and for structural and energy fidelity in **e** and **f**.

The outcome of exact and closest matching demonstrates that incorporating landscape-specific data into training datasets generally enhances the overlap between DFT and MLIP KTNs, improving both structures (lower number of missing minima - panel **a**, lower RMSD - panel **e**) and

a # Missing Minima**b # Missing TS****c # Additional Minima****d # Additional TS****e Structural Similarity of Closest Min. Matches****f Rel. Energy RMSE of Closest Min. Matches**

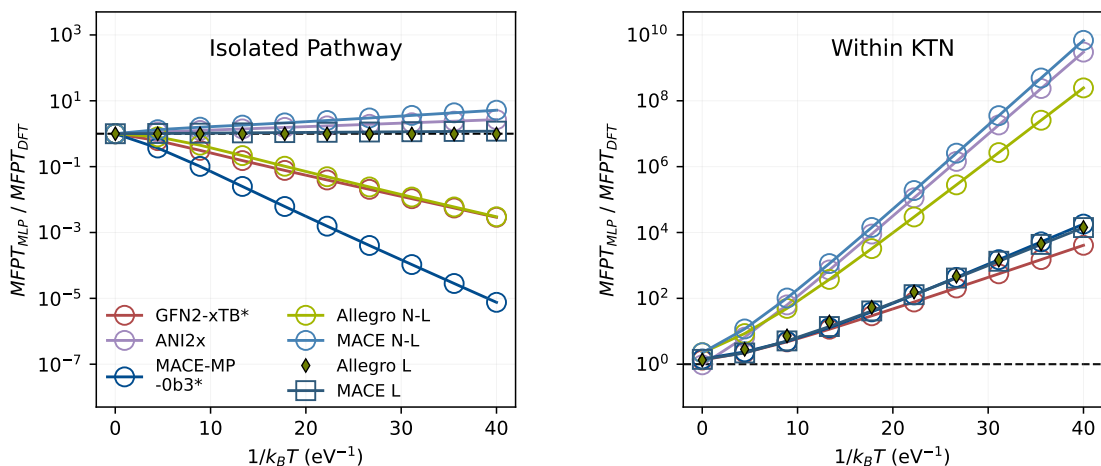
energetics (lower relative energy RMSE - panel **f**). Nevertheless, the challenge of fully reproducing KTNs remains unsolved, especially for transition states (panel **b**). Additionally, panels **c**, **d** show that all of the models predict large numbers of unmatched stationary points. These include MLIP minima and transition states with different structures to their DFT counterparts, and MLIP TSs with matched geometry but different connecting minima. The presence of additional stationary points is a common problem for empirical potentials.⁹⁴ For the purpose of studying landscapes where the DFT KTN is not known *a priori*, it is essential that one chooses a model with the best trade-off between maximizing exact matches and minimizing artifacts.

The substantial number of successful minima matches and low RMSD values observed for models trained on data of different DFT functionals (AIMNet2, MACE-MP-0b3, SO3LR) confirms that the stationary points remain consistent across functional choices for these six molecules. This



observation helps to validate our suggested benchmarking approach.

The case of salicylic acid illustrates general trends clearly (Fig. 5) when we overlay the MLIP networks on their corresponding DFT counterparts. Missing minima and transition state matches are highlighted in red, revealing gaps in MLIP landscape coverage, while additional (unmatched) stationary points are omitted. Most models generate minima structures which are close to their DFT counterparts, as seen by the overlapping small and large nodes. The materials-oriented MACE-MP-0b3 struggles to predict the correct structures for minima 1 and 5. We attribute this to the lack of molecular data in its training. The inclusion of landscape data in Allegro and MACE training datasets produced marked improvements in stationary point matching alongside reduced energy errors, as anticipated from previous work.⁹⁵ The systematic difficulty in physically correct transition state discovery using MLIP energy landscapes emerges as a key problem, with successful transition state searches proving challenging across all tested architectures. Of course, if specific stationary points are not found we cannot be certain that some analogue does not exist on the MLIP landscape; it is possible that some searches do not converge, or they are attracted to different stationary points. However, since the same strategies were followed, and the MLIP searches were more extensive than for the more computationally demanding DFT calculations, the missing structures and corresponding paths must be harder to find, if they exist. We mention that we have tested various different parameter choices for searching the MLIP landscapes, but observed little variation in the numbers of matched stationary points (Fig. S17). The presence of unphysical structures is unambiguous.

Mean First Passage Time Comparison For Pathway 3 → 5 → 4

As a final test we computed mean first passage times (MFPT) between nodes within these networks, using the graph transformation procedure.^{96–98} The MFPT provides a physically meaningful and succinct metric for the overlap between DFT and MLIP KTNs, which includes all the pathways with appropriate weights, whether physical or not. It also depends on the accuracy with which the DFT stationary points are reproduced. Here, we would expect the MFPT to increase significantly if the transition state corresponding to the most favorable pathway for the DFT landscape is not in the MLIP database. In that case, the rate would reflect the highest barrier on the alternative MLIP path. The MFPT could also be too small if the path on the MLIP landscape is too low, and we provide a numerical analysis of the sensitivity of the MFPT to landscape misspecification in Fig. S18. Since the rates depend exponentially on barriers, a proper account of kinetics requires a faithful account of energetics and network connections (with extension to vibrational effects explored in the Supplementary Information, Fig. S9). This test therefore corresponds to a demanding and important challenge.

In Fig. 6 we present MFPTs for a pathway that is well described by most models (with the exception of NequIP), corresponding to the nodes 3→5→4 (c.f. Fig. 5 and Fig. 1). We compute MFPTs in the regime where an unbranched pathway is extracted from the KTN (left panel), and where the pathway is embedded in the rest of the network (right panel). For ease of presentation we only show the best and worst results for foundational models (therefore we disregard SO3LR and AIMNet2), and concentrate on the comparison between N-L and L MACE and Allegro models.

L models perform significantly better than N-L counterparts, with the low-temperature results for the pathway embedded within the KTN leading to a factor of 10^4 and 10^6 improvement in the MFPT in the case of Allegro and MACE, respectively. GFN2-xTB has a low-temperature offset of

10^3 in both the embedded and unbranched single path case, while MACE-MP-0b3 leads to better MFPT results in the embedded case.

III. DISCUSSION

Machine learning potentials offer computational speed that facilitates longer simulations and more extensive exploration of potential energy surfaces compared to traditional quantum chemical methods. However, this improved sampling capacity raises important questions about the ability of these models to accurately reproduce the correct organization of the global energy landscape, and how training data composition influences this ability. Kinetic transition networks provide a concise description of the potential energy surface organization, and their faithful reproduction is essential for computing reliable transition rates. While comprehensive landscape searches remain largely computationally intractable for density functional theory calculations, machine learning potentials make such sampling feasible, creating an opportunity to rigorously test their global accuracy.

We address this challenge by introducing benchmarks for KTN reproduction, providing a framework to validate machine learning potential performance in regions crucial for both thermodynamics and kinetics. We computed complete reference KTNs for six molecules from the rMD17 dataset using hybrid DFT, creating the Landscape17 dataset containing all minima and transition states for ethanol, malonaldehyde, paracetamol, azobenzene, salicylic acid, and aspirin. Using this benchmark, we systematically evaluated three categories of models: foundational models (ANI2x, AIMNet2, MACE-MP-0b3, SO3LR), system-specific models (Allegro, MACE, NequIP), and semi-empirical methods (GFN2-xTB).

To isolate the effects of training data composition on landscape reproduction, we compared two distinct approaches for system-specific models: training exclusively on molecular dynamics data from rMD17 (N-L models) versus augmenting MD data with approximate steepest-descent paths from transition states (L models). This comparison enables us to validate performance beyond force and energy errors, while directly probing how kinetically relevant training data affects potential energy surface representation.

The results reveal a systematic tendency for all machine learning potentials to overestimate the number of stationary points: MLIPs frequently generate unphysical conformations, despite basin-hopping steps constrained to only dihedral rotations. This behavior reflects substantial artifactual structure in machine learning potential energy surfaces compared to the DFT reference, a phenomenon that can be quantitatively analyzed using the landscape exploration methodology

proposed here. It is noteworthy that the GFN2-xTB framework exhibited markedly different behavior, with no unphysical conformations arising during exploration.

The incorporation of landscape-specific data into system-specific model training had two main effects: while increasing the total number of discovered stationary points (including both physically meaningful and spurious configurations), it simultaneously enhanced model performance across multiple metrics. These augmented models exhibit superior accuracy in reproducing DFT-calculated energies and forces at stationary points, while also generating KTNs with improved rates and overall fidelity to reference DFT landscapes. These results suggest that incorporating kinetically relevant training data can effectively improve the representation of potential energy surface curvature in machine learning models.

Nevertheless, achieving perfect correspondence with DFT reference data remains elusive across all the models we have tested. This becomes particularly pronounced for transition state prediction, where MLIP searches produced fewer than half of the DFT transition states with matching structure and connectivity. The persistent challenges in transition state path identification suggest fundamental limitations in current machine learning potential architectures and training methodologies, even for the set of simple, canonical molecules that constitute the Landscape17 dataset. These results do not prove that the missing transition state paths are absent, but we can say that they were not located by the systematic search protocol employed successfully for other potentials. Either way, there is an interesting question to address, in terms of whether the expected TS paths are not supported by the machine learning fits, or why they are somehow elusive.

Based on these results, we suggest that reproducing full conformational landscapes of small molecules could serve as a useful benchmark for evaluating future machine learning potentials. This assessment framework provides a test of model fidelity beyond the usual energy and force prediction metrics, directly probing the ability to capture the features that are essential for kinetic modeling and the tendency of MLIPs to generate stable unphysical structures.

In broader terms, we envision several routes to extend this project. Firstly, we acknowledge the essential need for accurate data and draw attention to recent works on developing wavefunction-based databases and MLIPs for accurate transition state characterization.⁹⁹ The Landscape17 dataset presented here contains rearrangements that proceed largely via dihedral rotation, a type of motion common to many reactive steps, and applied to some of the most common functional groups. However, increased generalizability could be achieved through addition of selected molecules with greater chemical diversity. Secondly, in practical applications, it would be beneficial to understand how to efficiently fine-tune foundational models for the tasks presented here and add the tests to

open-source testing arenas.^{32,100} Finally, the benchmark would be extended to include generative approaches to structure prediction.⁷⁵

IV. METHODS

A. DFT landscapes

The energy landscape framework provides a comprehensive approach to mapping surface topography through the identification and characterization of stationary points.^{78,84,85,101} These are atomic configurations at which the gradient vanishes and we focus on local minima and the transition states that connect them, which are distinguished by their Hessian eigenvalue spectrum. Local minima exhibit only positive and zero Hessian eigenvalues, indicating that any displacement of internal coordinates increases the energy. Transition states are defined as first-order saddle points with exactly one negative eigenvalue, corresponding to a local maximum along the reaction coordinate, with positive curvature in the orthogonal eigendirections (aside from the zero eigenvalues corresponding to overall rotation and translation).¹⁰²

These stationary points can be represented by weighted graphs, known as kinetic transition networks (KTNs),^{58–60} where minima serve as nodes and edges connect minima that are directly linked by transition states. Appropriate post-processing using standard tools of statistical mechanics and unimolecular rate theory enables efficient computation of observable thermodynamic and kinetic properties within well-defined approximations.³⁹ In particular, the explicit inclusion of transition states, which are more difficult to characterize using standard molecular dynamics, allows for assessment of global kinetics and comparison of MLIP landscapes with the DFT reference.^{98,103}

The reference potential energy landscapes were computed using density functional theory with the ω B97x hybrid-energy exchange correlation functional and a 6-31G(d) basis set within Psi4.¹⁰⁴ These settings are consistent with the ones used to generate the ANI2x training data.¹⁴ We applied tight energy and density convergence criteria (E_CONVERGENCE and D_CONVERGENCE) of 10^{-9} Hartree and 10^{-9} a.u., along with extremely fine integration grids (100 radial and 770 spherical points) and a restricted Kohn-Sham reference.

We validated the convergence, grid, and spin settings against published data from rMD17, using the appropriate functional and basis set: PBE/def2-SVP. We achieved energies and forces within 0.1 meV/atom and 5 meV/Å respectively, well within the standard acceptable resolution of 1 meV/atom and 10 meV/Å.

For benchmarking energies of AIMNet2, MACE-MP-0b3 and SO3LR we recomputed the DFT KTN stationary points with the same functionals (and dispersion correction) as detailed in the original publications, specifically: ω B97m-D3BJ, PBE and PBE0-D3BJ. Following AIMNet2, we used the localized def2-TZVPP basis set for all three cases, although we note that the DFT calculations for the MACE-MP-0b3 and SO3LR training datasets originally used plane wave basis sets. While using different functionals might yield different stationary points compared to our ω B97x calculations, we expect the results to be very similar for these small molecules, as confirmed by our comparison results (Fig. 4). Future work could include reconverging the stationary points using these alternative functionals.

Local minima were first collected from the basin-hopping global optimization runs,^{81–83} an approach that has been employed successfully for diverse molecular and abstract landscapes.¹⁰⁵ The basin-hopping exploration employed random angular perturbations applied to flexible dihedral angles, as identified by the Atomic Simulation Environment package.¹⁰⁶ We employ the MMFF94¹⁰⁷ force field to prevent atomic clashes. The surveys used 50 basin-hopping steps with an accept/reject Metropolis condition equivalent to a temperature of 100 K. The convergence criteria required either the maximum force component to fall below 10^{-3} eV/Å or the relative energy change between steps to drop below 10^7 eV multiplied by machine precision.

DFT transition state searches were performed between each minimum and its three nearest neighbors, determined from the Euclidean distance with optimal alignment via the MINPERMDIST routine.^{108,109} This procedure minimises the distance with respect to translation, rotation, permutation and, additionally, we include the inversion operation. The alignment is not deterministic when permutations are included, and employs a shortest augmenting path algorithm¹¹⁰ inside an iterative loop.¹⁰⁸ Distinct stationary points were distinguished by a root mean square distance threshold greater than 0.3 Å and energy differences exceeding 10^{-3} eV. Only one representative permutation-inversion isomer was retained for each minimum and transition state.

Transition states were located using a two-step protocol starting with a nudged elastic band (NEB)^{111–113} calculation. Initial pathways between minima were generated through linear interpolation in internal coordinates after endpoint alignment. For endpoint alignment, extra care has to be taken for highly symmetric molecules, such as malonaldehyde and azobenzene. The NEB algorithm optimizes these interpolations of 20 images with a 50 eV/Å spring constant and convergence criterion of 10^{-2} eV/Å for the maximum force component. This double-ended phase of the calculation only needs to converge sufficiently to identify the local maxima in the profile, which are taken as starting points for accurate refinement using hybrid eigenvector-following.^{114,115} Here the

smallest non-zero eigenvalue and the corresponding eigenvector are obtained using a variational method, with minimisation in all orthogonal directions. Convergence required RMS forces below 3×10^{-2} eV/Å. Transition states were verified through Hessian analysis, confirming exactly one negative eigenvalue. Only transition states with barriers below 1 eV were retained. We do not assume that a path links the two original end minima from the NEB phase, even when there is only a single transition state in the profile. The connectivity of each transition state is always established from the corresponding steepest descent pathways. In general, there may be multiple transition states between two minima, and there may be gaps in the connection profile. The two-phase procedure is applied until a complete discrete path is obtained, using the missing connection algorithm to propose new pairs of minima for additional searches.¹¹⁶

Each transition state connects two local minima, identified through minimization from perturbed transition state geometries. Perturbations of approximately $0.3\mathbf{y}$ ($0.6\mathbf{y}$ for flatter modes, such as the cis-cis transition state in azobenzene) were applied along parallel and antiparallel directions to the normalized eigenvector, \mathbf{y} , corresponding to the negative eigenvalue, followed by LBFGS minimization. All intermediate configurations during minimization were stored to map the approximate steepest-descent paths from transition states to connected minima, excluding initial configurations with force components exceeding 2 eV/Å.

Complete KTNs were constructed for six molecules from the rMD17 dataset:^{33,34} ethanol, malonaldehyde, salicylic acid, azobenzene, paracetamol, and aspirin. These molecules were selected as they contain multiple distinct isomers, enabling meaningful landscape analysis. All calculations utilized the TopSearch Python package,⁸⁶ (or OPTIM program⁸⁷, for azobenzene, see the Supplementary Information I). The resulting compilation of minima, transition states, and approximate steepest-descent pathways constitutes the Landscape17 dataset. Furthermore, we include the Hessian eigenspectrum at the stationary points for reference.

Mean first passage times were computed from the KTNs using the graph transformation approach,^{96–98} as implemented in the PyGT package.¹¹⁷ The individual rates for each elementary transition between minima directly connected by a transition state were computed using harmonic transition state theory with entropic effects set to unity.^{64,65,118,119} The same choice is made consistently to compare the results from the MLIP landscapes and the reference DFT KTN. More accurate elementary transition rates could be used instead, including methods based on explicit dynamics and associated path sampling.^{49–52,54,120} However, the harmonic transition state theory approach is a common choice for studies that treat the global dynamics of a KTN,^{57,121–123} and is appropriate for the comparison that we require here. We compute the rates between the states

corresponding to the nodes 3-4 in Fig. 5. For the unbranched path, we retain only nodes 3, 5 and 4 in the network, removing direct connections between nodes 3-4.

B. Training MLIPs

We trained and evaluated a range of machine learning potentials: Nequip,¹¹ MACE,¹² Allegro,¹³ ANI2x,¹⁴ AIMNet2,¹⁵ MACE-MP-0b3 (medium),²¹ SO3LR;²⁵ but also a semi-empirical method: GFN2-xTB⁷⁹ (via DFTB+).¹²⁴ For detailed architecture descriptions, we refer readers to the original publications. Among these, ANI2x, AIMNet2, MACE-MP-0b3, SO3LR, and GFN2-xTB are transferable potentials that can be used without retraining, while MACE, Allegro, and NequIP are architectures requiring training for each application. We conducted ANI2x and AIMNet2 evaluations on single CPUs, with all other models were evaluated on single NVIDIA A100 GPUs.

We train models both with and without the incorporation of landscape data. In both cases, the custom models were trained from scratch using the hyperparameters specified in their respective publications. For the non-landscape (rMD17) dataset, we used the same structures in a 950/50 test/train split, maintaining consistency with the original papers. For the landscape models we additionally included 40% of the pathway data from Landscape17 for the corresponding molecules (up to 500 structures total), reserving 10% for validation. The remaining 60% of pathway configurations (up to 1000 structures) served as additional test data. For ethanol and malonaldehyde we add fewer than 40% data points, in line with their low number of atoms. Training details and model performance metrics are provided in Secs. SI and SIII.

C. MLIP landscapes

We generated MLIP landscapes using the methodology described in section IV A, but using two nearest neighbors for transition state searching in the first instance. For each MLIP, we created the final KTN by initiating the landscape generation process from 20 different starting structures taken from rMD17. Each starting point produced a separate KTN, which we subsequently merged while eliminating permutation-inversion isomers according to the same similarity criteria used for the reference landscapes. The Supplementary Information contains a flowchart illustrating this process (Fig. S10). Upon completion of the first iteration of MLIP KTNs, we extended them by running transition state searches between pairs of third nearest neighbors and re-running failed TS searches from the first instance (see Fig. S17).

D. Landscape analysis

We evaluated the quality of the MLIP landscapes with respect to the DFT references using various metrics detailed below and in the Supplementary Information section V, using the methodology outlined in Figs. S11, S12 and S13.

A significant limitation of current MLIPs is their tendency to produce unphysical molecular structures. To identify these unphysical structures, we compare each MLIP stationary point against the initial bonding framework. Changes in the adjacency matrix indicate inappropriate bond formation or breaking, which should not occur under the simple dihedral angle rotations that we implement. We detect these issues by monitoring changes in SMILES representations (via RDKit¹²⁵) and hydrogen atom coordination numbers to identify spurious multiple bonding.

As described above, application of the energy landscape framework to a given potential energy surface produces a KTN.⁵⁸⁻⁶⁰ Comparing these networks allows us to evaluate the similarity between potential energy surfaces from different methods, particularly in both low-energy (minima) and higher-energy (transition state) regions. Here, we describe our methodology for quantifying similarities between two networks: a reference DFT-generated landscape and one produced by an MLIP.

Our comparison process begins by removing unphysical structures from the MLIP networks and checking for duplicate structures. We then conduct the comparison along two tracks, as mentioned in the Results section. The first track identifies *exact matches* between DFT and MLIP minima and transition states, while the second track computes similarities between the *closest pairs* of minima using less stringent matching criteria.

Both approaches start by computing similarity matrices between minima in both networks using the MINPERMDIST routine^{108,109} to compute RMSD values. We then match minima using a greedy algorithm that pairs each DFT minimum with its closest MLIP counterpart (we note that Flowchart S13 shows an easier computational alternative for finding exact matches that is in fact equivalent to computing the similarity matrix). For exact matching, we add the requirement that matches must be within 0.3 Å RMSD. Successful matches meet this criterion; failed matches do not. The number of additional minima (shown in Fig. 4 a) is calculated by subtracting successful matches from the total MLIP minima count.

For transition state comparisons, the closest-match approach uses the same algorithm as for minima. The exact-match approach is more stringent, requiring both a 0.3 Å RMSD threshold and that the transition states connect minima that are themselves exactly matched between the DFT

and MLIP KTNs. This is essential for rates calculations (see Supplementary Information section VI). These results appear in Fig. 4 **b**.

In almost all cases, MLIP networks contain more physically-correct minima than the DFT counterparts, resulting in consistent closest match counts across models for each molecule (equal to the DFT minima count). This observation enables meaningful comparison of average RMSD and energy RMSE across models (Fig. 4 **c** and **d**). However, for transition states the counts vary significantly between models and there are often fewer MLIP than DFT physically-correct transition states (Fig. 3 **b**), preventing reliable cross-model comparison.

DATA AVAILABILITY

The Landscape17 dataset, which was generated and analyzed during the current study, is available in the Figshare repository at DOI: <https://doi.org/10.6084/m9.figshare.29949230>.⁸⁰

All DFT data described in this work was generated using the TopSearch Python package (v0.0.3),⁸⁶ available at <https://github.com/IBM/topography-searcher>. The MLIP and GFN2-xTB landscapes were produced using the forked branch available at: https://github.com/VladCarare/topography-searcher/tree/mlp_run, while the analysis employed the forked branch: <https://github.com/VladCarare/topography-searcher/tree/analysis>. An example of how to use the generation and analysis code is available at <https://github.com/VladCarare/mlp-landscapes/tree/main>, along with the production MLP landscapes.

ACKNOWLEDGMENTS

This work was performed using resources provided by the Cambridge Service for Data Driven Discovery (www.csd3.cam.ac.uk), provided by Dell EMC and Intel using Tier-2 funding from the Engineering and Physical Sciences Research Council (capital grant EP/T022159/1), and DiRAC funding from the Science and Technology Facilities Council (www.dirac.ac.uk). VC acknowledges the computational resources obtained through the University of Cambridge EPSRC Core Equipment Award (EP/X034712/1) and EPSRC IAA award number G116766.

We thank Gábor Csányi for useful discussions.

AUTHOR CONTRIBUTIONS

LD, FT and EPK conceived the project. LD, VC, JM and DJW generated the DFT reference data. VC and FT trained the MLIPs. VC generated the MLIP landscapes. VC, FT, LD, JM and EPK performed the data analysis. VC, LD and FT wrote the original draft. All authors read, edited and approved the final manuscript.

COMPETING INTERESTS

All authors declare no financial or non-financial competing interests.

-
- [1] Pyzer-Knapp, E. O. & Curioni, A. Advancing biomolecular simulation through exascale HPC, AI and quantum computing. *Curr. Opin. Struct. Biol.* **87**, 102826 (2024).
 - [2] Deringer, V. L., Caro, M. A. & Csányi, G. Machine lLearning interatomic potentials as emerging tools for materials science. *Advanced Materials* **31**, 1902765 (2019).
 - [3] Unke, O. T. *et al.* Machine learning force fields. *Chem. Rev.* **121**, 10142–10186 (2021).
 - [4] Behler, J. Four generations of high-dimensional neural network potentials. *Chem. Rev.* **121**, 10037–10072 (2021).
 - [5] Thiemann, F. L., O'Neill, N., Kapil, V., Michaelides, A. & Schran, C. Introduction to machine learning potentials for atomistic simulations. *J. Phys.: Condens. Matter* **37**, 073002 (2025).
 - [6] Schütt, K. T., Sauceda, H. E., Kindermans, P.-J., Tkatchenko, A. & Müller, K.-R. SchNet — a deep learning architecture for molecules and materials. *J. Chem. Phys.* **148**, 241722 (2018).
 - [7] Smith, J. S., Isayev, O. & Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* **8**, 3192–3203 (2017).
 - [8] Qiao, Z., Welborn, M., Anandkumar, A., Manby, F. R. & Miller III, T. F. OrbNet: deep learning for quantum chemistry using symmetry-adapted atomic-orbital features. *J. Chem. Phys.* **153**, 124111 (2020).
 - [9] Unke, O. T. *et al.* SpookyNet: learning force fields with electronic degrees of freedom and nonlocal effects. *Nat. Commun.* **12**, 7273 (2021).
 - [10] Gasteiger, J., Groß, J. & Günnemann, S. Directional message passing for molecular graphs. *arXiv* (2022).
 - [11] Batzner, S. *et al.* E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat. Commun.* **13**, 2453 (2022).
 - [12] Batatia, I., Kovács, D. P., Simm, G. N. C., Ortner, C. & Csányi, G. MACE: higher order equivariant message passing neural networks for fast and accurate force fields. *NeurIPS* **35** (2022).

- [13] Musaelian, A. *et al.* Learning local equivariant representations for large-scale atomistic dynamics. *Nat. Commun.* **14**, 579 (2023).
- [14] Devereux, C. *et al.* Extending the applicability of the ANI deep learning molecular potential to sulfur and halogens. *J. Chem. Theory Comput.* **16**, 4192–4202 (2020).
- [15] Anstine, D., Zubatyuk, R. & Isayev, O. AIMNet2: a neural network potential to meet your neutral, charged, organic, and elemental-organic needs. *Chem. Sci.* **16**, 10228–10244 (2025).
- [16] Bartók, A. P., Payne, M. C., Kondor, R. & Csányi, G. Gaussian Approximation Potentials: the accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.* **104**, 136403 (2010).
- [17] Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.* **134**, 074106 (2011).
- [18] Drautz, R. Atomic cluster expansion for accurate and transferable interatomic potentials. *Phys. Rev. B* **99**, 014104 (2019).
- [19] Chmiela, S., Sauceda, H. E., Poltavsky, I., Müller, K.-R. & Tkatchenko, A. sGDML: Constructing accurate and data efficient molecular force fields using machine learning. *Computer Physics Communications* **240**, 38–45 (2019).
- [20] Huo, H. & Rupp, M. Unified representation of molecules and crystals for machine learning. *Mach. Learn.: Sci. Tech.* **3**, 045017 (2022).
- [21] Batatia, I. *et al.* A foundation model for atomistic materials chemistry. *arXiv* (2023).
- [22] Rhodes, B. *et al.* Orb-v3: atomistic simulation at scale. *arXiv* (2025).
- [23] Fu, X. *et al.* Learning smooth and expressive interatomic potentials for physical property prediction. *arXiv* (2025).
- [24] Kovács, D. *et al.* MACE-OFF: short-range transferable machine learning force fields for organic molecules. *J. Am. Chem. Soc.* **147**, 17598–17611 (2025).
- [25] Kabylda, A. *et al.* Molecular simulations with a pretrained neural network and universal pairwise force fields. *ChemRxiv* (2025).
- [26] Frank, T., Unke, O., Müller, K.-R. & Chmiela, S. A Euclidean transformer for fast and stable machine learned force fields. *Nat. Commun.* **15**, 6539 (2024).
- [27] Wood, B. M. *et al.* UMA: a family of universal models for atoms. *arXiv* (2025).
- [28] Deng, B. *et al.* CHGNet: pretrained universal neural network potential for charge-informed atomistic modeling. *Nat. Mach. Intell.* **5**, 1031–1041 (2023).
- [29] Eastman, P. *et al.* SPICE, a dataset of drug-like molecules and peptides for training machine learning potentials. *Sci. Data* **10**, 11 (2023).
- [30] Shuaibi, D. S. L. M. *et al.* The Open Molecules 2025 (OMol25) dataset, evaluations, and models. *arXiv* (2025).
- [31] Ramakrishnan, R., Dral, P. O., Rupp, M. & von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **1**, 140022 (2014).

- [32] Riebesell, J. *et al.* A framework to evaluate machine learning crystal stability predictions. *Nat. Mach. Intell.* **7**, 1–12 (2025).
- [33] Chmiela, S. *et al.* Machine learning of accurate energy-conserving molecular force fields. *Sci. Adv.* **3**, e1603015 (2017).
- [34] Christensen, A. S. & von Lilienfeld, O. A. On the role of gradients for machine learning of molecular energies and forces. *Mach. Learn.: Sci. Tech.* **1**, 045018 (2020).
- [35] Pelaez, R. P. *et al.* TorchMD-Net 2.0: fast neural network potentials for molecular simulations. *J. Chem. Theory Comput.* **20**, 4076–4087 (2024).
- [36] Hartmann, C., Banisch, R., Sarich, M., Badowski, T. & Schütte, C. Characterization of rare events in molecular dynamics. *Entropy* **16**, 350–376 (2014).
- [37] Staub, R., Gantzer, P., Harabuchi, Y., Maeda, S. & Varnek, A. Challenges for kinetics predictions via neural network potentials: a Wilkinson’s catalyst case. *Molecules* **28**, 4477 (2023).
- [38] Shenoy, N. *et al.* Role of structural and conformational diversity for machine learning potentials. *arXiv* (2023).
- [39] Swinburne, T. D., Kannan, D., Sharpe, D. J. & Wales, D. J. Rare events and first passage time statistics from the energy landscape. *J. Chem. Phys.* **153**, 134115 (2020).
- [40] Kong, L. & Bryce, R. A. Discriminating high from low energy conformers of druglike molecules: an assessment of machine learning potentials and quantum chemical methods. *ChemPhysChem* **26**, e202400992 (2025).
- [41] Hukushima, K. & Nemoto, K. Exchange Monte Carlo method and application to spin glass simulations. *J. Phys. Soc. Japan* **65**, 1604 (1996).
- [42] Tesi, M. C., van Rensburg, E. J. J., Orlandini, E. & Whittington, S. G. Monte Carlo study of the interacting self-avoiding walk model in three dimensions. *J. Stat. Phys.* **82**, 155 (1996).
- [43] Hansmann, U. H. E. Parallel tempering algorithm for conformational studies of biological molecules. *Chem. Phys. Lett.* **281**, 140–150 (1997).
- [44] Sugita, Y. & Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **314**, 141–151 (1999).
- [45] Nakajima, N., Nakamura, H. & Kidera, A. Multicanonical ensemble generated by molecular dynamics simulation for enhanced conformational sampling of peptides. *J. Phys. Chem. B* **101**, 817–824 (1997).
- [46] Wang, F. & Landau, D. P. Efficient, multiple-range random walk algorithm to calculate the density of states. *Phys. Rev. Lett.* **86**, 2050 (2001).
- [47] Torrie, G. M. & Valleau, J. P. Monte Carlo free energy estimates using non-Boltzmann sampling: application to the sub-critical Lennard-Jones fluid. *Chem. Phys. Lett.* **28**, 578 (1974).
- [48] Laio, A. & Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci. USA* **99**, 12562–12566 (2002).
- [49] Bolhuis, P. G., Chandler, D., Dellago, C. & Geissler, P. L. Transition path sampling: throwing ropes over rough mountain passes, in the dark. *Annu. Rev. Phys. Chem.* **53**, 291–318 (2002).

- [50] van Erp, T. S., Moroni, D. & Bolhuis, P. G. A novel path sampling method for the calculation of rate constants. *J. Chem. Phys.* **118**, 7762–7774 (2003).
- [51] Faradjian, A. K. & Elber, R. Computing time scales from reaction coordinates by milestoning. *J. Chem. Phys.* **120**, 10880–10889 (2004).
- [52] Allen, R. J., Frenkel, D. & ten Wolde, P. R. Forward flux sampling-type schemes for simulating rare events: efficiency analysis. *J. Chem. Phys.* **124**, 194111 (2006).
- [53] Chodera, J. D. *et al.* Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *J. Chem. Phys.* **126**, 155101 (2007).
- [54] Dellago, C. & Bolhuis, P. G. Transition path sampling and other advanced simulation techniques for rare events. *Adv. Polymer Sci.* **221**, 167–233 (2009).
- [55] Pande, V. S., Beauchamp, K. & Bowman, G. R. Everything you wanted to know about Markov State Models but were afraid to ask. *Methods* **52**, 99–105 (2010).
- [56] Prinz, J. H. *et al.* Markov models of molecular kinetics: generation and validation. *J. Chem. Phys.* **134**, 174105 (2011).
- [57] Swinburne, T. D. & Perez, D. Self-optimized construction of transition rate matrices from accelerated atomistic simulations with Bayesian uncertainty quantification. *Phys. Rev. Materials* **2**, 053802 (2018).
- [58] Noé, F. & Fischer, S. Transition networks for modeling the kinetics of conformational change in macromolecules. *Curr. Opin. Struct. Biol.* **18**, 154–162 (2008).
- [59] Prada-Gracia, D., Gómez-Gardenes, J., Echenique, P. & Falo, F. Exploring the free energy landscape: from dynamics to networks and back. *PLoS Comput. Biol.* **5**, e1000415 (2009).
- [60] Wales, D. J. Energy landscapes: some new horizons. *Curr. Opin. Struct. Biol.* **20**, 3–10 (2010).
- [61] Wales, D. J. Discrete path sampling. *Mol. Phys.* **100**, 3285–3306 (2002).
- [62] Wales, D. J. Some further applications of discrete path sampling to cluster isomerization. *Mol. Phys.* **102**, 891–908 (2004).
- [63] Murrell, J. N. & Laidler, K. J. Symmetries of activated complexes. *Trans. Faraday. Soc.* **64**, 371–377 (1968).
- [64] Forst, W. *Theory of Unimolecular Reactions* (Academic Press, New York, 1973).
- [65] Laidler, K. J. *Chemical Kinetics* (Harper & Row, New York, 1987).
- [66] van Kampen, N. G. *Stochastic processes in physics and chemistry* (North-Holland, Amsterdam, 1981).
- [67] Bowman, G. R., Beauchamp, K. A., Boxer, G. & Pande, V. S. Progress and challenges in the automated construction of Markov state models for full protein systems. *J. Chem. Phys.* **131**, 124101 (2009).
- [68] Lane, T. J., Bowman, G. R., Beauchamp, K., Voelz, V. A. & Pande, V. S. Markov state model reveals folding and functional dynamics in ultra-long MD trajectories. *J. Am. Chem. Soc.* **133**, 18413–18419 (2011).
- [69] Schreiner, M., Bhowmik, A., Vegge, T., Busk, J. & Winther, O. Transition1x – a dataset for building generalizable reactive machine learning potentials. *Sci. Data* **9**, 779 (2022).

- [70] de Oca Zapiain, D. M. *et al.* Training data selection for accuracy and transferability of interatomic potentials. *Npj Comput. Mater.* **8**, 189 (2022).
- [71] Kulichenko, M. *et al.* Data generation for machine learning interatomic potentials and beyond. *Chem. Rev.* **124**, 13681–13714 (2024).
- [72] Niblett, S. P., Kourtis, P., Magdău, I.-B., Grey, C. P. & Csányi, G. Transferability of data sets between machine-learned interatomic potential algorithms. *J. Chem. Inf. Model.* **21**, 6096–6112 (2025).
- [73] Kim, S., Woo, J. & Kim, W. Y. Diffusion-based generative AI for exploring transition states from 2D molecular graphs. *Nat. Commun.* **15**, 341 (2024).
- [74] Choi, S. Prediction of transition state structures of gas-phase chemical reactions via machine learning. *Nat. Commun.* **14**, 1168 (2023).
- [75] Duan, C. *et al.* Optimal transport for generating transition states in chemical reactions. *Nat. Mach. Intell.* **7**, 615–626 (2025).
- [76] Kuryla, D., Csányi, G., van Duin, A. C. T. & Michaelides, A. Efficient exploration of reaction pathways using reaction databases and active learning. *J. Chem. Phys.* **162** (2025).
- [77] Wales, D. J. *Energy Landscapes* (Cambridge University Press, Cambridge, UK, 2003).
- [78] Wales, D. J. Exploring energy landscapes. *Ann. Rev. Phys. Chem.* **69**, 401–425 (2018).
- [79] Bannwarth, C., Ehlert, S. & Grimme, S. GFN2-xTB – an accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *J. Chem. Theory Comput.* **15**, 1652–1671 (2019).
- [80] Cărare, V. *et al.* Landscape17. figshare 10.6084/m9.figshare.29949230 (2025).
- [81] Li, Z. & Scheraga, H. A. Monte Carlo-minimization approach to the multiple-minima problem in protein folding. *Proc. Natl. Acad. Sci. USA* **84**, 6611–6615 (1987).
- [82] Wales, D. J. & Doye, J. P. K. Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms. *J. Phys. Chem. A* **101**, 5111–5116 (1997).
- [83] Wales, D. J. & Scheraga, H. A. Global optimization of clusters, crystals and biomolecules. *Science* **285**, 1368–1372 (1999).
- [84] Joseph, J. A., Röder, K., Chakraborty, D., Mantell, R. G. & Wales, D. J. Exploring biomolecular energy landscapes. *Chem. Commun.* **53**, 6974–6988 (2017).
- [85] Röder, K., Joseph, J. A., Husic, B. E. & Wales, D. J. Energy landscapes for proteins: from single funnels to multifunctional systems. *Adv. Theory Simul.* **2**, 1800175 (2019).
- [86] Dicks, L. & Pyzer-Knapp, E. O. TopSearch: a Python package for topographical analysis of machine learning models and physical systems. *J. Open Source Soft.* **20**, 317–330 (2024).
- [87] Wales, D. J. OPTIM (2025).
- [88] McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Soft.* **3**, 861 (2018).
- [89] Nourse, J. G. Self-inverse and nonself-inverse degenerate isomerizations. *J. Am. Chem. Soc.* **102**, 4883 (1980).

- [90] Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**, 31–36 (1988).
- [91] de Souza, V. K., Stevenson, J. D., Niblett, S. P., Farrell, J. D. & Wales, D. J. Defining and quantifying frustration in the energy landscape: applications to atomic and molecular clusters, biomolecules, jammed and glassy systems. *J. Chem. Phys.* **146** (2017).
- [92] Dicks, L., Graff, D. E., Jordan, K. E., Coley, C. W. & Pyzer-Knapp, E. O. A physics-inspired approach to the understanding of molecular representations and models. *Mol. Syst. Des. Eng.* **9**, 449–455 (2024).
- [93] Wilson, M. P., Pyzer-Knapp, E. O., Galichet, N. & Dicks, L. Refining embeddings with fill-tuning: data-efficient generalised performance improvements for materials foundation models. *arXiv* (2025).
- [94] Furman, D. & Wales, D. J. Transforming the accuracy and numerical stability of ReaxFF reactive force fields. *J. Phys. Chem. Lett.* **10**, 7215–7223 (2019).
- [95] Csányi, G., Morgan, J. W. R. & Wales, D. J. Global analysis of energy landscapes for materials modeling: a test case for C₆₀. *J. Chem. Phys.* **159**, 104107 (2023).
- [96] Trygubenko, S. A. & Wales, D. J. Kinetic analysis of discrete path sampling stationary point databases. *Mol. Phys.* **104**, 1497–1507 (2006).
- [97] Trygubenko, S. A. & Wales, D. J. Graph transformation method for calculating waiting times in markov chains. *J. Chem. Phys.* **124**, 234110 (2006).
- [98] Sharpe, D. J. & Wales, D. J. Nearly reducible finite markov chains: theory and algorithms. *J. Chem. Phys.* **155**, 140901 (2021).
- [99] Allen, A. E. A. *et al.* Reactive Chemistry at Unrestricted Coupled Cluster Level: High-throughput Calculations for Training Machine Learning Potentials. *arXiv* (2025).
- [100] Zhao, Q. *et al.* Harnessing Machine Learning to Enhance Transition State Search with Interatomic Potentials and Generative Models. *Advanced Science* **12** (2025).
- [101] Niroomand, M. P., Dicks, L., Pyzer-Knapp, E. O. & Wales, D. J. Insights into machine learning models from chemical physics: an energy landscapes approach (EL for ML). *Digital Discovery* **3**, 637–648 (2024).
- [102] Murrell, J. N. & Laidler, K. J. Symmetries of activated complexes. *Trans. Faraday Soc.* **64**, 371–377 (1968).
- [103] Woods, E. J. & Wales, D. J. Analysis and interpretation of first passage time distributions featuring rare events. *Phys. Chem. Chem. Phys.* **26**, 1640–1657 (2024).
- [104] Smith, D. G. A. *et al.* Psi4 1.4: Open-source software for high-throughput quantum chemistry. *J. Chem. Phys.* **152**, 184108 (2020).
- [105] Dicks, L. & Wales, D. J. Exploiting sequence-dependent rotamer information in global optimization of proteins. *J. Phys. Chem. B* **126**, 8381–8390 (2022).
- [106] Larsen, A. H. *et al.* The atomic simulation environment – a Python library for working with atoms. *J. Phys. Condens. Matter* **29**, 273002 (2017).

- [107] Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *Journal of Computational Chemistry* **17**, 490–519 (1996).
- [108] Wales, D. J. & Carr, J. M. Quasi-continuous interpolation scheme for pathways between distant configurations. *J. Chem. Theory Comput.* **8**, 5020–5034 (2012).
- [109] Griffiths, M., Niblett, S. P. & Wales, D. J. Optimal alignment of structures for finite and periodic systems. *J. Chem. Theory Comput.* **13**, 4914–4931 (2017).
- [110] Jonker, R. & Volgenant, A. A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing* **38**, 325 (1987).
- [111] Jónsson, H., Mills, G. & Jacobsen, K. W. Nudged elastic band method for finding minimum energy paths of transitions. In *Classical and Quantum Dynamics in Condensed Phase Simulations*, chap. 16, 385–404 (World Scientific, Singapore, 1998).
- [112] Henkelman, G., Uberuaga, B. P. & Jónsson, H. A climbing image nudged elastic band method for finding saddle points and minimum energy paths. *J. Chem. Phys.* **113**, 9901–9904 (2000).
- [113] Henkelman, G. & Jónsson, H. Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points. *J. Chem. Phys.* **113**, 9978–9985 (2000).
- [114] Munro, L. J. & Wales, D. J. Defect migration in crystalline silicon. *Phys. Rev. B* **59**, 3969–3980 (1999).
- [115] Kumeda, Y., Munro, L. J. & Wales, D. J. Transition states and rearrangement mechanisms from hybrid eigenvector-following and density functional theory. application to C₁₀H₁₀ and defect migration in crystalline silicon. *Chem. Phys. Lett.* **341**, 185–194 (2001).
- [116] Carr, J. M., Trygubenko, S. A. & Wales, D. J. Finding pathways between distant local minima. *J. Chem. Phys.* **122**, 234903 (2005).
- [117] Swinburne, T. D. & Wales, D. J. Defining, calculating, and converging observables of a kinetic transition network. *J. Chem. Theory Comput.* **16**, 2661–2679 (2020).
- [118] Eyring, H. The activated complex and the absolute rate of chemical reactions. *Chem. Rev.* **17**, 65 (1935).
- [119] Evans, M. G. & Polanyi, M. Some applications of the transition state method to the calculation of reaction velocities, especially in solution. *Trans. Faraday Soc.* **31**, 875 (1935).
- [120] van Erp, T. S. & Bolhuis, P. G. Elaborating transition interface sampling methods. *J. Comput. Phys.* **205**, 157–181 (2005).
- [121] Sørensen, M. R. & Voter, A. F. Temperature-accelerated dynamics for simulation of infrequent events. *J. Chem. Phys.* **112**, 9599–9606 (2000).
- [122] Xiao, P., Wu, Q. & Henkelman, G. Basin constrained κ -dimer method for saddle point finding. *J. Chem. Phys.* **141**, – (2014).
- [123] Sharia, O. & Henkelman, G. Analytic dynamical corrections to transition state theory. *New J. Phys.* **18**, 013023 (2016).

- [124] Hourahine, B. *et al.* DFTB+, a software package for efficient approximate density functional theory based atomistic simulations. *J. Chem. Phys.* **152**, 124101 (2020).
- [125] Landrum, G. *et al.* rdkit/rdkit: 2023_09_6 (q3 2023) release (2024).
- [126] Stukowski, A. Visualization and analysis of atomistic simulation data with OVITO – the open visualization tool. *Model. Simul. Mater. Sci. Eng.* **18**, 015012 (2009).

ARTICLE IN PRESS

FIG. 1: Contrasting **Landscape** with **Non-Landscape** data. Non-Landscape data is sampled from MD trajectories, while Landscape data is sampled from approximate paths between transition states and the connected minima. **a)** Sketch of the distributions resulting from the two sampling schemes. **b)** Energy histogram comparison for test sets of Non-Landscape and Landscape structures of salicylic acid. The former data was collected from an MD run at 500 K (rMD17³⁴). **c)** UMAP⁸⁸ projections of the structurally-averaged MACE¹² descriptors of three salicylic acid datasets, showing clear distinction between Landscape and Non-Landscape datasets. **d)** Visual representation of the DFT KTN of salicylic acid. Minima are given by nodes and transition states by edges connecting the nodes. Self-loops and multi-edges are permitted, representing transition states between degenerate rearrangements and alternative transition states, respectively. The colors of the nodes correspond to the energy of the corresponding minima relative to the global minimum. We highlight the structures along the 3-5-4 pathway. Images were rendered using OVITO.¹²⁶

FIG. 2: Model performance on the stationary points of the DFT landscapes. **a)** Relative energy errors of MLIPs on minima, transition states, and an rMD17 test set, for models trained on Non-Landscape data (N-L models) and the combination of Non-Landscape & Landscape data (L models). Values for the N-L test set and for ANI2x are shown for reference. Relative energy errors are obtained by dividing the root mean squared errors by the standard deviations of the ground truth energy distributions. The results are averaged over all six molecules. **b)** Number of DFT minima and transition states where the maximum force components are correctly predicted to be smaller than 10^{-3} eV/Å and 10^{-2} eV/Å respectively, by the corresponding model. **c)** Energy errors parity plot for predictions of Non-Landscape and Landscape MACE models on a Landscape test set. Error tables for all the other molecules and models are presented in the Supplementary Information section III.

FIG. 3: Physicality of MLIP energy landscapes. **a)** Cumulative counts of physical and unphysical minima and **b)** transition states resulting from energy landscape exploration using the corresponding N-L or L models. *Non-physicality* of a structure is defined by a difference in its SMILES representation from that of the global minimum conformation of the respective molecule, or by the presence of a multiply-bonded hydrogen atom. **c)** Examples of unphysical configurations for the aspirin molecule. Images were rendered using OVITO.¹²⁶

FIG. 4: DFT-MLIP KTNs comparison. **a)–d)** *exact* and **e), f)** *closest* DFT matching statistics for each MLIP KTN. The exact match counts the number of MLIP stationary points that can (Matched) and cannot (Additional) be matched to within an RMSD of 0.3 \AA to DFT analogues. The number of Missing stationary points is computed by subtracting the count of matches from the number of DFT stationary points. The closest match computes the RMSD and relative energy RMSE of closest MLIP-KTN minima pairs. TS pairs are not considered because of the inconsistent number of pairings between models. DFT supports a total of 28 minima and 67 transition states. The asterisk indicates that energy and forces of the DFT KTN were reevaluated with the DFT functional matching the training set of the corresponding MLIP.

FIG. 5: KTNs of MLIPs superimposed on the reference DFT for salicylic acid. Large circles represent DFT minima and solid black lines indicate transition states. Distances between DFT minima are scaled quadratically with the energy barrier between them. Small overlaid circles represent MLIP minima, with gray contours indicating a matched pair of DFT-MLIP minima (according to an RMSD threshold of 0.3 \AA), and red indicating a missing MLIP structure. Matched pairs are linked by gray lines, of distance proportional to the RMSD between minima at optimal alignment. Black dashed lines represent transition states in the MLIP KTN which can be mapped to transition states in the DFT KTN. The correspondence requires successful MLIP-DFT mapping for the connected minima, as well as structure similarity of 0.3 \AA between the TS. A red dashed line represent a lack of such a TS. For DFT minima, the colors indicate the energy difference of the node with respect to the global minimum (node 0). For MLIP minima, the colors track the energy difference to their corresponding matched DFT minimum, with white implying a low error.

FIG. 6: Mean first passage times between nodes 3 and 4 (c.f. Fig. 5) as computed from various MLIP KTNs and reported with respect to the DFT values. In the left panel we show the case where the path 3-5-4 is isolated from the rest of the network, while in the right panel we retain the entire KTN.