Learning Beyond the Endpoints: Benchmarking and Learning Transition-State-Informed Reaction Representations

Anonymous authors
Paper under double-blind review

ABSTRACT

Representation learning is central to machine learning for chemistry, yet learning from reactant and product structures alone leaves information about the intervening transformation underexplored. We investigate how much transition-state (TS) information benefits reaction representations and how these benefits can be transferred to predictors that use only reactants and products at deployment. We introduce Transition-State-Informed Representation Learning (TSIRL), a benchmark and learning framework that evaluates the value of direct TS access and its transfer to endpoint-only prediction. The benchmark comprises four audited reaction datasets with aligned reactant, product, and TS structures, fixed partitions, and five evaluation metrics for reaction barrier prediction. We further propose a simple and effective strategy: pretrain a shared molecular encoder on reactants, products, and TS structures with barrier supervision, then reuse the encoder in a standard endpoint-only predictor. We evaluate direct TS use and TS-informed pretraining across four datasets and six encoders, test reaction-scaffold generalization, extend TS exposure to five additional pretraining strategies, and compare against six established barrier-prediction baselines. TS information is useful in every evaluated task, while TS-informed pretraining consistently improves endpoint-only prediction, transfers to unseen reaction-scaffold families, generalizes across other pretraining strategies, and achieves the best performance among the evaluated baseline methods.

1 INTRODUCTION

Representation learning is a central route by which artificial intelligence turns scientific structure into predictive features. Molecular Representation Learning (MRL) is widely studied in numerous chemical and biological applications. These studies of how to represent a molecule have proceeded from two-dimensional molecular graphs (Duvenaud et al., 2015; Gilmer et al., 2017) to three-dimensional geometric encoders (Schütt et al., 2017; 2021) and then conformer ensembles (Zhu et al., 2024).

However, chemical reactions require representations that consider not only the related molecules but also the *reactions* that transform reactants to products. Existing reaction representation learning approaches have already used reactants and products information, i.e., the *endpoints*, such as using graph differences, condensed reaction graphs (CGR), or paired endpoint geometries (Wang et al., 2022a; Heid & Green, 2022; van Gerwen et al., 2024). Beyond the endpoints, the structural changes during the reaction process also matter for understanding the underlying chemistry. This motivates exploring *intermediate* structural information as a resource for reaction representation learning.

Fortunately, chemists regard the *transition state* (TS) as a representative transient structure that captures the underlying chemical transformation. Despite their usefulness, they are computationally expensive to obtain, while fast, scalable and precise TS determination remains an open challenge. Consequently, various approaches have been developed to tackle this problem. Methods such as TSNet and React-OT learn to generate TS geometries from reaction endpoints (Jackson et al., 2021; Duan et al., 2025), while another approach incorporates on-the-fly TS predictions into barrier estimation (Karwounopoulos et al., 2025). Although previously, researchers have attempted to construct and use TS, a comprehensive evaluation of the value of TS structures remains lacking.

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

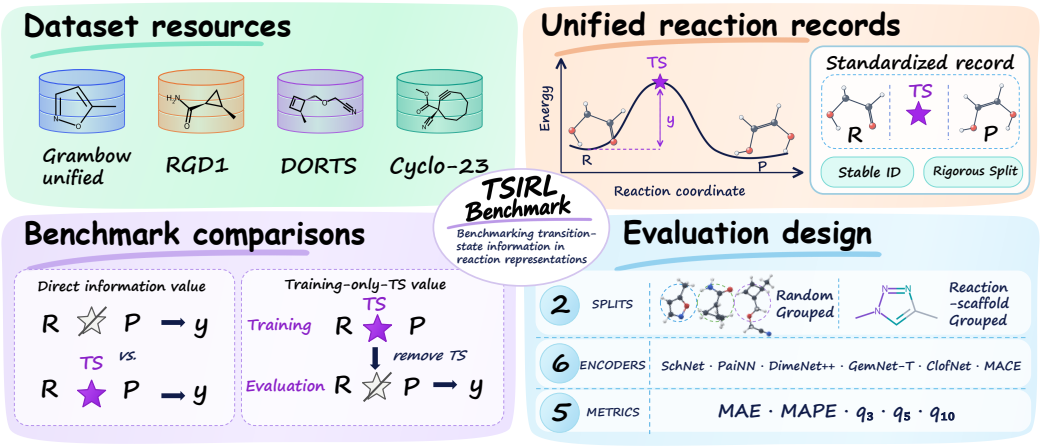


Figure 1: Overview of TSIRL. Four datasets provide aligned reactant (R), product (P), and transition-state (TS) structures with barrier labels. The benchmark distinguishes direct TS predictive value from benefits retained after training-time TS use, with grouped-random and reaction-scaffold splits, six geometric encoders, and five evaluation metrics.

In this scenario, we address this gap with TSIRL, a benchmark to evaluate *how much TS information improves prediction* when directly available and *how effectively its benefits transfer to endpoint-only models* across datasets and encoders. This benchmark combines audited reactant, product, and TS structures with barrier labels, fixed data partitions, and five evaluation metrics. Figure 1 summarizes the benchmark resources, comparisons, and evaluation design. This resource supports both direct prediction with different reaction views and evaluation of representations pretrained with TS structures before endpoint-only use.

We further propose a general learning strategy to use TS, i.e., the *encoder pretraining*. Reactants, products, and transition states are processed by a shared molecular encoder. We pretrain an RPTS predictor, transfer only its molecular encoder into a standard RP predictor, and fine-tune using endpoints and ground-truth labels. We evaluate this strategy along four dimensions: (1) its effectiveness in transferring TS information to endpoint-only reaction representations across datasets and encoders; (2) whether the transferred benefits persist for unseen reaction-scaffold families; (3) whether the benefits of TS-informed training extend to other representation-learning pretraining strategies; and (4) its predictive performance relative to established reaction-barrier prediction methods.

Our experiments show that encoder pretraining consistently improves endpoint-only barrier prediction across the evaluated datasets and encoders. These gains persist across all six encoders under a reaction-scaffold split that holds out depth-1 endpoint-edit families on Grambow Unified. The benefits of incorporating TS structures also extend to other representation-learning pretraining strategies. Our RPTS-pretrained ClofNet further achieves the lowest mean absolute error in comparison against the evaluated external barrier-prediction baselines on Grambow Unified.

Our contributions are:

- **A benchmark for TS-informed reaction representation learning.** It provides audited R/P/TS structures and barrier labels across four datasets, with fixed partitions and five evaluation metrics for studying training-time TS information and endpoint-only prediction.
- **A TS-informed encoder pretraining strategy.** We pretrain shared 3D molecular encoders using R/P/TS structures and transfer them into standard RP predictors, enabling endpoint-only barrier prediction without requiring or generating TS structures at inference.
- **A systematic evaluation of TS transfer and generalization.** We study TS-informed learning across datasets, encoder architectures, reaction-scaffold shifts, pretraining strategies, and established barrier-prediction baselines.

2 Related Work

Molecular encoders. Two-dimensional molecular models use graph convolutions and message passing (Duvenaud et al., 2015; Gilmer et al., 2017; Yang et al., 2019; Xiong et al., 2020). Geometric encoders represent distances in SchNet (Schütt et al., 2017), directions in DimeNet++ and GemNet (Gasteiger et al., 2020; 2021), scalar and vector features in PaiNN (Schütt et al., 2021), local frames in ClofNet and LEFTNet (Du et al., 2022; 2023), and higher-order interactions in MACE (Batatia et al., 2022). MARCEL evaluates learning from conformer ensembles under a common encoder interface (Zhu et al., 2024).

Molecular and reaction pretraining. Existing objectives include attribute masking (Hu et al., 2020), contrastive graph views in MolCLR (Wang et al., 2022b), 3D geometry in Uni-Mol (Zhou et al., 2023), and graph–geometry correspondence in GraphMVP and 3D Infomax (Liu et al., 2022; Stärk et al., 2022). The objectives used in our comparison draw on coordinate denoising (Zaidi et al., 2023), GeoSSL-DDM (Liu et al., 2023), equivariant masked position prediction (An et al., 2025), and C-FREE (Ariguib et al., 2026); their adapted variants are specified in Section 7.1. Reaction-aware pretraining also learns from reactant–product relationships (Wang et al., 2022a).

Endpoint-based barrier prediction. For an energy-based label, the difference between transition-state (TS) and reactant reference energies defines the forward *activation energy*, E_a. The corresponding free-energy barrier is $\Delta G^{\ddagger}$; Section 4 specifies the target used by each dataset. The magnitude of the barrier strongly influences reaction rates and helps compare competing pathways, prioritize chemical transformations, and accelerate automated reaction-mechanism generation (Grambow et al., 2020a; Heinen et al., 2021; Gallarati et al., 2021). Endpoint predictors use condensed reaction graphs (Heid & Green, 2022), 3D endpoint geometries in 3DReact and EquiReact (van Gerwen et al., 2024), differential reaction fingerprints (DRFP) (Probst et al., 2022), reaction-language fingerprints (RXNFP) (Schwaller et al., 2021), and graph models such as Graph2Ea (Vadaddi et al., 2024); other models add quantum-chemical descriptors (Stuyver & Coley, 2022). The reaction-vector benchmark compares molecular fingerprints (MFP) and $\mathrm{B}^2\mathrm{R}^2_\ell$ on computed barriers (Van Gerwen et al., 2024).

TS modeling and evaluation. TSNet and React-OT generate TS geometries from endpoints (Jackson et al., 2021; Duan et al., 2025); other pipelines generate a TS on the fly before barrier regression (Karwounopoulos et al., 2025). Obtaining a reliable TS commonly requires path or saddle-point searches and subsequent vibrational and intrinsic reaction-coordinate checks (Henkelman et al., 2000; Fukui, 1970). With a converged TS and reference reactant energies, the electronic barrier at that level of theory is already determined. Existing benchmarks examine endpoint models and data splits (De Landsheere et al., 2026) or TS generation on unseen chemistry (Darouich et al., 2026).

3 Problem Formulation

We study reaction representation learning with transition-state structures available during training. A reaction comprises reactant structures $R = \{R_i\}$, product structures $P = \{P_j\}$, a transition-state structure T, and a dataset-specific barrier label y. Each molecular structure contains atomic identities and three-dimensional coordinates. The deployable input is $x = (R, P)$; the additional training variable is $z = T$.

The learning goal is an endpoint representation $h_\theta(R, P)$ and a prediction head H_ψ such that $\widehat{y} = H_\psi(h_\theta(R, P))$ estimates the barrier. Pretraining may use (R, P, T, y), while downstream adaptation and inference use only (R, P, y) and (R, P), respectively. This setting permits information from reaction-path structures to influence the learned parameters without requiring a TS for a new prediction.

We evaluate two input views: RP contains the reactant and product endpoints, and RPTS additionally contains the TS. A predictor evaluated with a true TS measures the direct predictive value of that structure. An RP predictor initialized through TS-informed pretraining measures whether learning from that structure remains useful after it is removed. Reaction-barrier prediction provides the empirical task for studying this distinction.

4 DATASETS AND TASKS

We evaluate four datasets with distinct chemistry, reference calculations, and barrier definitions (Table 1). Comparisons are within dataset, not pooled across datasets.

Table 1: Audited benchmark scope. Counts denote usable R/P/TS reaction records in the evaluated view after project-side quality control and, for RGD1, single-channel Core selection. Barrier definitions and TS-validation standards remain dataset specific.

Dataset	Records	Target	Chemical and computational scope
Grambow Unified	11,961	$\Delta E^{\ddagger} + \Delta\mathrm{ZPE}$	Gas-phase CHNO reactions derived from GDB-7; high-level stationary-point refinement (Grambow et al., 2020b).
RGD1 v9 Core	93,815	$\Delta E^{\ddagger}$	CHNO model reactions generated with the Yet Another Reaction Program (YARP), an automated reaction-enumeration and transition-state-search workflow; electronic forward barriers from the corrected v9 release (Zhao et al., 2023b;a).
DORTS	8,142	$\Delta E^{\ddagger}_{\mathrm{SP}}$	Template-derived organic reaction paths; high-level single-point energy labels on validated low-level paths (Ren et al., 2026; Ren, 2025).
Cyclo-23-TS	5,249	$\Delta G^{\ddagger}$	Uncatalyzed $[3+2]$ cycloadditions in implicit water with solution-phase activation free energies (Stuyver et al., 2023).

RGD1. RGD1 contains CHNO model reactions with up to ten heavy atoms, constructed by truncating PubChem-derived reaction candidates around their reactive centers. YARP combines conformer sampling with GFN2-xTB growing-string searches, Berny TS optimization, single-imaginary-frequency screening, and intrinsic reaction-coordinate (IRC) calculations, followed by B3LYP-D3/TZVP refinement. High-level reaction assignment uses a geometry- and imaginary-mode-based classifier rather than full DFT IRC validation, and reactant and product components are optimized separately at the DFT level (Zhao et al., 2023b). The source paper reports 176,992 TS/channel records spanning 126,857 distinct reactions; our audited snapshot contains 176,898 label-complete records. The evaluated Core view retains 93,815 single-channel records with their original split assignments. We predict the released forward electronic barrier, $\mathrm{DE_F} = E_{\mathrm{TS}} - \sum_i E_{R_i}$, with audited reaction directions. ZPE and thermal corrections are not included in this target. We use the corrected Figshare version-9 release (Zhao et al., 2023a).

DORTS. DORTS was generated from 255 organic reaction templates spanning ten elements. Scaffold-based molecular design and GENiniTS-RS provide reaction candidates and template-based TS guesses, followed by GFN2-xTB TS optimization and IRC validation rather than GSM or NEB path searches. Normal-mode sampling expands the paths; the released DORTS-9K subset contains 9,017 reactions annotated with ωB97M-V/def2-TZVP single-point energies and forces (Ren et al., 2026). Our audited view retains 8,142 complete R/P/TS records and predicts $E^{\mathrm{SP}}_{\mathrm{TS}} - E^{\mathrm{SP}}_{R}$, without ZPE or thermal corrections. Because the TS geometries were optimized at GFN2-xTB rather than reoptimized as DFT saddle points, we treat the target as an *electronic single-point barrier proxy*.

Grambow Unified. The Grambow collection covers gas-phase CHNO reactions with unimolecular reactants sampled from GDB-7 (up to seven heavy atoms) and potentially multicomponent products. Single-ended growing-string searches propose pathways at B97-D3/def2-mSVP; R/P/TS stationary points are then reoptimized at ωB97X-D3/def2-TZVP with harmonic vibrational analysis. TS validation requires one imaginary frequency and checks that its normal mode matches the reaction's bond changes; IRC validation is not reported. Unlike the RGD1 electronic target, the released activation energy includes harmonic ZPE, $E_a = (E_{\mathrm{TS}} + \mathrm{ZPE}_{\mathrm{TS}}) - (E_R + \mathrm{ZPE}_R)$ (Grambow et al., 2020b). Our Unified view retains all 11,961 complete R/P/TS records from the original high-level release.

Cyclo-23-TS. Cyclo-23-TS contains 5,269 computational profiles for uncatalyzed $[3+2]$ cycloadditions (Stuyver et al., 2023). Reactant, product, and TS geometries and frequencies were computed at B3LYP-D3(BJ)/def2-SVP, followed by B3LYP-D3(BJ)/def2-TZVP single-point energies using the Solvation Model based on Density (SMD) with water at 298.15 K and a 1 M standard state. Its target is the activation free energy $\Delta G^{\ddagger} = G_{\mathrm{TS}} - \sum_i G_{R_i}$, where each species free energy combines the final single-point energy with the released thermal correction. We follow the source's alternative-reactant selection and retain an audited conservative view of 5,249 profiles after treating 20 product-connectivity cases as traceable QC exclusions. We do not claim an independent frequency or IRC revalidation of the source calculations.

Label fidelity and benchmark preparation. Because reference states, corrections, and TS-validation standards differ, absolute errors across datasets do not rank chemical difficulty or label accuracy. Preparation and label-fidelity audits are in Appendix A.

5 Benchmarking Reaction Representation Learning

5.1 Prediction tasks and evaluation metrics

The benchmark quantifies two complementary roles of TS information: its direct predictive value when the TS is added to reactant and product inputs, and its transferable value when the TS is available only during pretraining. Direct prediction compares RP and RPTS inputs, whereas transfer evaluation compares RP-pretrained and RPTS-pretrained encoders under RP-only adaptation and inference. All comparisons use matched records within each dataset and retain the corresponding dataset-specific barrier definition from Section 4.

For N test reactions, let $e_i = |\widehat{y}_i - y_i|$. We report five lower-is-better metrics:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} e_i, \qquad \mathrm{MAPE} = \frac{100}{N}\sum_{i=1}^{N} \frac{e_i}{|y_i|}, \tag{1}$$

$$q_\tau = \frac{100}{N}\sum_{i=1}^{N} \mathbf{1}[e_i > \tau], \qquad \tau \in \{3, 5, 10\}\ \mathrm{kcal/mol}. \tag{2}$$

MAE measures absolute error in kcal/mol and is the primary main-text metric. MAPE measures relative error as a percentage and is sensitive to labels near zero. The three q_τ values report the percentage of reactions exceeding each absolute-error tolerance, exposing tail behavior that an average alone can obscure. We choose thresholds of 3, 5, and 10 kcal/mol to provide chemically interpretable error scales. The 3 kcal/mol threshold is comparable to the barrier for internal rotation in ethane, approximately 5 kcal/mol is characteristic of an ordinary hydrogen bond, and 10 kcal/mol represents the energy scale of strong hydrogen bonds (Pitzer, 1983; Perrin & Nielson, 1997). These values serve only as reference scales for prediction errors and do not imply that each error corresponds to a specific physical interaction.The implementation uses the original labels without an epsilon in MAPE. Complete five-metric results accompany the main MAE comparisons in Appendix B.

5.2 Data partitions

Grouped-random partitions. The primary evaluation uses the same 80/10/10 partitions constructed over dataset-specific endpoint-pair groups. Within each dataset, an endpoint-pair group collects all records sharing the same reactant and product identities. Assigning the entire group to one partition prevents alternative transition states, conformers, or duplicate channels of the same endpoint transformation from leaking across training, validation, and test sets. This strategy does not enforce disjoint reactant molecules, scaffolds, or reaction families.

Depth-1 reaction-family partition. To evaluate transfer to unseen local reaction families, we additionally define a depth-1 endpoint-edit-family split. The family key includes atoms and bonds that change between mapped reactants and products, their immediate neighbors in either endpoint, and the connecting boundary bonds. Reactions connected by a shared family or same endpoint pair are assigned together, preventing family and endpoint-pair crossings. The construction is audited on all four dataset sources, while predictive evaluation is currently reported only for Grambow Unified.

Appendix A provides dataset-specific grouping rules, split counts, leakage checks, and machine-readable assignments.

6 A Simple TS-Informed Pretraining Strategy

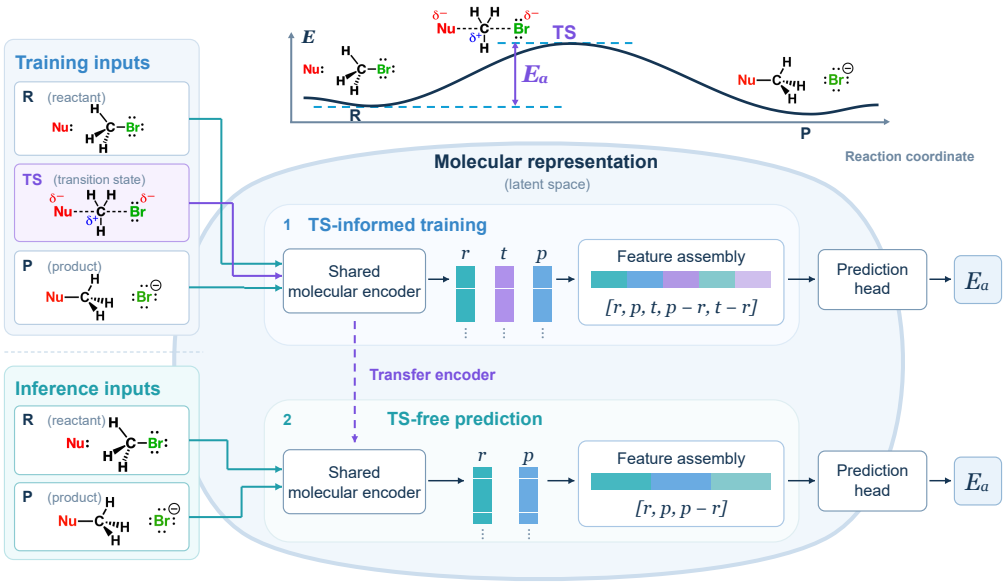


Figure 2: Overview of TS-informed training and TS-free prediction. During pretraining, a shared molecular encoder maps reactant, product, and transition-state structures into the RPTS reaction representation. The pretrained encoder is then transferred to an endpoint-only RP predictor; its adaptation and inference stages use reactant and product structures but no TS input.

6.1 Shared Molecular Encoding

Let E_θ map each molecular graph to a fixed-dimensional molecular embedding using a permutation-invariant readout over its atoms. At the reaction level, we sum the embeddings of all molecular components within each reaction side:

$$r = \sum_{R_i \in R} E_\theta(R_i), \qquad p = \sum_{P_j \in P} E_\theta(P_j), \qquad t = E_\theta(T). \tag{3}$$

The same encoder parameters are used for every state. This shared interface is the only architectural requirement of the method and allows the same chemical feature extractor to be reused after the TS input is removed.

6.2 RPTS Pretraining

The pretraining representation combines absolute state embeddings with endpoint- and TS-relative differences:

$$q_{\mathrm{RPTS}} = [r, p, t, p - r, t - r]. \tag{4}$$

A prediction head H_ϕ is trained with the dataset's ground-truth barrier labels,

$$(\theta^\star, \phi^\star) = \arg\min_{\theta,\phi} \sum_n \mathcal{L}\Big(H_\phi(q^{(n)}_{\mathrm{RPTS}}), y^{(n)}\Big). \tag{5}$$

This objective is task-aligned supervised pretraining: TS geometry is coupled to the downstream barrier target rather than introduced through an auxiliary reconstruction loss.

6.3 ENDPOINT-ONLY ADAPTATION

RPTS pretraining produces an encoder $E_{\theta^\star}$ and an RPTS prediction head $H_{\phi^\star}$. We discard $H_{\phi^\star}$ and use $E_{\theta^\star}$ to replace the encoder of a trained, same-seed RP parent, while retaining the parents RP head $H_{\psi^{\mathrm{RP}}}$ and other non-encoder parameters. The resulting endpoint-only model is initialized as

$$(\theta_0^{\mathrm{down}}, \psi_0^{\mathrm{down}}) = (\theta^\star, \psi^{\mathrm{RP}})$$

and jointly fine-tuned using

$$q_{\mathrm{RP}} = [r, p, p - r], \qquad \widehat{y} = H_\psi(q_{\mathrm{RP}}).$$

The full endpoint model, including the inherited RP head and the transferred encoder, is fine-tuned. We refer to this pipeline as **RPTS-pretrained**. The matched **RP-pretrained** control continues training from the complete RP parent without replacing its encoder. Both branches begin with identical non-encoder parameters and use the same downstream architecture, split, seeds, and adaptation budget. Their comparison tests the benefit of RPTS rather than RP encoder initialization under this shared adaptation protocol.

7 EXPERIMENTS

7.1 EXPERIMENTAL SETUP

Evaluation of TS information value and transfer. We test how much TS structures improve direct barrier prediction and whether encoder pretraining transfers this benefit to endpoint-only prediction across datasets and encoder architectures. Across four datasets, we evaluate SchNet-1M, PaiNN, DimeNet++, GemNet-T, ClofNet, and MACE using the shared reaction-state aggregation interface in Section 6. Directly trained RP and RPTS predictors measure the value of TS access for each dataset–encoder pair. To evaluate transfer, RP-pretrained continues the complete trained RP parent, while RPTS-pretrained replaces only its encoder with the same-seed RPTS encoder before adaptation. Both descendants share the RP architecture and initial non-encoder parameters and predict from endpoints alone. Direct RPTS prediction uses the true TS at inference.

Reaction-family generalization. We test whether the benefit of TS-informed pretraining persists for unseen local reaction families, beyond the grouped-random setting. We evaluate Grambow Unified on the depth-1 endpoint-edit-family split in Section 5, which separates reaction centers and their first-shell local environments across partitions. The split contains 9,569 training, 1,196 validation, and 1,196 test reactions. We repeat the RP, RPTS, RP-pretrained, and RPTS-pretrained comparisons for the six encoders in the primary matrix: SchNet-1M, PaiNN, DimeNet++, GemNet-T, ClofNet, and MACE. Although the shared family construction has been audited for all four dataset source populations, predictive generalization is evaluated only on Grambow.

Generalization across Pretraining Strategies. We test whether including TS structures also benefits other representation pretraining strategies, rather than only our barrier-supervised strategy. On Grambow Unified with ClofNet fixed, we pair RP and RPTS pretraining under CoordDenoise, GeoSSL-DDM, EMPP-l1, C-FREE-Norm, and Energy-Referenced. The first four use structural self-supervision; Energy-Referenced uses state-energy targets after training-only elemental-reference subtraction. Both branches within each objective use the same training reactions. We discard the pretext head and transfer the encoder into an RP predictor whose other parameters come from the same-seed RP parent. Table 5 compares RP and RPTS within each objective; adaptations and training budgets are specified in Appendix C.3.

Comparison with Established Barrier-Prediction Methods. We assess whether our encoder-pretraining strategy is competitive with established endpoint-only barrier predictors, beyond improvements over its own RP control. The Grambow comparison includes MFP and DRFP with random forests, $\mathrm{B}^2\mathrm{R}_\ell^2$ with kernel ridge regression, RXNFP, Graph2Ea/EGAT, and EquiReact(our implementation) (Section 2). All six methods use endpoint information without TS generation and are evaluated against the same splits. Training settings, checkpoint selection, and statistical reporting for the experiments are provided in Appendix C.

Table 2: Direct TS information across four datasets. RP and RPTS are three-seed mean test MAEs in kcal/mol, and $\Delta_{\mathrm{TS}} = M(\mathrm{RP}) - M(\mathrm{RPTS})$. Positive differences indicate lower error with direct TS access; RPTS uses a true TS at inference.

	Grambow Unified			RGD1 Core			DORTS			Cyclo-23-TS		
Encoder	RP	RPTS	Δ_{TS}	RP	RPTS	Δ_{TS}	RP	RPTS	Δ_{TS}	RP	RPTS	Δ_{TS}
SchNet-1M	11.698	4.287	7.411	11.084	4.575	6.509	5.208	3.865	1.343	2.827	2.281	0.546
PaiNN	10.826	3.588	7.238	11.777	4.460	7.317	5.648	3.889	1.760	2.706	2.116	0.590
DimeNet++	11.541	3.613	7.928	10.863	4.349	6.514	5.299	3.425	1.874	3.289	1.808	1.481
GemNet-T	11.943	3.708	8.235	11.301	4.185	7.116	5.093	3.565	1.528	2.539	1.768	0.771
ClofNet	8.591	3.477	5.115	9.657	4.115	5.541	4.776	3.424	1.352	2.760	1.963	0.797
MACE	9.181	3.172	6.009	10.243	4.141	6.103	4.965	3.245	1.721	2.395	1.338	1.057

Table 3: Endpoint-only adaptation across four datasets. RP-pretrained and RPTS-pretrained are three-seed mean test MAEs in kcal/mol, and $\Delta_{\mathrm{transfer}} = M(\mathrm{RP\text{-}pretrained}) - M(\mathrm{RPTS\text{-}pretrained})$. Positive differences indicate lower error after TS-informed pretraining; both predictors use only R/P at inference.

	Grambow Unified			RGD1 Core			DORTS			Cyclo-23-TS		
Encoder	RP-pre.	RPTS-pre.	$\Delta_{\mathrm{transfer}}$	RP-pre.	RPTS-pre.	$\Delta_{\mathrm{transfer}}$	RP-pre.	RPTS-pre.	$\Delta_{\mathrm{transfer}}$	RP-pre.	RPTS-pre.	$\Delta_{\mathrm{transfer}}$
SchNet-1M	11.569	10.674	0.894	11.083	10.501	0.582	5.053	4.662	0.391	2.813	2.693	0.120
PaiNN	10.703	9.875	0.828	11.815	11.395	0.420	5.612	5.308	0.304	2.674	2.561	0.113
DimeNet++	11.545	11.510	0.035	10.840	10.303	0.537	5.292	5.222	0.070	3.151	2.553	0.598
GemNet-T	11.758	11.177	0.581	11.285	10.822	0.463	5.070	4.869	0.201	2.575	2.425	0.150
ClofNet	8.614	8.050	0.564	9.634	9.509	0.125	4.737	4.427	0.310	2.693	2.663	0.030
MACE	9.073	8.443	0.629	10.353	10.047	0.306	4.941	4.583	0.358	2.384	2.312	0.073

7.2 Results and analysis

Observation 1: Direct TS access provides substantial predictive value. Table 2 compares RP and RPTS prediction across all 24 dataset–encoder pairs. Providing the true TS at inference lowers mean MAE in every case, showing that TS structures contain predictive information not fully captured by reactant and product endpoints alone.

Observation 2: TS-informed pretraining transfers part of this value to endpoint-only prediction. Table 3 shows that RPTS-pretrained models lower endpoint-only mean MAE in all 24 matched comparison. These transferred gains are consistently smaller than the corresponding direct-TS gains and are not statistically supported for every dataset–encoder pair. Beyond MAE, Appendix B reports the complete metric-level results and uncertainty estimates.

Table 4: Grambow mean MAE gains (kcal/mol) on grouped-random and unseen depth-1 reaction-family splits. Direct TS gain is RP minus RPTS; transfer gain is RP-pretrained minus RPTS-pretrained. Positive values favor TS use. The splits have different train/test assignments, so differences between their columns do not isolate an OOD penalty.

	Grouped-random transfer gain	Unseen-family direct TS gain	Unseen-family transfer gain
SchNet-1M	0.894	7.220	0.457
PaiNN	0.828	7.916	0.692
DimeNet++	0.035	9.099	0.757
GemNet-T	0.581	8.595	0.491
ClofNet	0.564	6.230	0.750
MACE	0.629	6.943	0.068

Observation 3: TS-informed transfer persists on unseen local reaction families in the Grambow stress test. The family key includes the R/P reaction-edit center and its first-shell local environment, so test reactions have unseen depth-1 endpoint-edit families rather than merely unseen endpoint-pair IDs. For all six encoders in Table 4, direct TS access and RPTS pretraining each lower mean MAE on this split. The transfer gain is supported by the hierarchical 95% interval

for SchNet-1M, whereas the smaller MACE gain crosses zero; mean improvement across encoders should therefore not be read as uniform statistical support.

Observation 4: TS exposure also lowers mean MAE under five additional pretraining objectives on Grambow Unified. Table 5 shows within-objective improvements of 0.004–0.160 kcal/mol, compared with 0.564 kcal/mol for the task-aligned ClofNet pair. The C-FREE-Norm change is negligible relative to seed variability, and its mean MAPE and all three tail-error rates worsen. Coord-Denoise and GeoSSL-DDM also have higher mean q_{10}, so the benefit depends on the objective and metric.

Table 5: Generality across pretraining objectives on Grambow Unified. Endpoint-only MAE (kcal/mol; lower is better).

Method	RP pretraining	RPTS pretraining
Task-aligned (ours)	8.614 ± 0.075	$\mathbf{8.050 \pm 0.329}$
CoordDenoise	7.871 ± 0.110	$\mathbf{7.711 \pm 0.171}$
GeoSSL-DDM	7.820 ± 0.181	$\mathbf{7.705 \pm 0.246}$
EMPP-l1	7.792 ± 0.084	$\mathbf{7.753 \pm 0.074}$
C-FREE-Norm	7.982 ± 0.066	$\mathbf{7.978 \pm 0.186}$
Energy-Referenced	7.665 ± 0.134	$\mathbf{7.594 \pm 0.267}$

All methods use ClofNet and RP-only downstream inputs. Entries are mean ± sample SD over three seeds. Bold marks the lower mean within each pair, not statistical significance. All rows inherit the same-seed RP parent outside the encoder. The task-aligned RP-pretraining entry is an RP-parent continuation control.

Table 6: Grambow Unified comparison with established barrier predictors. MAE in kcal/mol; lower is better.

Method	MAE
RPTS-pretrained (ours; ClofNet)	$\mathbf{8.050 \pm 0.329}$
MFP + RF	$\underline{10.870 \pm 0.003}$
DRFP + RF	13.600 ± 0.030
RXNFP	11.287 ± 0.336
$B^2R_\ell^2$ + KRR	$18.557^\dagger$
Graph2Ea / EGAT	13.350 ± 0.323
EquiReact (paper-spec)	12.106 ± 0.186

† One deterministic run; other rows use three seeds. Bold/underline mark the best/second mean in this table only. All methods are evaluated on the same 1,198 Grambow test reactions.

Observation 5: RPTS-pretrained ClofNet has the lowest mean MAE among the six external barrier predictors evaluated on Grambow Unified. In Table 6, it reaches 8.050 ± 0.329 kcal/mol, compared with 10.870 ± 0.003 kcal/mol for the strongest external baseline, MFP with a random forest. This is a comparison within that table, not a general state-of-the-art claim: Energy-Referenced RPTS in Table 5 reaches 7.594 ± 0.267 kcal/mol under a distinct adaptation protocol, and TS-generation pipelines are not evaluated.

8 Limitations and Conclusion

Limitations. The benchmark evaluates reaction representations through barrier prediction on four small-molecule datasets. It does not cover explicit homogeneous catalysts, heterogeneous surfaces, or biomacromolecular reactions in protein environments. The evidence therefore does not establish transfer to those systems or to other reaction-learning tasks. TS acquisition remains a training-data cost, and deployment assumes three-dimensional endpoint structures. Electronic, zero-point-corrected, and free-energy barriers retain their dataset-specific meanings. Predictive evaluation on unseen depth-1 reaction families is currently limited to Grambow Unified.

Positive matched gains demonstrate useful changes in the learned representation without identifying the specific geometric or electronic features transferred.

Conclusion. We introduced TSIRL, a benchmark for separating the direct predictive value of TS structures from benefits retained in endpoint-only reaction representations. Encoder pretraining improves mean endpoint-only MAE across all four datasets and six encoders, while additional pretraining objectives show objective- and metric-dependent benefits on Grambow Unified. The mean transfer gain also remains positive for all six encoders on unseen depth-1 reaction families, although its statistical support varies by encoder. These findings support training-time TS structures as a resource for reaction representation learning, evaluated here through barrier prediction.

AI USE STATEMENT

We used generative AI tools to assist with drafting and editing the manuscript, improving the clarity of the presentation, and preparing the LaTeX typesetting and layout. We also used generative AI tools to review research code and identify potential implementation issues or inconsistencies.

We reviewed all AI-assisted text, formatting, and code suggestions before use. Generative AI models did not directly generate any experimental data or reported numerical results. We take responsibility for the final content of the manuscript and for all code, analyses, and results reported in this work.

REFERENCES

Junyi An, Chao Qu, Yun-Fei Shi, XinHao Liu, Qianwei Tang, Fenglei Cao, and Yuan Qi. Equivariant masked position prediction for efficient molecular representation. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/7ab7073a147f0a4ee5c76995800d8f14-Abstract-Conference.html.

Boshra Ariguib, Mathias Niepert, and Andrei Manolache. Learning the neighborhood: Contrast-free multimodal self-supervised molecular graph pretraining. In *International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=Vbj1jSaqTo.

Ilyes Batatia, Dávid Péter Kovács, Gregor N. C. Simm, Christoph Ortner, and Gábor Csányi. MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. In *Advances in Neural Information Processing Systems*, volume 35, pp. 11423–11436, 2022. URL https://papers.neurips.cc/paper_files/paper/2022/hash/4a36c3c51af11ed9f34615b81edb5bbc-Abstract-Conference.html.

Samir Darouich, Jacob W. Toney, Weiliang Luo, Johannes Kästner, Mathias Niepert, and Heather J. Kulik. Robust generative transition-state models for unseen chemistry. *Nature Computational Science*, 2026. doi: 10.1038/s43588-026-01034-5.

Jasper De Landsheere, Anton Zamyatin, Johannes Karwounopoulos, and Esther Heid. ChemTorch: A deep learning framework for benchmarking and developing chemical reaction property prediction models. *Journal of Chemical Information and Modeling*, 66(5):2434–2442, 2026. doi: 10.1021/acs.jcim.5c02645.

Weitao Du, He Zhang, Yuanqi Du, Qi Meng, Wei Chen, Nanning Zheng, Bin Shao, and Tie-Yan Liu. SE(3) equivariant graph neural networks with complete local frames. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5583–5608, 2022. URL https://proceedings.mlr.press/v162/du22e.html.

Weitao Du, Yuanqi Du, Limei Wang, Dieqiao Feng, Guifeng Wang, Shuiwang Ji, Carla Gomes, and Zhi-Ming Ma. A new perspective on building efficient and expressive 3D equivariant graph neural networks. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2304.04757.

Chenru Duan, Guan-Horng Liu, Yuanqi Du, Tianrong Chen, Qiyuan Zhao, Haojun Jia, Carla P. Gomes, Evangelos A. Theodorou, and Heather J. Kulik. Optimal transport for generating transition states in chemical reactions. *Nature Machine Intelligence*, 7(4):615–626, 2025. doi: 10.1038/s42256-025-01010-0.

David Duvenaud, Dougal Maclaurin, Jorge Aguilera-Iparraguirre, Rafael Gómez-Bombarelli, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P. Adams. Convolutional networks on graphs for learning molecular fingerprints. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL https://papers.nips.cc/paper_files/paper/2015/hash/f9be311e65d81a9ad8150a60844bb94c-Abstract.html.

Kenichi Fukui. Formulation of the reaction coordinate. *The Journal of Physical Chemistry*, 74(23):4161–4163, 1970. doi: 10.1021/j100717a029.

Simone Gallarati, Raimon Fabregat, Rubén Laplaza, Sinjini Bhattacharjee, Matthew D. Wodrich, and Clemence Corminboeuf. Reaction-based machine learning representations for predicting the enantioselectivity of organocatalysts. *Chemical Science*, 12(20):6879–6889, 2021. doi: 10.1039/D1SC00482D.

Johannes Gasteiger, Shankari Giri, Johannes T. Margraf, and Stephan Günnemann. Fast and uncertainty-aware directional message passing for non-equilibrium molecules. In *Machine Learning for Molecules Workshop at NeurIPS*, 2020. URL https://arxiv.org/abs/2011.14115.

Johannes Gasteiger, Florian Becker, and Stephan Günnemann. GemNet: Universal directional graph neural networks for molecules. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://arxiv.org/abs/2106.08903.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1263–1272, 2017. URL https://proceedings.mlr.press/v70/gilmer17a.html.

Colin A. Grambow, Lagnajit Pattanaik, and William H. Green. Deep learning of activation energies. *The Journal of Physical Chemistry Letters*, 11(8):2992–2997, 2020a. doi: 10.1021/acs.jpclett.0c00500.

Colin A. Grambow, Lagnajit Pattanaik, and William H. Green. Reactants, products, and transition states of elementary chemical reactions based on quantum chemistry. *Scientific Data*, 7(1):137, 2020b. doi: 10.1038/s41597-020-0460-4.

Esther Heid and William H. Green. Machine learning of reaction properties via learned representations of the condensed graph of reaction. *Journal of Chemical Information and Modeling*, 62(9):2101–2110, 2022. doi: 10.1021/acs.jcim.1c00975.

Stefan Heinen, Guido Falk von Rudorff, and O. Anatole von Lilienfeld. Toward the design of chemical reactions: Machine learning barriers of competing mechanisms in reactant space. *The Journal of Chemical Physics*, 155(6):064105, 2021. doi: 10.1063/5.0059742.

Graeme Henkelman, Blas P. Uberuaga, and Hannes Jónsson. A climbing image nudged elastic band method for finding saddle points and minimum energy paths. *The Journal of Chemical Physics*, 113(22):9901–9904, 2000. doi: 10.1063/1.1329672.

Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1905.12265.

Riley Jackson, Wenyuan Zhang, and Jason Pearson. TSNet: Predicting transition state structures with tensor field networks and transfer learning. *Chemical Science*, 12(29):10022–10040, 2021. doi: 10.1039/D1SC01206A.

Johannes Karwounopoulos, Jasper De Landsheere, Leonard Galustian, Tobias Jechtl, and Esther Heid. Graph-based prediction of reaction barrier heights with on-the-fly prediction of transition states. *Digital Discovery*, 4(11):3208–3216, 2025. doi: 10.1039/D5DD00240K.

Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. Pre-training molecular graph representation with 3D geometry. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=xQUe1pOKPam.

Shengchao Liu, Hongyu Guo, and Jian Tang. Molecular geometry pretraining with SE(3)-invariant denoising distance matching. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=CjTHVo1dvR.

Charles L. Perrin and Jennifer B. Nielson. "Strong" Hydrogen Bonds in Chemistry and Biology. *Annual Review of Physical Chemistry*, 48:511–544, 1997. doi: 10.1146/annurev.physchem.48.1.511.

Russell M. Pitzer. The barrier to internal rotation in ethane. *Accounts of Chemical Research*, 16(6): 207–210, 1983. doi: 10.1021/ar00090a004.

Daniel Probst, Philippe Schwaller, and Jean-Louis Reymond. Reaction classification and yield prediction using the differential reaction fingerprint DRFP. *Digital Discovery*, 1:91–97, 2022. doi: 10.1039/D1DD00006C.

Kaipai Ren. Data. Zenodo, 2025. URL https://doi.org/10.5281/zenodo.17141108. DORTS dataset, published 17 September 2025.

Kaipai Ren, Kun Tang, Yujing Zhao, Lei Zhang, Jian Du, Qingwei Meng, and Qilei Liu. Reactive machine learning potential for accelerating transition state search in organic synthesis. *Nature Communications*, 17(1):6253, 2026. doi: 10.1038/s41467-026-72945-0.

Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper_files/paper/2017/hash/303ed4c69846ab36c2904d3ba8573050-Abstract.html.

Kristof Schütt, Oliver Unke, and Michael Gastegger. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9377–9388, 2021. URL https://proceedings.mlr.press/v139/schutt21a.html.

Philippe Schwaller, Daniel Probst, Alain C. Vaucher, Vishnu H. Nair, David Kreutter, Teodoro Laino, and Jean-Louis Reymond. Mapping the space of chemical reactions using attention-based neural networks. *Nature Machine Intelligence*, 3:144–152, 2021. doi: 10.1038/s42256-020-00284-w.

Hannes Stärk, Dominique Beaini, Gabriele Corso, Prudencio Tossou, Christian Dallago, Stephan Günnemann, and Pietro Lió. 3D Infomax improves GNNs for molecular property prediction. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 20479–20502, 2022. URL https://proceedings.mlr.press/v162/stark22a.html.

Thijs Stuyver and Connor W. Coley. Quantum chemistry-augmented neural networks for reactivity prediction: Performance, generalizability, and explainability. *The Journal of Chemical Physics*, 156(8):084104, 2022. doi: 10.1063/5.0079574.

Thijs Stuyver, Kjell Jorner, and Connor W. Coley. Reaction profiles for quantum chemistry-computed [3 + 2] cycloaddition reactions. *Scientific Data*, 10:66, 2023. doi: 10.1038/s41597-023-01977-8.

Sai Mahit Vadaddi, Qiyuan Zhao, and Brett M. Savoie. Graph to activation energy models easily reach irreducible errors but show limited transferability. *The Journal of Physical Chemistry A*, 128(13), 2024. doi: 10.1021/acs.jpca.3c07240.

Puck van Gerwen, Ksenia R. Briling, Charlotte Bunne, Vignesh Ram Somnath, Ruben Laplaza, Andreas Krause, and Clemence Corminboeuf. 3DReact: Geometric deep learning for chemical reactions. *Journal of Chemical Information and Modeling*, 64(15):5771–5785, 2024. doi: 10.1021/acs.jcim.4c00104.

Puck Van Gerwen, Ksenia R. Briling, Yannick Calvino Alonso, Malte Franke, and Clemence Corminboeuf. Benchmarking machine-readable vectors of chemical reactions on computed activation barriers. *Digital Discovery*, 3(5):932–943, 2024. doi: 10.1039/D3DD00175J.

Hongwei Wang, Weijiang Li, Xiaomeng Jin, Kyunghyun Cho, Heng Ji, Jiawei Han, and Martin D. Burke. Chemical-reaction-aware molecule representation learning. In *International Conference on Learning Representations*, 2022a. URL https://blender.cs.illinois.edu/paper/moleculerepresentation2022.pdf.

Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3):279–287, 2022b. doi: 10.1038/s42256-022-00447-x.

Zhaoping Xiong, Dingyan Wang, Xiaohong Liu, Feisheng Zhong, Xiaozhe Wan, Xutong Li, Zhaojun Li, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, and Mingyue Zheng. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of Medicinal Chemistry*, 63(16):8749–8760, 2020. doi: 10.1021/acs.jmedchem.9b00959.

Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, Andrew Palmer, Volker Settels, Tommi Jaakkola, Klavs Jensen, and Regina Barzilay. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59(8):3370–3388, 2019. doi: 10.1021/acs.jcim.9b00237.

Sheheryar Zaidi, Michael Schaarschmidt, James Martens, Hyunjik Kim, Yee Whye Teh, Alvaro Sanchez-Gonzalez, Peter Battaglia, Razvan Pascanu, and Jonathan Godwin. Pre-training via denoising for molecular property prediction. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=tYIMtogyee.

Qiyuan Zhao, Brett Savoie, Michael Woulfe, Sai Mahit Vaddadi, Lawal A. Ogunfowora, and Sanjay Garimella. RGD1-CNHO Database. figshare, version 9, 2023a. URL https://doi.org/10.6084/m9.figshare.21066901.v9. Dataset, released 26 November 2023.

Qiyuan Zhao, Sai Mahit Vaddadi, Michael Woulfe, Lawal A. Ogunfowora, Sanjay S. Garimella, Olexandr Isayev, and Brett M. Savoie. Comprehensive exploration of graphically defined reaction spaces. *Scientific Data*, 10(1):145, 2023b. doi: 10.1038/s41597-023-02043-z.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-Mol: A universal 3D molecular representation learning framework. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6K2RM6wVqKu.

Yanqiao Zhu, Jeehyun Hwang, Keir Adams, Zhen Liu, Bozhao Nan, Brock Stenfors, Yuanqi Du, Jatin Chauhan, Olaf Wiest, Olexandr Isayev, Connor Coley, Yizhou Sun, and Wei Wang. Learning over molecular conformer ensembles: Datasets and benchmarks. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/7eb6233e02f7d9efbb84acd839a996fb-Abstract-Conference.html.

A Dataset Descriptions and Audits

The four datasets play the same benchmark role but differ in chemistry, reference state, TS validation, and barrier definition. Consequently, all model comparisons and paired differences are computed within a dataset. The counts below refer to the audited views used by this work rather than every record in the corresponding source release.

A.1 Grambow Unified

Grambow Unified contains gas-phase CHNO reactions derived from GDB-7. Single-ended growing-string searches propose pathways, followed by ωB97X-D3/def2-TZVP stationary-point refinement and harmonic analysis. The retained target includes the released harmonic zero-point-energy correction. The benchmark view contains 11,961 complete R/P/TS records (Grambow et al., 2020b).

A.2 RGD1 v9 Core

RGD1 contains CHNO model reactions constructed around reactive centers. Its YARP workflow combines conformer sampling, low-level pathway and IRC checks, and B3LYP-D3/TZVP refinement. The corrected v9 manifest contains 176,898 label-complete channel records. Our experiments use its 93,815-record single-channel Core subset without reassigning its frozen splits. The target is the forward electronic barrier (Zhao et al., 2023b;a).

A.3 DORTS

DORTS begins from 255 organic reaction templates and uses low-level TS optimization and IRC validation before high-level single-point annotation. Our target is the released TS-minus-reactant electronic single-point energy, without a ZPE or thermal correction. After completeness and neutral closed-shell endpoint audits, the benchmark view contains 8,142 complete R/P/TS records. We construct reaction-level grouped partitions rather than reuse the source conformer allocation (Ren et al., 2026; Ren, 2025).

A.4 Cyclo-23-TS

Cyclo-23-TS contains solution-phase $[3 + 2]$ cycloaddition profiles. The target is the published activation free energy relative to separated reactants at 298.15 K and a 1 M standard state. Following the source alternative-reactant rule and a conservative connectivity and label-consistency audit, the benchmark view contains 5,249 records (Stuyver et al., 2023). The fixed record list and audit implementation will accompany the code release.

A.5 Energy-label distributions

Figure 3 shows the paired barrier and reaction- energy distributions used to audit the four sources. The axes retain the dataset-specific energy definitions; the panels should not be interpreted as measurements at a common electronic-structure level.

A.6 Grouped splits and leakage control

The nominal train/validation/test allocation is 80/10/10. Table 7 reports the actual evaluated populations, not rounded percentages. A direct recount of the frozen manifests finds unique record IDs and no group crossing in any of the four evaluated views.

Dataset-specific grouping and allocation. Grambow and Cyclo-23-TS use directed canonical reactant–product endpoint-pair groups. DORTS uses direction-invariant endpoint-pair groups after removing atom maps and retaining explicit hydrogen and stereochemical information. RGD1 links source reaction identities that share a direction-invariant endpoint pair and allocates the resulting connected groups together. Each group is assigned as a unit to one partition. For RGD1, grouping is performed before the single-channel Core subset is selected. Fixed assignments and the construction scripts will accompany the code release.

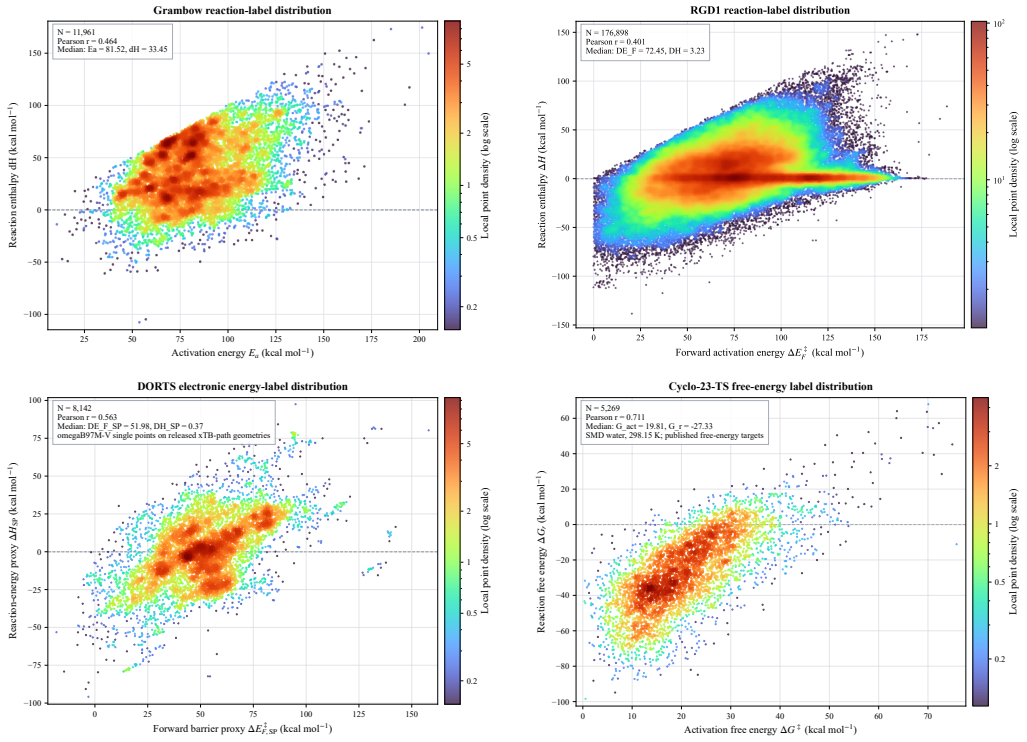


Figure 3: Dataset-specific joint distributions of activation barriers and reaction-energy quantities. From upper left: Grambow Unified, RGD1, DORTS, and Cyclo-23-TS. The RGD1 panel describes the full audited source population, not solely the single-channel Core subset used for prediction.

Table 7: Frozen evaluated partitions and grouping audit. "Crossing" counts groups represented in more than one partition, not chemically related reactions.

Dataset	Train	Validation	Test	Groups	Crossing
Grambow Unified	9,604	1,159	1,198	11,961	0
RGD1 v9 Core	75,053	9,381	9,381	93,815	0
DORTS	6,496	819	827	8,066	0
Cyclo-23-TS	4,182	525	542	5,237	0

Identity and reconstruction materials. The release will provide the fixed reaction identifiers, group assignments, partitions, population filters, and leakage checks needed to reconstruct the evaluated populations. Exact file identifiers and integrity metadata are kept with the machine-readable release rather than reproduced in the paper.

What the leakage checks do and do not establish. No stored endpoint-pair group crosses partitions. This does not establish reactant-molecule, reaction-family, or scaffold disjointness. In particular, the Cyclo audit finds 651 reactant-pair groups spanning partitions despite zero directed endpoint-pair crossings. The split is therefore not an OOD evaluation. Pretraining inputs are restricted to outer-training records, and normalization uses training-only statistics; validation selects checkpoints, not test performance. The separate Grambow depth-1 endpoint-edit-family split assigns 9,569/1,196/1,196 records to train/validation/test, with no family, endpoint-pair, or leakage-component crossing. It supports the Grambow-only unseen-family evaluation in Table 4; the other datasets have not been evaluated under this split.

B COMPLETE RESULTS

The following tables report the complete manuscript-scope benchmark and paired pretraining results for the six encoders evaluated in the main text. RP and RPTS quantify the value of direct TS access. RP-pretrained and RPTS-pretrained share the endpoint-only architecture and inference contract. The intended intervention replaces only the initial encoder; the Cyclo-23-TS DimeNet++ optimization exception is explicitly documented in Appendix C.2. RPTS-pretrained uses R/P/TS during pretraining, not TS-only input. In every table, SchNet denotes the approximately one-million-parameter encoder; the earlier small-SchNet runs are retained separately as a capacity ablation.

Table 8: Grambow Unified benchmark and paired pretraining results over three seeds. MAE is in kcal/mol; all other metrics are percentages; lower is better. RPTS is a non-deployable oracle, whereas RP-pretrained and RPTS-pretrained use only reactants and products at inference. SchNet denotes SchNet-1M.

Encoder	Role	Method	MAE ↓	MAPE ↓	q_3 ↓	q_5 ↓	q_{10} ↓
SchNet	Benchmark	RP	11.698 ± 0.160	14.684 ± 0.129	82.053 ± 2.871	69.839 ± 3.134	45.910 ± 2.628
	Benchmark	RPTS	4.287 ± 0.166	5.215 ± 0.223	43.962 ± 3.294	26.906 ± 2.463	10.462 ± 0.482
	Pretraining	RP-pretrained	11.569 ± 0.294	14.394 ± 0.512	79.104 ± 2.382	67.891 ± 2.451	43.378 ± 2.809
	Pretraining	RPTS-pretrained	10.674 ± 0.171	13.198 ± 0.187	76.127 ± 2.107	62.493 ± 1.822	38.203 ± 1.064
PaiNN	Benchmark	RP	10.826 ± 0.204	13.459 ± 0.344	78.464 ± 1.530	65.748 ± 1.551	40.178 ± 1.383
	Benchmark	RPTS	3.588 ± 0.047	4.378 ± 0.058	39.733 ± 0.871	22.371 ± 0.465	6.427 ± 0.364
	Pretraining	RP-pretrained	10.703 ± 0.262	13.322 ± 0.267	78.742 ± 0.923	65.665 ± 0.960	39.677 ± 0.960
	Pretraining	RPTS-pretrained	9.875 ± 0.072	12.113 ± 0.199	75.598 ± 1.503	61.658 ± 1.466	35.198 ± 1.112
DimeNet++	Benchmark	RP	11.541 ± 0.063	14.244 ± 0.138	81.608 ± 0.394	70.006 ± 0.632	43.489 ± 1.688
	Benchmark	RPTS	3.613 ± 0.237	4.340 ± 0.323	37.257 ± 3.299	20.924 ± 2.339	7.679 ± 0.765
	Pretraining	RP-pretrained	11.545 ± 0.209	14.370 ± 0.362	82.332 ± 0.475	69.950 ± 1.448	44.018 ± 1.187
	Pretraining	RPTS-pretrained	11.510 ± 0.567	14.217 ± 0.630	79.744 ± 2.096	67.724 ± 2.796	43.406 ± 2.005
GemNet-T	Benchmark	RP	11.943 ± 0.240	14.895 ± 0.409	81.775 ± 1.760	69.866 ± 1.753	45.131 ± 0.675
	Benchmark	RPTS	3.708 ± 0.076	4.558 ± 0.122	39.149 ± 1.612	22.983 ± 0.960	8.013 ± 0.083
	Pretraining	RP-pretrained	11.758 ± 0.265	14.660 ± 0.358	81.080 ± 1.133	68.809 ± 1.935	44.380 ± 1.628
	Pretraining	RPTS-pretrained	11.177 ± 0.029	14.042 ± 0.060	80.551 ± 0.963	68.753 ± 0.824	42.988 ± 0.805
ClofNet	Benchmark	RP	8.591 ± 0.070	10.579 ± 0.050	72.426 ± 0.900	56.149 ± 1.089	29.410 ± 0.568
	Benchmark	RPTS	3.477 ± 0.039	4.201 ± 0.061	40.262 ± 0.877	21.425 ± 0.386	5.871 ± 0.592
	Pretraining	RP-pretrained	8.614 ± 0.075	10.565 ± 0.048	71.341 ± 1.112	55.760 ± 0.796	29.577 ± 0.537
	Pretraining	RPTS-pretrained	8.050 ± 0.329	9.870 ± 0.287	68.809 ± 0.757	52.365 ± 2.271	26.739 ± 2.153
MACE	Benchmark	RP	9.181 ± 0.073	11.409 ± 0.107	72.899 ± 0.337	58.097 ± 1.312	32.526 ± 0.811
	Benchmark	RPTS	3.172 ± 0.029	3.867 ± 0.065	36.505 ± 1.598	18.114 ± 1.066	4.814 ± 0.556
	Pretraining	RP-pretrained	9.073 ± 0.245	11.330 ± 0.395	72.593 ± 0.460	57.067 ± 1.451	32.304 ± 1.183
	Pretraining	RPTS-pretrained	8.443 ± 0.306	10.482 ± 0.426	69.254 ± 0.757	54.480 ± 2.038	29.048 ± 2.624

Table 9: Grambow Unified direct TS information and paired TS-aware pretraining gains. For each lower-is-better metric, $\Delta_{\mathrm{TS}} = M(\mathrm{RP}) - M(\mathrm{RPTS})$ and $\Delta_{\mathrm{transfer}} = M(\text{RP-pretrained}) - M(\text{RPTS-pretrained})$. Intervals are 95% paired-record and hierarchical seed–record bootstrap intervals for $\Delta_{\mathrm{transfer}}$ (20,000 draws). SchNet denotes SchNet-1M.

Encoder	Metric	Δ_{TS}	$\Delta_{\mathrm{transfer}}$	Paired 95% CI	Hier. 95% CI
SchNet	MAE	7.411 ± 0.320	0.894 ± 0.333	$[0.585, 1.202]$	$[0.497, 1.300]$
	MAPE	9.469 ± 0.351	1.195 ± 0.499	$[0.810, 1.581]$	$[0.683, 1.799]$
	q_3	38.091 ± 4.086	2.977 ± 2.796	$[1.308, 4.674]$	$[-0.250, 5.815]$
	q_5	42.933 ± 4.592	5.398 ± 2.975	$[3.422, 7.373]$	$[2.003, 8.542]$
	q_{10}	35.448 ± 3.094	5.175 ± 2.208	$[3.339, 7.012]$	$[2.755, 7.902]$
PaiNN	MAE	7.238 ± 0.234	0.828 ± 0.326	$[0.623, 1.037]$	$[0.511, 1.207]$
	MAPE	9.082 ± 0.386	1.209 ± 0.466	$[0.923, 1.503]$	$[0.746, 1.733]$
	q_3	38.731 ± 2.223	3.144 ± 2.161	$[1.475, 4.869]$	$[0.723, 5.899]$
	q_5	43.378 ± 1.139	4.007 ± 1.343	$[2.198, 5.815]$	$[1.864, 6.093]$
	q_{10}	33.751 ± 1.499	4.480 ± 0.869	$[2.977, 5.982]$	$[2.866, 6.093]$
DimeNet++	MAE	7.928 ± 0.294	0.035 ± 0.497	$[-0.260, 0.327]$	$[-0.536, 0.529]$
	MAPE	9.903 ± 0.441	0.153 ± 0.378	$[-0.244, 0.544]$	$[-0.365, 0.650]$
	q_3	44.352 ± 3.225	2.588 ± 1.920	$[0.835, 4.313]$	$[0.139, 5.064]$
	q_5	49.082 ± 2.968	2.226 ± 2.952	$[0.306, 4.146]$	$[-0.835, 5.732]$
	q_{10}	35.810 ± 2.087	0.612 ± 2.132	$[-1.252, 2.504]$	$[-1.920, 3.339]$
GemNet-T	MAE	8.235 ± 0.316	0.581 ± 0.287	$[0.245, 0.921]$	$[0.205, 0.964]$
	MAPE	10.338 ± 0.526	0.618 ± 0.362	$[0.159, 1.075]$	$[0.138, 1.120]$

Continued on next page

Encoder	Metric	Δ_{TS}	$\Delta_{\mathrm{transfer}}$	Paired 95% CI	Hier. 95% CI
	q_3	42.627 ± 2.966	0.529 ± 1.670	$[-1.280, 2.337]$	$[-1.753, 2.782]$
	q_5	46.884 ± 2.367	0.056 ± 2.267	$[-2.059, 2.142]$	$[-2.643, 3.033]$
	q_{10}	37.117 ± 0.748	1.391 ± 2.279	$[-0.668, 3.506]$	$[-1.280, 4.313]$
ClofNet	MAE	5.115 ± 0.106	0.564 ± 0.361	$[0.342, 0.788]$	$[0.196, 0.961]$
	MAPE	6.378 ± 0.095	0.694 ± 0.304	$[0.408, 0.982]$	$[0.323, 1.079]$
	q_3	32.165 ± 1.672	2.532 ± 1.856	$[0.612, 4.452]$	$[0.111, 5.092]$
	q_5	34.725 ± 1.448	3.395 ± 2.280	$[1.419, 5.370]$	$[0.668, 6.288]$
	q_{10}	23.539 ± 0.083	2.838 ± 1.881	$[1.224, 4.480]$	$[0.473, 4.953]$
MACE	MAE	6.009 ± 0.052	0.629 ± 0.226	$[0.370, 0.893]$	$[0.319, 0.955]$
	MAPE	7.542 ± 0.051	0.847 ± 0.427	$[0.504, 1.195]$	$[0.369, 1.370]$
	q_3	36.394 ± 1.593	3.339 ± 0.521	$[1.419, 5.232]$	$[1.363, 5.315]$
	q_5	39.983 ± 0.742	2.588 ± 1.881	$[0.556, 4.591]$	$[0.028, 5.259]$
	q_{10}	27.713 ± 0.383	3.255 ± 1.559	$[1.503, 4.981]$	$[1.029, 5.398]$

Table 10: DORTS benchmark and paired pretraining results over three seeds. MAE is in kcal/mol; all other metrics are percentages; lower is better. RPTS is a non-deployable oracle, whereas RP-pretrained and RPTS-pretrained use only reactants and products at inference. SchNet denotes SchNet-1M.

Encoder	Role	Method	MAE ↓	MAPE ↓	q_3 ↓	q_5 ↓	q_{10} ↓
SchNet	Benchmark	RP	5.208 ± 0.029	32.347 ± 5.374	52.116 ± 0.736	33.817 ± 0.985	12.414 ± 0.458
	Benchmark	RPTS	3.865 ± 0.098	24.088 ± 4.264	41.677 ± 1.478	22.975 ± 1.280	6.771 ± 0.527
	Pretraining	RP-pretrained	5.053 ± 0.059	27.519 ± 3.065	50.423 ± 0.605	32.205 ± 1.143	12.092 ± 0.484
	Pretraining	RPTS-pretrained	4.662 ± 0.044	25.236 ± 0.940	46.836 ± 1.401	29.061 ± 1.097	10.842 ± 1.056
PaiNN	Benchmark	RP	5.648 ± 0.214	29.806 ± 4.430	55.905 ± 1.187	39.097 ± 2.491	14.672 ± 1.124
	Benchmark	RPTS	3.889 ± 0.147	21.994 ± 2.193	42.120 ± 1.174	22.249 ± 0.913	6.892 ± 0.967
	Pretraining	RP-pretrained	5.612 ± 0.185	30.661 ± 4.038	55.623 ± 1.033	39.097 ± 2.721	14.470 ± 1.241
	Pretraining	RPTS-pretrained	5.308 ± 0.036	29.073 ± 5.651	54.897 ± 0.554	35.349 ± 0.891	12.777 ± 0.252
DimeNet++	Benchmark	RP	5.299 ± 0.079	27.645 ± 1.370	52.640 ± 2.014	34.301 ± 1.014	13.865 ± 1.223
	Benchmark	RPTS	3.425 ± 0.088	16.560 ± 3.351	35.268 ± 0.729	18.501 ± 0.436	5.401 ± 0.349
	Pretraining	RP-pretrained	5.292 ± 0.075	28.461 ± 2.939	52.640 ± 2.014	33.656 ± 1.847	13.503 ± 0.559
	Pretraining	RPTS-pretrained	5.222 ± 0.037	26.767 ± 1.134	52.600 ± 0.673	33.938 ± 1.223	12.495 ± 0.425
GemNet-T	Benchmark	RP	5.093 ± 0.174	23.863 ± 3.267	51.270 ± 0.640	32.285 ± 0.913	12.374 ± 0.916
	Benchmark	RPTS	3.565 ± 0.081	17.692 ± 1.496	37.283 ± 1.878	19.629 ± 0.858	6.489 ± 0.304
	Pretraining	RP-pretrained	5.070 ± 0.217	23.796 ± 1.639	49.053 ± 0.777	32.124 ± 0.666	12.656 ± 1.282
	Pretraining	RPTS-pretrained	4.869 ± 0.288	25.592 ± 1.399	50.020 ± 2.467	31.318 ± 2.833	11.649 ± 1.651
ClofNet	Benchmark	RP	4.776 ± 0.036	27.870 ± 5.104	50.141 ± 1.097	31.237 ± 0.489	10.077 ± 0.425
	Benchmark	RPTS	3.424 ± 0.101	19.258 ± 4.210	39.057 ± 2.056	19.629 ± 1.573	4.837 ± 0.846
	Pretraining	RP-pretrained	4.737 ± 0.030	29.474 ± 5.644	50.584 ± 1.282	30.713 ± 0.320	9.754 ± 0.559
	Pretraining	RPTS-pretrained	4.427 ± 0.053	24.620 ± 3.974	46.957 ± 1.573	27.973 ± 0.503	9.311 ± 0.726
MACE	Benchmark	RP	4.965 ± 0.107	26.409 ± 2.743	51.915 ± 1.762	33.253 ± 1.998	11.165 ± 0.787
	Benchmark	RPTS	3.245 ± 0.026	23.114 ± 3.015	35.429 ± 0.363	17.009 ± 1.211	4.877 ± 0.389
	Pretraining	RP-pretrained	4.941 ± 0.081	28.269 ± 3.513	53.366 ± 0.823	32.487 ± 1.458	11.245 ± 0.990
	Pretraining	RPTS-pretrained	4.583 ± 0.067	24.906 ± 5.741	50.141 ± 0.140	31.520 ± 0.458	10.157 ± 0.527

Table 11: DORTS direct TS information and paired TS-aware pretraining gains. For each lower-is-better metric, $\Delta_{\mathrm{TS}} = M(\mathrm{RP}) - M(\mathrm{RPTS})$ and $\Delta_{\mathrm{transfer}} = M(\mathrm{RP\text{-}pretrained}) - M(\mathrm{RPTS\text{-}pretrained})$. Intervals are 95% paired-record and hierarchical seed–record bootstrap intervals for $\Delta_{\mathrm{transfer}}$ (20,000 draws). SchNet denotes SchNet-1M.

Encoder	Metric	Δ_{TS}	$\Delta_{\mathrm{transfer}}$	Paired 95% CI	Hier. 95% CI
SchNet	MAE	1.343 ± 0.071	0.391 ± 0.046	$[0.235, 0.550]$	$[0.247, 0.533]$
	MAPE	8.259 ± 4.230	2.284 ± 2.602	$[0.090, 5.798]$	$[-0.647, 6.733]$
	q_3	10.439 ± 1.463	3.587 ± 2.003	$[1.532, 5.683]$	$[0.927, 6.288]$
	q_5	10.842 ± 0.458	3.144 ± 0.673	$[1.330, 4.917]$	$[1.370, 4.877]$
	q_{10}	5.643 ± 0.503	1.249 ± 1.531	$[0.040, 2.459]$	$[-0.443, 3.063]$
PaiNN	MAE	1.760 ± 0.357	0.304 ± 0.156	$[0.124, 0.484]$	$[0.112, 0.513]$
	MAPE	7.812 ± 4.196	1.588 ± 7.445	$[-0.315, 4.074]$	$[-8.191, 10.224]$
	q_3	13.785 ± 2.332	0.726 ± 0.484	$[-1.572, 2.983]$	$[-1.290, 2.741]$
	q_5	16.848 ± 3.389	3.748 ± 2.150	$[1.854, 5.643]$	$[1.249, 6.489]$
	q_{10}	7.779 ± 2.046	1.693 ± 1.033	$[0.403, 2.983]$	$[0.161, 3.225]$
DimeNet++	MAE	1.874 ± 0.066	0.070 ± 0.058	$[-0.149, 0.291]$	$[-0.129, 0.270]$
	MAPE	11.085 ± 2.301	1.694 ± 3.875	$[0.127, 3.666]$	$[-2.116, 6.906]$

Continued on next page

Encoder	Metric	Δ_{TS}	$\Delta_{\mathrm{transfer}}$	Paired 95% CI	Hier. 95% CI
	q_3	17.372 ± 1.304	0.040 ± 1.432	$[-2.338, 2.418]$	$[-2.580, 2.701]$
	q_5	15.800 ± 1.443	-0.282 ± 1.282	$[-2.378, 1.814]$	$[-2.499, 2.015]$
	q_{10}	8.464 ± 1.108	1.008 ± 0.252	$[-0.484, 2.499]$	$[-0.403, 2.378]$
GemNet-T	MAE	1.528 ± 0.140	0.201 ± 0.074	$[0.030, 0.372]$	$[0.030, 0.369]$
	MAPE	6.171 ± 1.922	-1.796 ± 3.002	$[-6.561, 1.064]$	$[-7.389, 2.154]$
	q_3	13.986 ± 1.940	-0.967 ± 1.814	$[-3.105, 1.129]$	$[-3.668, 1.612]$
	q_5	12.656 ± 0.070	0.806 ± 2.168	$[-1.048, 2.620]$	$[-2.015, 3.305]$
	q_{10}	5.885 ± 0.666	1.008 ± 0.369	$[-0.161, 2.177]$	$[-0.202, 2.217]$
ClofNet	MAE	1.352 ± 0.134	0.310 ± 0.073	$[0.141, 0.478]$	$[0.142, 0.481]$
	MAPE	8.612 ± 9.298	4.854 ± 4.524	$[-2.313, 16.819]$	$[-1.155, 14.889]$
	q_3	11.084 ± 3.107	3.628 ± 1.227	$[1.330, 5.925]$	$[1.088, 6.167]$
	q_5	11.608 ± 1.486	2.741 ± 0.252	$[0.846, 4.636]$	$[0.886, 4.635]$
	q_{10}	5.240 ± 0.970	0.443 ± 0.916	$[-0.605, 1.532]$	$[-0.927, 1.733]$
MACE	MAE	1.721 ± 0.133	0.358 ± 0.057	$[0.124, 0.606]$	$[0.152, 0.567]$
	MAPE	3.294 ± 0.359	3.363 ± 3.930	$[-0.130, 9.441]$	$[-1.970, 11.912]$
	q_3	16.485 ± 1.417	3.225 ± 0.875	$[0.645, 5.844]$	$[0.685, 5.764]$
	q_5	16.243 ± 3.156	0.967 ± 1.166	$[-1.331, 3.345]$	$[-1.411, 3.305]$
	q_{10}	6.288 ± 0.838	1.088 ± 1.486	$[-0.403, 2.580]$	$[-0.846, 2.902]$

Table 12: RGD1 v9 Core benchmark and paired pretraining results over three seeds. MAE is in kcal/mol; all other metrics are percentages; lower is better. RPTS is a non-deployable oracle, whereas RP-pretrained and RPTS-pretrained use only reactants and products at inference. SchNet denotes SchNet-1M.

Encoder	Role	Method	MAE ↓	MAPE ↓	q_3 ↓	q_5 ↓	q_{10} ↓
SchNet	Benchmark	RP	11.084 ± 0.301	27.249 ± 0.860	80.009 ± 0.565	67.370 ± 1.075	42.305 ± 1.386
	Benchmark	RPTS	4.575 ± 0.083	12.009 ± 0.342	50.769 ± 1.090	30.665 ± 0.891	9.960 ± 0.414
	Pretraining	RP-pretrained	11.083 ± 0.293	27.544 ± 1.318	79.810 ± 0.732	67.448 ± 0.967	42.280 ± 1.441
	Pretraining	RPTS-pretrained	10.501 ± 0.273	26.421 ± 1.051	78.790 ± 1.057	65.725 ± 1.337	40.166 ± 1.699
PaiNN	Benchmark	RP	11.777 ± 0.072	28.741 ± 0.653	81.200 ± 0.657	70.028 ± 0.105	45.333 ± 0.411
	Benchmark	RPTS	4.460 ± 0.046	11.900 ± 0.255	48.623 ± 0.770	29.396 ± 0.874	9.700 ± 0.105
	Pretraining	RP-pretrained	11.815 ± 0.063	28.860 ± 0.933	81.555 ± 0.133	69.691 ± 0.187	45.592 ± 0.200
	Pretraining	RPTS-pretrained	11.395 ± 0.185	27.062 ± 0.313	80.567 ± 0.382	68.873 ± 0.257	43.951 ± 0.694
DimeNet++	Benchmark	RP	10.863 ± 0.073	26.694 ± 0.163	80.023 ± 0.157	67.630 ± 0.689	42.060 ± 0.609
	Benchmark	RPTS	4.349 ± 0.228	11.521 ± 0.652	48.236 ± 1.804	28.302 ± 2.076	9.047 ± 1.092
	Pretraining	RP-pretrained	10.840 ± 0.037	27.107 ± 0.647	79.842 ± 0.234	67.111 ± 0.064	41.719 ± 0.091
	Pretraining	RPTS-pretrained	10.303 ± 0.430	24.622 ± 1.521	78.972 ± 1.289	66.112 ± 1.782	39.616 ± 1.963
GemNet-T	Benchmark	RP	11.301 ± 0.080	27.721 ± 0.093	80.389 ± 0.280	68.081 ± 0.391	43.204 ± 0.722
	Benchmark	RPTS	4.185 ± 0.054	11.393 ± 0.159	45.454 ± 0.423	26.156 ± 0.606	8.659 ± 0.171
	Pretraining	RP-pretrained	11.285 ± 0.082	27.581 ± 0.729	80.322 ± 0.316	67.996 ± 0.309	43.069 ± 0.302
	Pretraining	RPTS-pretrained	10.822 ± 0.209	26.018 ± 0.672	79.192 ± 0.705	66.542 ± 0.233	41.243 ± 0.791
ClofNet	Benchmark	RP	9.657 ± 0.035	22.994 ± 0.357	77.842 ± 0.204	64.290 ± 0.297	37.182 ± 0.330
	Benchmark	RPTS	4.115 ± 0.017	10.247 ± 0.315	48.204 ± 0.305	27.335 ± 0.264	7.483 ± 0.119
	Pretraining	RP-pretrained	9.634 ± 0.099	22.540 ± 0.480	78.073 ± 0.380	63.810 ± 0.471	36.968 ± 0.409
	Pretraining	RPTS-pretrained	9.509 ± 0.081	22.097 ± 0.967	77.522 ± 0.246	63.749 ± 0.089	35.991 ± 0.612
MACE	Benchmark	RP	10.243 ± 0.090	24.668 ± 0.293	79.327 ± 0.133	66.244 ± 0.429	39.701 ± 0.512
	Benchmark	RPTS	4.141 ± 0.023	9.970 ± 0.227	47.674 ± 0.541	27.211 ± 0.198	7.473 ± 0.150
	Pretraining	RP-pretrained	10.353 ± 0.092	24.413 ± 0.859	79.508 ± 0.402	66.581 ± 0.144	40.369 ± 0.531
	Pretraining	RPTS-pretrained	10.047 ± 0.121	24.360 ± 0.331	78.894 ± 0.571	65.800 ± 0.532	39.079 ± 0.549

Table 13: RGD1 v9 Core direct TS information and paired TS-aware pretraining gains. For each lower-is-better metric, $\Delta_{\mathrm{TS}} = M(\mathrm{RP}) - M(\mathrm{RPTS})$ and $\Delta_{\mathrm{transfer}} = M(\text{RP-pretrained}) - M(\text{RPTS-pretrained})$. Intervals are 95% paired-record and hierarchical seed–record bootstrap intervals for $\Delta_{\mathrm{transfer}}$ (20,000 draws). SchNet denotes SchNet-1M.

Encoder	Metric	Δ_{TS}	$\Delta_{\mathrm{transfer}}$	Paired 95% CI	Hier. 95% CI
SchNet	MAE	6.509 ± 0.306	0.582 ± 0.422	$[0.472, 0.692]$	$[0.167, 0.980]$
	MAPE	15.240 ± 0.669	1.123 ± 1.254	$[0.263, 2.093]$	$[-0.330, 2.296]$
	q_3	29.240 ± 1.279	1.020 ± 1.674	$[0.377, 1.649]$	$[-0.650, 2.672]$
	q_5	36.705 ± 1.040	1.723 ± 1.903	$[1.006, 2.420]$	$[-0.185, 3.550]$
	q_{10}	32.346 ± 1.286	2.114 ± 2.654	$[1.450, 2.779]$	$[-0.462, 4.662]$
PaiNN	MAE	7.317 ± 0.105	0.420 ± 0.206	$[0.320, 0.521]$	$[0.215, 0.630]$
	MAPE	16.842 ± 0.856	1.799 ± 1.233	$[1.064, 2.560]$	$[0.570, 3.133]$

Continued on next page

Encoder	Metric	Δ_{TS}	$\Delta_{\mathrm{transfer}}$	Paired 95% CI	Hier. 95% CI
	q_3	32.576 ± 1.179	0.988 ± 0.512	$[0.384, 1.581]$	$[0.252, 1.727]$
	q_5	40.632 ± 0.903	0.817 ± 0.077	$[0.160, 1.471]$	$[0.181, 1.457]$
	q_{10}	35.632 ± 0.417	1.642 ± 0.826	$[1.002, 2.281]$	$[0.679, 2.605]$
DimeNet++	MAE	6.514 ± 0.284	0.537 ± 0.433	$[0.434, 0.640]$	$[0.202, 1.014]$
	MAPE	15.173 ± 0.775	2.485 ± 2.161	$[1.657, 3.590]$	$[0.510, 4.836]$
	q_3	31.788 ± 1.927	0.871 ± 1.143	$[0.242, 1.496]$	$[-0.227, 2.164]$
	q_5	39.328 ± 2.748	0.998 ± 1.768	$[0.302, 1.698]$	$[-0.519, 2.949]$
	q_{10}	33.014 ± 1.648	2.104 ± 2.015	$[1.450, 2.750]$	$[0.345, 4.236]$
GemNet-T	MAE	7.116 ± 0.128	0.463 ± 0.290	$[0.363, 0.564]$	$[0.206, 0.760]$
	MAPE	16.328 ± 0.222	1.563 ± 0.717	$[0.949, 2.282]$	$[0.670, 2.492]$
	q_3	34.936 ± 0.702	1.130 ± 1.019	$[0.508, 1.741]$	$[0.085, 2.263]$
	q_5	41.925 ± 0.918	1.453 ± 0.519	$[0.778, 2.121]$	$[0.664, 2.253]$
	q_{10}	34.545 ± 0.634	1.826 ± 0.961	$[1.215, 2.445]$	$[0.835, 2.946]$
ClofNet	MAE	5.541 ± 0.021	0.125 ± 0.132	$[0.024, 0.225]$	$[-0.032, 0.278]$
	MAPE	12.747 ± 0.445	0.443 ± 1.024	$[-0.062, 0.946]$	$[-0.651, 1.605]$
	q_3	29.638 ± 0.496	0.551 ± 0.626	$[-0.096, 1.212]$	$[-0.306, 1.439]$
	q_5	36.954 ± 0.373	0.060 ± 0.514	$[-0.679, 0.792]$	$[-0.821, 0.913]$
	q_{10}	29.698 ± 0.214	0.977 ± 0.678	$[0.313, 1.649]$	$[0.085, 1.905]$
MACE	MAE	6.103 ± 0.105	0.306 ± 0.191	$[0.199, 0.413]$	$[0.093, 0.493]$
	MAPE	14.698 ± 0.425	0.053 ± 1.062	$[-1.009, 0.911]$	$[-1.367, 1.122]$
	q_3	31.653 ± 0.644	0.615 ± 0.173	$[-0.028, 1.244]$	$[-0.046, 1.272]$
	q_5	39.033 ± 0.417	0.782 ± 0.675	$[0.071, 1.503]$	$[-0.178, 1.691]$
	q_{10}	32.228 ± 0.573	1.290 ± 0.936	$[0.608, 1.986]$	$[0.185, 2.349]$

Table 14: Cyclo-23-TS benchmark and paired pretraining results over three seeds. MAE is in kcal/mol; all other metrics are percentages; lower is better. RPTS is a non-deployable oracle, whereas RP-pretrained and RPTS-pretrained use only reactants and products at inference. SchNet denotes SchNet-1M.

Encoder	Role	Method	MAE ↓	MAPE ↓	q_3 ↓	q_5 ↓	q_{10} ↓
SchNet	Benchmark	RP	2.827 ± 0.091	19.288 ± 0.394	34.010 ± 1.503	15.068 ± 0.648	3.137 ± 0.320
	Benchmark	RPTS	2.281 ± 0.030	15.274 ± 0.379	25.215 ± 1.820	9.779 ± 0.845	1.968 ± 0.426
	Pretraining	RP-pretrained	2.813 ± 0.062	19.175 ± 0.110	33.395 ± 0.923	15.191 ± 0.768	3.014 ± 0.107
	Pretraining	RPTS-pretrained	2.693 ± 0.038	18.466 ± 0.358	31.980 ± 0.746	13.715 ± 0.746	2.645 ± 0.282
PaiNN	Benchmark	RP	2.706 ± 0.038	19.275 ± 0.371	31.119 ± 0.929	14.084 ± 0.947	2.522 ± 0.213
	Benchmark	RPTS	2.116 ± 0.009	13.949 ± 0.232	21.771 ± 1.122	7.995 ± 0.384	1.722 ± 0.107
	Pretraining	RP-pretrained	2.674 ± 0.016	19.293 ± 0.168	32.042 ± 0.768	14.514 ± 1.016	2.583 ± 0.185
	Pretraining	RPTS-pretrained	2.561 ± 0.047	18.394 ± 0.383	29.766 ± 1.335	12.854 ± 1.409	2.214 ± 0.185
DimeNet++	Benchmark	RP	3.289 ± 0.106	21.873 ± 1.511	42.005 ± 1.547	21.218 ± 1.464	3.752 ± 1.081
	Benchmark	RPTS	1.808 ± 0.671	11.708 ± 4.073	17.466 ± 11.399	6.581 ± 5.329	1.415 ± 0.852
	Pretraining	RP-pretrained	3.151 ± 0.092	20.940 ± 0.900	39.422 ± 1.397	19.988 ± 0.910	3.506 ± 0.000
	Pretraining	RPTS-pretrained	2.553 ± 0.579	17.812 ± 3.663	30.750 ± 11.422	12.362 ± 7.033	2.276 ± 0.593
GemNet-T	Benchmark	RP	2.539 ± 0.058	17.819 ± 0.525	30.566 ± 1.612	12.239 ± 1.491	2.276 ± 0.213
	Benchmark	RPTS	1.768 ± 0.061	11.630 ± 0.193	16.851 ± 1.714	5.105 ± 0.384	0.800 ± 0.213
	Pretraining	RP-pretrained	2.575 ± 0.060	18.038 ± 0.451	31.242 ± 1.065	12.485 ± 0.593	2.276 ± 0.107
	Pretraining	RPTS-pretrained	2.425 ± 0.035	16.875 ± 0.123	28.782 ± 0.804	11.439 ± 0.665	2.091 ± 0.107
ClofNet	Benchmark	RP	2.760 ± 0.099	18.371 ± 0.452	31.734 ± 1.576	14.699 ± 1.491	2.706 ± 0.699
	Benchmark	RPTS	1.963 ± 0.045	12.141 ± 0.267	19.311 ± 1.016	8.118 ± 1.027	0.861 ± 0.282
	Pretraining	RP-pretrained	2.693 ± 0.056	18.351 ± 0.493	31.181 ± 0.665	13.469 ± 1.510	2.460 ± 0.464
	Pretraining	RPTS-pretrained	2.663 ± 0.097	17.549 ± 0.507	31.058 ± 2.343	13.100 ± 1.027	2.337 ± 0.213
MACE	Benchmark	RP	2.395 ± 0.045	16.783 ± 0.308	26.261 ± 1.503	11.316 ± 0.832	2.460 ± 0.213
	Benchmark	RPTS	1.338 ± 0.017	8.378 ± 0.329	8.364 ± 0.648	1.968 ± 0.746	0.185 ± 0.000
	Pretraining	RP-pretrained	2.384 ± 0.048	16.617 ± 0.406	25.830 ± 1.441	11.501 ± 1.016	2.583 ± 0.369
	Pretraining	RPTS-pretrained	2.312 ± 0.029	16.184 ± 0.163	25.215 ± 0.699	10.271 ± 0.464	1.968 ± 0.107

Table 15: Cyclo-23-TS direct TS information and paired TS-aware pretraining gains. For each lower-is-better metric, $\Delta_{\mathrm{TS}} = M(\mathrm{RP}) - M(\mathrm{RPTS})$ and $\Delta_{\mathrm{transfer}} = M(\mathrm{RP\text{-}pretrained}) - M(\mathrm{RPTS\text{-}pretrained})$. Intervals are 95% paired-record and hierarchical seed–record bootstrap intervals for $\Delta_{\mathrm{transfer}}$ (20,000 draws). SchNet denotes SchNet-1M.

Encoder	Metric	Δ_{TS}	$\Delta_{\mathrm{transfer}}$	Paired 95% CI	Hier. 95% CI
SchNet	MAE	0.546 ± 0.093	0.120 ± 0.092	$[0.039, 0.199]$	$[0.014, 0.239]$
	MAPE	4.014 ± 0.772	0.709 ± 0.370	$[0.100, 1.335]$	$[-0.047, 1.469]$

Continued on next page

Encoder	Metric	Δ_{TS}	$\Delta_{\mathrm{transfer}}$	Paired 95% CI	Hier. 95% CI
	q_3	8.795 ± 1.714	1.415 ± 1.503	$[-0.738, 3.506]$	$[-0.984, 3.998]$
	q_5	5.289 ± 0.213	1.476 ± 1.441	$[0.000, 2.952]$	$[-0.431, 3.444]$
	q_{10}	1.169 ± 0.648	0.369 ± 0.369	$[-0.123, 0.861]$	$[-0.123, 0.984]$
PaiNN	MAE	0.590 ± 0.041	0.113 ± 0.051	$[0.050, 0.178]$	$[0.021, 0.207]$
	MAPE	5.327 ± 0.411	0.899 ± 0.374	$[0.432, 1.389]$	$[0.162, 1.729]$
	q_3	9.348 ± 1.309	2.276 ± 1.973	$[0.492, 4.059]$	$[-0.615, 5.289]$
	q_5	6.089 ± 1.330	1.661 ± 2.312	$[0.554, 2.829]$	$[-0.923, 4.244]$
	q_{10}	0.800 ± 0.282	0.369 ± 0.320	$[-0.123, 0.923]$	$[-0.123, 1.107]$
DimeNet++	MAE	1.481 ± 0.762	0.598 ± 0.670	$[0.470, 0.729]$	$[-0.117, 1.107]$
	MAPE	10.165 ± 5.467	3.129 ± 4.341	$[1.881, 4.267]$	$[-1.417, 6.813]$
	q_3	24.539 ± 12.945	8.672 ± 12.795	$[5.904, 11.439]$	$[-4.797, 18.696]$
	q_5	14.637 ± 6.416	7.626 ± 7.619	$[5.535, 9.717]$	$[-0.369, 14.084]$
	q_{10}	2.337 ± 1.591	1.230 ± 0.593	$[0.492, 2.030]$	$[0.000, 2.399]$
GemNet-T	MAE	0.771 ± 0.114	0.150 ± 0.085	$[0.064, 0.238]$	$[0.017, 0.280]$
	MAPE	6.188 ± 0.494	1.163 ± 0.524	$[0.458, 1.986]$	$[0.205, 2.227]$
	q_3	13.715 ± 3.261	2.460 ± 1.569	$[0.308, 4.613]$	$[-0.738, 5.597]$
	q_5	7.134 ± 1.503	1.046 ± 1.200	$[-0.492, 2.645]$	$[-1.107, 3.260]$
	q_{10}	1.476 ± 0.369	0.185 ± 0.000	$[-0.431, 0.861]$	$[-0.554, 0.984]$
ClofNet	MAE	0.797 ± 0.055	0.030 ± 0.146	$[-0.085, 0.145]$	$[-0.172, 0.222]$
	MAPE	6.229 ± 0.232	0.802 ± 0.884	$[-0.079, 1.741]$	$[-0.581, 2.152]$
	q_3	12.423 ± 0.852	0.123 ± 2.585	$[-2.583, 2.829]$	$[-4.244, 4.121]$
	q_5	6.581 ± 0.648	0.369 ± 2.496	$[-1.292, 2.091]$	$[-2.952, 3.506]$
	q_{10}	1.845 ± 0.976	0.123 ± 0.533	$[-0.677, 0.923]$	$[-0.923, 1.230]$
MACE	MAE	1.057 ± 0.032	0.073 ± 0.077	$[0.002, 0.145]$	$[-0.046, 0.189]$
	MAPE	8.405 ± 0.273	0.433 ± 0.245	$[-0.113, 0.975]$	$[-0.387, 1.214]$
	q_3	17.897 ± 0.976	0.615 ± 2.130	$[-1.415, 2.645]$	$[-2.768, 3.813]$
	q_5	9.348 ± 1.536	1.230 ± 0.593	$[-0.062, 2.583]$	$[-0.492, 3.014]$
	q_{10}	2.276 ± 0.213	0.615 ± 0.384	$[0.246, 1.046]$	$[-0.062, 1.292]$

Grambow unseen-family results. Table 16 reports the six encoders in the primary matrix on the separately trained depth-1 endpoint-edit-family split. These descriptive results use three seeds and 1,196 test reactions. The historical small-SchNet and LEFTNet variants are outside the six-encoder scope of the present manuscript. The family key was fixed before model results were available.

Table 16: Grambow depth-1 unseen-family test MAE in kcal/mol, mean $\pm$ sample SD over three seeds. RPTS uses TS at inference; both pretrained branches predict from R/P only.

Encoder	RP	RPTS	RP-pretrained	RPTS-pretrained
SchNet-1M	13.610 ± 0.115	6.390 ± 0.069	13.614 ± 0.347	13.158 ± 0.203
PaiNN	13.165 ± 0.125	5.248 ± 0.041	13.188 ± 0.173	12.497 ± 0.078
DimeNet++	14.279 ± 0.108	5.180 ± 0.084	13.987 ± 0.292	13.230 ± 0.058
GemNet-T	14.110 ± 0.257	5.516 ± 0.023	14.046 ± 0.195	13.555 ± 0.103
ClofNet	11.142 ± 0.051	4.912 ± 0.132	11.054 ± 0.135	10.304 ± 0.129
MACE	11.886 ± 0.273	4.943 ± 0.067	11.877 ± 0.233	11.810 ± 0.345

Additional tail-error comparisons. Figure 4 complements the MAE, MAPE, and q_3 comparisons tabulated above. Mean q_5 decreases in 23 of 24 dataset–encoder pairs, and mean q_{10} decreases in all 24. The exception is DimeNet++ on DORTS, whose q_5 increases by 0.28 percentage points. These counts concern mean differences, not statistical significance; the preceding tables report the corresponding confidence intervals.

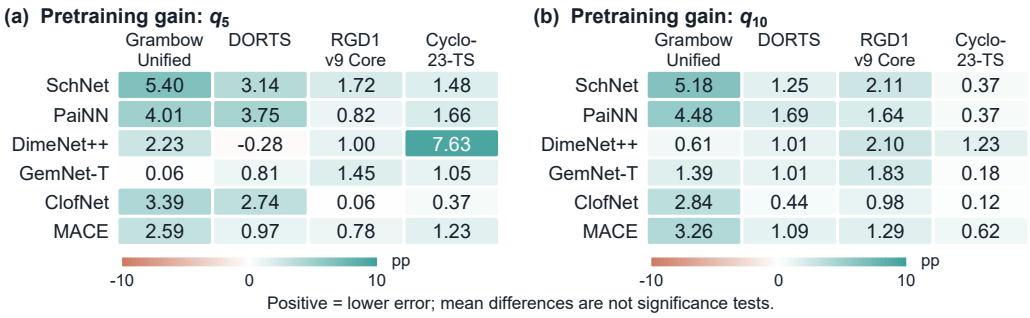


Figure 4: Additional matched endpoint-only pretraining gains for **(a)** q_5 and **(b)** q_{10}, complementing the main-text MAE comparison in Table 3. Values are three-seed mean differences, $M(\text{RP-pretrained}) - M(\text{RPTS-pretrained})$, in percentage points. Positive values indicate fewer predictions with absolute errors exceeding 5 or 10 kcal/mol, respectively. The panels share a colour scale; colour does not indicate statistical significance.

C Implementation and Reproducibility

C.1 Metric definitions

The five metrics are defined in Section 5.1. MAE uses kcal/mol; MAPE and q_3, q_5, q_{10} are percentages. All metrics are lower-is-better, and q_τ uses a strict greater-than threshold. Means and sample standard deviations are reported over seeds 0–2. $B^2R_\ell^2$ with kernel ridge regression has one deterministic run. Paired-record bootstrap resamples reaction IDs after seed averaging; the hierarchical analysis resamples seeds and paired records within each sampled seed. For the primary grouped-random comparisons, we report 95% intervals for paired MAE differences. The Grambow family-split table reports descriptive three-seed means and sample standard deviations. For SchNet-1M and MACE, we additionally report hierarchical seed-and-group intervals; the remaining four encoders are not assigned inferential claims on this split.

C.2 Optimization and checkpoint selection

SchNet denotes the approximately one-million-parameter configuration. Models minimize target-standardized mean squared error using training-only normalization statistics. Checkpoints are selected by minimum validation MAE; test metrics are never used for checkpoint selection. RP and RPTS views and their pretrained descendants share data and seeds. Recorded epoch limits and optimization exceptions are detailed below. Optimizer state is not transferred into endpoint-only adaptation. Equal epoch budgets do not imply equal floating-point cost because RPTS pretraining encodes one additional molecular state.

Recorded optimization settings. Table 17 reports settings recovered from the selected runs. The available adaptation records specify AdamW, weight decay 10^{-4}, hidden dimension 128, head dropout 0.2, a 2,000-epoch limit and early-stopping patience 200. These are epoch limits, not a claim that every run trained for 2,000 epochs. The table reports settings supported by archived run records rather than substituting current defaults. Parameter counts include the encoder, state normalization and RP head, exclude buffers, and may differ across datasets with different element vocabularies. They must not be read as encoder-only counts.

Geometric neighborhoods. The retained encoder implementations specify distance cutoffs of 10 Å (SchNet-1M), 12 Å (PaiNN), 5 Å (DimeNet++), 6 Å (GemNet-T), 6.5 Å (ClofNet), and 5 Å (MACE). These are implementation-level settings. Scheduler histories are not used as comparison factors, so we report the recorded initial learning rates without inferring unrecorded scheduler states.

Optimization exception. The accepted Cyclo-23-TS DimeNet++ bundle combines original and recovery runs. Its RPTS parent uses learning rates 10^{-3} and 5×10^{-5} across seeds. For endpoint adaptation, seed 2 uses 5×10^{-5} for RP-pretrained but 10^{-3} for RPTS-pretrained; seeds 0 and 1

Table 17: Recorded endpoint-adaptation settings. Batch sizes count reactions; parameter counts are total trainable RP predictors, not encoder-only sizes. Ranges retain run differences; SchNet always means SchNet-1M.

Dataset	Encoder	Batch	Learning rate	Parameters
Grambow	SchNet	256	0.001	1,133,185
Grambow	PaiNN	128	0.0001	1,371,265
Grambow	DimeNet++	256	0.001	2,038,151
Grambow	GemNet-T	64	0.0001	2,013,778
Grambow	ClofNet	256	0.001	666,178
Grambow	MACE	128	0.001	1,066,113
DORTS	SchNet	256	0.001	1,138,945
DORTS	PaiNN	128	0.0001	1,377,025
DORTS	DimeNet++	256	0.001	2,043,911
DORTS	GemNet-T	64	0.0001	2,019,538
DORTS	ClofNet	256	0.001	671,938
DORTS	MACE	128	0.001	4,095,873
RGD1 Core	SchNet	256	0.001	1,133,185
RGD1 Core	PaiNN	128	0.0001	1,371,265
RGD1 Core	DimeNet++	256	0.001	2,038,151
RGD1 Core	GemNet-T	64	0.0001	2,013,778
RGD1 Core	ClofNet	256	0.001	666,178
RGD1 Core	MACE	128	0.001	1,066,113
Cyclo-23-TS	SchNet	256	0.001	1,136,641
Cyclo-23-TS	PaiNN	32	0.0001	1,374,721
Cyclo-23-TS	DimeNet++	16	5×10^{-5}–0.001	2,041,607
Cyclo-23-TS	GemNet-T	16	0.0001	2,017,234
Cyclo-23-TS	ClofNet	256	0.001	669,634
Cyclo-23-TS	MACE	8	0.001	2,883,969

use 5×10^{-5} for both branches. The reported aggregate is retained, but is not a strictly learning-rate-matched attribution of TS exposure. Equal epoch limits also do not imply equal pretraining compute.

Checkpoint lineage and evidence boundary. For each dataset, encoder and seed, the RP-parent validation checkpoint supplies the complete initialization of RP-pretrained. RPTS-pretrained retains its non-encoder parameters and replaces the encoder using the same-seed RPTS parent. The initialization audit verifies that the two adapted models differ only in the transferred encoder. The resulting endpoint model is selected by validation MAE. Detailed checkpoint lineage and integrity metadata are supplied with the machine-readable release rather than reproduced here.

C.3 Pretraining-objective adaptations

The Grambow objective comparison uses ClofNet and the same 1,198 frozen test reactions. Pretext heads are discarded before endpoint adaptation; only the encoder replaces that of the same-seed RP parent, whose remaining parameters initialize the downstream model. Method sources are cited in Section 2; the adaptations below do not claim numerical parity with their original architectures. The two branches use the complete RP or RPTS outer-training pool without compute matching.

CoordDenoise and GeoSSL-DDM. Coordinate denoising uses isotropic Gaussian coordinate noise with $\sigma = 0.04$ Å and predicts the injected noise. GeoSSL-DDM instead uses multiscale pair-distance scores, with coordinate noise of 0.3 Å, 50 distance-noise levels from 10 to 0.01 Å, and weighting exponent 0.2. Both use 100 pretraining epochs under the frozen Grambow protocol.

EMPP-l1 and C-FREE-Norm. EMPP-l1 masks one atom per graph by physical deletion and predicts radial and directional distributions from the retained neighbourhood. The ClofNet adaptation is restricted to angular order $l_{\max} = 1$ with 128 radial bins and 20 pretraining epochs; it is not the original higher-order architecture. C-FREE-Norm uses a contrast-free online/target-encoder ob-

jective with normalized prediction and target vectors. Its geometric ego/node-complement views and ClofNet backbone differ from the original multimodal model and its paper-defined edge-complement views. We use 200 pretraining epochs and an EMA coefficient scheduled from 0.995 to 1.0.

Energy-Referenced. Energy-Referenced supervises absolute electronic-plus-ZPE state energies after a training-only elemental-reference subtraction. Element-reference coefficients and residual normalization are fitted on outer-train R/P states, then shared by both branches. The scalar energy head is discarded at transfer, and inference requires no energy labels. This control differs from the task-aligned strategy in its state-energy rather than barrier supervision, not in whether the downstream head is inherited from the RP parent.