Adaptive Transition State Refinement with Learned Equilibrium Flows

Samir Darouich
Institute of Theoretical Chemistry / Artificial Intelligence
University of Stuttgart
samir.darouich@theochem.uni-stuttgart.de

Vinh Tong
Institute of Artificial Intelligence
University of Stuttgart
vinh.tong@ki.uni-stuttgart.de

Tanja Bien
Institute of Artificial Intelligence
University of Stuttgart
tanja.bien@ki.uni-stuttgart.de

Johannes Kästner
Institute of Theoretical Chemistry
University of Stuttgart
kaestner@theochem.uni-stuttgart.de

Mathias Niepert
Institute of Artificial Intelligence
University of Stuttgart
mathias.niepert@ki.uni-stuttgart.de

Abstract

Identifying transition states (TSs), the high-energy configurations that molecules pass through during chemical reactions, is essential for understanding and designing chemical processes. However, accurately and efficiently identifying these states remains one of the most challenging problems in computational chemistry. In this work, we introduce a new generative AI approach that improves the quality of initial guesses for TS structures. Our method can be combined with a variety of existing techniques, including both machine learning models and fast, approximate quantum methods, to refine their predictions and bring them closer to chemically accurate results. Applied to TS guesses from a state-of-the-art machine learning model, our approach reduces the median structural error to just 0.088 Å and lowers the median absolute error in reaction barrier heights to 0.79 kcal mol^{-1}. When starting from a widely used tight-binding approximation, it increases the success rate of locating valid TSs by 41% and speeds up high-level quantum optimization by a factor of three. By making TS searches more accurate, robust, and efficient, this method could accelerate reaction mechanism discovery and support the development of new materials, catalysts, and pharmaceuticals.

1 Introduction

The transition state (TS) plays a central role in elucidating reaction mechanisms and understanding the microkinetic behavior of chemical processes (Truhlar et al., 1996; Peng et al., 2016; Dewyer et al., 2018; von Lilienfeld et al., 2020; Unsleber and Reiher, 2020; Nandy et al., 2021; Jorner et al., 2021). A detailed knowledge of the underlying kinetics enables the rational design of catalysts, synthetic routes, and functional materials, driving progress toward more efficient, sustainable, and innovative chemical processes (Taylor et al., 2023; Chacko et al., 2024). Computationally, a TS corresponds to a first-order saddle point on the potential energy surface (PES). Classical algorithms exist to locate TSs (Jónsson et al., 1998), but when paired with high-level electronic structure methods, such as density functional

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for Science: The Reach and Limits of AI for Scientific Discovery.

theory (DFT) (Mardirossian and Head-Gordon, 2017), they become computationally prohibitive. This presents a major bottleneck for the scalable discovery of reaction mechanisms.

To mitigate this, machine-learned potentials offer an efficient surrogate for the PES, enabling faster TS search (Yuan et al., 2024). As an alternative to iterative search algorithms, deep learning methods have emerged that aim to directly predict transition state structures. These approaches vary in form, from models that infer TS distance matrices (Choi, 2023) to generative frameworks that attempt to learn the joint distribution over reactants, products, and TSs (Duan et al., 2023). While generative models are promising, they can struggle to resolve fine-grained geometric details, sometimes producing unphysical features such as atomic collisions or distorted bond lengths (Peng et al., 2023; Williams and Inala, 2024; Vost et al., 2025; Wohlwend et al., 2025; Galustian et al., 2025). In contrast, TS guesses from approximate quantum chemical methods, like tight-binding, tend to be physically plausible, but can systematically deviate from DFT-level structures (Rasmussen and Jensen, 2020). In both scenarios, the predicted TS structures function as low-fidelity approximations that, while providing valuable initial estimates for reaction exploration, may require further refinement to achieve the accuracy needed for quantitatively reliable kinetic analysis.

To address this gap, we introduce Adaptive Equilibrium Flow Matching (AEFM), a structure-only refinement method that transforms low-fidelity TS guesses, regardless of their origin, into high-accuracy transition state geometries. AEFM learns to invert noise-injected perturbations of reference TS structures using a novel time-independent form of variational flow matching (VFM) (Eijkelboom et al., 2024). The model operates by predicting integration steps that iteratively refine the structure, converging toward a fixed-point solution. By additionally respecting the symmetry inherent in molecular structures, AEFM introduces a $\mathrm{SE}(3)$-equivariant method that facilitates robust inference, adaptable to the quality of the initial TS structure. To further improve the chemical realism of refined structures, we incorporate a physics-inspired bond-based loss that guides the model toward physically plausible geometries.

AEFM is particularly suited for high-throughput settings, where efficient and reliable refinement is essential to handle large numbers of candidates. Additionally, it benefits in-depth mechanistic studies by reducing the need for costly TS optimization steps. When used in conjunction with React-OT, a state-of-the-art ML-based model, AEFM reduces the median root-mean-square deviation (RMSD) of predicted TS structures to 0.088 Å and achieves a median absolute error in barrier heights of just 0.793 kcal mol^{-1}, a 27% improvement over React-OT alone. Incorporating a physics-inspired bond-length loss further enhances structural realism, with the bonded distance distribution of AEFM-refined samples aligning 35% more closely to the ground truth distribution from the Transition1x dataset (Schreiner et al., 2022a).

AEFM introduces several methodological innovations to enable efficient and accurate refinement of TS structures. Unlike standard FM, which relies on a time-dependent vector field and fixed integration schedules, AEFM learns a time-independent equilibrium flow field that supports adaptive fixed-point inference. To promote chemically realistic outputs, a physics-inspired bond-length loss that penalizes implausible bond distortions is incorporated.

2 Related work

Classical approaches. Algorithms for locating TSs fall into two broad categories, single-ended (Banerjee et al., 1985; Baker, 1986; Henkelman and Jónsson, 1999) and double-ended methods (Jónsson et al., 1998; Henkelman et al., 2000; Peters et al., 2004). Single-ended methods refine an initial 3D structure using gradient and sometimes Hessian information, while double-ended approaches construct a continuous path between reactant and product geometries to locate the TS along this path. Although effective, both classes of methods require repeated energy and force evaluations, which become computationally demanding when applied with accurate quantum chemical techniques such as DFT, limiting their scalability to large systems or reaction networks.

Approximating the PES. Surrogate models, such as Gaussian process regressions (Pozun et al., 2012; Koistinen et al., 2016, 2017; Denzel and Kästner, 2018; Denzel et al., 2019; Garrido Torres et al., 2019; Heinen et al., 2022) or machine-learned interatomic potentials (Peterson, 2016; Schreiner et al., 2022b; Zhang et al., 2024; Wander et al., 2025; Yuan et al., 2024; Zhao et al., 2025), can approximate the PES, significantly accelerating TS searches when coupled with traditional optimization schemes.

However, these approaches require high-quality non-equilibrium data, particularly around the TS region, which limits their scalability (Yuan et al., 2024).

Predictive / generative models. Beyond surrogate-assisted optimization, other deep learning approaches aim to directly predict the transition state structure (Pattanaik et al., 2020; Jackson et al., 2021; Zhang et al., 2021; Choi, 2023). Many of these methods predict the TS distance matrix and then convert it into 3D coordinates. More recently, generative models have reframed TS prediction as a distribution learning problem, aiming to learn the distribution of TS geometries conditioned on given reactant and product structures (Makoś et al., 2021; Duan et al., 2023; Kim et al., 2024; Galustian et al., 2025; Duan et al., 2025; Hayashi et al., 2025). For instance, ReactDiff (Duan et al., 2023) models the joint distribution of reactant, TS, and product using denoising diffusion and inpainting to sample plausible TS candidates. Its successor, React-OT (Duan et al., 2025), leverages flow matching (FM) (Lipman et al., 2022; Liu et al., 2022) and optimal transport to improve generation accuracy and efficiency. Other models bypass the need for 3D input entirely by generating TS geometries directly from 2D molecular graphs (Kim et al., 2024; Galustian et al., 2025).

3 Background

Flow Matching. FM is a generative modeling approach that learns a transformation from a simple base distribution q_0 to a target distribution q_1. The base distribution q_0 is often referred to as the prior, and the target distribution q_1 as the data distribution.

To model this transformation, FM learns a continuous-time vector field $\mathbf{v}_\theta(\mathbf{x}_t, t)$. The point $\mathbf{x}_t$ lies along a predefined interpolation path between samples $\mathbf{x}_0 \sim q_0$ and $\mathbf{x}_1 \sim q_1$. Therefore, an optimal transport probability path (Tong et al., 2024) with the interpolation variable $t \in [0, 1]$ is defined as:

$$p_t(\mathbf{x} \mid \mathbf{x}_0, \mathbf{x}_1) = \mathcal{N}\left(\mathbf{x} \mid (1-t)\mathbf{x}_0 + t\mathbf{x}_1,\ \sigma_{\text{FM}}^2 \mathbf{I}\right), \tag{1}$$

leading to samples:

$$\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1 + \sigma_{\text{FM}}\boldsymbol{\epsilon}. \tag{2}$$

We set σ_{FM} to 0.5 in our experiment. The corresponding target velocity field is defined as:

$$\mathbf{v}_t(\mathbf{x}_t; \mathbf{x}_0, \mathbf{x}_1) = \frac{d\mathbf{x}_t}{dt} = \mathbf{x}_1 - \mathbf{x}_0. \tag{3}$$

In doing so, FM models how probability mass moves over time from the prior to the data distribution. To train the vector field $\mathbf{v}_\theta$, the squared error between the predicted velocity and the target velocity is minimized. The training objective, known as the FM loss, is given by:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{\mathbf{x}_0, \mathbf{x}_1, t}\left[\left\|\mathbf{v}_\theta(\mathbf{x}_t, t) - (\mathbf{x}_1 - \mathbf{x}_0)\right\|^2\right], \tag{4}$$

As an alternative loss formulation, the model $\phi_\theta(\mathbf{x}_t, t)$ can be trained to directly predict $\mathbf{x}_1$ at time t instead of the velocity, a strategy that has demonstrated improved performance in practice (Stark et al., 2024). This approach is commonly referred to as variational flow matching (VFM) (Eijkelboom et al., 2024). Once $\phi_\theta(\mathbf{x}_t, t)$ is trained, new samples can be generated by solving the following ordinary differential equation (ODE) forward in time:

$$\frac{d\mathbf{x}_t}{dt} = \frac{\phi_\theta(\mathbf{x}_t, t) - \mathbf{x}_t}{1-t}, \quad \mathbf{x}_0 \sim q_0. \tag{5}$$

This integration starts from a sample $\mathbf{x}_0$ drawn from the prior q_0, and produces a sample $\mathbf{x}_1 \sim q_1$ at time $t = 1$, using any black-box ODE solver. Later, we will make $\phi_\theta(\mathbf{x}_t, t)$ time-independent and use it to iteratively refine approximate solutions $\mathbf{x}^{k+1} = \phi_\theta(\mathbf{x}^k)$.

4 Adaptive Equilibrium Flow Matching

AEFM refines low-fidelity TS structures into high-quality geometries by learning a flow field that maps noisy initial guesses back to reference TSs, as shown in Figure 1. It builds on FM, but replaces time-dependent integration with a time-independent, equilibrium formulation. This allows for iterative, fixed-point refinement that adapts to the quality of each input, allocating more steps to less accurate guesses. Training is guided by optimal transport, using noise-scaled perturbations of accurate TSs to simulate typical low-fidelity errors, enabling generalization across different prior methods.

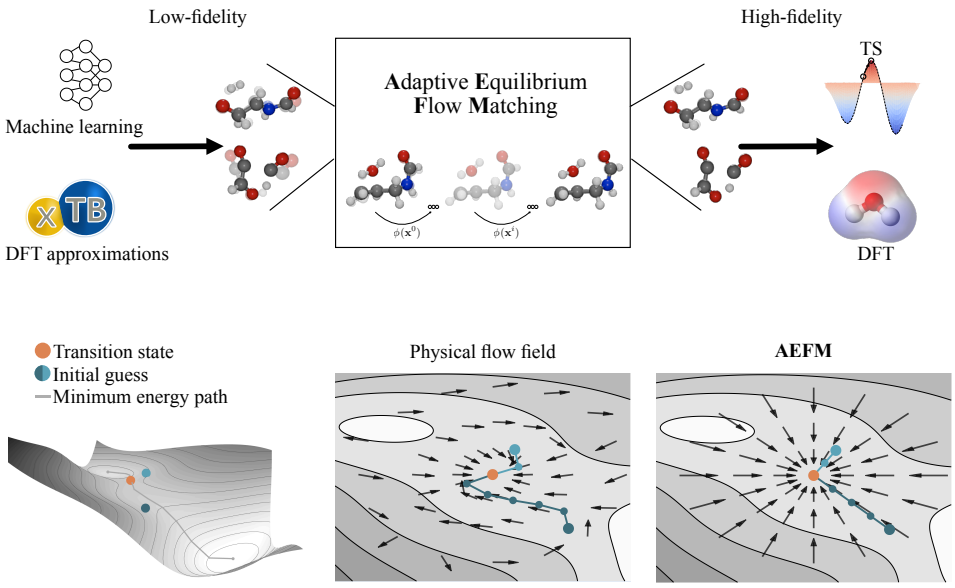

Figure 1: **AEFM pipeline for TS structure refinement. a** The input consists of low-fidelity TS samples, which may originate from various sources such as ML models or tight-binding approximations. These inputs are iteratively refined to produce high-fidelity, chemically valid TS geometries near the DFT level. **b** Comparison between actual physical flow and the one learned by AEFM on the Müller–Brown potential energy surface. Integrating the physical flow field requires multiple function evaluations, which can become computationally expensive with methods such as DFT. In contrast, AEFM learns a much simpler representation that captures the essential structure while requiring significantly fewer and more efficient evaluations.

Adaptive Prior. A central component of AEFM is its adaptive behavior, which arises from the formulation of the source distribution p_0 that we learn to map to the target distribution p_1. In our case, the target distribution is determined by the high-fidelity TSs from the Transition1x dataset (Schreiner et al., 2022a). Given a sample $\mathbf{x}_1 \sim p_1$, we define the corresponding source sample $\mathbf{x}_0 \sim p_0$ as a noisy perturbation of $\mathbf{x}_1$:

$$\mathbf{x}_0 = \mathbf{x}_1 + \sigma\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}). \tag{6}$$

The key parameter in this formulation is σ, which controls the extent to which the source distribution deviates from the target. We assume that the deviation of low-fidelity samples $\mathbf{x}_1^{\mathrm{w}}$ from their corresponding high-fidelity TSs can be modeled as Gaussian noise.

Under this assumption, we want to determine the noise scale σ such that the expected error from a Gaussian corruption process with variance σ^2 matches the expected error between low-fidelity and reference TSs. Specifically, we set

$$\mathbb{E}_{\mathbf{x}_1 \sim D, \mathbf{x}_0 \sim \mathcal{N}(\mathbf{x}_1, \sigma^2 I)} \left[\frac{\|\mathbf{x}_0 - \mathbf{x}_1\|^2}{N(\mathbf{x}_1)} \right] = \mathbb{E}_{(\mathbf{x}_1, \mathbf{x}_1^{\mathrm{w}}) \sim D} \left[\frac{\|\mathbf{x}_1^{\mathrm{w}} - \mathbf{x}_1\|^2}{N(\mathbf{x}_1)} \right], \tag{7}$$

where $N(\mathbf{x}_1)$ is the number of atoms involved. Since $\mathbf{x}_0 = \mathbf{x}_1 - \sigma\boldsymbol{\epsilon}$, the left-hand side simplifies to

$$\mathbb{E}_{\boldsymbol{\epsilon}} \left[\frac{\|\sigma\boldsymbol{\epsilon}\|^2}{N(\boldsymbol{\epsilon})} \right] = \frac{\sigma^2}{N(\boldsymbol{\epsilon})} \cdot 3N(\boldsymbol{\epsilon}) = 3\sigma^2. \tag{8}$$

Solving for σ, we obtain

$$\sigma = \left(\mathbb{E}_{(\mathbf{x}_1, \mathbf{x}_1^{\mathrm{w}}) \sim D} \left[\frac{\|\mathbf{x}_1 - \mathbf{x}_1^{\mathrm{w}}\|^2}{3N(\mathbf{x}_1)} \right] \right)^{1/2}. \tag{9}$$

Thus, σ can be calculated using the mean RMSD of the low-fidelity samples. The source distribution in our setup is designed to model the expected deviation of the low-fidelity predictions from the reference TS. It captures the distribution of typical errors observed in the low-fidelity method and

provides a learning signal during training. However, in contrast to the standard FM framework, we do not sample from the prior during inference. Instead, we start from the actual output of the low-fidelity model. As a result, the model learns from the prior during training, but at inference time, it needs to adapt to the specific error of each low-fidelity input. These errors can vary considerably, with some samples being very close to the true TS and others deviating more. Assigning a uniform time value of $t = 0$ to all such samples during inference, as done in conventional FM, may lead to over- or under-correction by the model. To address this, we remove explicit time conditioning during training, allowing the model to implicitly infer the quality of a given input $\mathbf{x}_t$. This helps the model estimate how far each sample is from the final prediction target. In practice, this behavior is encouraged through the use of a direct $\mathbf{x}_1$-prediction loss, as described earlier, while omitting time as an input to the network.

$$\mathcal{L}_{\text{AEFM}} = \mathbb{E}_{\mathbf{x}_0,\mathbf{x}_1,t}\left[\|\mathbf{x}_1 - \phi_\theta(\mathbf{x}_t)\|^2\right]. \tag{10}$$

Fixed-point inference. Since we omit the concept of time, we no longer integrate the ODE from Equation 5. Instead, we train a neural network ϕ_θ to directly predict the endpoint $\mathbf{x}_1$ of a dynamical process starting from an initial point $\mathbf{x}_0$. This formulation aligns with the perspective of VFM, where learning a velocity field that matches trajectories between $\mathbf{x}_0$ and $\mathbf{x}_1$ can be reinterpreted as minimizing a divergence between model and reference endpoint distributions. In our case, although we do not instantiate or evaluate $\mathbf{v}_\theta(\mathbf{x}_t, t)$ directly at test time, the network's prediction implicitly corresponds to the result of integrating such a field over time. In this sense, our model acts as a learned approximation of the ODE solution operator. To further refine predictions and ensure consistency with underlying dynamics, we employ a fixed-point iteration scheme at inference time:

$$\mathbf{x}^{k+1} = \phi_\theta(\mathbf{x}^k), \tag{11}$$

where the initial guess $\mathbf{x}^0$ is taken as the low-fidelity prediction, $\mathbf{x}_1^{\text{w}}$. Conceptually, this mirrors the inference procedure in Deep Equilibrium Models (Bai et al., 2019, 2021), where a neural network is iterated to convergence at test time to find a fixed point $\mathbf{x}^*$ satisfying $\mathbf{x}^* = f_\theta(\mathbf{x}^*)$. To perform the iteration, one may employ any fixed-point solver, such as Broyden's method (Broyden, 1965) or Anderson acceleration (Anderson, 1965). In this work, we use the latter, which enhances convergence by leveraging multiple previous iterates and their residuals to extrapolate a more accurate fixed point. Given m previous iterates $\mathbf{x}^{k-m}, \ldots, \mathbf{x}^k$ and corresponding residuals $\mathbf{g}(\mathbf{x}^i) = \phi_\theta(\mathbf{x}^i) - \mathbf{x}^i$, the method solves a least-squares problem to find coefficients $\boldsymbol{\alpha}$ such that the weighted sum of residuals $\sum_{i=0}^{m} \alpha_i \mathbf{g}(\mathbf{x}^{k-m+i})$ is minizimed. Given $\boldsymbol{\alpha}$, the next iterate is computed as:

$$\mathbf{x}^{k+1} = \beta \sum_{i=0}^{m} \alpha_i \phi_\theta(\mathbf{x}^{k-m+i}) + (1-\beta) \sum_{i=0}^{m} \alpha_i \mathbf{x}^{k-m+i}, \tag{12}$$

where $\beta \in [0, 1]$ is a damping parameter and $\sum_i \alpha_i = 1$.

Physical consistency loss. To address issues such as bond length inconsistencies and atomic clashes in generative models, we introduce an additional loss term focused on bonding (Peng et al., 2023; Williams and Inala, 2024; Wohlwend et al., 2025; Vost et al., 2025; Galustian et al., 2025). This is particularly important because the PES is highly sensitive to small geometric deviations. In some cases, accurately reproducing critical bond lengths is more important than minimizing the overall positional error. A prediction may yield a low RMSD while still introducing small but chemically significant distortions in key bonds, resulting in large energetic errors. To improve the chemical plausibility of generated structures, we compare the local environment of each atom within a cutoff radius r_{cut} to that of the corresponding atom in the ground truth structure, as shown in Figure 2.

$$\mathcal{L}_{\text{b}} = \mathbb{E}\left[\sum_{(i,j)\in\mathcal{B}(\mathbf{x}_1)} \frac{\left[d_{ij}(\phi_\theta(\mathbf{x}_t)) - d_{ij}(\mathbf{x}_1)\right]^2}{|\mathcal{B}(\mathbf{x}_1)|}\right] \tag{13}$$

$$\mathcal{B}(\mathbf{x}_1) := \{(i,j) \mid \|\mathbf{x}_{1,i} - \mathbf{x}_{1,j}\| < r_{\text{cut}}\} \tag{14}$$

with $d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|$ as the euclidian distance between atom i and j. Thus, the total loss used in training is:

$$\mathcal{L} = \mathcal{L}_{\text{AEFM}} + w_{\text{b}}\mathcal{L}_{\text{b}} \tag{15}$$

with w_{b} as a hyperparameter to weight the bond loss influence during training.

5 Experiments

To evaluate AEFM, we use the Transition1x dataset (Schreiner et al., 2022a), which contains climbing-image nudged elastic band (CI-NEB) (Henkelman et al., 2000) calculations performed with DFT (ωB97x/6-31G(d) (Ditchfield et al., 1971; Chai and Head-Gordon, 2008)) for 10,073 organic reactions encompassing diverse reaction types. These reactions were sampled from an enumeration of 1,154 reactants in the GDB7 dataset (Grambow et al., 2020), which includes molecules with up to 7 heavy atoms (C, N, and O) and a total of 23 atoms. We adopt the same random split as Duan et al. (Duan et al., 2023), using 9,000 reactions for training and 1,073 for testing.

5.1 Refining TS structures across fidelity scales

AEFM is applied to refine prior low-fidelity TS structures toward valid TS geometries at the target level of theory. To assess the quality of the refined structures, we evaluate both the RMSD of atomic positions and the absolute error in the reaction barrier.

Table 1: Performance of AEFM refinement. Structural and energetic errors of various low-fidelity TS guesses before and after refinement with AEFM. The refinement consistently reduces both mean and median deviations relative to the reference TS structures. In addition, average inference times per sample are reported, showing that AEFM introduces only negligible computational overhead.

Approach	RMSD (Å)		$\lvert\Delta E_{\text{TS}}\rvert$ (kcal mol^{-1})		Inference (s)
	Mean	Median	Mean	Median	
xTB CI-NEB	0.312	0.179	10.426	2.673	9.23
xTB CI-NEB + AEFM	0.250 (↓20%)	0.119 (↓34%)	6.204 (↓40%)	1.090 (↓59%)	+0.24
React-OT (xTB)	0.211	0.108	4.697	1.186	0.14
React-OT (xTB) + AEFM	0.214 (↑1%)	0.102 (↓6%)	4.153 (↓12%)	0.824 (↓31%)	+0.12
React-OT	0.183	0.092	3.405	1.092	0.14
React-OT + AEFM	0.188 (↑3%)	0.088 (↓4%)	3.341 (↓2%)	0.793 (↓27%)	+0.13
React-OT + AEFM [a]	0.176 (↓4%)	0.086 (↓7%)	3.158 (↓7%)	0.790 (↓27%)	+0.13

[a] For 26 reactions, a different intended TS was selected if the RMSD between the low-fidelity sample and this alternative TS was at least 30% lower than the RMSD to the originally intended TS.

To assess the effectiveness of AEFM, we consider React-OT (Duan et al., 2025) as the first low-fidelity source, a state-of-the-art generative model for TS prediction. React-OT achieves remarkable accuracy, producing samples with a mean RMSD of 0.18 Å and a median absolute error in barrier height of 1.092 kcal mol^{-1}. Applying AEFM to refine the React-OT samples yields a 27% improvement in the median barrier height error, requiring only 2 model calls in median and approximately 0.13 seconds per refinement on an Nvidia A40 GPU. Consequently, 69% of the TSs had a more accurate barrier height, achieving a median absolute error of 0.793 kcal mol^{-1}.

As a second low-fidelity source, we consider GFN2-xTB (Bannwarth et al., 2019), a tight-binding approximation that is commonly used as a starting point for elucidating reaction mechanisms. Tight-binding methods are approximately three orders of magnitude faster than DFT, enabling high-throughput reaction scans that would be otherwise computationally prohibitive. For the 1,073 test reactions, reactant and product geometries were first relaxed, followed by CI-NEB calculations using GFN2-xTB. Of these, 945 calculations converged successfully, yielding samples with a mean RMSD of 0.31 Å and a median absolute error in barrier height of 2.673 kcal mol^{-1}. Applying AEFM improves the median absolute error in barrier height by 59%, reducing it to 1.090 kcal mol^{-1}, while requiring only a median of 4 model calls. Analyzing the chemical accuracy of samples reveals that only 25% of the original GFN2-xTB-generated structures meet this threshold, whereas AEFM refinement increases this accuracy rate to 57%.

To reduce the computational cost of generating DFT-quality reactant and product structures, we follow Duan et al. (Duan et al., 2025) and employ React-OT directly on xTB-optimized geometries. This approach enables rapid TS generation without requiring expensive DFT-level optimization of endpoints. React-OT can be reliably applied to xTB-level structures, yielding a mean RMSD of 0.21 Å and a median absolute error in barrier height of 1.186 kcal mol^{-1}. Building on this, we apply AEFM to refine the resulting TS guesses further, reducing the median absolute error by an additional

31% with only a median of two model evaluations. The results of AEFM applied to each low-fidelity method are summarized in Table 1.

Physics-Informed loss improves chemical validity

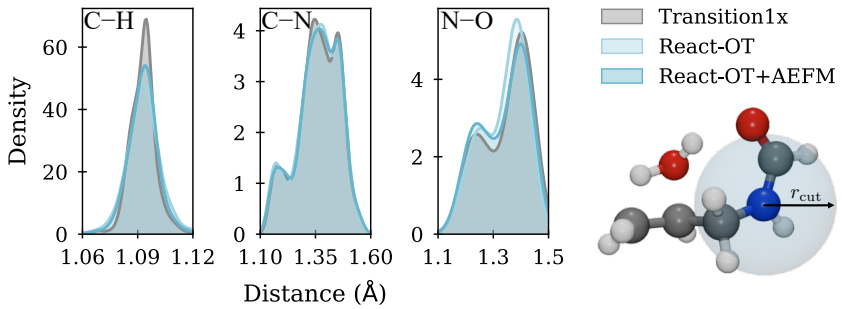

Figure 2: **Bond length distributions.** Distributions of C–H, C–N, and N–O bond lengths in the Transition1x dataset compared to those in the React-OT and AEFM-refined structures.

To assess the impact of the bond loss term, we compare AEFM's fine-tuning performance when including the term versus omitting it, using two representative low-fidelity sources, React-OT (Duan et al., 2025) and xTB (Bannwarth et al., 2019). For React-OT samples, incorporating the bond loss results in a 27% reduction in the median absolute error of barrier heights. In contrast, the same model without the bond loss achieves only a 3.5% improvement (Supplementary Table 6). To understand the source of this improvement, we analyze how the bond loss affects the model's ability to recover chemically plausible local structures. Specifically, we evaluate whether the refined structures better match the bond length distributions within a 2 Å neighborhood of each atom present in the dataset. Interactions are categorized as either bonded or non-bonded based on threshold distances (Supplementary Table 5). With the bond loss, the average similarity to the reference bond length distributions improves by 35.7% for bonded interactions and by 6% for non-bonded ones, as illustrated for selected bonds in Figure 2. Given that bonded interactions dominate the intramolecular potential energy landscape, enhancing their accuracy is critical for reliable energy predictions. This effect is even more pronounced for xTB samples, which exhibit larger deviations from the target distribution. Here, the bond loss leads to a 57% improvement in bonded interaction similarity and a 54% improvement for non-bonded ones (see Supplementary Table 7).

Moreover, average displacement metrics such as RMSD often fail to reflect meaningful changes in energy, underscoring their limited sensitivity, as shown in Supplementary Figure 3a and Supplementary Figure 4b. Notably, the fraction of samples that improve in both RMSD and energy is considerably smaller than the fraction that improve in energy alone (Supplementary Figure 3a). In line with this, the correlation between energetic and structural improvement is weak, with a Pearson coefficient of only 0.17 (Supplementary Figure 4b). A similarly weak relationship between RMSD and energy difference was also reported by Duan et al. (Duan et al., 2023). That highlights that generating realistic bond lengths in the refinement process is just as crucial as minimizing deviations in atomic positions. In many TS structures, the energetic accuracy is governed primarily by the reactive center. Consequently, even if the RMSD improves slightly for some atoms, introducing unrealistic bonds, such as excessively short ones, can severely degrade energetic similarity (Zhao et al., 2023). This effect is further illustrated by the distribution of C–H bond lengths, which, after refinement with AEFM, shows a 44% higher similarity to the dataset distribution compared to the original React-OT samples. While the refined C–H bond might not match the exact pose of the reference, its physically accurate length improves energetic similarity, even if the overall RMSD appears worse.

This observation relates to a broader challenge in molecular generative modeling, generating chemically consistent bond geometries (Peng et al., 2023; Williams and Inala, 2024; Vost et al., 2025; Wohlwend et al., 2025; Galustian et al., 2025). Several recent works have proposed solutions to mitigate this issue. For example, Boltz-1 (Wohlwend et al., 2025) biases generation toward low-energy configurations using physically inspired energy functions. While effective in diffusion-based generation schemes, this approach is incompatible with our fixed-point inference method, which does not

rely on stochastic sampling. Vost et al. (Vost et al., 2025) address the sensitivity of generative models to geometric distortions by augmenting training data with perturbed structures and conditioning the diffusion process on the distortion level. However, this requires training a diffusion model from pure Gaussian noise on distortion-conditioned data, whereas our method uses an adaptive prior. Williams et al. (Williams and Inala, 2024) propose a physics-informed diffusion model that decomposes the generative task into separate components for bonding, bending, torsion, and chirality, enabling more physically grounded predictions. This decomposition, however, depends on a specialized neural network architecture and limits the flexibility to choose general-purpose backbones. Finally, Falck et al. (Falck et al., 2025) analyze the influence of the noising schedule on the recovery of high-frequency features, such as precise bond lengths. While theoretically insightful, their analysis was not conducted in the context of molecular modeling.

These efforts highlight the importance of incorporating structural or energetic priors to improve the physical fidelity of generated molecules. In contrast to more complex solutions, AEFM addresses this issue with a simple yet effective bond loss term, which guides the model toward reproducing the bond distributions found in the underlying data.

Convergence analysis

As AEFM relies on fixed-point iteration, understanding its convergence behavior is critical. A standard indicator of local convergence is the Lipschitz constant L, which quantifies how sensitively the model output responds to input perturbations. In practice, however, this condition is often evaluated via the spectral radius $\rho(J_\phi)$, the largest absolute eigenvalue of the Jacobian J_ϕ at a given point $\mathbf{x}$. By Lyapunov's linearization theorem, the condition $\rho(J_\phi) < 1$ suffices for convergence in the absence of advanced solvers. However, as Bai et al. (Bai et al., 2021) point out, this requirement can be overly conservative in practice. Methods like Broyden's method (Broyden, 1965) or Anderson acceleration (Anderson, 1965) often succeed even when $\rho(J_\phi) < 1$, due to their ability to handle mild local non-contractive behavior. To assess aefm's convergence characteristics, Supplementary Figure 5e displays the evolution of $\rho(J_\phi)$ over refinement iterations on GFN2-xTB samples. Convergence is defined as the point where the RMSD between successive iterates falls below 0.01, as specified in Equation 16. If convergence is not achieved, inference is terminated after 100 iterations. The plot shows the median, along with the 25th and 75th percentiles, and overlays the cumulative convergence rate. Initially, the spectral radius drops sharply, reflecting strong local contractivity and rapid convergence. After iteration 4, the median convergence point begins to rise again. This increase does not signal failure but highlights that remaining unconverged samples tend to be more structurally complex and locally less stable. These more complicated cases dominate the later iterations, pushing the upper quantiles of $\rho(J_\phi)$ upward. Still, even in these regions, the 75th percentile remains below 1.3, indicating near-contractive dynamics. Out of 1073 React-OT and 945 xTB samples, only 6 and 3, respectively, failed to converge before reaching the iteration limit. Overall, AEFM achieves fast and stable convergence for the majority of samples, with early iterations characterized by low spectral radii and minimal computational overhead. Although convergence is slower for a few complex cases, they remain computationally manageable, with inference times not exceeding 1.6 seconds.

6 Discussion

AEFM addresses a core challenge in reaction mechanism elucidation by converting low-fidelity TS guesses into chemically accurate, DFT-quality structures with minimal computational cost. By learning a time-independent flow field, conditioned on a prior tailored to the systematic error distribution of approximate methods, AEFM provides a lightweight, physically informed correction mechanism that enhances fast TS generators like GFN2-xTB or React-OT.

The physics-informed loss steers predictions toward chemically meaningful structures, addressing a core limitation of generative approaches. This makes AEFM well-suited for complex systems such as catalysis or enzymatic reactions, where high-quality initial guesses are hard to obtain and subtle geometric features are essential. Incorporating higher-order terms like angles or torsions, along with adaptive cutoffs, could further boost accuracy and extend applicability.

Despite its robustness, AEFM is limited by the support of its training prior. For initial guesses that deviate substantially from typical training-time errors, performance degrades. One promising path forward involves a two-stage refinement strategy guided by model uncertainty. A specialized model,

trained on broader structural deviations, could be applied when the primary model signals high uncertainty, enabling robust treatment of more strongly perturbed inputs.

Finally, the principles behind AEFM generalize beyond TS refinement. In fields such as scientific machine learning, where coarse-grained simulations are used to accelerate predictions in high-dimensional systems, AEFM-like architectures could enhance the spatial and temporal resolution of neural PDE solvers. This may enable both more accurate forecasts and longer stable simulation horizons. Overall, AEFM offers a flexible and computationally efficient paradigm for lifting low-fidelity predictions to chemically and physically meaningful accuracy across a range of domains.

Acknowledgments and Disclosure of Funding

This research was funded by the Ministry of Science, Research and the Arts Baden-Wuerttemberg in the Artificial Intelligence Software Academy (AISA). We thank the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) for supporting this work by funding - EXC2075 – 390740016 under Germany's Excellence Strategy. Tanja was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2120/1 – 390831618. We acknowledge the support by the Stuttgart Center for Simulation Science (SimTech). The authors acknowledge support by the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant no INST 40/575-1 FUGG (JUSTUS 2 cluster). We thank Moritz René Schäfer and Anji Liu for insightful discussions throughout the development of this work. We are also grateful to Boshra Ariguib for her valuable input on improving the visual presentation of the manuscript.

References

Anderson, D. G. (1965). Iterative procedures for nonlinear integral equations. *Journal of the ACM (JACM)*, 12(4):547–560.

Bai, S., Kolter, J. Z., and Koltun, V. (2019). Deep Equilibrium Models. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Bai, S., Koltun, V., and Kolter, Z. (2021). Stabilizing Equilibrium Models by Jacobian Regularization. In *Proceedings of the 38th International Conference on Machine Learning*, pages 554–565. PMLR. ISSN: 2640-3498.

Baker, J. (1986). An algorithm for the location of transition states. *Journal of Computational Chemistry*, 7(4):385–395.

Banerjee, A., Adams, N., Simons, J., and Shepard, R. (1985). Search for stationary points on surfaces. *The Journal of Physical Chemistry*, 89(1):52–57.

Bannwarth, C., Ehlert, S., and Grimme, S. (2019). Gfn2-xtb—an accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *Journal of Chemical Theory and Computation*, 15(3):1652–1671. PMID: 30741547.

Broyden, C. G. (1965). A class of methods for solving nonlinear simultaneous equations. *Mathematics of computation*, 19(92):577–593.

Chacko, R., Gossler, H., Angeli, S., and Deutschmann, O. (2024). Interconnected digital solutions to accelerate modeling of the reaction kinetics in catalysis. *ChemCatChem*, 16(4):e202301355.

Chai, J.-D. and Head-Gordon, M. (2008). Systematic optimization of long-range corrected hybrid density functionals. *The Journal of chemical physics*, 128(8).

Choi, S. (2023). Prediction of transition state structures of gas-phase chemical reactions via machine learning. *Nature Communications*, 14(1):1168. Publisher: Nature Publishing Group.

Denzel, A., Haasdonk, B., and Kästner, J. (2019). Gaussian Process Regression for Minimum Energy Path Optimization and Transition State Search. *The Journal of Physical Chemistry A*, 123(44):9600–9611. Publisher: American Chemical Society.

Denzel, A. and Kästner, J. (2018). Gaussian Process Regression for Transition State Search. *Journal of Chemical Theory and Computation*, 14(11):5777–5786. Publisher: American Chemical Society.

Dewyer, A. L., Argüelles, A. J., and Zimmerman, P. M. (2018). Methods for exploring reaction space in molecular systems. *WIREs Computational Molecular Science*, 8(2):e1354.

Ditchfield, R., Hehre, W. J., and Pople, J. A. (1971). Self-consistent molecular-orbital methods. ix. an extended gaussian-type basis for molecular-orbital studies of organic molecules. *The Journal of Chemical Physics*, 54(2):724–728.

Du, Y., Wang, L., Feng, D., Wang, G., Ji, S., Gomes, C. P., Ma, Z.-M., et al. (2023). A new perspective on building efficient and expressive 3d equivariant graph neural networks. *Advances in neural information processing systems*, 36:66647–66674.

Duan, C., Du, Y., Jia, H., and Kulik, H. J. (2023). Accurate transition state generation with an object-aware equivariant elementary reaction diffusion model. *Nature Computational Science*, 3(12):1045–1055. Publisher: Nature Publishing Group.

Duan, C., Liu, G.-H., Du, Y., Chen, T., Zhao, Q., Jia, H., Gomes, C. P., Theodorou, E. A., and Kulik, H. J. (2025). Optimal transport for generating transition states in chemical reactions. *Nature Machine Intelligence*, 7(4):615–626. Publisher: Nature Publishing Group.

Eijkelboom, F., Bartosh, G., Andersson Naesseth, C., Welling, M., and van de Meent, J.-W. (2024). Variational flow matching for graph generation. *Advances in Neural Information Processing Systems*, 37:11735–11764.

Falck, F., Pandeva, T., Zahirnia, K., Lawrence, R., Turner, R., Meeds, E., Zazo, J., and Karmalkar, S. (2025). A fourier space perspective on diffusion models. *arXiv preprint arXiv:2505.11278*.

Galustian, L., Mark, K., Karwounopoulos, J., Kovar, M. P.-P., and Heid, E. (2025). GoFlow: Efficient Transition State Geometry Prediction with Flow Matching and E(3)-Equivariant Neural Networks.

Garrido Torres, J. A., Jennings, P. C., Hansen, M. H., Boes, J. R., and Bligaard, T. (2019). Low-Scaling Algorithm for Nudged Elastic Band Calculations Using a Surrogate Machine Learning Model. *Physical Review Letters*, 122(15):156001.

Grambow, C. A., Pattanaik, L., and Green, W. H. (2020). Reactants, products, and transition states of elementary chemical reactions based on quantum chemistry. *Scientific Data*, 7(1):137.

Hayashi, A., Takamoto, S., Li, J., Tsuboi, Y., and Okanohara, D. (2025). Generative Model for Constructing Reaction Path from Initial to Final States. *Journal of Chemical Theory and Computation*, 21(3):1292–1305. Publisher: American Chemical Society.

Heinen, S., von Rudorff, G. F., and von Lilienfeld, O. A. (2022). Transition state search and geometry relaxation throughout chemical compound space with quantum machine learning. *The Journal of Chemical Physics*, 157(22):221102.

Henkelman, G. and Jónsson, H. (1999). 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.

Henkelman, G., Uberuaga, B. P., and Jónsson, H. (2000). A climbing image nudged elastic band method for finding saddle points and minimum energy paths. *The Journal of Chemical Physics*, 113(22):9901–9904.

Hermes, E. D., Sargsyan, K., Najm, H. N., and Zádor, J. (2022). Sella, an open-source automation-friendly molecular saddle point optimizer. *Journal of Chemical Theory and Computation*, 18(11):6974–6988. PMID: 36257023.

Jackson, R., Zhang, W., and Pearson, J. (2021). TSNet: predicting transition state structures with tensor field networks and transfer learning. *Chemical Science*, 12(29):10022–10040. Publisher: The Royal Society of Chemistry.

Jónsson, H., Mills, G., and Jacobsen, K. W. (1998). Nudged elastic band method for finding minimum energy paths of transitions. In *Classical and quantum dynamics in condensed phase simulations*, pages 385–404. World Scientific.

Jorner, K., Tomberg, A., Bauer, C., Sköld, C., and Norrby, P.-O. (2021). Organic reactivity from mechanism to machine learning. *Nature Reviews Chemistry*, 5(4):240–255. Publisher: Nature Publishing Group.

Kim, S., Woo, J., and Kim, W. Y. (2024). Diffusion-based generative AI for exploring transition states from 2D molecular graphs. *Nature Communications*, 15(1):341. Publisher: Nature Publishing Group.

Koistinen, O.-P., Dagbjartsdóttir, F. B., Ásgeirsson, V., Vehtari, A., and Jónsson, H. (2017). Nudged elastic band calculations accelerated with Gaussian process regression. *The Journal of Chemical Physics*, 147(15):152720.

Koistinen, O.-P., Maras, E., Vehtari, A., and Jónsson, H. (2016). Minimum energy path calculations with Gaussian process regression. *Nanosystems: Physics, Chemistry, Mathematics*, pages 925–935. arXiv:1703.10423 [physics, stat].

Lam, Y.-h., Abramov, Y., Ananthula, R. S., Elward, J. M., Hilden, L. R., Nilsson Lill, S. O., Norrby, P.-O., Ramirez, A., Sherer, E. C., Mustakis, J., et al. (2020). Applications of quantum chemistry in pharmaceutical process development: Current state and opportunities. *Organic Process Research & Development*, 24(8):1496–1507.

Larsen, A. H., Mortensen, J. J., Blomqvist, J., Castelli, I. E., Christensen, R., Dułak, M., Friis, J., Groves, M. N., Hammer, B., Hargus, C., Hermes, E. D., Jennings, P. C., Jensen, P. B., Kermode, J., Kitchin, J. R., Kolsbjerg, E. L., Kubal, J., Kaasbjerg, K., Lysgaard, S., Maronsson, J. B., Maxson, T., Olsen, T., Pastewka, L., Peterson, A., Rostgaard, C., Schiøtz, J., Schütt, O., Strange, M., Thygesen, K. S., Vegge, T., Vilhelmsen, L., Walter, M., Zeng, Z., and Jacobsen, K. W. (2017). The atomic simulation environment—a python library for working with atoms. *Journal of Physics: Condensed Matter*, 29(27):273002.

Lipman, Y., Chen, R. T. Q., Ben-Hamu, H., Nickel, M., and Le, M. (2022). Flow Matching for Generative Modeling. In *The Eleventh International Conference on Learning Representations*.

Liu, X., Gong, C., and Liu, Q. (2022). Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. In *The Eleventh International Conference on Learning Representations*.

Makoś, M. Z., Verma, N., Larson, E. C., Freindorf, M., and Kraka, E. (2021). Generative adversarial networks for transition state geometry prediction. *The Journal of Chemical Physics*, 155(2):024116.

Mardirossian, N. and Head-Gordon, M. (2017). Thirty years of density functional theory in computational chemistry: an overview and extensive assessment of 200 density functionals. *Molecular physics*, 115(19):2315–2372.

Nandy, A., Duan, C., Taylor, M. G., Liu, F., Steeves, A. H., and Kulik, H. J. (2021). Computational Discovery of Transition-metal Complexes: From High-throughput Screening to Machine Learning. *Chemical Reviews*, 121(16):9927–10000. Publisher: American Chemical Society.

Neese, F. (2022). Software update: the orca program system, version 5.0. *WIRES Comput. Molec. Sci.*, 12(1):e1606.

Pattanaik, L., Ingraham, J. B., Grambow, C. A., and Green, W. H. (2020). Generating transition states of isomerization reactions with deep learning. *Physical Chemistry Chemical Physics*, 22(41):23618–23626. Publisher: The Royal Society of Chemistry.

Peng, Q., Duarte, F., and Paton, R. S. (2016). Computing organic stereoselectivity–from concepts to quantitative calculations and predictions. *Chemical Society Reviews*, 45(22):6093–6107.

Peng, X., Guan, J., Liu, Q., and Ma, J. (2023). MolDiff: Addressing the Atom-Bond Inconsistency Problem in 3D Molecule Diffusion Generation. In *Proceedings of the 40th International Conference on Machine Learning*, pages 27611–27629. PMLR. ISSN: 2640-3498.

Peters, B., Heyden, A., Bell, A. T., and Chakraborty, A. (2004). A growing string method for determining transition states: Comparison to the nudged elastic band and string methods. *The Journal of Chemical Physics*, 120(17):7877–7886.

Peterson, A. A. (2016). Acceleration of saddle-point searches with machine learning. *The Journal of Chemical Physics*, 145(7):074106.

Pozun, Z. D., Hansen, K., Sheppard, D., Rupp, M., Müller, K.-R., and Henkelman, G. (2012). Optimizing transition states via kernel-based machine learning. *The Journal of Chemical Physics*, 136(17):174101.

Rasmussen, M. H. and Jensen, J. H. (2020). Fast and automatic estimation of transition state structures using tight binding quantum chemical calculations. *PeerJ Physical Chemistry*, 2:e15.

R.Domingo, L. (2014). A new C–C bond formation model based on the quantum chemical topology of electron density. *RSC Advances*, 4(61):32415–32428. Publisher: Royal Society of Chemistry.

Schreiner, M., Bhowmik, A., Vegge, T., Busk, J., and Winther, O. (2022a). Transition1x - a dataset for building generalizable reactive machine learning potentials. *Scientific Data*, 9(1):779. Publisher: Nature Publishing Group.

Schreiner, M., Bhowmik, A., Vegge, T., Jørgensen, P. B., and Winther, O. (2022b). NeuralNEB—neural networks can find reaction paths fast. *Machine Learning: Science and Technology*, 3(4):045022.

Stark, H., Jing, B., Barzilay, R., and Jaakkola, T. (2024). Harmonic Self-Conditioned Flow Matching for joint Multi-Ligand Docking and Binding Site Design. In *Proceedings of the 41st International Conference on Machine Learning*, pages 46468–46494. PMLR. ISSN: 2640-3498.

Taylor, C. J., Pomberger, A., Felton, K. C., Grainger, R., Barecka, M., Chamberlain, T. W., Bourne, R. A., Johnson, C. N., and Lapkin, A. A. (2023). A brief introduction to chemical reaction optimization. *Chemical Reviews*, 123(6):3089–3126.

Tong, A., FATRAS, K., Malkin, N., Huguet, G., Zhang, Y., Rector-Brooks, J., Wolf, G., and Bengio, Y. (2024). Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*. Expert Certification.

Truhlar, D. G., Garrett, B. C., and Klippenstein, S. J. (1996). Current Status of Transition-State Theory. *The Journal of Physical Chemistry*, 100(31):12771–12800. Publisher: American Chemical Society.

Unsleber, J. P. and Reiher, M. (2020). The Exploration of Chemical Reaction Networks. *Annual Review of Physical Chemistry*, 71(Volume 71, 2020):121–142. Publisher: Annual Reviews.

von Lilienfeld, O. A., Müller, K.-R., and Tkatchenko, A. (2020). Exploring chemical compound space with quantum-based machine learning. *Nature Reviews Chemistry*, 4(7):347–358. Publisher: Nature Publishing Group.

Vost, L., Chenthamarakshan, V., Das, P., and Deane, C. M. (2025). Improving structural plausibility in diffusion-based 3D molecule generation via property-conditioned training with distorted molecules. *Digital Discovery*, 4(4):1092–1099. Publisher: RSC.

Wander, B., Shuaibi, M., Kitchin, J. R., Ulissi, Z. W., and Zitnick, C. L. (2025). Cattsunami: Accelerating transition state energy calculations with pretrained graph neural networks. *ACS Catalysis*, 15(7):5283–5294.

Williams, D. C. and Inala, N. (2024). Physics-Informed Generative Model for Drug-like Molecule Conformers. *Journal of Chemical Information and Modeling*, 64(8):2988–3007. Publisher: American Chemical Society.

Wohlwend, J., Corso, G., Passaro, S., Getz, N., Reveiz, M., Leidal, K., Swiderski, W., Atkinson, L., Portnoi, T., Chinn, I., Silterra, J., Jaakkola, T., and Barzilay, R. (2025). Boltz-1 Democratizing Biomolecular Interaction Modeling. Pages: 2024.11.19.624167 Section: New Results.

Yuan, E. C.-Y., Kumar, A., Guan, X., Hermes, E. D., Rosen, A. S., Zádor, J., Head-Gordon, T., and Blau, S. M. (2024). Analytical ab initio hessian from a deep learning potential for transition state optimization. *Nature Communications*, 15(1):8865. Publisher: Nature Publishing Group.

Zhang, J., Lei, Y.-K., Zhang, Z., Han, X., Li, M., Yang, L., Yang, Y. I., and Gao, Y. Q. (2021). Deep reinforcement learning of transition states. *Physical Chemistry Chemical Physics*, 23(11):6888–6895. Publisher: The Royal Society of Chemistry.

Zhang, S., Makoś, M. Z., Jadrich, R. B., Kraka, E., Barros, K., Nebgen, B. T., Tretiak, S., Isayev, O., Lubbers, N., Messerly, R. A., et al. (2024). Exploring the frontiers of condensed-phase chemistry with a general reactive machine learning potential. *Nature Chemistry*, 16(5):727–734.

Zhao, Q., Anstine, D. M., Isayev, O., and Savoie, B. M. (2023). δ 2 machine learning for reaction property prediction. *Chemical Science*, 14(46):13392–13401.

Zhao, Q., Han, Y., Zhang, D., Wang, J., Zhong, P., Cui, T., Yin, B., Cao, Y., Jia, H., and Duan, C. (2025). Harnessing machine learning to enhance transition state search with interatomic potentials and generative models. *Advanced Science*, n/a(n/a):e06240.

A Additional details on AEFM

To respect molecular symmetries, such as rotation, translation, and atom index permutation, we employ the SE(3)-equivariant LEFTNet (Du et al., 2023) architecture as the backbone of our model. The cutoff radius in the physics-informed loss is set to 2 Å, based on the longest equilibrium bond lengths typically observed in C, N, O, and H chemistry, with an added margin to accommodate extended bond distances that may arise in transition state structures (R.Domingo, 2014). The weight of the loss is fixed to 1.0. During inference, the fixed-point iteration is terminated once the RMSD between successive iterates falls at or below a threshold of 0.01:

$$\frac{\|\mathbf{x}^{k+1} - \mathbf{x}^k\|}{\sqrt{N(\mathbf{x}^k)}} \leq 0.01 \tag{16}$$

If the convergence criterion is not satisfied, inference is terminated after a maximum of 100 iterations. The damping parameter β is set to 1.0 and the history size m to 5, based on a hyperparameter search. Table 2 summarizes the model hyperparameters and training configurations.

Table 2: Hyperparameters and training configurations.

(a) Model hyperparameters

Parameter	Value
Message passing layers	6
Equivariant readout layers	1
Hidden features	196
Radial basis functions	96
Cutoff radius	10 Å
Learning rate	1e-3
Batch size	64

(b) Training settings

Method	σ	Epochs
xTB CI-NEB	0.19	1000
React-OT (xTB)	0.12	600
React-OT	0.11	600

B Additional experiments

B.1 Performance summary

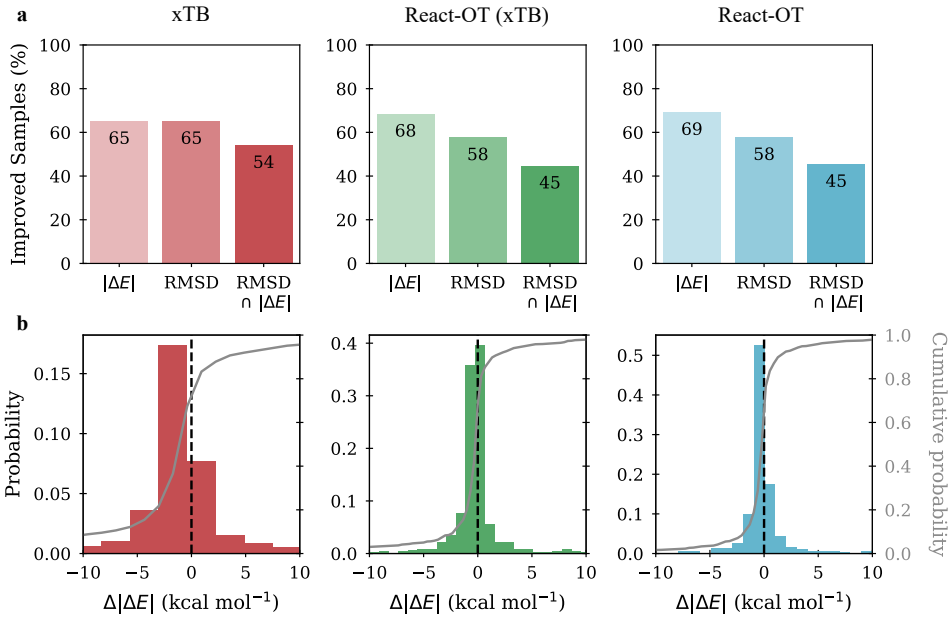

Figure 3: **Performance summary of AEFM across diverse low-fidelity sources. a** Percentage of test samples showing improvement in energy difference $|\Delta E|$ relative to the reference TS (irrespective of RMSD), in RMSD (irrespective of energy), and in both RMSD and energy difference (RMSD $\cap$ $|\Delta E|$). **b** Histogram (colored, left y-axis) and cumulative distribution (grey, right y-axis) of the change in energy difference between the low-fidelity and AEFM fine-tuned samples, measured relative to the reference TS. Negative $|\Delta E|$ values indicate that the refined samples are energetically improved.

B.2 Understanding refinement dynamics

To further investigate the performance of AEFM, we conduct a detailed analysis across diverse scenarios, aiming to better understand the factors influencing its strengths and limitations.

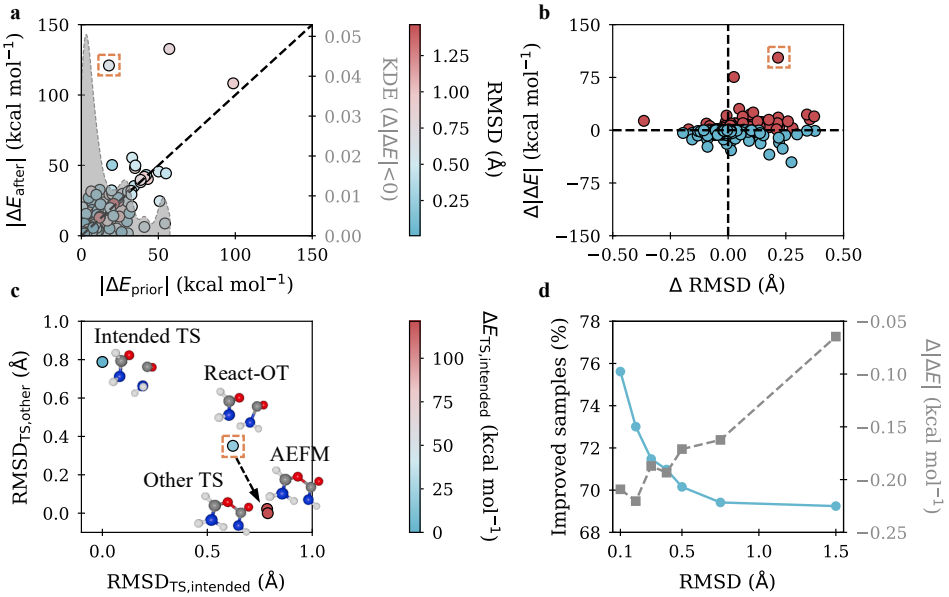

Figure 4: **Relationship between energetic and structural changes in AEFM refinements, with a focus on outliers and correlation trends. a** Energetic differences of AEFM-refined structures versus initial React-OT predictions on the left y-axis. Points below the diagonal line indicate improved agreement with the reference TS, while points above reflect increased deviation. On the right y-axis, the KDE of improvement weighted by the improvement magnitude is shown. Additionally, an outlier (top left) shows a nearly sixfold increase in error after fine-tuning. **b** Energetic vs. geometric changes resulting from the application of AEFM. The bottom-left quadrant indicates improvements in both structural and energetic similarity, while the bottom-right quadrant reflects improved energy alignment accompanied by reduced structural similarity. Blue points indicate an energetic improvement, while red points correspond to increased dissimilarity. **c** Structural analysis of the outlier. The x-axis shows RMSD to the intended TS, and the y-axis shows RMSD to an alternative, structurally similar TS. Displayed are the initial React-OT prediction, the fine-tuned sample, and both TS structures. **d** Improvement rate (left y-axis in blue) and mean reduction in energy error (right y-axis in grey) as a function of the initial React-OT RMSD.

A first aspect we examine is the asymmetry in the distribution of barrier height errors, which is particularly evident for refined samples generated using React-OT as prior. Supplementary Figure 4a shows pre- and post-refinement energetic errors, where points below the bisecting line indicate improvement. An illustrative outlier contributing to the skewed mean is shown in Supplementary Figure 4c. For the particular reaction we consider four TS, the reference (intended) TS, the React-OT prediction, its fine-tuned version obtained via AEFM, and an alternative TS associated with a different but structurally similar reaction. The plot illustrates the structural deviation, measured as RMSD, to the intended TS on the y-axis and to the alternative TS on the x-axis, while the marker color encodes the relative energy with respect to the intended TS. The original React-OT prediction deviates notably from the intended TS, with an RMSD of 0.632 Å and an energy difference of 17.904 kcal mol^{-1}. After fine-tuning, the sample shifts further away from the intended TS, reaching an RMSD of 0.793 Å and a significantly larger energy difference of 120.993 kcal mol^{-1}. At first glance, this might appear to be a failure of the optimization process. However, comparison with the alternative TS reveals a different picture, the fine-tuned structure is nearly identical to this other TS, exhibiting an RMSD of just 0.048 Å and an energy deviation of merely 0.256 kcal mol^{-1}. This behavior is explained by the initial proximity of the React-OT sample to the alternative TS, with an RMSD of 0.359 Å compared to the intended TS. Since AEFM operates purely on structural refinement and is trained on perturbed TS geometries without access to reactant-product context, it interprets the input as a noisy version of the alternative TS and converges accordingly. To further analyse this effect, all React-OT samples were compared with similar other TS. To ensure that the alternative TSs are meaningfully closer to the sample, we only retain cases in which the RMSD to the alternative TS is at least 30%

lower than the RMSD to the originally intended TS. The mean RMSD is now improved by 7% and the absolute energetic error by 5% compared to the initial analysis of fine-tuned samples. This example highlights an essential characteristic of the approach, in the absence of explicit reaction context, AEFM fine-tunes samples toward structurally and energetically valid TSs, which may not always correspond to the originally intended reaction. Such behavior is typical for surface walking algorithms, where the target is to find any nearby viable TS given an initial guess structure (Banerjee et al., 1985; Baker, 1986; Henkelman and Jónsson, 1999).

A key element influencing the performance of AEFM is the quality of the initial guess. Figure 4c illustrates this by showing the percentage of energetically improved samples along the left y-axis, and the corresponding mean energy improvement along the right y-axis, both plotted against increasing RMSD thresholds applied to the initial React-OT samples. At each threshold, only those samples with an initial RMSD below the given value are included in the statistics. The results show a clear trend, with both the likelihood and magnitude of improvement being higher at lower RMSD thresholds. Specifically, for samples with RMSD below 0.2 Å, 73% of the reactions show an energetic improvement after fine-tuning, with a mean improvement of 0.15 kcal mol^{-1}. In contrast, at higher thresholds, we have 69% improved reactions and a mean energetic improvement of 0.06 kcal mol^{-1}.

B.3 Quantum chemical validation

While the combination of tight-binding methods or generative models with AEFM enables fast and robust high-throughput TS screening, a full quantum mechanical treatment remains essential for detailed mechanistic studies (Peng et al., 2016; von Lilienfeld et al., 2020; Unsleber and Reiher, 2020; Nandy et al., 2021; Jorner et al., 2021). In such cases, transition states must be refined using saddle point optimization at the DFT level (Lam et al., 2020). These optimizations typically require multiple evaluations of forces or even full Hessians, making them computationally demanding, even for small molecules (Koistinen et al., 2017; Yuan et al., 2024).

To highlight the practical impact of AEFM on downstream applications, we evaluate its effect on the chemical validity of TS structures and the efficiency of DFT-based TS optimizations. For a representative set of 100 reactions, we compare three key metrics, namely the fraction of valid TS structures (a), identified by exactly one imaginary frequency in the Hessian, the convergence rate of DFT TS optimizations (b), and the number of optimization steps required (c). Each metric is assessed for both the raw input structures and the corresponding AEFM-refined samples. AEFM incurs minimal overhead, typically requiring only 2 to 5 model evaluations depending on the quality of the initial guess, as seen in Supplementary Figure 5d. In contrast, full DFT optimizations are significantly more expensive. Applied to GFN2-xTB initial guesses, AEFM increases the fraction of valid TS structures from 27% to 68%, a 41% absolute improvement (Supplementary Figure 5a). Moreover, AEFM improves the overall convergence rate of TS optimizations from 91% to 99% (Supplementary Figure 5b), further underscoring its robustness. Lastly, AEFM reduces the median number of DFT optimization steps by 10, corresponding to a threefold acceleration of CPU hours needed in the refinement process (Supplementary Figure 5c, Supplementary Table 2).

To compute the electronic energy of samples, we use ORCA5.0.4 (Neese, 2022) in combination with ASE (Larsen et al., 2017) at the same level of theory as the Transition1x dataset (Schreiner et al., 2022a) was generated with ωB97x/6-31G(d) (Ditchfield et al., 1971; Chai and Head-Gordon, 2008). To generate the GFN2-xTB (Bannwarth et al., 2019) TS guesses, CI-NEB (Henkelman et al., 2000) using ASE and the python interface tblite. For the CI-NEB computations, the same protocol is used as for Transition1x generation. The NEB calculation is first run until the maximum force perpendicular to the path falls below a threshold of 0.5 eVÅ^{-1}. Subsequently, the CI-NEB refinement continues until convergence, defined as a maximum perpendicular force below 0.05 eVÅ^{-1} or a maximum of 500 iterations. Reactions that do not meet this criterion are considered not converged. For TS optimization, the Sella package (Hermes et al., 2022) using the P-RFO (Banerjee et al., 1985; Baker, 1986) algorithm along with the ASE ORCA calculator is run until the maximum force of 0.001 eVÅ^{-1} is achieved with a maximum number of 300 iterations. Numerical Hessians are computed using finite central difference method with an δ of 0.01Å.

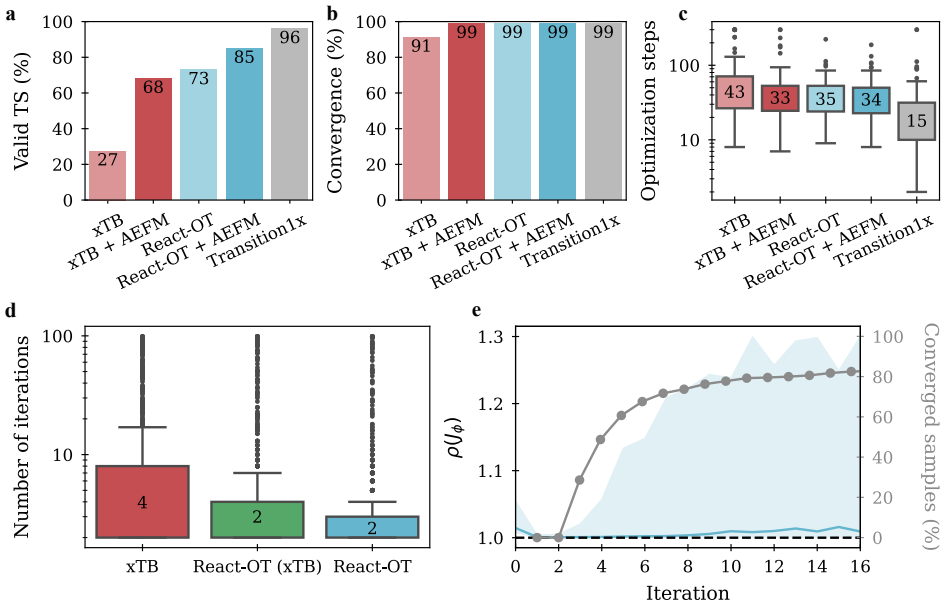

Figure 5: **Chemical validation and fixed-point convergence analysis.** **a** Fraction of valid TS structures, defined by the presence of exactly one imaginary frequency in the Hessian. **b** Convergence rate of DFT TS optimizations. **c** Boxplot of DFT optimization steps required to reach a converged TS structure. **d** Number of iterations required by AEFM to reach a fixed point. Convergence is defined by an RMSD below 0.01 between successive iterates; otherwise, inference is terminated after 100 iterations. **e** Spectral radius of the model's Jacobian with respect to the input structure, shown as median (solid line) and interquartile range (shaded region) over iterations (left y-axis). The percentage of converged samples is plotted on the right y-axis. Contractive behavior ensuring convergence occurs when $\rho(J_\phi) < 1.0$, while advanced solvers still succeed beyond this threshold.

Table 3: Total CPU hours required for TS refinement using the p-RFO algorithm implemented in the Sella package with 48 CPU cores.

Method	CPU Hours
xTB CI-NEB	1430
xTB CI-NEB + AEFM	506
React-OT	455
React-OT + AEFM	439

C Ablation studies

C.1 Comparison to flow matching

An alternative to AEFM is to apply flow matching (FM) using a Gaussian prior and the transition states as the target distribution. The initial time from which the ODE in FM is integrated is inferred based on the mean RMSD between the low-fidelity structures and the reference transition states, using the definition of the intermediate interpolants $\mathbf{x}_t$. Specifically, t_0 is chosen such that the RMSD between the time interpolant $\mathbf{x}_t$ and the target $\mathbf{x}_1$, $\|\mathbf{x}_t - \mathbf{x}_1\|/N$, matches the average RMSD of the low-fidelity source. For GFN2-xTB samples, this yields $t_0 = 0.87$, and for React-OT samples, $t_0 = 0.93$. Table 4 reports the corresponding performance. These results are significantly worse than those obtained with AEFM, which can be attributed to the fact that FM must learn the flow field from a Gaussian prior, making the task considerably more complex compared to AEFM. Furthermore, to ensure a fair comparison with AEFM, the source and target molecules are not aligned, resulting in a non-linear vector field that is harder to integrate and leads to less accurate structures.

Table 4: Performance using FM for refinement.

Approach	RMSD (Å)		$\lvert\Delta E_{\text{TS}}\rvert$ (kcal mol^{-1})	
	Mean	Median	Mean	Median
xTB CI-NEB	0.312	0.179	10.426	2.673
xTB CI-NEB + FM	0.439	0.310	91.732	88.953
React-OT	0.183	0.092	3.405	1.092
React-OT + FM	0.252	0.167	34.723	34.265

C.2 Data efficiency

To evaluate the data efficiency of AEFM, the model was trained using subsets of 2000, 4000, 6000, 8000, and all 9000 training samples. Each trained model was then applied to the GFN2-xTB samples, and the resulting mean and median energetic differences to the ground truth transition states are compared in Supplementary Figure 6. The results reveal a clear trend of decreasing energetic difference with increasing training data. Notably, using only 4000 training samples, less than half of the whole dataset, already achieves a 24.6% reduction in mean absolute error in barrier height, compared to the 40% reduction obtained using the full 9000 samples.

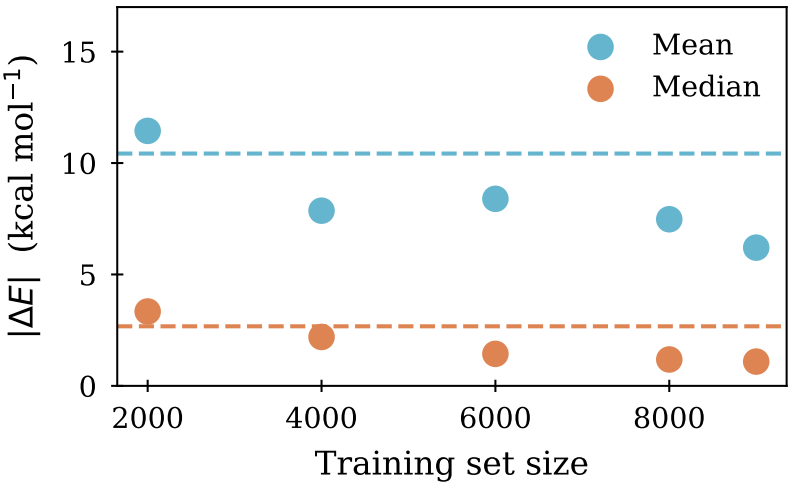

Figure 6: **Data efficiency of AEFM.** Mean and median energy errors for models trained with different sample sizes. Dashed lines show corresponding errors from GFN2-xTB.

C.3 Physics-Informed loss

Supplementary Table 6 shows the Wasserstein-1 distance for each bonded interaction using the thresholds defined in Supplementary Table 5. Combining React-OT with AEFM results in consistently lower Wasserstein-1 distances across nearly all bonded and non-bonded interactions, indicating improved agreement with the underlying Transition1x dataset.

Table 5: Threshold distances used to determine bonded atom pairs. To categorize into bonded and non-bonded, an additional margin of 0.1 Å is added on top of the threshold values.

	C–C	C–H	C–N	C–O	H–H	H–N	H–O	N–N	N–O
Bond Threshold (Å)	1.54	1.09	1.47	1.43	0.74	1.01	0.96	1.45	1.40

Table 6: Wasserstein-1 distance to bond distribution inherent in the test samples of the Transition1x dataset for different bonded (bd) and non-bonded (nbd) atom pairs (lower is better).

Method	C–C	C–H	C–N	C–O	H–H	H–N	H–O	N–N	N–O
xTB CI-NEB (bd)	0.0041	0.0058	0.0097	0.0065	0.0151	0.0046	0.0029	0.0085	0.0113
xTB CI-NEB + AEFM (bd)	0.0032	0.0015	0.0021	0.0018	0.0543	0.0063	0.0140	0.0116	0.0099
React-OT (bd)	0.0022	0.0019	0.0033	0.0019	0.0124	0.0035	0.0059	0.0080	0.0142
React-OT + AEFM (bd)	0.0011	0.0011	0.0023	0.0014	0.0081	0.0034	0.0061	0.0059	0.0110
xTB CI-NEB (nbd)	0.0067	0.0289	0.0192	0.0129	0.0082	0.0461	0.0913	–	0.0697
xTB CI-NEB + AEFM (nbd)	0.0046	0.0097	0.0138	0.0234	0.0027	0.0324	0.0416	0.0981	0.0466
React-OT (nbd)	0.0074	0.0068	0.0128	0.0115	0.0037	0.0088	0.0205	0.1957	0.0413
React-OT + AEFM (nbd)	0.0061	0.0065	0.0106	0.0122	0.0034	0.0139	0.0184	0.0851	0.0518

Table 7: Performance using AEFM without bond loss for refinement.

Approach	RMSD (Å)		$\lvert\Delta E_{\text{TS}}\rvert$ (kcal mol^{-1})	
	Mean	Median	Mean	Median
xTB CI-NEB	0.312	0.179	10.426	2.673
xTB CI-NEB + AEFM (w_{b}=0.0)	0.249	0.103	10.405	1.518
React-OT	0.183	0.092	3.405	1.092
React-OT + AEFM (w_{b}=0.0)	0.186	0.083	3.750	1.056

D Metrics

The RMSD for molecules is determined by first aligning the molecules $\mathbf{x}_1$ and $\mathbf{x}_2$ using the Kabsch algorithm and then computing:

$$\begin{aligned}\text{RMSD}(\mathbf{x}_1, \mathbf{x}_2) &= \sqrt{\frac{\sum_{i=1}^{N} \|\mathbf{x}_{1,i} - \mathbf{x}_{2,i}\|^2}{N}} \\ &= \sqrt{\frac{\sum_{i=1}^{N} \sum_{j\in\{x,y,z\}} (x_{1,i,j} - x_{2,i,j})^2}{N}}\end{aligned} \tag{17}$$

with N denoting the number of atoms. Note that this definition differs from the one used in React-OT, where the RMSD is normalized by $3N$ instead. To access the difference in barrier height, the electronic energy V of each sample TS structure is computed and the MAE is defined as

$$\text{MAE} = \frac{1}{M} \sum_{i}^{M} |V(\mathbf{x}_i) - V(\hat{\mathbf{x}}_i)| \tag{18}$$

with $\hat{\mathbf{x}}_i$ as the predicted TS and $\mathbf{x}_i$ as the corresponding database TS and M as the total number of samples. To compare the distribution of bond lengths in the predicted structures with those in the reference data, we use the Wasserstein-1 distance. Given two one-dimensional empirical distributions p and q over bond lengths with cumulative distribution functions P and Q, respectively, the Wasserstein-1 distance is defined as:

$$W_1(p, q) = \int_{-\infty}^{\infty} |P(x) - Q(x)| \, dx. \tag{19}$$

The Wasserstein-1 distance is computed separately for each bond type in the dataset and subsequently averaged across all types. As a metric that quantifies the minimal effort required to transform one distribution into another, it is particularly well-suited for capturing differences in geometric structure distributions, such as bond lengths.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes. The abstract states the primary improvements achieved by AEFM, and the introduction summarizes the methodological contributions (equilibrium flow formulation, physics-informed bond loss, fixed-point refinement) and scope. These claims are supported by the experiments and analyses presented in the main text and the Supplementary Information.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The discussion section outlines the main limitations of the current AEFM implementation and suggests directions for future improvement. In addition, ablation studies on data efficiency and the impact of the physics-informed loss provide further insight into the method's constraints.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The Methods section provides detailed derivations of all equations and clearly states the assumptions underlying the theoretical results.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: A link to a public GitHub repository is provided, including a README file that details the training and inference procedures.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility.

In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: A link to a public GitHub repository is provided, including a README file that details the training and inference procedures. Furthermore, a link to zenodo for the pretrained models and datasets is provided.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The Supplementary Information contains an extensive table listing all hyperparameters used for training and inference.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We report mean and median metrics to summarize model performance. Although we did not include multiple random seeds in the main experiments, we conducted additional tests training models with different seeds and observed consistent results. Because our method is largely deterministic with minimal variability, traditional error bars or statistical significance tests are not applicable.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The GPU hardware and the number of CPU nodes are reported in both the main text and the Supplementary Information.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in this paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The introduction discusses the broader impact of improved transition state prediction for advancing sustainable chemistry and accelerating reaction discovery. Potential negative societal impacts are minimal given the foundational nature of the work, though misuse in designing harmful compounds is conceivable.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The work focuses on a specialized scientific model for transition state prediction in chemistry, which has no foreseeable high-risk misuse potential. Therefore, safeguards are not applicable.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All prior work and datasets are cited accordingly.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Pretrained models and low-fidelity datasets are publicly available on Zenodo, accompanied by thorough documentation. Additionally, a detailed README on GitHub guides users on how to utilize these assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.