nature computational science

Supplementary information

https://doi.org/10.1038/s43588-026-01034-5

Robust generative transition-state models for unseen chemistry

In the format provided by the authors and unedited

Contents

Supplementary Section 1 Architecture

To verify that the LEFTNet architecture retains its expressiveness under the newly introduced atomic embeddings, we first perform controlled ablation studies on the original Transition1x dataset. Specifically, we fine-tune a vanilla React-OT model equipped with the new embedding scheme, replacing the original one-hot encoding of atom types (H, C, N, O) with learnable atomic number embeddings augmented by selected electronic properties. This modification results in a median structural RMSD of 0.096 Å and a median energetic difference of 1.123 kcal mol^{-1}, compared to 0.092 Å and 1.092 kcal mol^{-1} for the original React-OT model, indicating that the new representation preserves the predictive performance of the architecture. In a second ablation, we evaluate the impact of the aggregation mechanism within LEFTNet. The original implementation employs sum aggregation over a fully connected graph connecting reactant, transition state, and product representations. We additionally test a variant using mean aggregation, motivated by observed numerical instabilities in larger molecular systems, leading to a median structural RMSD of 0.099 Å and a median energetic difference of 1.232 kcal mol^{-1}.

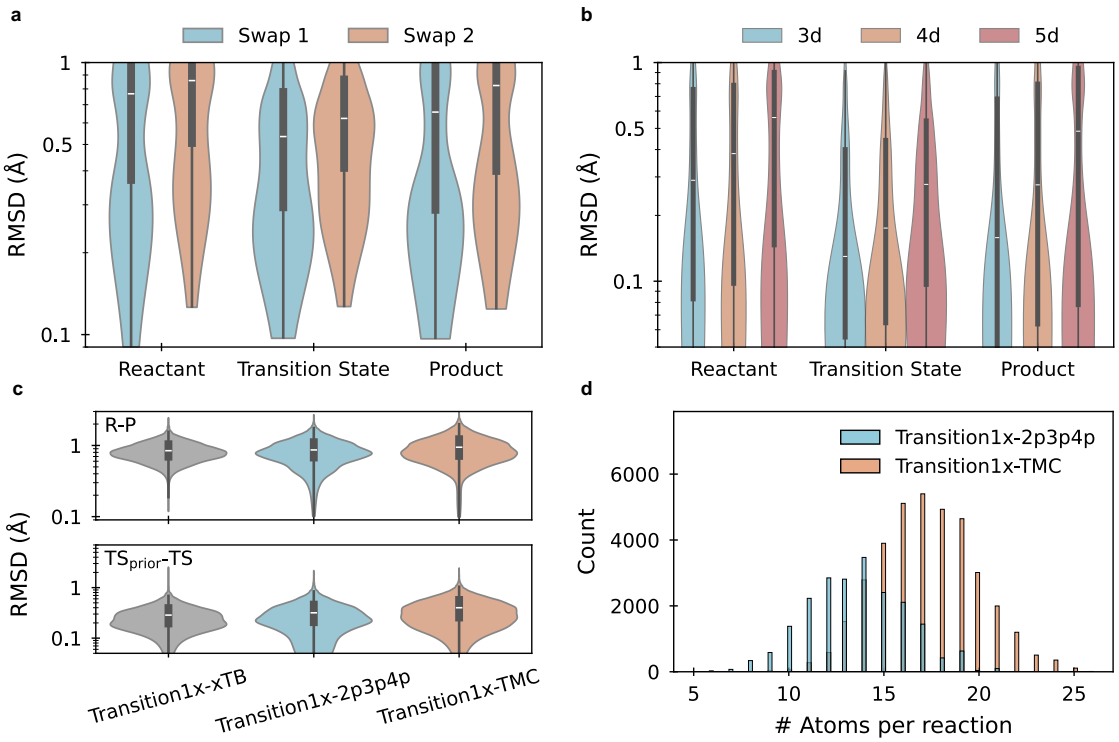
Supplementary Section 2 Datasets

The Transition1x-2p3p4p dataset comprises not only reactions generated through atomic substitutions, but also the corresponding source reactions from Transition1x optimized at the GFN2-xTB level. Including the original HCNO reactions in the training set improves the performance of the fine-tuned React-OT model, yielding a median RMSD of 0.14 Å and a median energetic deviation of 2.97 kcal mol^{-1}, compared to a model fine-tuned exclusively on swap type 1 and 2 reactions, which attains a median RMSD of 0.17 Å and a median energetic deviation of 4.75 kcal mol^{-1}. This improvement can be attributed to the increased training set size, which comprises 18594 reactions when including both Transition1x-xTB and swapped reactions, compared to 12396 reactions when only swapped reactions are used. This effect is not observed for the Transition1x-TMC dataset, which already contains more than 36000 reactions. A React-OT model trained on 32633 Transition1x-TMC reactions achieves a median RMSD of 0.24 Å and a median energetic deviation of 6.68 kcal mol^{-1}. Adding 6198 Transition1x-xTB reactions to the training set does not alter performance, as the resulting model attains identical median RMSD and energetic deviation values. This indicates that, in this regime, additional organic reactions do not further influence model behavior. Based on these observations, all Transition1x-2p3p4p-trained models include both the swapped and organic reactions, whereas Transition1x-TMC-trained models are trained exclusively on TMC-substituted reactions.

As both Transition1x-2p3p4p and Transition1x-TMC are derived from Transition1x, it is instructive to quantify how their reaction trajectories deviate from the original Transition1x-xTB references (Supplementary Figure 1a). For the computation of RMSDs between Transition1x-xTB and Transition1x-TMC structures, the added TMC substituent is removed prior to alignment so that only the common organic part is compared. A pronounced deviation is observed for the reactant and product structures in Transition1x-2p3p4p. To elucidate this effect, we further analyze the number of formed and broken bonds in Transition1x-2p3p4p relative to Transition1x-xTB (Supplementary Table 1). Approximately 40% of reactions involve the swapped atom directly, and both Transition1x-2p3p4p subsets exhibit a globally reduced reactivity, consistent with the stabilizing influence of heavier substituents.

For Transition1x-TMC, deviations from the Transition1x-xTB reference increase systematically with the period of the substituted TMC atom (Supplementary Figure 1b). As an additional structural analysis, Supplementary Figure 1c compares (i) the RMSD between reactant and product, and (ii) the RMSD between the linear prior and the target TS structure across the Transition1x-xTB, Transition1x-2p3p4p, and Transition1x-TMC datasets. The reactant-product RMSD distributions for Transition1x-2p3p4p and Transition1x-TMC both exhibit extended low-RMSD tails, indicating a larger proportion of reactions involving relatively minor structural rearrangements. In contrast, the RMSD between the linear prior and the TS structure reflects the intrinsic difficulty of the dataset. Here, Transition1x-TMC yields the largest deviations, confirming that it represents the most challenging setting, with the linear prior lying farthest from the true TS and thereby posing a more demanding learning problem. Finally, Supplementary Figure 1d shows the distribution of the number of atoms participating in each reaction, highlighting that Transition1x-TMC consists of larger molecular systems with correspondingly more degrees of freedom.

GFN2-xTB provides a practical balance between computational efficiency and accuracy. Supplementary



Supplementary Figure 1: **Statistics for Transition1x-2p3p4p and Transition1x-TMC. a** and **b** Shown is the RMSD of the reactant, TS, and product structures of the Transition1x-2p3p4p and Transition1x-TMC datasets with respect to their corresponding source structures from Transition1x-xTB. **c** RMSD distributions for reactant–to–product structures, as well as the prior TS guess used in React-OT to the true TS structures, for Transition1x-xTB, Transition1x-2p3p4p, and Transition1x-TMC. **d** Distribution of the number of atoms participating in the reactions for Transition1x-2p3p4p and Transition1x-TMC ($n_{\text{Transition1x-2p3p4p}} = 13948$, $n_{\text{Transition1x-TMC}} = 35684$, $n_{\text{Transition1x-xTB}} = 9703$).

Table 2 reports the average wall time required to process one reaction, including both the P-RFO optimization and the subsequent IRC calculations. A Transition1x-TMC reaction is found to require nearly three times the computational time of a Transition1x-2p3p4p reaction, which can be attributed to the larger molecular sizes involved (Transition1x-2p3p4p: 13.5 atoms on average; Transition1x-TMC: 17.3 atoms on average). For reference, Supplementary Table 2 also includes DFT timings for a set of 100 Transition1x-TMC reactions, illustrating that GFN2-xTB is more than 700-fold faster than the corresponding DFT calculations.

Supplementary Section 3 Additional analysis

Supplementary Section 3.1 Similar reactions

To assess the transferability of React-OT to previously unseen elements while preserving comparable reaction mechanisms, we identify structurally similar reactions between the altered datasets, Transition1x-2p3p4p or Transition1x-TMC, and their corresponding Transition1x-xTB source reactions. A reaction is considered structurally similar if, for each structure along the reaction path, reactant, TS, and product, the RMSD to the original Transition1x-xTB structure is below 0.2 Å. Under this criterion, 40 reactions from Transition1x-2p3p4p qualify as similar, comprising 36 swap type 1 reactions and 4 swap type 2 reactions. For Transition1x-TMC, 197 distinct Transition1x-xTB reactions have at least one structurally similar Transition1x-TMC counterpart. However, not every TM leads to the same set of similar reactions, these 197 source reactions give rise to a total of 928 Transition1x-TMC reactions, with the number of similar reactions varying across metals. Supplementary Table 3 summarizes the performance of the vanilla React-OT model on the structurally

Supplementary Table 1: **Reactivity metrics for Transition1x-xTB and Transition1x-2p3p4p.** Average number of reactive atoms, formed bonds, and broken bonds per reaction for Transition1x-xTB and the two Transition1x-2p3p4p swap types, along with the percentage of reactions in which the swapped atom itself participates in the reactive center.

Type	# Reactive atoms	# Formed bonds	# Broken bonds	Swapped atom in reactive center (%)
Transition1x-xTB	3.52	1.13	1.54	-
Swap 1	2.23	0.72	0.97	40.6
Swap 2	2.19	0.77	0.92	42.2

Supplementary Table 2: **Timing for dataset creation for Transition1x-2p3p4p and Transition1x-TMC using P-RFO and IRC.** Average wall-clock time per reaction for the combined P-RFO TS optimization and IRC calculation at the GFN2-xTB level for Transition1x-2p3p4p and Transition1x-TMC, along with the corresponding DFT timing for a subset of Transition1x-TMC reactions, illustrating the computational speed-up of GFN2-xTB relative to DFT.

Type	Dataset	Time per reaction (s)
GFN2-xTB	Transition1x-2p3p4p	18.3
GFN2-xTB	Transition1x-TMC	46.5
DFT	Transition1x-TMC	33174.7

similar Transition1x-2p3p4p and Transition1x-TMC reactions, alongside its performance on the corresponding Transition1x-xTB source reactions. The pseudo-reaction pretrained React-OT models are also evaluated, demonstrating improved accuracy in both Transition1x-2p3p4p and Transition1x-TMC settings.

Supplementary Table 3: **Structural and energetic errors of Transition1x-2p3p4p and Transition1x-TMC on similar reactions, compared with HCNO-counterpart Transition1x-xTB errors.** Mean and median RMSD, DMAE, and TS energetic error (ΔE_{TS}) of the vanilla React-OT model evaluated on structurally similar reaction subsets. The number of reactions in each subset is also reported.

Approach	Dataset	RMSD (Å)		DMAE (Å)		$\lvert\Delta E_{\mathrm{TS}}\rvert$ (kcal mol^{-1})		# Reactions
		Mean	Median	Mean	Median	Mean	Median	
Vanilla React-OT	Transition1x-xTB	0.06	0.04	0.03	0.02	1.19	0.54	40
Vanilla React-OT	Transition1x-2p3p4p	0.21	0.18	0.11	0.09	177.51	130.66	40
Vanilla React-OT[a]	Transition1x-2p3p4p	0.11	0.10	0.06	0.06	21.43	18.21	40
Vanilla React-OT	Transition1x-xTB	0.08	0.05	0.12	0.10	1.74	0.65	197
Vanilla React-OT	Transition1x-TMC	0.42	0.39	0.27	0.26	2396.83	1648.49	928
Vanilla React-OT[a]	Transition1x-TMC	0.26	0.19	0.12	0.10	18.96	10.43	928

[a] Pretraining with 2500 pseudo-reactions of the corresponding dataset.

Supplementary Section 3.2 Behavior on novel elements

To obtain a more detailed geometric assessment, we analyze the Wasserstein-1 distance between the empirical bonded-distance distributions and those generated by React-OT. While React-OT predicts bond distances between familiar organic elements with good accuracy, bonds involving previously unseen elements are often assigned lengths similar to their organic analogues, resulting in broad and frequently unphysical distributions with a tendency toward unrealistically short values. Supplementary Table 4 summarizes these distances for the vanilla, fine-tuned, and self-supervised pretrained React-OT models. A clear separation is observed be-

tween HCNO-only bonds, which yield consistently low Wasserstein-1 distances, and bonds involving novel elements. This behavior is further illustrated in Supplementary Figure 2. Panel (b) compares the empirical bonded-distance distributions with those predicted by the vanilla React-OT model for substituted carbon and nitrogen derivatives in the Transition1x-2p3p4p dataset, while panels (c) and (d) show the corresponding distributions for bonds involving carbon and nitrogen in the Transition1x-TMC dataset. Across all cases, React-OT accurately reproduces the bonded-distance distributions for bond types present in the Transition1x-xTB reference data. In contrast, for bond types absent from the training set, the model exhibits pronounced uncertainty, generating broad and frequently unphysical distance distributions that extend toward unrealistically short bond lengths.

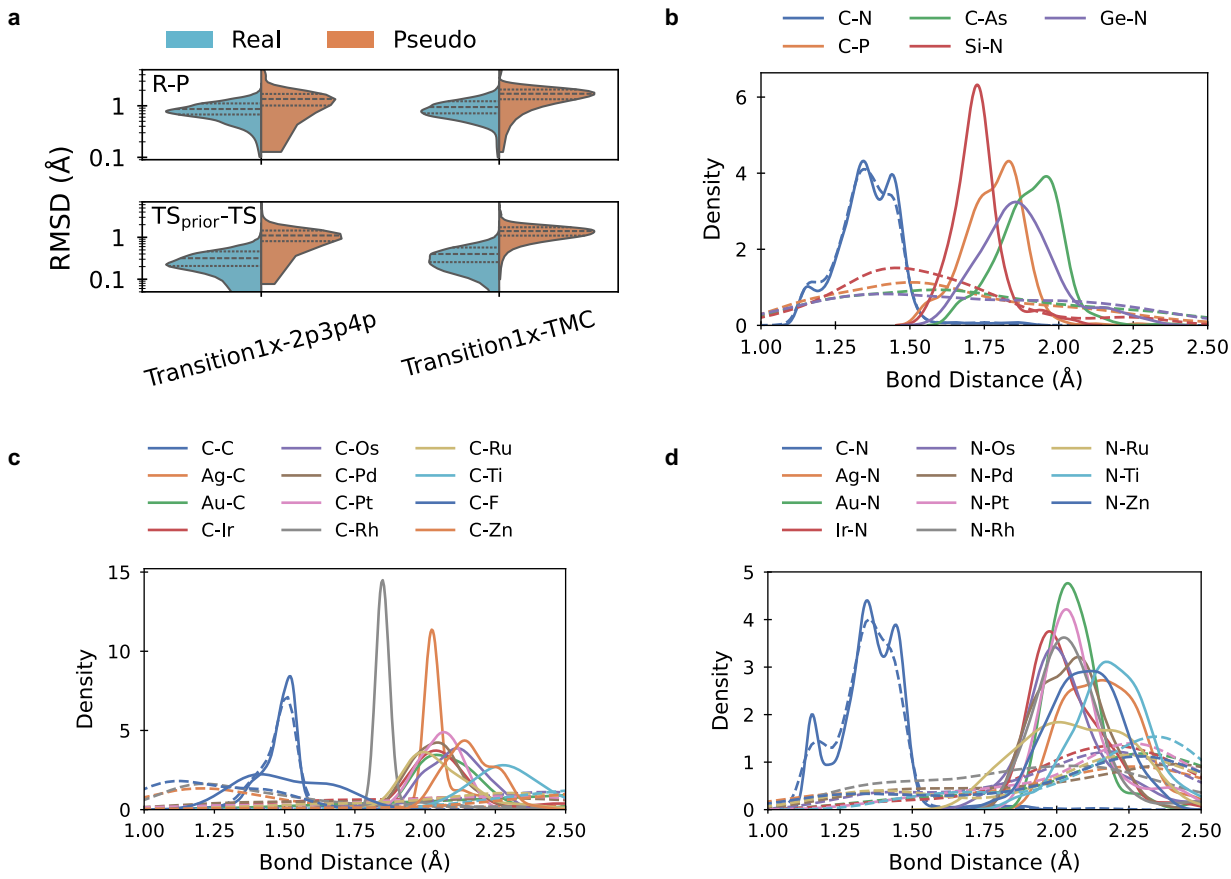
A post hoc relaxation of React-OT–predicted structures, in which atomic positions are adjusted to match the mean bond lengths of the corresponding bond types observed in the dataset, reinforces this conclusion by showing that most extreme errors originate from these unphysical distances. For the Transition1x-2p3p4p benchmark, the median Wasserstein-1 distance decreases from 0.29 Å to 0.07 Å. A similar reduction is seen for the Transition1x-TMC benchmark, where the median Wasserstein-1 distance drops from 0.30 Å to 0.07 Å (Supplementary Table 4). Although adjusting bond lengths to match the dataset mean significantly improves local bond statistics, it only marginally affects overall structural accuracy and does not resolve the fundamental limitations of the predictions.

Supplementary Table 4: **Wasserstein-1 distance between the empirical and the model-generated bond distributions.** Mean and median Wasserstein-1 distance between reference and React-OT-generated bond-length distributions for Transition1x-2p3p4p and Transition1x-TMC, computed over all bond types and separately over HCNO-only bonds.

Approach	Dataset	Wasserstein-1 distance all elements (Å)		Wasserstein-1 distance HCNO elements (Å)	
		Mean	Median	Mean	Median
Vanilla React-OT	Transition1x-2p3p4p	0.32	0.29	0.04	0.01
Vanilla React-OT[a]	Transition1x-2p3p4p	0.22	0.16	0.03	0.01
Vanilla React-OT[b]	Transition1x-2p3p4p	0.07	0.07	0.06	0.06
Vanilla React-OT	Transition1x-TMC	0.33	0.30	0.03	0.02
Vanilla React-OT[a]	Transition1x-TMC	0.14	0.10	0.02	0.01
Vanilla React-OT[b]	Transition1x-TMC	0.07	0.07	0.07	0.07
Fine-tuned React-OT	Transition1x-2p3p4p	0.12	0.02	0.03	0.01
Fine-tuned React-OT[a]	Transition1x-2p3p4p	0.12	0.02	0.02	0.01
Fine-tuned React-OT	Transition1x-TMC	0.08	0.04	0.02	0.01
Fine-tuned React-OT[a]	Transition1x-TMC	0.07	0.03	0.02	0.01

[a] Pretraining with 2500 pseudo-reactions of the corresponding dataset.
[b] Post hoc relaxation to empirical mean bond length.

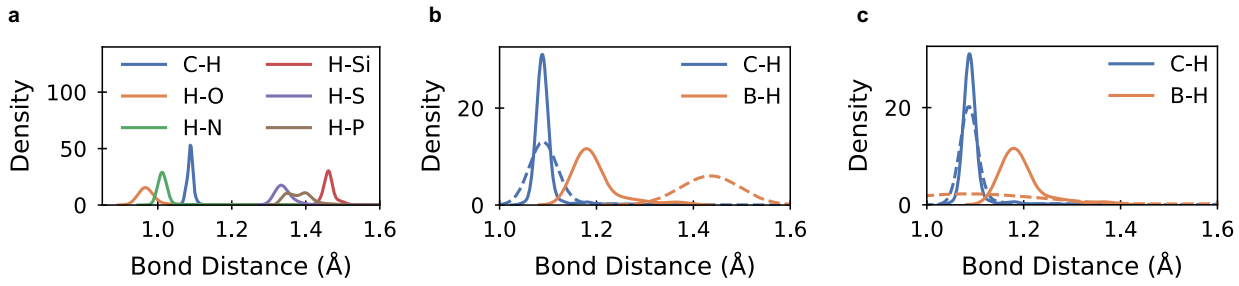
Supplementary Figure 3 shows the comparison of the empirical and predicted bonded distribution for the vanilla and Transition1x-2p3p4p fine-tuned React-OT model evaluated on the Transition1x-OOD benchmark. The fine-tuned model overestimates the B–H bond lengths. This behavior can be traced back to the bond-length distributions present in the Transition1x-2p3p4p training data, where bonds between hydrogen and newly introduced elements such as Si, S, and P consistently peak between 1.3 and 1.5 Å (Supplementary Figure 3a). As a result, the fine-tuned model predicts a broadened B–H distribution centered around 1.45Å, rather than the empirical peak near 1.2 Å (Supplementary Figure 3b). By contrast, the vanilla model produces an even broader distribution with a weak maximum near the organic C–H bond length of approximately 1.1Å (Supplementary Figure 3c), which incidentally leads to a lower overall RMSD. Crucially, this apparent structural agreement of the vanilla model comes at the expense of physical plausibility. Its broader bond-length distributions include a substantial fraction of unphysically short bonds, which is reflected in significantly worse energetic accuracy. Despite achieving lower RMSDs, the vanilla React-OT model exhibits a median energetic error of 27.85 kcal mol^{-1}, compared to 21.54 kcal mol^{-1} for the Transition1x-2p3p4p-



Supplementary Figure 2: **Comparison of RMSD distributions for pseudo- and real-reactions and bonded distributions from the vanilla React-OT model. a** RMSD distributions of reactant–product pairs and of the linear prior TS guess relative to the reference TS structures, shown for Transition1x-2p3p4p and Transition1x-TMC, comparing pseudo- and real-reactions. ($n_{\text{Transition1x-2p3p4p, real}} = 13948$, $n_{\text{Transition1x-TMC, real}} = 35684$, $n_{\text{Transition1x-2p3p4p/TMC, pseudo}} = 2500$) **b** Bond distributions for swapped derivatives of carbon and nitrogen in Transition1x-2p3p4p. **c** Bond distributions involving carbon in Transition1x-TMC. **d** Bond distributions involving nitrogen in Transition1x-TMC. Empirical distributions are shown as solid lines and model-generated distributions as dashed lines.

fine-tuned model (Supplementary Table 9). Consistently, evaluation across all bond types confirms inferior distributional agreement for the vanilla model, with a median Wasserstein-1 distance of 0.09Å, relative to 0.05 Å for the fine-tuned counterpart. A similar trend is observed for AEFM. The Transition1x-2p3p4p fine-tuned model exhibits a modest improvement over the vanilla version, reducing the median RMSD from 0.26 Å to 0.23 Å, while retaining the same fundamental limitations in extrapolation behavior. A comparable pattern is observed for the catalytic benchmarks (Supplementary Table 9). Nevertheless, when focusing only on the reactive center (Rh, coordinated ethylene, and adsorbed H_2), Transition1x-TMC pretraining yields a slight improvement, reducing the RMSD to 0.17 Å compared to 0.20 Å for the vanilla model, suggesting improved modeling of the metal coordination environment. However, generalization to chemically novel ligands remains limited, particularly for phosphorus-containing species.

The performance summary of AEFM on the Transition1x-2p3p4p and Transition1x-TMC benchmark is provided in Supplementary Table 5, showing the same qualitative failure modes as React-OT. This is further highlighted in Supplementary Figure 4, showing a detailed performance of the vanilla AEFM model on Transition1x-2p3p4p.



Supplementary Figure 3: **Comparison of empirical (solid) and generated (dashed) bond distributions for vanilla and fine-tuned React-OT model. a** Empirical bond distributions for a subset of hydrogen containing bonds in Transition1x-2p3p4p. **b** Empirical and predicted distribution for C-H and B-H bonds for fine-tuned React-OT model. **c** Empirical and predicted distribution for C-H and B-H bonds for vanilla React-OT model.

Supplementary Table 5: **Structural and energetic errors for variants of AEFM.** Mean and median RMSD, DMAE, and TS energetic error (ΔE_{TS}) of the vanilla and fine-tuned AEFM model on Transition1x-2p3p4p and Transition1x-TMC.

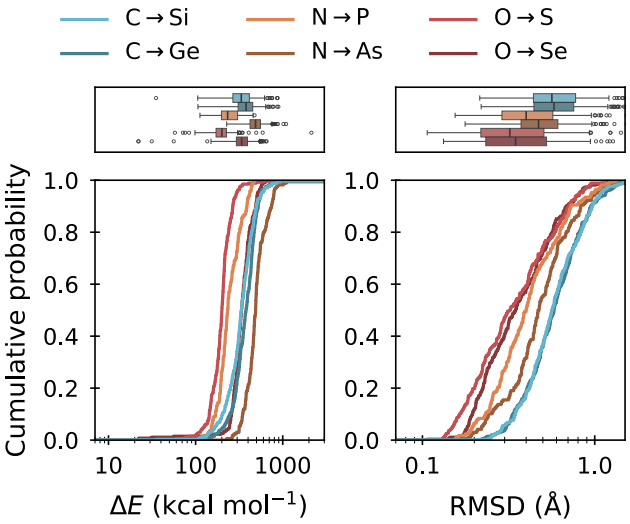
Approach	Dataset	RMSD (Å)		DMAE (Å)		$\lvert\Delta E_{\mathrm{TS}}\rvert$ (kcal mol^{-1})	
		Mean	Median	Mean	Median	Mean	Median
Vanilla AEFM	Transition1x-2p3p4p	0.52	0.47	0.28	0.25	353.68	322.44
Vanilla AEFM	Transition1x-TMC	1.11	1.08	0.67	0.61	2050.31	1568.24
Fine-tuned AEFM	Transition1x-2p3p4p	0.28	0.16	0.11	0.06	8.22	0.93
Fine-tuned AEFM	Transition1x-TMC	0.38	0.27	0.15	0.10	9.19	1.48

Supplementary Section 3.3 Data dependency

To assess the data dependency of React-OT, we train models on varying fractions of the available altered reaction datasets. Supplementary Figure 5 shows the relationship between the energetic error, the structural similarity, and the amount of training data used. The case of zero additional data corresponds to the vanilla React-OT model trained solely on Transition1x-xTB, whereas increasing percentages indicate the fraction of Transition1x-2p3p4p or Transition1x-TMC data incorporated during fine-tuning. Solid lines denote the mean error, and dashed lines the corresponding median. A consistent and monotonic reduction in error is observed as the amount of training data increases.

Supplementary Section 4 Pseudo-reactions for self-supervised pretraining

OAReactDiff is used as the pretraining framework in our two-stage pipeline because it models the joint distribution of the full reaction, thereby enabling the model to learn from reactant, product, and TS-like structures simultaneously. In contrast, React-OT can be viewed as a training variant of the same underlying architecture, where the model is optimized using a flow matching objective that maps a predefined prior distribution, given by linear interpolations between reactants and products, to the TS. In this formulation, the learning signal is primarily governed by the prior, while reactant and product structures serve as conditioning information. Importantly, OAReactDiff and React-OT share an identical network architecture and differ only in their training objective (diffusion versus flow matching), making them naturally compatible within a unified two-stage training strategy. Empirically, this difference in training objective translates into a more informative structural learning signal when using OAReactDiff for pretraining. To assess this, we compare against an alternative setup in which pretraining is performed using the React-OT flow matching objective



Supplementary Figure 4: **Failure modes of the vanilla AEFM model on reactions involving novel chemistry.** Cumulative probability distributions of structural and energetic errors for the Transition1x-2p3p4p benchmark, illustrating the sharp performance degradation introduced by single-atom substitutions ($n_{\mathrm{C} \to \mathrm{Si/Ge}} = 340$, $n_{\mathrm{N} \to \mathrm{P/As}} = 168$, $n_{\mathrm{O} \to \mathrm{S/Se}} = 268$).

instead of the OAReactDiff diffusion loss. The comparison is conducted on the Transition1x-2p3p4p dataset using 2500 pseudo-reactions for pretraining, followed by fine-tuning on 10% of the Transition1x-2p3p4p dataset, while keeping all other training parameters identical. We find that OAReactDiff pretraining leads to improved downstream performance, yielding a median RMSD of 0.17 Å and a median energetic error of 3.87 kcal mol^{-1}. In contrast, pretraining with the React-OT objective results in a higher median RMSD of 0.185 Å and a larger median energetic error of 5.67 kcal mol^{-1}.

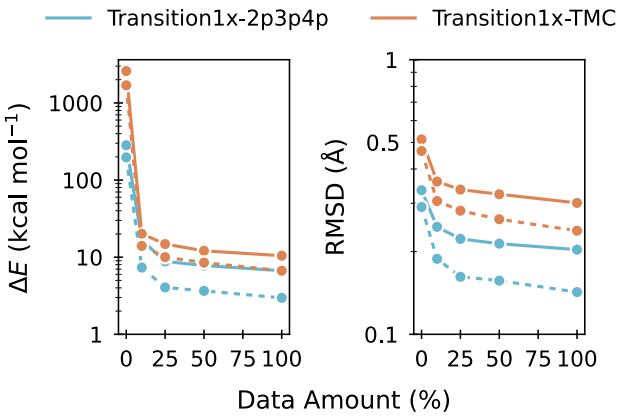
To evaluate whether the generated pseudo-reactions exhibit structural overlap with their corresponding source reactions, that is, the reactions from which the minima were extracted, we compute the structural similarity between each structure in the pseudo-reaction and each structure in the associated source trajectory. The results are reported in Supplementary Tables 6–8. Notably, even the smallest similarity values remain above 1 Å, indicating that the pseudo-reactions do not share meaningful structural overlap with real reactions and therefore constitute a distinct pretraining signal.

Supplementary Figure 6 presents the relationship between energetic and structural error as a function of the amount of fine-tuning data, comparing the vanilla model to models pretrained on 500 and 2500 pseudo-reactions for (a) Transition1x-2p3p4p and (b) Transition1x-TMC. For Transition1x-2p3p4p, pretraining on 2500 pseudo-reactions combined with fine-tuning on only 25% of the data yields performance comparable to a vanilla React-OT model trained on the full dataset. A similar trend is observed for Transition1x-TMC, although the effect of pretraining is less pronounced than in the Transition1x-2p3p4p case.

Supplementary Table 6: **Similarity of pseudo-reactants to source reactions.** Mean and median RMSD between the pseudo-reactant structure used in self-supervised pretraining and the reference reactant, TS, and product structures of the corresponding source reaction, for Transition1x-2p3p4p and Transition1x-TMC.

Dataset	RMSD($\mathrm{R_{pseudo}}$, $\mathrm{R_{ref.}}$) (Å)		RMSD($\mathrm{R_{pseudo}}$, $\mathrm{TS_{ref.}}$) (Å)		RMSD($\mathrm{R_{pseudo}}$, $\mathrm{P_{ref.}}$) (Å)	
	Mean	Median	Mean	Median	Mean	Median
Transition1x-2p3p4p	1.92	1.37	1.89	1.33	1.24	1.28
Transition1x-TMC	1.84	1.68	1.83	1.64	1.65	1.69

Supplementary Table 9 summarizes the impact of pretraining on the ability to elucidate chemically novel reactions. In particular, self-supervised pretraining on 100 pseudo-reactions from the Transition1x-OOD



Supplementary Figure 5: **Data ablation.** Correlation of energetic error and RMSD for Transition1x-2p3p4p and Transition1x-TMC as a function of the amount of training data. The case of no additional training data corresponds to the vanilla React-OT model trained on Transition1x-xTB. Solid lines indicate the mean, while dashed lines indicate the median.

Supplementary Table 7: **Similarity of pseudo-TSs to source reactions.** Mean and median RMSD between the pseudo-TS structure used in self-supervised pretraining and the reference reactant, TS, and product structures of the corresponding source reaction, for Transition1x-2p3p4p and Transition1x-TMC.

Dataset	RMSD(TS_{pseudo}, $R_{ref.}$) (Å)		RMSD(TS_{pseudo}, $TS_{ref.}$) (Å)		RMSD(TS_{pseudo}, $P_{ref.}$) (Å)	
	Mean	Median	Mean	Median	Mean	Median
Transition1x-2p3p4p	1.91	1.36	1.89	1.33	1.85	1.35
Transition1x-TMC	1.84	1.67	1.83	1.64	1.83	1.66

dataset yields clear improvements over the vanilla model trained solely on Transition1x-xTB data. In contrast, pretraining on Transition1x-2p3p4p alone does not improve performance on these novel systems, underscoring the need for system-specific pretraining strategies. For the catalytic systems, both the vanilla and the pretrained model are trained on 50 target reactions from the respective dataset, while the pretrained model is first trained on 100 pseudo-reactions of the corresponding dataset. Comparing the median energetic errors, shows improvements of 44% and 33% for the Pt- and Rh-catalyzed system respectively. In contrast, pretraining on Transition1x-TMC, which includes Rh and Pt systems but does not contain the specific ligands of interest, yields comparable performance after fine-tuning on the 50 target reactions. For example, in the Pt-catalyzed case, the Transition1x-TMC-pretrained model achieves a median RMSD of 0.15 Å and a median energetic error of 5.68 kcal mol^{-1}. Notably, this is similar to the performance obtained using only 100 pseudo-reactions for pretraining, despite Transition1x-TMC comprising more than 30,000 reactions. This effect becomes even more pronounced in the Transition1x-OOD setting. Pseudo-reaction pretraining on the system of interest, without any targeted fine-tuning, yields a median RMSD of 0.14 Å and a median energetic error of 9.22 kcal mol^{-1}. In contrast, pretraining on Transition1x-2p3p4p, which does not include the novel elements fluorine and boron, results in substantially worse performance, with a median RMSD of 0.19 Å and a median energetic error of 21.54 kcal mol^{-1}. These comparisons indicate that the observed improvements cannot be attributed to increased data volume or chemical diversity alone. Instead, they highlight that pseudo-reaction pretraining provides a more targeted and data-efficient learning signal by leveraging equilibrium structures from the system of interest. This enables the model to acquire element- and environment-specific structural information even when large, relevant reaction datasets are unavailable.

Supplementary Table 8: **Similarity of pseudo-products to source reactions.** Mean and median RMSD between the pseudo-product structure used in self-supervised pretraining and the reference reactant, TS, and product structures of the corresponding source reaction, for Transition1x-2p3p4p and Transition1x-TMC.

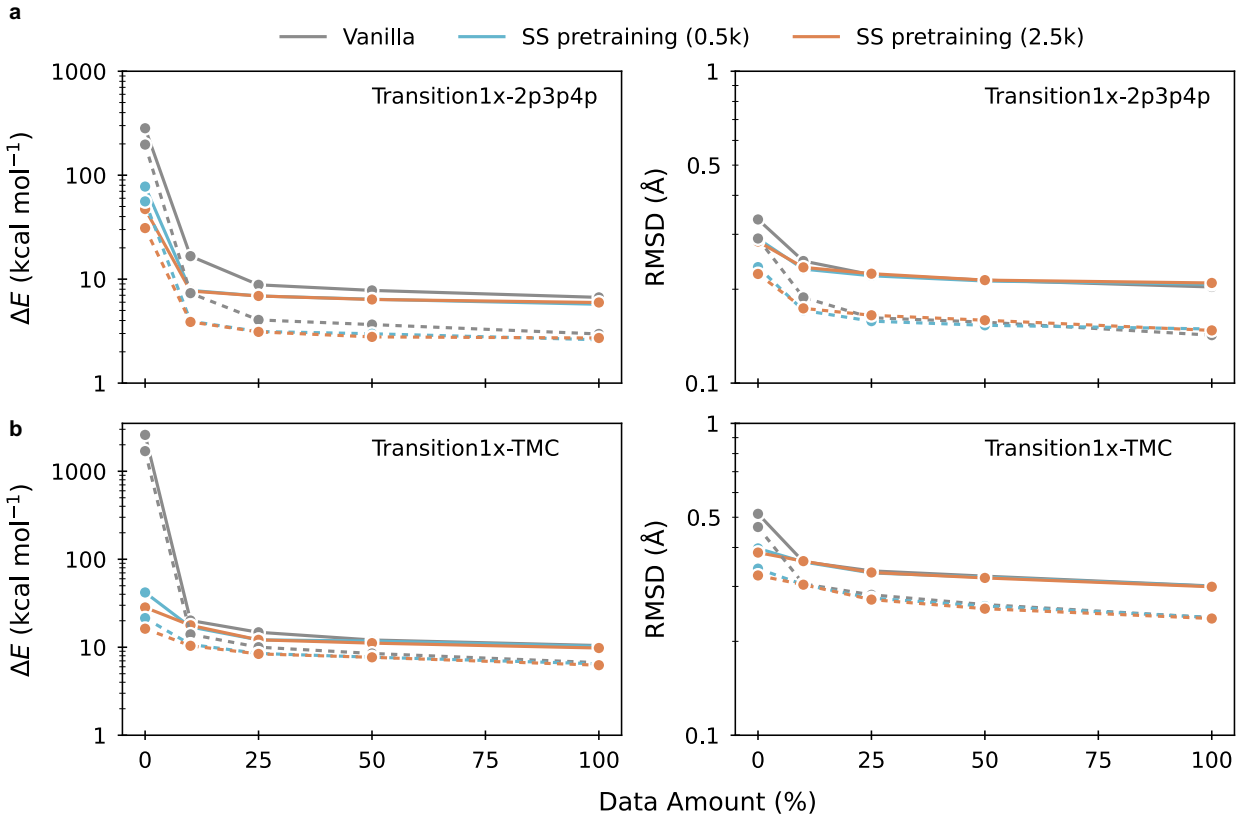
Dataset	RMSD(P_{pseudo}, $R_{ref.}$) (Å)		RMSD(P_{pseudo}, $TS_{ref.}$) (Å)		RMSD(P_{pseudo}, $P_{ref.}$) (Å)	
	Mean	Median	Mean	Median	Mean	Median
Transition1x-2p3p4p	1.18	1.24	1.23	1.23	1.24	1.28
Transition1x-TMC	1.63	1.66	1.65	1.68	1.65	1.69

Supplementary Section 5 Quantum chemical validation

Supplementary Table 10 compares the energetic errors obtained when evaluating model predictions with GFN2-xTB and with DFT, where the respective functional and basis set are indicated as superscripts. In all cases, the energetic error is larger at the DFT level, indicating that GFN2-xTB is less sensitive to certain geometric deviations. The Transition1x-2p3p4p and Transition1x-TMC models are fine-tuned on 100% of their respective datasets, while the model used for the catalytic sets is pretrained on Transition1x-TMC and subsequently fine-tuned on 50 target reactions.

Supplementary Table 11 reports the structural differences between GFN2-xTB TS geometries and their DFT-optimized counterparts, with energetic differences evaluated consistently at the DFT level. For Transition1x-2p3p4p, GFN2-xTB already provides accurate geometries, requiring only minimal structural adjustments upon DFT optimization (median RMSD of 0.16 Å). In contrast, both Transition1x-TMC and the Pt-catalyzed system exhibit larger deviations, reflecting the worse reliability of GFN2-xTB for these reactions.

Supplementary Table 12 compares the React-OT–generated TS structures directly with their corresponding DFT reference TSs, demonstrating that the model achieves reasonably high accuracy, with median energetic errors below 20 kcal mol^{-1}. Finally, Supplementary Table 13 reports the results obtained after DFT optimization of the model-generated structures to valid TSs. Across all evaluated datasets, more than 75% of generated geometries converge to the correct TS and thereby to the intended reaction pathway, demonstrating the practical utility of the generated structures for downstream quantum-chemical applications.



Supplementary Figure 6: **Data ablation including self-supervised pseudo-reaction pretraining. a**, **b** Correlation between energetic error and RMSD for Transition1x-2p3p4p and Transition1x-TMC as a function of training set size. The case without additional training data corresponds to a React-OT model trained on Transition1x-xTB only, while color denotes whether the model underwent prior self-supervised pretraining using pseudo-reactions constructed from the corresponding dataset. Solid lines indicate mean values, and dashed lines indicate medians.

Supplementary Table 9: **Structural and energetic errors on additional benchmark systems.** Mean and median RMSD, DMAE, and TS energetic error (ΔE_{TS}) for the Transition1x-OOD, Pt-catalyzed, and Rh-catalyzed benchmarks, comparing vanilla and fine-tuned React-OT and AEFM models with variants incorporating self-supervised pretraining on pseudo-reactions or on Transition1x-TMC.

Approach	Dataset	RMSD (Å)		DMAE (Å)		$\lvert\Delta E_{\mathrm{TS}}\rvert$ (kcal mol^{-1})	
		Mean	Median	Mean	Median	Mean	Median
Vanilla React-OT	Transition1x-OOD	0.21	0.15	0.11	0.08	41.32	27.85
Vanilla React-OT[a]	Transition1x-OOD	0.20	0.14	0.09	0.06	16.87	9.22
Vanilla React-OT[b]	Transition1x-OOD	0.25	0.19	0.12	0.10	26.65	21.54
Vanilla AEFM	Transition1x-OOD	0.32	0.26	0.14	0.11	26.70	16.93
Vanilla AEFM[c]	Transition1x-OOD	0.30	0.23	0.13	0.08	23.28	14.08
Vanilla React-OT	Pt-catalyzed	0.35	0.33	0.22	0.20	684.57	412.46
Vanilla React-OT[a]	Pt-catalyzed	0.32	0.29	0.17	0.16	185.69	130.76
Vanilla React-OT[d]	Pt-catalyzed	0.34	0.30	0.18	0.17	58.38	36.52
Vanilla AEFM	Pt-catalyzed	0.80	0.75	0.49	0.43	4928.35	931.56
Vanilla AEFM[e]	Pt-catalyzed	0.31	0.25	0.17	0.13	44.05	20.92
Fine-tuned React-OT	Pt-catalyzed	0.20	0.16	0.09	0.08	43.19	9.04
Fine-tuned React-OT[a]	Pt-catalyzed	0.19	0.15	0.08	0.07	11.42	5.06
Fine-tuned React-OT[d]	Pt-catalyzed	0.20	0.15	0.09	0.07	9.67	5.68
Vanilla React-OT	Rh-catalyzed	0.56	0.24	0.43	0.16	2065.93	374.98
Vanilla React-OT[a]	Rh-catalyzed	0.20	0.19	0.18	0.17	203.16	164.66
Vanilla React-OT[d]	Rh-catalyzed	0.35	0.32	0.24	0.22	230.82	191.10
Vanilla AEFM	Rh-catalyzed	0.95	0.88	0.60	0.57	3198.51	1827.39
Vanilla AEFM[e]	Rh-catalyzed	0.39	0.37	0.25	0.24	172.57	150.99
Fine-tuned React-OT	Rh-catalyzed	0.06	0.04	0.03	0.02	38.24	3.65
Fine-tuned React-OT[a]	Rh-catalyzed	0.05	0.03	0.03	0.02	26.53	2.45
Fine-tuned React-OT[d]	Rh-catalyzed	0.05	0.03	0.02	0.02	7.24	1.04

[a] OAReactDiff self-supervised pretraining with 100 pseudo-reactions of the corresponding dataset.
[b] React-OT pretraining with Transition1x-2p3p4p.
[c] AEFM pretraining with Transition1x-2p3p4p.
[d] React-OT pretraining with Transition1x-TMC.
[e] AEFM pretraining with Transition1x-TMC.

Supplementary Table 10: **Comparison of energetic error evaluated with GFN2-xTB and DFT.** Mean and median TS energetic error (ΔE_{TS}) of fine-tuned React-OT predictions computed with GFN2-xTB versus single-point DFT for Transition1x-2p3p4p, Transition1x-TMC, and the Pt- and Rh-catalyzed benchmarks, with the percentage increase at the DFT level indicated in parentheses.

Approach	Dataset	$\lvert\Delta E_{\mathrm{TS}}^{\mathrm{GFN2\text{-}xTB}}\rvert$ (kcal mol^{-1})		$\lvert\Delta E_{\mathrm{TS}}^{\mathrm{DFT}}\rvert$ (kcal mol^{-1})	
		Mean	Median	Mean	Median
Fine-tuned React-OT	Transition1x-2p3p4p [a]	6.41	2.77	8.00 (↑25%)	3.39 (↑22%)
Fine-tuned React-OT	Transition1x-TMC [a]	10.37	6.81	12.78 (↑23%)	7.79 (↑14%)
Fine-tuned React-OT	Pt-catalyzed[b]	9.67	5.68	10.87 (↑12%)	6.89 (↑21%)
Fine-tuned React-OT	Rh-catalyzed[c]	7.24	1.04	7.75 (↑7%)	1.21 (↑16%)

[a] DFT functional/basis set: ωB97x/def2-SVP
[b] DFT functional/basis set: PBE0 D3BJ/def2-SVP
[c] DFT functional/basis set: ωB97x-D3/def2-SVP

Supplementary Table 11: **Statistics of energetic and structural difference after DFT TS optimization for datasets.** Mean and median RMSD and TS energetic error (ΔE_{TS}) between GFN2-xTB TS structures and their DFT-optimized counterparts for Transition1x-2p3p4p, Transition1x-TMC, and the Pt-catalyzed benchmark, along with the number of converged structures and the mean and median number of optimization steps required.

Dataset	RMSD (Å)		$\lvert\Delta E_{\mathrm{TS}}\rvert$ (kcal mol^{-1})		# Converged TS	# Optimization steps	
	Mean	Median	Mean	Median		Mean	Median
Transition1x-2p3p4p [a]	0.27	0.16	5.80	3.31	194/200	61	43
Transition1x-TMC [a]	0.46	0.35	13.93	7.97	99/100	100	80
Pt-catalyzed[b]	0.42	0.33	7.19	6.19	44/50	190	128

[a] DFT functional/basis set: ωB97x/def2-SVP
[b] DFT functional/basis set: PBE0 D3BJ/def2-SVP

Supplementary Table 12: **Comparison of energetic and structural difference of React-OT models compared to DFT TS references.** Mean and median RMSD and TS energetic error (ΔE_{TS}) between fine-tuned React-OT-generated TS structures and the corresponding DFT reference TSs for Transition1x-2p3p4p, Transition1x-TMC, and the Pt-catalyzed benchmark, including the variant pretrained on 1500 DFT-level pseudo-reactions of Transition1x-TMC.

Approach	Dataset	RMSD (Å)		$\lvert\Delta E_{\mathrm{TS}}\rvert$ (kcal mol^{-1})	
		Mean	Median	Mean	Median
Fine-tuned React-OT	Transition1x-2p3p4p [a]	0.36	0.26	12.17	7.67
Fine-tuned React-OT	Transition1x-TMC [a]	0.52	0.47	21.09	16.34
Fine-tuned React-OT	Transition1x-TMC [a,c]	0.52	0.42	20.11	15.48
Fine-tuned React-OT	Pt-catalyzed[b]	0.49	0.46	20.48	16.99

[a] DFT functional/basis set: ωB97x/def2-SVP
[b] DFT functional/basis set: PBE0 D3BJ/def2-SVP
[c] Self-supervised pretraining with 1500 pseudo-reactions at DFT level of theory.

Supplementary Table 13: **Comparison of DFT refined React-OT samples compared to their DFT TS reference.** Mean and median RMSD between fine-tuned React-OT-generated TS structures, subsequently refined at the DFT level, and their corresponding DFT reference TS, for Transition1x-2p3p4p, Transition1x-TMC, and the Pt-catalyzed benchmark. Also reported are the percentage of structures matching the reference TS (RMSD below 0.05 Å), the number of converged optimizations, and the mean and median number of optimization steps required.

Approach	Dataset	RMSD (Å)		Same (%)	# Converged TS	# Optimization steps	
		Mean	Median			Mean	Median
Fine-tuned React-OT	Transition1x-2p3p4p [a]	0.20	0.00	75	188/200	74	51
Fine-tuned React-OT	Transition1x-TMC [a]	0.14	0.01	82	93/100	96	85
Fine-tuned React-OT	Pt-catalyzed[b]	0.15	0.01	76	37/50	162	100

[a] DFT functional/basis set: ωB97x/def2-SVP
[b] DFT functional/basis set: PBE0 D3BJ/def2-SVP