TREK: BENCHMARKING REACTION-LEVEL TRANSITION-STATE ENSEMBLE RECOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurate transition-state generation for a known reactant–product pair does not establish recovery of the alternatives associated with a reaction. Measuring this broader capability requires a benchmark whose unit is the reaction and whose output is a set. We introduce **TREK** (Transition-state Recovery for Ensemble Kinetics), a benchmark for budgeted TS-ensemble recovery from complete retained reactant and product conformer sets. TREK organizes 3,493 reactions, 10,865 candidate endpoint pairs, and 19,647 xTB-derived proxy channels into two controlled tasks: known-pair one-shot generation and reaction-level recovery with target-blind pair selection and a shared proposal budget. Pair-aware one-to-one scoring measures distinct reference recovery, while a stagewise decomposition attributes the recovery gap to selection, capacity, and geometry/diversity. Across five generators, leading models achieve 94.0–95.6% pair-hit rates on held-out reactions at 0.50 Å, alongside 74.3–76.1% complete-reference-set recovery in the ensemble task. TREK further identifies a tolerance-dependent shift from geometric precision to output capacity, quantifies the coverage gained per additional proposal, and isolates 38 shared zero-hit reactions enriched in internal endpoint rearrangement. Frozen DFT and bidirectional IRC audits provide complementary evidence for individual reference candidates and conformationally distinct connections. TREK establishes a common experimental basis for developing generators that recover reaction-level structural diversity, connecting accurate individual proposals to budget-efficient ensemble coverage.

1 INTRODUCTION

Transition states (TSs) characterize saddle regions of the potential energy surface that connect reactant and product states, providing a structural basis for understanding reaction mechanisms and kinetics (Truhlar et al., 1996; E & Vanden-Eijnden, 2010; Unsleber & Reiher, 2020; Dewyer et al., 2018). Conventional TS search methods, including NEB (Henkelman & Jónsson, 2000) and GSM (Zimmerman, 2013), locate saddle points through iterative energy- and force-based optimization, often at substantial computational cost. Recent learning-based approaches instead directly predict or generate TS geometries from reactant–product (R/P) structures (Duan et al., 2023; 2025; Luo et al., 2026; Tuo et al., 2026; Darouich et al., 2026). Despite substantial methodological progress, these models are predominantly evaluated in a pair-conditioned setting: given a predefined R/P geometry pair, a model is assessed by whether it can recover an associated reference TS.

However, pair-conditioned evaluation captures only a local view of the TS landscape of a reaction. A reaction can involve multiple reactant and product conformations, yielding multiple plausible endpoint combinations and distinct multiple TS structures or reaction channels. Prior studies have shown that conformational sampling and mechanistic exploration can reveal multiple TS candidates for the same reaction, with different geometries and activation barriers (Zhao et al., 2022; Choi, 2023; Zhao et al., 2023; Schmid et al., 2026). These methods highlight the importance of moving beyond a single predefined R/P pair and considering TS multiplicity at the reaction level. Yet existing learning-based benchmarks (Schreiner et al., 2022) predominantly evaluate individual R/P–TS instances rather than how well a model recovers the collection of distinct TS alternatives associated with a reaction. Consequently, strong pair-conditioned accuracy does not necessarily imply broad reaction-level coverage, leaving multi-TS recovery without a systematic evaluation framework.

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.

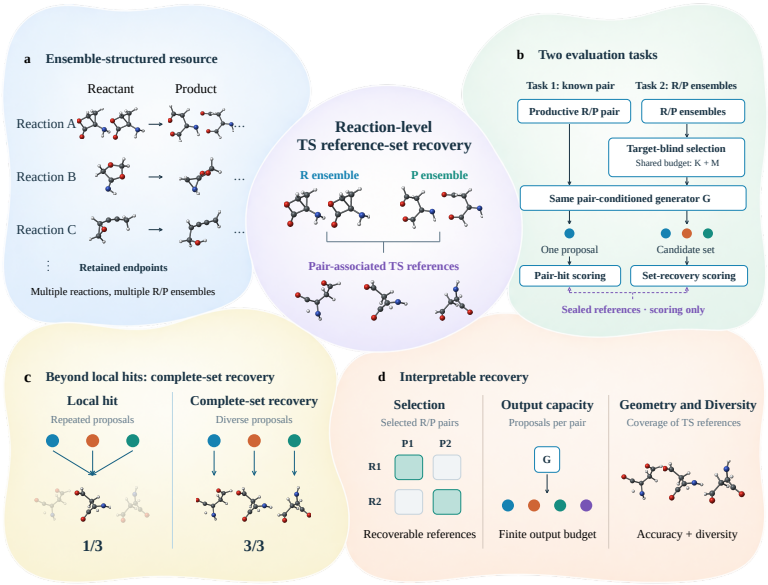

Figure 1: **From an accurate individual TS to reaction-level ensemble recovery.** TREK asks how much of a reaction's reference TS set an inference system can recover from retained endpoint ensembles under a finite proposal budget. Known-pair generation measures a local hit; ensemble recovery additionally requires selecting endpoint pairs and covering distinct targets.

To address this gap, we introduce **TREK** (**T**ransition-state **R**ecovery for **E**nsemble **K**inetics), a benchmark that formulates reaction-level TS-ensemble recovery as an explicit, budgeted learning problem (Figure 1). TREK combines retained R/P conformer ensembles, productive-pair evidence, and structured xTB references across 3,493 reactions, comprising 10,865 candidate endpoint pairs and 19,647 reaction-unique proxy channels. Rather than asking only whether a generator can recover a TS for a supplied R/P pair, TREK evaluates how much of a reaction's reference set can be recovered from its endpoint ensemble under a fixed proposal budget. Two controlled tasks connect these pair-conditioned generation and reaction-level ensemble recovery within a unified evaluation framework.

The benchmark provides three capabilities that make this question actionable. First, a common target-blind selector and reaction-shared budget place existing pair-conditioned generators within the same ensemble-recovery system. Second, pair-aware one-to-one matching rewards additional reference coverage rather than repeated discovery of the same target. Third, a stagewise decomposition identifies whether recovery is governed by pair selection, output capacity, or geometric precision and diversity. Frozen DFT and bidirectional IRC audits (Fukui, 1981b) complement the structural reference inventory with evidence at both individual-candidate and small-ensemble scales.

A study of five generators demonstrates that reaction-level evaluation reveals capabilities not captured by pair-conditioned accuracy alone. The leading models achieve 94.0–95.6% known-pair hit rates on held-out reactions at 0.50 Å, while complete-reference-set recovery under the ensemble protocol is 74.3–76.1%. These metrics are complementary success criteria: the former measures a local hit, and the latter requires simultaneous coverage of a reaction's alternatives. TREK turns their distinction into concrete design guidance. For RobustTSGen (Darouich et al., 2026), the dominant recovery bottleneck changes from geometry/diversity at 0.20 Å to pair-specific output capacity at 0.50 Å; budget curves expose where additional proposals are most productive. Cross-generator coverage further identifies complementary discoveries and a focused set of 38 shared zero-hit reactions associated with larger internal endpoint rearrangements. Our contributions are fourfold:

1. We introduce TREK, a reaction-level multi-TS benchmark, which organizes retained endpoint ensembles, productive-pair relations, and xTB TS proxies within a unified framework. TREK further establishes an explicit hierarchy between endpoint-conditioned targets and TS geometry families, with DFT/IRC evidence supporting the reference construction.
2. We establish a controlled ensemble-recovery protocol. Two complementary tasks bridge known-pair generation and budgeted reaction-level recovery from ensemble inputs. A

target-blind selector and shared allocation rules support direct comparisons across five existing generators.

3. We develop a set-level explanatory evaluation framework. Pair-aware one-to-one matching measures distinct reference recovery, while stagewise decomposition attributes failures to pair selection, proposal allocation/capacity, or geometry/diversity, enabling reaction-level coverage and completeness to evaluated beyond pair-level hit accuracy.
4. We provide empirical design insights into reaction-level TS recovery. Cross-generator analyses uncover a tolerance-dependent shift in recovery bottlenecks, a systematic coverage–budget trade-off, and complementary reference coverage, highlighting distinct opportunities in geometry, proposal allocation, and generation diversity.

2 Related Work

Transition State Prediction and Benchmarking Learning-based TS prediction has predominantly been studied at the pair level, where a model predicts or generates a TS structure for a predefined reactant–product (R/P) geometry pair. Early approaches directly predict TS geometries from individual R/P pairs (Jackson et al., 2021; Choi, 2023), while datasets such as Transition1x (Schreiner et al., 2022) have enabled increasingly expressive generative models, including diffusion-, transport-, and flow-based approaches (Duan et al., 2023; Kim et al., 2024; Duan et al., 2025; Galustian et al., 2025; Luo et al., 2026; Tuo et al., 2026; Darouich et al., 2026; Xu et al., 2026; Qiu et al., 2026). Despite substantial methodological progress, evaluation in this setting typically treats each predefined R/P pair as an independent instance with a designated TS reference.

Beyond pair-level prediction, recent studies have increasingly considered multiple TS structures associated with the same reaction. Conformational sampling has revealed multiple endpoint configurations and corresponding TS structures (Choi, 2023; Zhao et al., 2022; 2023), with RGD1 providing multiple TS conformations discovered through such sampling (Zhao et al., 2023). More recently, racerTS explicitly benchmarks TS conformer-ensemble generation from an input TS structure and reaction-center information (Schmid et al., 2026). These advances establish the importance of TS multiplicity, but do not address a complementary reaction-level problem: recovering a structured set of TS references directly from reactant and product conformational ensembles under a finite generation budget. TREK targets this setting by jointly evaluating endpoint-pair selection, proposal allocation, and set-level TS recovery, while decomposing unrecovered references into selection, allocation/capacity, and geometry/diversity failures.

3 TREK Benchmark

TREK couples a reaction-level reference resource with two controlled tasks and pair-aware set scoring. The benchmark measures *how much reference support is recovered under a finite proposal budget*, then attributes the remaining gap to specific stages of inference.

3.1 Problem Formulation

TREK extends pair-conditioned TS evaluation to reaction-level set recovery over conformational endpoint ensembles. For reaction i, let $\mathcal{R}_i$ and $\mathcal{P}_i$ denote the retained reactant and product conformer sets, with $\mathcal{E}_i = \mathcal{R}_i \times \mathcal{P}_i$ defining the candidate endpoint-pair set and $\mathcal{C}_i$ the reaction-unique set of xTB proxy channels. A concrete endpoint-pair datapoint is denoted by $E = (R, P) \in \mathcal{E}_i$, where $R \in \mathcal{R}_i$ and $P \in \mathcal{P}_i$. The relation $\mathcal{L}_i \subseteq \mathcal{E}_i \times \mathcal{C}_i$ links productive pairs to their supported TS proxy channels. Given the endpoint ensemble and a finite proposal budget, the objective is to recover as many distinct members of $\mathcal{C}_i$ as possible. TREK evaluates this objective through two complementary tasks: known-pair generation (Task 1 in Section 3.3) and reaction-level recovery with target-blind pair selection and a shared proposal budget (Task 2 in Section 3.3).

3.2 Reaction-Level Reference Construction

Reference construction. Starting from 5,000 Transition1x reactions (Schreiner et al., 2022), we expand the endpoint support through CREST conformer exploration (Pracht et al., 2020b), followed by GFN2-xTB path searches and saddle refinement (Bannwarth et al., 2019). Physical acceptance checks retain TS candidates and their supporting endpoint geometries. The resulting resource contains 3,493 reactions, 5,167 R endpoints, and 5,213 P endpoints. Across reactions, the resulting

Table 1: **TREK inventory by Transition1x source split.** Candidate pairs comprise observed productive pairs and unknown cross-combinations.

Split	Reactions	Candidate pairs	Productive pairs	Unknown pairs	Reference channels
Train	3,317	10,194	5,212	4,982	18,487
Validation	67	299	130	169	569
Test	109	372	183	189	591
Total	**3,493**	**10,865**	**5,525**	**5,340**	**19,647**

candidate sets $\mathcal{E}_i$ contain 10,865 endpoint pairs, including 5,525 observed productive pairs and 5,340 unobserved cross-combinations (Figure 2).

Retaining endpoints from accepted transitions ensures that every observed productive pair is available before pair selection. The inference challenge is therefore to identify useful combinations within a complete retained ensemble, rather than to reconstruct endpoints removed by a separate conformer-truncation step. The unobserved cross-combinations enlarge this decision space without revealing which pairs have historical TS support. Because absence from the frozen search inventory does not imply chemical non-productivity, these pairs are treated as unknown rather than negative.

A productive endpoint pair need not map to a single reference channel: the frozen search can identify multiple TS targets for the same pair, while the same target can be supported by multiple pairs. TREK therefore contains 19,747 pair–channel links connecting the productive pairs to 19,647 reaction-unique xTB proxy channels. These proxy channels preserve the endpoint context under which each transition is observed and define the primary recovery targets. To distinguish endpoint-conditioned multiplicity from structural multiplicity, we additionally group geometrically equivalent proxy targets into 11,174 geometry-only TS families. Multiplicity remains substantial at both levels: 3,482 of 3,493 reactions and 4,633 of 5,525 productive pairs contain multiple proxy channels, while 2,273 reactions and 2,854 productive pairs still contain multiple geometry families. Thus, the multiplicity captured by TREK reflects not only alternative endpoint contexts but also distinct TS structures.

Table 1 summarizes the resulting inventory according to the original Transition1x source split, including the composition of productive and unknown candidate pairs and the corresponding reference-channel coverage. The retained candidate space preserves all observed productive pairs while also including cross-combinations without known TS support, providing a unified reaction-level inventory for TS recovery under conformational diversity. These source splits are preserved throughout benchmark construction, with the held-out test split used for the generalization evaluation described in Section 4.

Physical audit. Because the large-scale reference inventory is constructed at the xTB level, we assess its physical plausibility using frozen higher-level audits. In a 100-candidate sample evaluated with ORCA/B97-3c (Neese, 2022b; Brandenburg et al., 2018b), optimization and frequency analysis retain 96 candidates, of which 75 satisfy the strict acceptance criteria after bidirectional IRC validation (Fukui, 1981b; Gonzalez & Schlegel, 1989a). A complementary eight-reaction multi-round audit canonicalizes 96 complete DFT/IRC certificates into 11 stable basin-resolved connections. These audits support the physical plausibility of reference construction; the large-scale benchmark measures structural recovery of xTB proxies, not DFT certification of every target. Full protocols and representative IRC profiles appear in Appendix E and Figure 5.

3.3 Two Evaluation Tasks

TREK defines two complementary tasks that progressively introduce the challenges of reaction-level TS recovery. Task 1 provides a controlled pair-conditioned baseline, whereas Task 2 evaluates recovery from the full retained endpoint ensemble under target-blind pair selection and a shared reaction-level proposal budget.

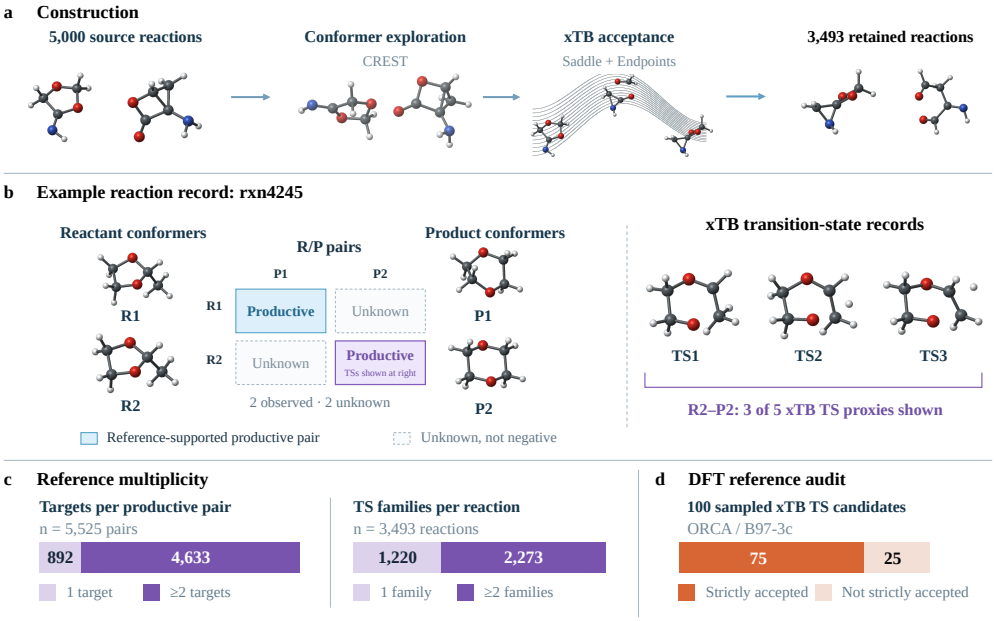

Figure 2: **TREK organizes conformational support into a reaction-level benchmark.** (a) Construction retains 3,493 of 5,000 source reactions after conformer exploration and xTB acceptance. (b) Reaction rxn4245 distinguishes two productive endpoint pairs from two unknown cross-pairs; three of the five xTB TS proxies associated with R2–P2 are shown. (c) Multiplicity is reported across 5,525 productive pairs and 3,493 reactions, with proxy targets distinguished from geometry families. (d) The frozen ORCA/B97-3c audit yields 75 strict passes among 100 candidates.

Task 1: known-pair single-proposal generation. For each observed productive pair $E \in \mathcal{E}_i$, the generator receives the atomic identities and R/P coordinates and emits a single TS proposal. A pair is a hit if this proposal matches any reference channel associated with E through $\mathcal{L}_i$ under the specified structural threshold. By providing the productive pair and allowing only one proposal, Task 1 isolates pair-conditioned TS generation from pair selection and cross-pair proposal allocation.

Task 2: reaction-level recovery under a shared budget. Given the retained endpoint ensembles $\mathcal{R}_i$ and $\mathcal{P}_i$, the objective of Task 2 is to recover as many distinct reference channels in $\mathcal{C}_i$ as possible under a finite reaction-level proposal budget. Productive-pair labels, reference TS geometries, channel/path identities, and acceptance counts are withheld during inference.

A common target-blind selector ranks candidate pairs in $\mathcal{E}_i$ by increasing atom-mapped endpoint RMSD, with deterministic tie-breaking, and selects $K_i = \min(8, |\mathcal{E}_i|)$ pairs. Each selected pair receives one initial proposal, followed by $M = 16$ additional proposals shared across the reaction. The total proposal budget is therefore $B_i = K_i + M \leq 24$. The additional proposals are allocated round-robin over the ranked pairs. Thus, M denotes a shared reaction-level top-up budget, not a separate budget assigned to each pair. For example, when $K_i = 4$ and $M = 16$, each of the four selected pairs receives five proposals (Figure 3c). Failed generation attempts consume their assigned slots and are not replaced. Budget-prefix analyses use $M \in \{0, 4, 8, 16\}$.

TREK fixes the selector and allocator across generators, providing a common reaction-level inference protocol without access to target multiplicity. The initial pass exposes every selected pair at least once, while the shared additional budget permits repeated proposals to explore alternative channels supported by the same pair. Full allocation rules and a secondary novelty-adaptive allocation policy are provided in Appendix C.

Together, the two tasks evaluate complementary capabilities: Task 1 isolates TS generation from a known productive pair, whereas Task 2 jointly evaluates target-blind pair selection, proposal allocation, and reaction-level reference-channel recovery. Section 3.4 defines the set-level matching and recovery metrics used to evaluate these outputs.

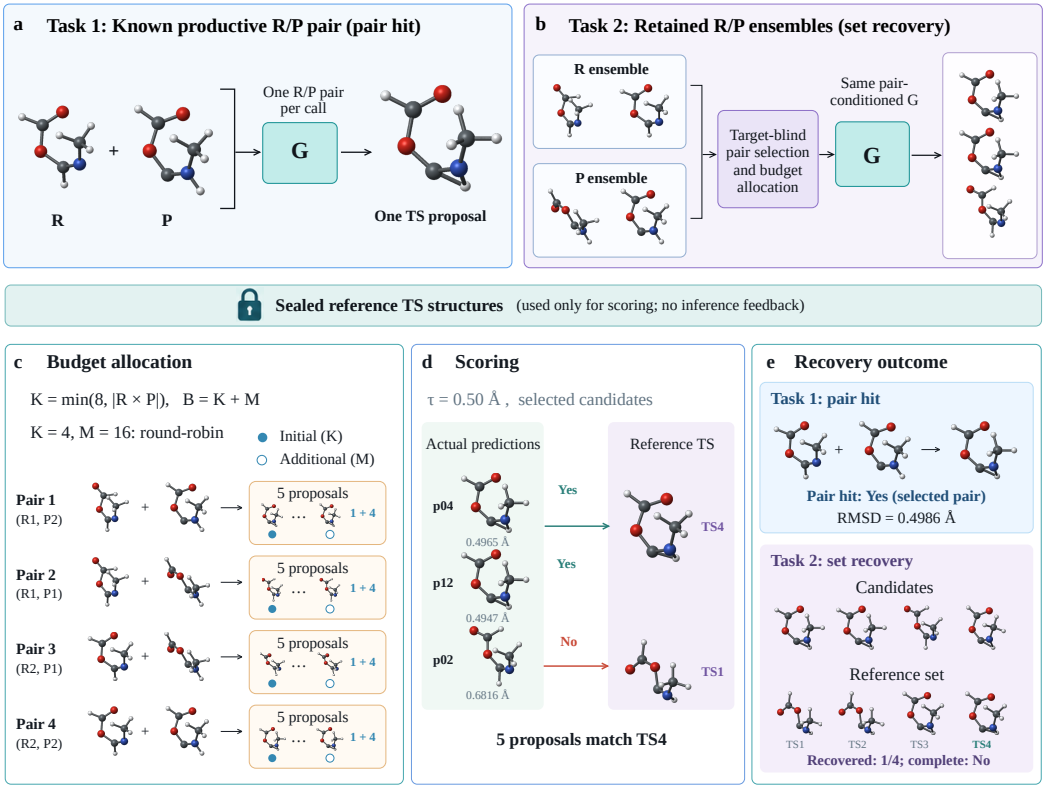

Figure 3: **A common inference wrapper separates local generation from set recovery.** (a,b) The same pair-conditioned generator receives a known productive pair or pairs selected from retained R/P ensembles. References remain sealed during inference. (c) With $K = 4$ and $M = 16$, round-robin allocation assigns one initial and four additional proposals to each selected pair, giving 20 proposals in total. (d) At 0.50 Å, multiple proposals match TS4, but one-to-one matching credits that reference once. (e) The illustrated Task 2 output recovers one of four references, whereas Task 1 achieves a pair hit. Panels (c–e) use the held-out reaction rxn1825 with RobustTSGen; displayed candidates are selected examples from the full output set.

3.4 Set-Level Scoring and Error Decomposition

Set-level scoring. A proposal generated from endpoint pair E can match a reference channel $C \in \mathcal{C}_i$ only if $(E, C) \in \mathcal{L}_i$ and its TS RMSD to the channel reference is at most a structural threshold τ. We compute TS RMSD using all atoms with equal-element permutation after proper alignment and without global reflection. After collapsing exact duplicate proposal geometries, we construct a proposal–channel bipartite graph from all admissible matches and compute a maximum-cardinality one-to-one matching. Each proposal can therefore recover at most one reference channel, and each channel is credited at most once. Among maximum-cardinality matchings, total RMSD is minimized as a secondary objective.

Let $m_i(\tau)$ denote the number of reference channels recovered by the matching for reaction i, and let N denote the number of reactions. Overall reference-channel recovery is measured by

$$\mathrm{Recall}_{\mathrm{micro}}(\tau) = \frac{\sum_i m_i(\tau)}{\sum_i |\mathcal{C}_i|}. \tag{1}$$

We additionally report *macro recall*, the unweighted mean of per-reaction recall over the N reactions, and *complete recovery*, the fraction of reactions for which every reference channel is recovered. These metrics distinguish aggregate coverage, reaction-balanced recovery, and complete-set recovery.

Error decomposition. To identify where recovery is lost in Task 2, let s_i denote the number of reference channels supported by the selected endpoint pairs, and let u_i denote the maximum number

of these channels that can be covered by the available pair-specific proposal slots when geometry is ignored. By construction,

$$|\mathcal{C}_i| \geq s_i \geq u_i \geq m_i(\tau).$$

The unrecovered reference channels can therefore be decomposed exactly as

$$|\mathcal{C}_i| - m_i(\tau) = \underbrace{|\mathcal{C}_i| - s_i}_{\text{selection}} + \underbrace{s_i - u_i}_{\text{allocation/capacity}} + \underbrace{u_i - m_i(\tau)}_{\text{geometry/diversity}}. \tag{2}$$

The three terms distinguish channels lost because their supporting pairs are not selected, channels that remain reachable but cannot be covered by the available pair-specific proposal slots, and channels that remain recoverable in principle but are not matched within the structural threshold.

4 Results

4.1 Experimental setup

Baseline models. We evaluate five representative generation-based TS methods: OA-ReactDiff (Duan et al., 2023), React-OT (Duan et al., 2025), TS-DFM (Luo et al., 2026), MolGEN (Tuo et al., 2026), and RobustTSGen (Darouich et al., 2026), using fixed checkpoints and inference configurations (Appendix C). All five are pair-conditioned generators evaluated under both TREK tasks. TREK preserves the original Transition1x (Schreiner et al., 2022) source partition throughout benchmark construction. The held-out evaluation uses the source-test subset, comprising 109 reactions, 183 productive pairs, and 591 reference channels (Table 1), while full-inventory analyses cover all 3,493 reactions and are reported descriptively.

Task instantiation. For Task 1, each observed productive endpoint pair receives one TS proposal. For Task 2, we use the common Top-8 selector and round-robin allocation defined in Section 3.3, with $M = 16$ additional proposals per reaction for the primary comparison. Budget-prefix analyses use $M \in \{0, 4, 8, 16\}$. All generators share the same pair-selection and allocation rules, and failed generation attempts consume their assigned slots.

Common evaluation protocol. Across the full 3,493-reaction inventory and five generators, Task 1 comprises 27,625 proposals, while the primary Task 2 setting ($M = 16$) allocates 318,955 proposal slots. Each generator retains its native pair-conditioned interface and numerical procedure within TREK's common coordinate and scoring interface. The comparison therefore controls proposal-slot budgets rather than requiring a common architecture, solver, or wall-clock cost. Structural recovery is evaluated at $\tau \in \{0.20, 0.50\}$ Å using the pair-aware matching and recovery metrics defined in Section 3.4. Reaction-bootstrap intervals and Holm-corrected paired comparisons are reported in Appendix D.2; boldface denotes a point-estimate maximum, not statistical significance.

4.2 Task 1: Known-Pair Generation Is Nearly Saturated

At 0.50 Å, RobustTSGen, React-OT, and TS-DFM achieve held-out pair-hit rates of 95.6%, 94.0%, and 94.0%, respectively (Table 2). Their pairwise differences are not statistically resolved after Holm correction. At the stricter 0.20 Å threshold, the corresponding rates are 69.9%, 74.9%, and 68.3%, distinguishing geometric precision within this leading group. These results establish a strong local-generation baseline while showing that model ordering depends on the matching tolerance. Full point estimates and all-pair hit rates are reported in Appendix D.1.

A pair hit establishes one local match; even an all-pair hit requires only one match for each productive pair, not recovery of its full TS set. Across the full inventory, one proposal per pair imposes a pair–channel-link recall ceiling of $5{,}525/19{,}747 = 27.98\%$, even for a perfect generator. This bound uses pair–channel links, not the reaction-unique denominator of Task 2. TREK therefore distinguishes access to a suitable structure from coverage of the alternatives associated with a reaction.

4.3 Task 2: Reaction-Level Ensemble Recovery Remains Incomplete

Under Top-8 and round-robin $M = 16$, RobustTSGen, React-OT, and TS-DFM achieve test micro recalls of 81.7%, 80.7%, and 81.2% at 0.50 Å, with complete-reference-set recovery of 75.2%,

Table 2: **Local hits and ensemble recovery across the full inventory and held-out test.** All: 3,493 reactions, 5,525 productive pairs, and 19,647 reference channels. Test: the held-out subset of 109 reactions, 183 pairs, and 591 channels. Values are percentages; All includes Test and is descriptive, not a held-out estimate. Task 2 uses Top-8 and round-robin $M = 16$. Recall is micro recall; all-pair hit requires a hit for every productive pair, whereas Complete requires every reference channel. **Bold** marks the highest displayed value within each set, not significance; 95% intervals are in Appendix D.2.

		Task 1: known pair			Task 2: ensemble recovery		
Generator	Set	Pair hit @0.20 Å	Pair hit @0.50 Å	All-pair hit @0.50 Å	Recall @0.20 Å	Recall @0.50 Å	Complete @0.50 Å
RobustTSGen	All	**73.0**	**96.1**	**94.4**	**53.8**	**83.4**	**79.6**
	Test	69.9	**95.6**	**94.5**	51.6	**81.7**	75.2
React-OT	All	72.7	94.6	92.4	**53.8**	82.0	78.4
	Test	**74.9**	94.0	91.7	**56.5**	80.7	74.3
TS-DFM	All	71.2	93.6	91.2	52.0	80.7	77.7
	Test	68.3	94.0	90.8	51.1	81.2	**76.1**
MolGEN	All	34.4	52.7	44.2	49.1	71.4	67.6
	Test	33.9	50.3	37.6	45.7	68.0	65.1
OA-ReactDiff	All	23.2	39.2	33.0	33.5	55.3	51.5
	Test	30.1	41.5	33.9	37.6	58.0	51.4

Evaluated scale: five generators; 27,625 Task 1 proposals and 318,955 Task 2 slots. Both thresholds score the same outputs.

74.3%, and 76.1%, respectively (Table 2). RobustTSGen has the highest micro-recall point estimate, while TS-DFM has the highest completeness; these leading models are not statistically separated after Holm correction. The distinction between high pair-hit rates and lower complete-set recovery exposes the additional demand of covering a reaction's alternatives together. The full five-model comparison and reaction-macro results appear in Appendix D.1.

Pair coverage and within-pair depth also differ. Over the full inventory at 0.20 Å, MolGEN hits the largest fraction of productive pairs (74.46%) but recovers 49.15% of reference channels. Finding a TS for many endpoint pairs therefore does not establish coverage of the target sets attached to them. Pair-aware one-to-one scoring rewards proposals that add an unmatched target, rather than repeated hits to an already recovered structure. This makes useful diversity measurable through reference coverage, rather than coordinate dispersion alone.

Proposal budgets expose a complementary allocation opportunity. For RobustTSGen over all 3,493 reactions, 0.50 Å recall rises from 26.46% to 62.88%, 75.24%, and 83.39% as M increases from 0 to 4, 8, and 16. The first four top-ups per reaction provide the largest gain, with diminishing marginal yield thereafter. Appendix D.4 gives proposal counts, marginal yields, and the novelty-adaptive comparison. These frozen-output comparisons measure coverage per proposal, not wall-time speedup.

4.4 Error Analysis: Where Does Ensemble Recovery Fail?

Figure 4 summarizes the stagewise accounting (a), shared failures (b), and illustrative geometric discrepancies (c,d).

The bottleneck shifts with tolerance. For RobustTSGen, geometry/diversity accounts for 76.93% of unrecovered channels at 0.20 Å, whereas output capacity accounts for 54.35% at 0.50 Å. At the latter tolerance, 93.34% of geometry-free addressable targets are recovered. With selection and allocation fixed, Equation 2 distinguishes geometric precision from allocation opportunity; denominators and detailed counts appear in Appendix D.4.

Generators provide complementary coverage. At 0.50 Å, the five-model union recovers 18,078 of 19,647 channels and completes 3,082 reactions, including 133 completed by no individual generator. This is complementarity across five budgets, not an equal-cost gain. All productive pairs of the 38 shared zero-hit reactions survive selection; 37 also remain unsuccessful in Task 1, ruling out Top-8 exclusion (Appendix D.5).

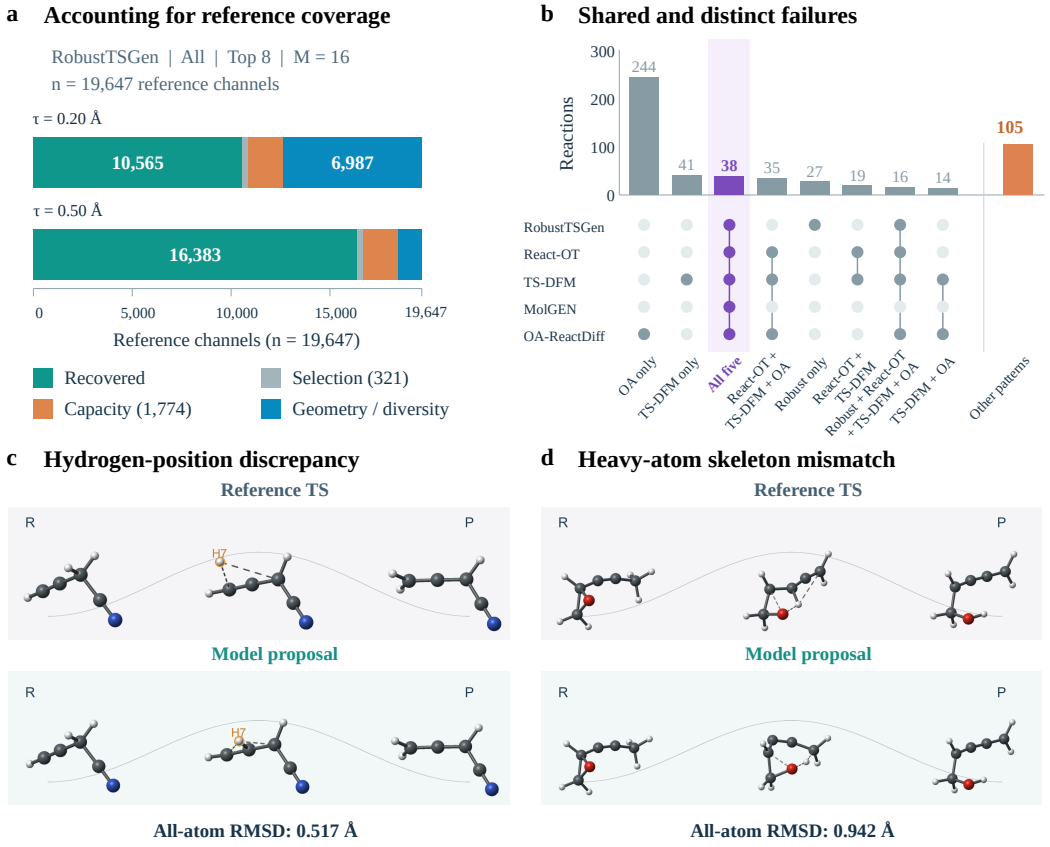

Figure 4: **TREK diagnoses the opportunities for better ensemble recovery.** (a) RobustTSGen recovery and stagewise losses over 19,647 references. (b) Zero-hit intersections at 0.50 Å: 539 reactions are missed by at least one generator and 38 by all five; remaining patterns are aggregated. (c,d) Illustrative hydrogen-position and heavy-atom skeleton discrepancies, with all-atom RMSDs of 0.517 and 0.942 Å, respectively. Paths are schematic, not IRC energy profiles.

Shared failures motivate targeted geometry diagnostics. The 38 shared failures have greater median endpoint RMSD (0.940 versus 0.763 Å) and distance-matrix change (0.679 versus 0.540 Å); both survive Holm correction. With only one candidate pair in 36 cases, this associates failure with rearrangement rather than pair-search ambiguity. Full feature tests and reference-observability strata are in Appendix D.5; these associations do not establish a single failure mechanism.

5 Conclusion

TREK establishes reaction-level TS-ensemble recovery as a budgeted benchmark with structured endpoint inputs, pair-aware set scoring, and physically audited reference construction. Its two tasks connect accurate generation for a known pair to the recovery of distinct alternatives across a reaction. The five-generator study shows how this perspective yields actionable conclusions: matching tolerance changes the balance between geometric precision and output capacity, proposal budgets define a measurable coverage trade-off, and cross-model recovery reveals both complementary discoveries and concentrated geometric challenges. These findings position ensemble coverage, allocation, and useful proposal diversity as explicit objectives for the next generation of TS generators. TREK provides the reference resource and evaluation protocol needed to measure progress toward that goal.

AI USE STATEMENT

We used large language model tools to assist with manuscript drafting and language editing, literature searches and reference checks, and LaTeX typesetting, table preparation, and figure preparation. Additionally, AI assistance included writing scripts for document consistency checks and figure export. The authors reviewed the AI-assisted text and artifacts and verified the numerical results reported in the manuscript against the original experimental records. The authors take full responsibility for the final manuscript, including its scientific claims, references, code, and other associated artifacts.

REFERENCES

Christoph Bannwarth, Sebastian Ehlert, and Stefan Grimme. Gfn2-xtb-an accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *Journal of chemical theory and computation*, 15(3):1652–1671, 10 2018. ISSN 1549-9618. doi: 10.1021/acs.jctc.8b01176. URL https://pubs.acs.org/jctcce/article/15/3/1652/975266/GFN2-xTB-An-Accurate-and-Broadly-Parametrized-Self.

Christoph Bannwarth, Sebastian Ehlert, and Stefan Grimme. GFN2-xTB: An accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *Journal of Chemical Theory and Computation*, 15(3):1652–1671, 2019. doi: 10.1021/acs.jctc.8b01176. URL https://doi.org/10.1021/acs.jctc.8b01176.

Junwei Lucas Bao and Donald G. Truhlar. Variational transition state theory: theoretical framework and recent developments. *Chemical Society reviews*, 46(24):7548–7596, 12 2017. ISSN 0306-0012. doi: 10.1039/c7cs00602k. URL https://pubs.rsc.org/cs/article/46/24/7548-7596/1245879.

Andrew Behn, Paul M. Zimmerman, Alexis T. Bell, and Martin Head-Gordon. Efficient exploration of reaction paths via a freezing string method. *The Journal of chemical physics*, 135(22):224108–224108, 12 2011. ISSN 0021-9606. doi: 10.1063/1.3664901. URL https://pubs.aip.org/jcp/article/135/22/224108/190831/Efficient-exploration-of-reaction-paths-via-a.

Jan Gerit Brandenburg, Christoph Bannwarth, Andreas Hansen, and Stefan Grimme. B97-3c: A revised low-cost variant of the b97-d density functional method. *The Journal of chemical physics*, 148(6):064104–064104, 2 2018a. ISSN 0021-9606. doi: 10.1063/1.5012601. URL https://pubs.aip.org/jcp/article/148/6/064104/196461/B97-3c-A-revised-low-cost-variant-of-the-B97-D.

Jan Gerit Brandenburg, Christoph Bannwarth, Andreas Hansen, and Stefan Grimme. B97-3c: A revised low-cost variant of the B97-D density functional method. *The Journal of Chemical Physics*, 148(6):064104, 2018b. doi: 10.1063/1.5012601. URL https://doi.org/10.1063/1.5012601.

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

Edward E Cureton. Rank-biserial correlation. *Psychometrika*, 21(3):287–290, 1956.

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*, 8 2026. ISSN 2662-8457. doi: 10.1038/s43588-026-01034-5. URL https://www.nature.com/articles/s43588-026-01034-5.

Amanda L. Dewyer, Alonso J. Argüelles, and Paul M. Zimmerman. Methods for exploring reaction space in molecular systems. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 8(2), 3 2018. ISSN 1759-0876. doi: 10.1002/wcms.1354. URL https://wires.onlinelibrary.wiley.com/doi/10.1002/wcms.1354.

Chenru Duan, Yuanqi Du, Haojun Jia, and Heather J Kulik. Accurate transition state generation with an object-aware equivariant elementary reaction diffusion model. *Nature computational science*, 3(12):1045–1055, 2023.

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, 4 2025. ISSN 2522-5839. doi: 10.1038/s42256-025-01010-0. URL https://www.nature.com/articles/s42256-025-01010-0.

Weinan E and Eric Vanden-Eijnden. Transition-path theory and path-finding algorithms for the study of rare events. *Annual review of physical chemistry*, 61(1):391–420, 3 2010. ISSN 0066-426X. doi: 10.1146/annurev.physchem.040808.090412. URL https://www.annualreviews.org/doi/10.1146/annurev.physchem.040808.090412.

Henry Eyring. The activated complex in chemical reactions. *Journal of Chemical Physics*, 3(2):107–115, 2 1935a. ISSN 0021-9606. doi: 10.1063/1.1749604. URL https://pubs.aip.org/jcp/article/3/2/107/203352/The-Activated-Complex-in-Chemical-Reactions.

Henry. Eyring. The activated complex and the absolute rate of chemical reactions. *Chemical Reviews*, 17(1):65–77, 8 1935b. ISSN 0009-2665. doi: 10.1021/cr60056a006. URL https://pubs.acs.org/doi/abs/10.1021/cr60056a006.

Ronald A Fisher. On the interpretation of χ 2 from contingency tables, and the calculation of p. *Journal of the royal statistical society*, 85(1):87–94, 1922.

Kenichi Fukui. The path of chemical reactions - the irc approach. *Accounts of Chemical Research*, 14(12):363–368, 12 1981a. ISSN 0001-4842. doi: 10.1021/ar00072a001. URL https://pubs.acs.org/doi/abs/10.1021/ar00072a001.

Kenichi Fukui. The path of chemical reactions—the irc approach. *Acc. Chem. Res*, 14(12):471–476, 1981b.

Leonard Galustian, Konstantin Mark, Johannes Karwounopoulos, Maximilian P-P Kovar, and Esther Heid. Goflow: efficient transition state geometry prediction with flow matching and e (3)-equivariant neural networks. *Digital Discovery*, 4(12):3492–3501, 2025.

Carlos Gonzalez and H Bernhard Schlegel. An improved algorithm for reaction path following. *The Journal of Chemical Physics*, 90(4):2154–2161, 1989a.

Carlos Gonzalez and H. Bernhard Schlegel. An improved algorithm for reaction path following. *Journal of Chemical Physics*, 90(4):2154–2161, 2 1989b. ISSN 0021-9606. doi: 10.1063/1.456010. URL https://pubs.aip.org/jcp/article/90/4/2154/463846/An-improved-algorithm-for-reaction-path.

Stefan Grimme. Exploration of chemical compound, conformer, and reaction space with metadynamics simulations based on tight-binding quantum chemical calculations. *Journal of chemical theory and computation*, 15(5):2847–2862, 2 2019. ISSN 1549-9618. doi: 10.1021/acs.jctc.9b00143. URL https://pubs.acs.org/jctcce/article/15/5/2847/975528/Exploration-of-Chemical-Compound-Conformer-and.

Graeme Henkelman and Hannes Jónsson. Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points. *The Journal of chemical physics*, 113(22):9978–9985, 2000.

Graeme Henkelman, Blas P. Uberuaga, and Hannes Jónsson. A climbing image nudged elastic band method for finding saddle points and minimum energy paths. *Journal of Chemical Physics*, 113(22):9901–9904, 11 2000. ISSN 0021-9606. doi: 10.1063/1.1329672. URL https://pubs.aip.org/jcp/article/113/22/9901/185050/A-climbing-image-nudged-elastic-band-method-for.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pp. 65–70, 1979.

Peter Hänggi, Peter Talkner, and Michal Borkovec. Reaction-rate theory: fifty years after kramers. *Reviews of Modern Physics*, 62(2):251–341, 4 1990. ISSN 0034-6861. doi: 10.1103/RevModPhys.62.251. URL https://link.aps.org/doi/10.1103/RevModPhys.62.251.

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.

W. Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 32(5):922–923, 9 1976. ISSN 0567-7394. doi: 10.1107/S0567739476001873. URL https://journals.iucr.org/paper?S0567739476001873.

Seonghwan Kim, Jeheon Woo, and Woo Youn Kim. Diffusion-based generative ai for exploring transition states from 2d molecular graphs. *Nature Communications*, 15(1):341, 2024.

H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 3 1955. ISSN 0028-1441. doi: 10.1002/nav.3800020109. URL https://onlinelibrary.wiley.com/doi/10.1002/nav.3800020109.

Joel Lehman and Kenneth O. Stanley. Abandoning objectives: Evolution through the search for novelty alone. *Evolutionary Computation*, 19(2):189–223, 6 2011. ISSN 1063-6560. doi: 10.1162/EVCO_a_00025. URL https://direct.mit.edu/evco/article/19/2/189-223/1365.

Yufei Luo, Xiang Gu, and Jian Sun. Generative flow model on distance geometry for predicting transition states of chemical reactions. *Nature Communications*, 17(1), 6 2026. ISSN 2041-1723. doi: 10.1038/s41467-026-74101-0. URL https://www.nature.com/articles/s41467-026-74101-0.

H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18(1):50–60, 3 1947. ISSN 0003-4851. doi: 10.1214/aoms/1177730491. URL http://projecteuclid.org/euclid.aoms/1177730491.

Frank Neese. Software update: The orca program system—version 5.0. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 12(5), 3 2022a. ISSN 1759-0876. doi: 10.1002/wcms.1606. URL https://wires.onlinelibrary.wiley.com/doi/10.1002/wcms.1606.

Frank Neese. Software update: The ORCA program system—version 5.0. *WIREs Computational Molecular Science*, 12(5):e1606, 2022b. doi: 10.1002/wcms.1606. URL https://doi.org/10.1002/wcms.1606.

Frank Neese, Frank Wennmohs, Ute Becker, and Christoph Riplinger. The orca quantum chemistry program package. *The Journal of chemical physics*, 152(22):224108–224108, 6 2020. ISSN 0021-9606. doi: 10.1063/5.0004608. URL https://pubs.aip.org/jcp/article/152/22/224108/1061982/The-ORCA-quantum-chemistry-program-package.

Quan Nguyen and Roman Garnett. Nonmyopic multiclass active search with diminishing returns for diverse discovery. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 5231–5249. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/nguyen23d.html.

S. Nikolić, N. Trinajstić, Z. Mihalić, and S. Carter. On the geometric-distance matrix and the corresponding structural invariants of molecular systems. *Chemical Physics Letters*, 179(1-2):21–28, 4 1991. ISSN 0009-2614. doi: 10.1016/0009-2614(91)90285-H. URL https://linkinghub.elsevier.com/retrieve/pii/000926149190285H.

Philipp Pracht, Fabian Bohle, and Stefan Grimme. Automated exploration of the low-energy chemical space with fast quantum chemical methods. *Physical chemistry chemical physics : PCCP*, 22(14):7169–7192, 2 2020a. ISSN 1463-9076. doi: 10.1039/c9cp06869d. URL https://pubs.rsc.org/cp/article/22/14/7169-7192/640618.

Philipp Pracht, Fabian Bohle, and Stefan Grimme. Automated exploration of the low-energy chemical space with fast quantum chemical methods. *Physical Chemistry Chemical Physics*, 22:7169–7192, 2020b. doi: 10.1039/C9CP06869D. URL https://doi.org/10.1039/C9CP06869D.

Philipp Pracht, Stefan Grimme, Christoph Bannwarth, Fabian Bohle, Sebastian Ehlert, Gereon Feldmann, Johannes Gorges, Marcel Müller, Tim Neudecker, Christoph Plett, Sebastian Spicher, Pit Steinbach, Patryk A. Wesołowski, and Felix Zeller. Crest—a program for the exploration of low-energy molecular chemical space. *The Journal of Chemical Physics*, 160(11), 3 2024. ISSN 0021-9606. doi: 10.1063/5.0197592. URL https://pubs.aip.org/jcp/article/160/11/114110/3278084/CREST-A-program-for-the-exploration-of-low-energy.

Jiangjie Qiu, Wentao Li, Honghao Chen, Leyi Zhao, and Xiaonan Wang. Adsorbflow: Energy-conditioned deterministic transport for catalyst screening in artificial intelligence-driven materials-reaction-process design. *Chinese Journal of Chemical Engineering*, 2026. ISSN 1004-9541. doi: https://doi.org/10.1016/j.cjche.2026.08.026. URL https://www.sciencedirect.com/science/article/pii/S1004954126004027.

Stefan P. Schmid, Henrik Seng, Thibault Kläy, and Kjell Jorner. Rapid generation of transition-state conformer ensembles via constrained distance geometry. *Journal of Chemical Information and Modeling*, 66(5):2777–2790, 2 2026. ISSN 1549-9596. doi: 10.1021/acs.jcim.5c02794. URL https://pubs.acs.org/doi/10.1021/acs.jcim.5c02794.

Mathias Schreiner, Arghya Bhowmik, Tejs Vegge, Jonas Busk, and Ole Winther. Transition1x-a dataset for building generalizable reactive machine learning potentials. *Scientific Data*, 9(1):779, 2022.

Donald G Truhlar, Bruce C Garrett, and Stephen J Klippenstein. Current status of transition-state theory. *The Journal of physical chemistry*, 100(31):12771–12800, 1996.

Ping Tuo, Jiale Chen, and Ju Li. Flow matching for reaction pathway generation. *Nature Communications*, 17(1), 7 2026. ISSN 2041-1723. doi: 10.1038/s41467-026-75654-w. URL https://www.nature.com/articles/s41467-026-75654-w.

Jan P. Unsleber and Markus Reiher. The exploration of chemical reaction networks. *Annual Review of Physical Chemistry*, 71(1):121–142, 2 2020. ISSN 0066-426X. doi: 10.1146/annurev-physchem-071119-040123. URL https://www.annualreviews.org/doi/10.1146/annurev-physchem-071119-040123.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 6 1927. ISSN 0162-1459. doi: 10.1080/01621459.1927.10502953. URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1927.10502953.

Li-Cheng Xu, Junyi An, Weiqi Liu, Yun-Fei Shi, Chong-Lei Ji, Fenglei Cao, and Yuan Qi. Automated transition state generation for mechanistic exploration in organic synthesis. *Nature Communications*, 8 2026. ISSN 2041-1723. doi: 10.1038/s41467-026-77230-8. URL https://www.nature.com/articles/s41467-026-77230-8.

Qiyuan Zhao, Hsuan-Hao Hsu, and Brett M Savoie. Conformational sampling for transition state searches on a computational budget. *Journal of Chemical Theory and Computation*, 18(5):3006–3016, 2022.

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

Paul Zimmerman. Reliable transition state searches integrated with the growing string method. *Journal of chemical theory and computation*, 9(7):3043–3050, 2013.

A REFERENCE INVENTORY AND DATA PARTITION

Table 3 gives the source-preserving partition, following the reaction-structured organization of the source data (Schreiner et al., 2022). A reaction is the grouping unit throughout construction, evaluation, and resampling. The retained endpoint sets reproduce the geometries supporting the accepted R/P–TS records; all combinations of these endpoints are made available to the selector. The 5,525 observed productive pairs are therefore included in the complete candidate input. Candidate pairs with no observed reference are retained as unknown combinations.

Table 3: TREK inventory by source split. "Channels" denotes reaction-unique xTB proxies; "TS families" denotes reaction-level geometry families.

Split	Reactions	Candidate pairs	Productive pairs	Channels	TS families
Train	3,317	10,194	5,212	18,487	10,498
Validation	67	299	130	569	345
Test	109	372	183	591	331
Total	3,493	10,865	5,525	19,647	11,174

A.1 CONSTRUCTION AND REFERENCE IDENTITY

The source cohort contains 5,000 reactions drawn from Transition1x (Schreiner et al., 2022), a large-scale reactive dataset constructed from DFT reaction pathways and transition-state-region configurations. Starting from the atom-mapped reactant and product geometries, we use CREST (Grimme, 2019; Pracht et al., 2020a; 2024) to expand the endpoint conformational ensembles and GFN2-xTB (Bannwarth et al., 2018) for efficient semiempirical evaluation during pathway exploration. Candidate reaction paths are searched between retained endpoint conformers and subsequently refined toward saddle-point structures, following the general principle of double-ended reaction-path and transition-state searches (Henkelman & Jónsson, 2000; Zimmerman, 2013; Henkelman et al., 2000; Behn et al., 2011)

High-confidence acceptance requires consistency with standard transition-state criteria: a candidate must exhibit first-order saddle character, with a single imaginary vibrational mode corresponding to motion along the intended reaction coordinate, and its positive and negative displacements must commit toward the expected reactant and product basins (Fukui, 1981a; Gonzalez & Schlegel, 1989b). We implement two complementary acceptance profiles to establish endpoint identity. The standard profile emphasizes heavy-atom endpoint consistency, whereas, when heavy-atom displacement is small, a hydrogen-sensitive profile additionally compares exact atom-mapped bond changes to disambiguate endpoints. Accepted R/P–TS records define the retained endpoint geometries used by TREK: the candidate input is the complete support of these accepted records rather than a newly resampled or independently truncated CREST pool.

Let $\mathcal{R}_i = \{R_{ia}\}_{a=1}^{n_i^R}$ and $\mathcal{P}_i = \{P_{ib}\}_{b=1}^{n_i^P}$ share atomic numbers and atom correspondence. Then $\mathcal{E}_i = \mathcal{R}_i \times \mathcal{P}_i$, and the reference relation $\mathcal{L}_i \subseteq \mathcal{E}_i \times \mathcal{C}_i$ associates pairs with proxy channels and their TS geometries X_c. The frozen inventory satisfies

$$\sum_i |\mathcal{E}_i| = 10{,}865, \qquad \sum_i |\mathcal{L}_i| = 19{,}747, \qquad \sum_i |\mathcal{C}_i| = 19{,}647. \tag{3}$$

Candidate pairs are counted by summing each reaction's Cartesian product, not by multiplying globally summed endpoint counts. A channel supported by several pairs contributes several links but one reaction-level target.

Proxy identity combines endpoint-family information, TS geometry family, atom mapping, and acceptance profile. Geometry-only grouping is reported separately; neither repeated acceptance nor membership in different proxy contexts alone establishes a different reaction mechanism. There are 5,167 retained R endpoints and 5,213 retained P endpoints; 5,525 candidate pairs have observed support and 5,340 remain unknown. The inventory includes up to 43 proxy targets per reaction.

Counting units. The target hierarchy contains 19,747 pair–channel links, 19,647 reaction-unique channels, 12,284 pair–geometry-family associations, and 11,174 reaction-level geometry families. Under the channel definition, 3,482 reactions and 4,633 productive pairs have multiple targets. Under the geometry-family definition, 2,273 reactions and 2,854 productive pairs have multiple families. A channel can retain a different endpoint or mapping context while sharing a TS geometry family with another channel. The primary matching uses the frozen channel definition consistently across all generators.

B GEOMETRIC SCORING AND ASSIGNMENT

For two geometries with n atoms, we use per-atom RMSD,

$$d(X, Y) = \min_{Q,t,\pi} \left(\frac{1}{n} \sum_{a=1}^{n} \|X_a - (QY_{\pi(a)} + t)\|_2^2 \right)^{1/2}, \tag{4}$$

where $Q \in SO(3)$ is a proper rotation, t is a translation, and π preserves atomic elements. The implementation searches permutations exhaustively when the number of equal-element permutations is below 10,000; otherwise, it uses the pinned genetic order matcher with a Hungarian fallback (Kuhn, 1955). The reported distance follows the per-atom normalization in Equation 4; it includes hydrogens and does not permit a global reflection. The permutation search is an operational approximation to the minimum for large permutation spaces.

Pair selection instead uses atom-mapped Kabsch alignment (Kabsch, 1976), with the same atom correspondence for R and P. Candidate pairs are ranked by this distance, with stable identifiers breaking ties. During scoring, the canonical retained-geometry index maps candidate pair identities to the observed productive pairs. This exact index is distinct from the element-permutation search used to compare TS shapes. Proposal–channel edges respect the indexed pair relation before applying the TS tolerance.

For reaction-level recovery, proposals with identical geometry hashes are collapsed before a maximum-cardinality assignment, with total RMSD as the secondary objective. Each reference channel and each retained proposal can be used at most once. The allocation ceiling u_i is computed by the same type of matching between reference channels and pair-specific proposal slots, omitting the geometric constraint. This handles channels linked to multiple pairs without double counting. The ceiling is an opportunity bound for the fixed allocation; the difference $u_i - m_i(\tau)$ measures unused opportunity due to the generated structures.

C GENERATOR CONFIGURATIONS AND ALLOCATION

The frozen runs use the official model sources and checkpoints through separate inference adapters. OA-ReactDiff (Duan et al., 2023) fixes the R/P fragments and generates the TS fragment with 150 diffusion timesteps, five resamplings, and jump length five. React-OT (Duan et al., 2025) uses the midpoint ODE solver with the configured NFE set to ten and absolute/relative tolerances of 0.01. RobustTSGen uses the CREST-medium pretrained, Transition1x-xTB fine-tuned checkpoint with the same midpoint/NFE-ten ODE configuration. TS-DFM (Luo et al., 2026) uses midpoint integration with step size 0.05 and weighted L-BFGS distance-to-coordinate reconstruction. MolGEN (Tuo et al., 2026) uses its recorded `dopri5` sampling configuration with `inference_steps=50`. These are the configurations evaluated in TREK; proposal calls preserve each generator's native inference behavior.

All models receive atom-mapped endpoint coordinates and atomic numbers. A common proper alignment places the endpoints in the model coordinate frame. Each call returns one coordinate proposal for its selected pair. The comparison equalizes proposal slots; model-specific numerical integration and coordinate reconstruction remain part of each generator. The one-shot task contains 27,625 emitted proposals across five models. The round-robin $M = 16$ task allocates 63,791 slots per model. The inference record retains model revision, checkpoint identity, pair identity, draw index, and seed alongside every result.

C.1 Round-robin allocation and inference boundary

For selected-pair rank $j \in \{1, \ldots, K_i\}$, the reaction-level budget and allocation are

$$B_i(M) = K_i + M, \qquad n_{ij}(M) = 1 + \left\lfloor \frac{M}{K_i} \right\rfloor + \mathbf{1}[j \leq M \bmod K_i]\,. \tag{5}$$

Here $K_i = \min(8, |\mathcal{E}_i|)$ and $M \in \{0, 4, 8, 16\}$. All selected pairs receive one initial proposal; top-ups cycle in rank order. For $K_i = 3$ and $M = 4$, the counts are $3, 2, 2$. Thus M is shared across the reaction, not multiplied by the number of pairs. Failed slots consume their allocation, and budget prefixes use the same underlying proposal sequence.

Only coordinates, atomic identities, selected-pair ranks, and generated outputs are exposed to the inference wrapper. Productive-pair assignments, reference TSs, channel identities, path records, and historical acceptance counts remain sealed until scoring. The wrapper reads reaction-level ensembles; the five generators retain their pair-conditioned interfaces. This isolates their ensemble-recovery behavior under the same selector and allocation rather than attributing native ensemble-input architectures to the models.

C.2 Novelty-adaptive allocation

Motivated by novelty- and diversity-aware search (Lehman & Stanley, 2011; Nguyen & Garnett, 2023), the novelty-adaptive policy maintains a reaction-level set of previously accepted novel proposals. Both allocation policies expose one initial proposal per selected pair. Round-robin then cycles through the ranked pairs for exactly M additional slots. The novelty-adaptive policy maintains a reaction-level set of previously accepted novel proposals. A new proposal is novel if its atom-mapped, properly aligned RMSD exceeds 0.10 Å from every member of that set. Nonfinite coordinates and atom pairs closer than 0.45 times the sum of their covalent radii fail the validity check. Two consecutive non-novel or invalid outcomes retire a pair from further allocation. Remaining pairs continue sharing the top-up budget; allocation stops if all pairs retire. This rule uses generated coordinates only.

The policies are evaluated as target-blind sequential exposures of a frozen proposal pool. Consequently, the budget curves measure proposals consumed by the policy. The lower exposure count of the adaptive policy characterizes its potential call savings; it is not a measurement of reduced wall time for constructing the original pool.

D Additional quantitative results

D.1 Full-inventory and held-out point estimates

Tables 4 and 5 report point estimates for the full inventory and source-test subset. **All** includes train, validation, and test; **Test** is the held-out subset within All, not a separate disjoint cohort. Each percentage is an observed summary of the fixed evaluated outputs; reaction-sampling uncertainty is reported separately in Appendix D.2. The main-text Table 2 combines six metrics across both tasks and sets. The task-specific tables below provide separate views and additionally report reaction-macro recall.

Local precision and all-pair coverage. At 0.20 Å, the Task 1 test pair-hit rates of RobustTSGen, React-OT, and TS-DFM are 69.9%, 74.9%, and 68.3%, respectively. React-OT has the largest point estimate at this stricter threshold, whereas RobustTSGen has the largest at 0.50 Å. Their test all-pair hit rates at 0.50 Å are 94.5%, 91.7%, and 90.8%, respectively. These outcomes establish reliable access to a nearby reference for many supplied pairs, but do not imply identical geometric precision or recovery of every target attached to those pairs. An all-pair hit requires one successful proposal for each productive pair, not a complete TS set. The statistical comparisons in Appendix D.2 distinguish point-estimate ordering from resolved model differences.

Reaction weighting and model profiles. Under Top-8 and round-robin $M = 16$ at 0.50 Å, RobustTSGen, React-OT, and TS-DFM attain test reaction-macro recalls of 87.5%, 87.0%, and 88.1%,

Table 4: **Task 1: known-pair generation.** Full-inventory and held-out hit rates (%); one proposal per productive pair. Bold marks the highest displayed point estimate within each set, not statistical significance. Test intervals are in Appendix D.2.

Generator	Set	Pair hit @0.20 Å	Pair hit @0.50 Å	All-pair hit @0.50 Å
RobustTSGen	All	**73.0**	**96.1**	**94.4**
	Test	69.9	**95.6**	**94.5**
React-OT	All	72.7	94.6	92.4
	Test	**74.9**	94.0	91.7
TS-DFM	All	71.2	93.6	91.2
	Test	68.3	94.0	90.8
MolGEN	All	34.4	52.7	44.2
	Test	33.9	50.3	37.6
OA-ReactDiff	All	23.2	39.2	33.0
	Test	30.1	41.5	33.9

All: 3,493 reactions / 5,525 productive pairs (train, validation, and test). **Test:** the 109-reaction / 183-pair held-out subset. **Evaluated scale:** 27,625 proposals across five generators.

Table 5: **Task 2: reaction-level ensemble recovery.** Top-8 selection with round-robin $M = 16$; 63,791 slots per generator. Values are percentages. Bold marks the highest displayed point estimate within each set, not statistical significance. Test intervals are in Appendix D.2.

Generator	Set	Micro recall @0.20 Å	Micro recall @0.50 Å	Macro recall @0.50 Å	Complete @0.50 Å
RobustTSGen	All	**53.8**	**83.4**	**90.0**	**79.6**
	Test	51.6	**81.7**	87.5	75.2
React-OT	All	**53.8**	82.0	88.8	78.4
	Test	**56.5**	80.7	87.0	74.3
TS-DFM	All	52.0	80.7	87.8	77.7
	Test	51.1	81.2	**88.1**	**76.1**
MolGEN	All	49.1	71.4	83.9	67.6
	Test	45.7	68.0	81.4	65.1
OA-ReactDiff	All	33.5	55.3	68.9	51.5
	Test	37.6	58.0	70.7	51.4

All: 3,493 reactions / 19,647 reference channels (train, validation, and test). **Test:** the 109-reaction / 591-channel held-out subset. **Evaluated scale:** 318,955 proposals across five generators.

respectively, compared with micro recalls of 81.7%, 80.7%, and 81.2%. Macro recall gives each reaction equal weight, while micro recall gives each reference channel equal weight; their difference reflects the distinct weighting of reactions with different target-set sizes. Complete recovery additionally requires all targets of a reaction to be recovered together. In the same setting, MolGEN achieves 68.0% test micro recall and 65.1% completeness, while OA-ReactDiff achieves 58.0% and 51.4%. The five-model comparison therefore describes distinct recovery profiles under a common selector, allocation, and scoring interface.

Productive-pair breadth and within-pair depth. Over the full inventory at 0.20 Å, MolGEN hits 74.46% of productive pairs, the largest fraction among the evaluated generators, while recovering 49.15% of reference channels. A system can distribute first hits across many pairs yet repeatedly cover a narrow subset of their associated targets. Conversely, a proposal that adds an unmatched target increases set recall even if its productive pair was already hit. Pair-aware one-to-one assignment makes this distinction explicit: useful diversity contributes to reference coverage, rather than merely increasing coordinate dispersion. The novelty-adaptive diagnostics in Appendix D.4 examine this distinction under sequential proposal exposure.

D.2 Held-out confidence intervals

Tables 6 and 7 accompany the full point estimates in Tables 4 and 5, including the held-out metrics selected for Table 2. Each Estimate row is the observed percentage on the fixed test set; its separately labeled 95% CI row reports a reaction-bootstrap interval. We resample 109 whole reactions with replacement for each of 2,000 replicates, carrying all pairs, channels, and frozen model outcomes belonging to each sampled reaction together. Each metric is recomputed on the resampled cohort, including its denominator. This preserves the dependence among multiple targets within a reaction.

Table 6: **Task 1: held-out estimates and reaction-bootstrap 95% intervals.** Test cohort: 109 reactions and 183 productive pairs. Values are percentages; estimates and intervals occupy separately labeled rows.

Generator	Statistic	Pair hit @0.20 Å	Pair hit @0.50 Å	All-pair hit @0.50 Å
RobustTSGen	Estimate	69.9	95.6	94.5
	95% CI	[62.1, 77.3]	[90.9, 98.9]	[89.9, 98.2]
React-OT	Estimate	74.9	94.0	91.7
	95% CI	[66.7, 82.1]	[89.5, 97.5]	[86.2, 96.3]
TS-DFM	Estimate	68.3	94.0	90.8
	95% CI	[59.4, 76.8]	[89.9, 97.4]	[84.4, 96.3]
MolGEN	Estimate	33.9	50.3	37.6
	95% CI	[26.9, 41.0]	[42.5, 58.1]	[28.4, 46.8]
OA-ReactDiff	Estimate	30.1	41.5	33.9
	95% CI	[23.2, 37.2]	[33.3, 49.4]	[24.8, 43.1]

The intervals describe uncertainty associated with reaction sampling conditional on the fixed generated outputs; they do not represent repeated model runs or variability across generation seeds. All rows in these two supplementary tables refer to Test, which is a subset of All rather than an independent evaluation cohort.

Paired comparisons use 10,000 reaction-level sign permutations, with Holm correction over the ten model pairs within each task, metric, and threshold. At 0.50 Å, the three pairwise comparisons among RobustTSGen, React-OT, and TS-DFM all have Holm-adjusted $p = 1.0$ for Task 1 pair hit and for Task 2 micro recall and completeness. These tests do not resolve within-group differences on the evaluated cohort; they are not tests establishing model equivalence. The frozen comparison record uses the same 109 reaction clusters and 2,000 bootstrap replicates as the interval tables.

Table 7: **Task 2: held-out estimates and reaction-bootstrap 95% intervals.** Top-8, round-robin $M = 16$; 109 test reactions and 591 reference channels. Values are percentages; estimates and intervals occupy separately labeled rows.

Generator	Statistic	Micro recall @0.20 Å	Micro recall @0.50 Å	Macro recall @0.50 Å	Complete @0.50 Å
RobustTSGen	Estimate	51.6	81.7	87.5	75.2
	95% CI	[44.0, 59.9]	[75.1, 88.0]	[82.4, 92.2]	[67.0, 82.6]
React-OT	Estimate	56.5	80.7	87.0	74.3
	95% CI	[48.4, 65.3]	[74.1, 87.1]	[81.7, 91.7]	[66.1, 82.6]
TS-DFM	Estimate	51.1	81.2	88.1	76.1
	95% CI	[43.2, 59.5]	[74.7, 87.8]	[83.1, 92.3]	[68.8, 83.5]
MolGEN	Estimate	45.7	68.0	81.4	65.1
	95% CI	[37.7, 54.5]	[60.1, 77.1]	[75.7, 86.9]	[56.0, 74.3]
OA-ReactDiff	Estimate	37.6	58.0	70.7	51.4
	95% CI	[30.6, 45.3]	[50.6, 66.4]	[64.2, 77.1]	[42.2, 60.6]

D.3 Whole-inventory performance

Table 8 supplies the descriptive aggregate underlying the dataset-wide analyses. The supplementary point-estimate tables align full-inventory and held-out results in adjacent rows for each generator (Tables 4 and 5). The link recall in Task 1 has a different denominator from the pair-hit rate and from reaction-unique channel recall in Task 2.

Table 8: Whole-inventory results (percent). Each cell gives @0.20 / @0.50 Å. Task 1 pair hit and link recall use 5,525 pairs and 19,747 links. Task 2 recall and completeness use 19,647 channels and 3,493 reactions.

	Task 1		Task 2: Top-8, $M = 16$	
Generator	Pair hit	Link recall	Channel recall	Complete recovery
RobustTSGen	72.98 / 96.13	20.42 / 26.90	53.77 / 83.39	47.09 / 79.56
React-OT	72.67 / 94.61	20.33 / 26.47	53.80 / 82.02	46.38 / 78.41
TS-DFM	71.19 / 93.63	19.92 / 26.20	52.01 / 80.69	45.23 / 77.73
MolGEN	34.41 / 52.69	9.63 / 14.74	49.15 / 71.45	43.52 / 67.62
OA-ReactDiff	23.17 / 39.19	6.48 / 10.96	33.53 / 55.32	29.49 / 51.50

D.4 Selector and budget curves

The minimum-endpoint-RMSD selector retains 62.01%, 78.19%, 91.29%, 97.92%, and 99.73% of observed productive pairs at Top-1, 2, 4, 8, and 16 over the full inventory. The corresponding test recalls are 58.47%, 75.41%, 90.16%, 97.27%, and 100%. At Top-8, minimum distance-matrix change retains 96.27% overall, stable random selection 88.24%, maximum distance-matrix change 81.45%, and maximum endpoint RMSD 80.36%. The minimum-RMSD choice offers high observed-support recall with a small selected set. These comparisons measure retention of observed productive pairs; unobserved combinations remain unknown, not negative examples.

Coverage per additional proposal. Table 9 reports the proposal counts, recovered channels, and marginal yields for the shared-budget sweep. The first four top-up proposals per reaction recover 7,156 extra channels. Increasing M from 4 to 8 yields 173.7 new channels per thousand additional proposals; increasing it from 8 to 16 yields 57.3, adding 1,601 channels for 27,944 proposals. The novelty-adaptive policy exposes 23,377 RobustTSGen proposals, 36.65% of the primary

Table 9: RobustTSGen budget scaling over all 3,493 reactions at 0.50 Å. Complete reactions require all reference channels; marginal yield is newly recovered channels per 1,000 additional proposals relative to the preceding budget.

M	Proposals	Recovered channels	Recall (%)	Complete reactions	Marginal yield
0	7,903	5,199	26.46	12	—
4	21,875	12,355	62.88	1,936	512.2
8	35,847	14,782	75.24	2,470	173.7
16	63,791	16,383	83.39	2,779	57.3

round-robin allocation, with 60.54% recall. For comparison, round-robin at $M = 4$ exposes 21,875 proposals and reaches 62.88% recall. Thus, coordinate-novelty stopping does not identify the most productive coverage frontier in this comparison. This is a coverage–proposal trade-off measured through sequential exposure of the frozen pool, not a measured wall-time speedup.

Coordinate novelty and useful target discovery differ. OA-ReactDiff marks 99.6% of its round-robin outputs as novel under the rule in Appendix C.2 and attains 55.32% channel recall, whereas RobustTSGen attains 83.39% recall with 11.9% marked novel. An allocator therefore needs a signal of coverage-relevant diversity rather than coordinate novelty alone.

Tolerance-dependent attribution. The selected 7,903 pairs support 19,326 of 19,647 reference channels. With the primary allocation, 17,552 channels are addressable when geometry is ignored; RobustTSGen recovers 93.34% of this opportunity at 0.50 Å. Table 10 holds the proposals and pair-specific slots fixed across tolerances: selection and capacity losses remain 321 and 1,774, while geometry/diversity loss changes from 6,987 to 1,169. The percentages are shares of the unrecovered channels, not of the full inventory. The larger capacity share at the broader tolerance therefore does not indicate that the allocation itself became worse.

Table 10: Stagewise loss for RobustTSGen, Top-8/$M = 16$, over the complete reference inventory. Parentheses give each stage's share of unrecovered channels.

Stage	@0.20 Å	@0.50 Å
Selection	321 (3.53%)	321 (9.83%)
Allocation/cardinality	1,774 (19.53%)	1,774 (54.35%)
Geometry/diversity	6,987 (76.93%)	1,169 (35.81%)
Total unrecovered	9,082	3,264

Reaction-profile stratification. For the hydrogen-sensitive profile, the RobustTSGen Task 1 pair-hit rate is 97.32%, compared with 95.22% for the standard profile, whereas Task 2 recalls are 80.41% and 84.74%. The larger pair and target multiplicities of the hydrogen-sensitive group create additional allocation demand: ready access to one structure need not imply complete recovery of its reaction-level alternatives.

D.5 Hard-core descriptors and reference observability

Complementarity and exact failure intersections. The five-model union at 0.50 Å covers 18,078 of 19,647 reference channels and completes 3,082 of 3,493 reactions. Of these complete reactions, 133 require complementary discoveries from multiple generators. This is a five-budget diagnostic, not a comparison against a single model at equal cost. Thirty nonempty exact zero-hit intersection patterns cover 539 reactions; the other 2,954 receive a hit from each model. The all-five intersection contains 38 reactions, all with their productive pairs retained at Top-8. Of these, 37 also receive no hit from any generator in Task 1, and the cohort comprises 37 source-training reactions and one source-test reaction.

The 2,954-reaction count describes agreement on obtaining at least one hit, not complete-set recovery. Exact zero-hit intersections separate model-specific from shared failures, whereas recovered-target unions measure complementary coverage. The common-failure cohort is defined by these fixed models, budgets, and reference matches, rather than by an assertion of intrinsic chemical impossibility.

Endpoint rearrangement and feature tests. For a mapped endpoint pair with distance matrices D_R and D_P, we define an internal-coordinate descriptor based on pairwise interatomic distances, which provide a rigid-motion-invariant representation of molecular geometry (Nikolić et al., 1991),

$$\Delta_D(R, P) = \left[\frac{1}{n^2} \sum_{a,b=1}^{n} \left((D_R)_{ab} - (D_P)_{ab} \right)^2 \right]^{1/2} . \tag{6}$$

We summarize both endpoint RMSD and Δ_D by the median over the observed productive pairs of a reaction. Continuous feature comparisons between the 38 common failures and the other 3,455 reactions use two-sided Mann–Whitney tests (Mann & Whitney, 1947); categorical comparisons use Fisher's exact test (Fisher, 1922). Holm correction is applied across the inferential feature comparisons (Holm, 1979). Median productive-endpoint RMSD is 0.940 versus 0.763 Å (Holm-adjusted $p = 0.0117$), and median distance-matrix change is 0.679 versus 0.540 Å (adjusted $p = 6.84 \times 10^{-5}$). Molecular size and reference multiplicity are not significantly enriched after correction. The rank-biserial effect is 0.316 for endpoint RMSD and 0.436 for distance-matrix change (Cureton, 1956). Single-pair reactions account for 36/38 common failures versus 2,392/3,455 other reactions (odds ratio 8.00; adjusted $p = 0.0046$).

Table 11 reports all feature tests, including nonsignificant comparisons and separate outcome checks. Geometry-derived bond descriptors are heuristic summaries, not independently certified mechanism labels.

Reference observability. Because transition-state search success can depend strongly on the sampled reaction conformations (Zhao et al., 2022), historical search support supplies another descriptive view of target observability. For RobustTSGen at 0.50 Å, channels observed once during reference construction have 74.5% recall, while those observed at least 65 times have 92.7% recall. The intermediate strata of 2–4, 5–16, and 17–64 observations have recalls of 81.9%, 88.1%, and 82.9%, respectively. These counts describe reference-search observability. The benchmark weights each reference channel once, independently of its discovery frequency, and keeps the historical counts sealed from inference. Repeated recovery of frequently observed targets therefore cannot substitute for coverage of less frequently observed alternatives.

Table 11: **Complete shared-failure feature comparison.** The hard cohort contains 38 reactions; the comparison contains the other 3,455. Continuous summaries are medians and effects are rank-biserial correlations (positive means larger in the hard cohort). Categorical summaries are counts and effects are odds ratios. $p^\dagger$ is Holm-adjusted across inferential feature tests; the two outcome checks report unadjusted p and are not explanatory hypotheses. All diagnostics use the full inventory.

Feature	Hard	Other	Effect	$p^\dagger$
Continuous features: medians; rank-biserial effects				
Atom count	12	14	-0.226	0.1742
Heavy-atom count	6	6	-0.080	1.0000
Candidate pairs	1	1	-0.259	0.0114
Productive pairs	1	1	-0.262	0.0104
Proxy channels	3	4	-0.201	0.3044
TS geometry families	2	2	-0.080	1.0000
Proxies / geometry family	1.917	2	-0.195	0.3094
Mean observations / channel	17.833	28.167	-0.253	0.0951
Singleton-channel fraction	0	0	-0.033	1.0000
Proxy energy spread (mEh)	0.520	0.609	-0.147	0.7197
Endpoint RMSD (Å)	0.940	0.763	0.316	0.0117
Distance-matrix change (Å)	0.679	0.540	0.436	6.84×10^{-5}
Max. productive-pair rank	1	1	-0.248	0.0168
Changed bonds (heuristic)	3	3	0.228	0.1040
Heavy changed bonds (heuristic)	1	1	0.163	0.5767
Categorical features: counts; odds ratios				
One candidate pair	36/38	2392/3455	7.999	0.0046
Standard (c75) profile	32/38	2458/3455	2.163	0.7197
Source-test membership	1/38	108/3455	0.838	1.0000
Any heavy-bond change (heuristic)	32/38	2485/3455	2.082	0.7197
Outcome checks, not explanatory tests: medians; rank-biserial effects				
Task 1 best-model RMSD (Å)	0.610	0.072	0.994	5.00×10^{-26}
Task 1 median RMSD (Å)	0.703	0.155	0.966	1.10×10^{-24}

E FROZEN DFT AUDIT DETAILS

E.1 RANDOM-100 AUDIT AND REPRESENTATIVE CONNECTIONS

A frozen set of 100 reaction-level xTB candidates is evaluated with ORCA/B97-3c (Neese et al., 2020; Neese, 2022a; Brandenburg et al., 2018a). Optimization/frequency checks retain 96 candidates, and 75 satisfy the strict acceptance criteria after bidirectional IRC validation (Fukui, 1981b; Gonzalez & Schlegel, 1989a). The strict pass rate over the full frozen sample is 75.0% (Wilson 95% interval (Wilson, 1927): 65.7–82.5%). Within this random-100 audit, all 96 Opt/Freq survivors complete bidirectional IRC; 82 satisfy the profile-aware physical criteria, including 75 in the strict production profiles. ORCA 5.0.3 (Neese et al., 2020) uses B97-3c (Brandenburg et al., 2018b), neutral singlet states, local OptTS/NumFreq optimization with a maximum of 60 iterations, and bidirectional IRC for converged structures with exactly one meaningful imaginary mode. The standard profile applies its fixed heavy-atom endpoint rule; the hydrogen-sensitive profile uses exact atom-mapped bond topology. The audit includes 62 standard-profile candidates, 23 hydrogen-sensitive candidates, and 15 diagnostic-profile candidates. Strict pass counts are 54/62, 21/23, and 0/15, respectively; seven diagnostic candidates pass the physical checks but are outside the strict inventory. Thus the full frozen sample yields 75/100 strict passes and the two strict profiles yield 75/85.

Representative DFT/IRC connections. Four strict-pass ORCA/B97-3c cases connect audited TS structures to terminal R/P geometries through explicit energy profiles (Figure 5). Their forward

electronic peak-to-terminal differences span 18.2–66.0 kcal mol^{-1}. These profiles illustrate the subset-level physical evidence, rather than serving as large-scale benchmark scores.

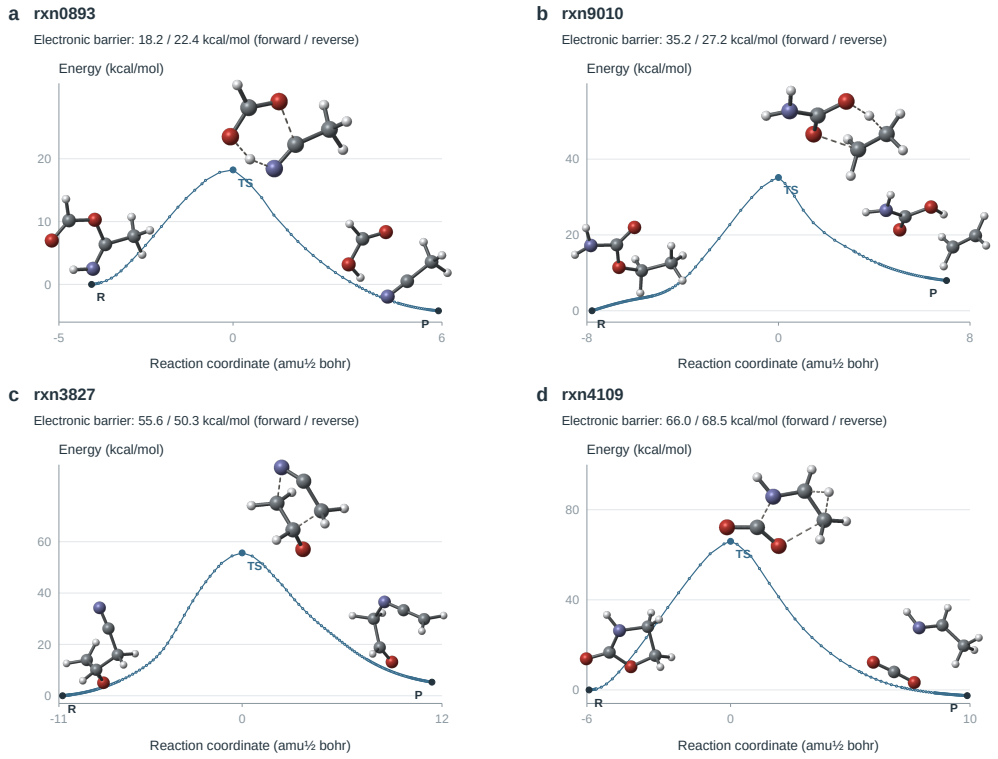

Figure 5: **DFT/IRC trajectories provide explicit saddle-to-endpoint evidence.** Strict-pass ORCA/B97-3c examples: (a) rxn0893, (b) rxn9010, (c) rxn3827, and (d) rxn4109. Energies are relative to each R terminal frame; forward/reverse barriers are electronic peak-to-terminal differences. Vertical scales are panel-specific. The coordinate is reconstructed mass-weighted arc length after sequential rigid alignment, centered at the TS peak, not the native integration coordinate. Insets show R/P terminal frames and enlarged TS structures in a common view. Dashed links mark geometric connectivity changes. The terminal frames are not separately optimized minima; thermal corrections are not included.

E.2 Multi-round pilot and ensemble evidence

Multi-round pilot. The multi-round pilot freezes eight reactions and 16 endpoint pairs before the reference-extension outcomes. Two of its 147 candidate attempts reuse random-100 evidence; the other 145 are new attempts. The 96 complete certificates from these pilot attempts canonicalize to 11 channels using endpoint-basin and TS-geometry equivalence at 0.30 Å, with heavy-atom RMSD for the standard profile and all-atom RMSD for the hydrogen-sensitive profile. Reactions rxn2556, rxn6192, rxn6194, and rxn8551 contain three, two, two, and two channels, respectively; rxn0893 and rxn6187 contain one each. Each of these 11 channels belongs to a distinct canonical R/P basin pair. Within each reaction the multiple channels share the mapped endpoint bond topologies, providing evidence for conformationally distinct connections (Zhao et al., 2022).

Ensemble-closure criterion. Protocol-level ensemble closure requires coverage of the frozen pair support, at least two physical channels, compatible endpoint and TS thermochemistry at 298.15, 400, and 500 K, and two complete exploration rounds with no newly discovered channel. Reactions rxn6192, rxn6194, and rxn8551 meet these criteria. This operational closure describes the frozen search and stationary-state evidence. The DFT audit supports the physical interpretation of reference construction; the large-scale model comparisons use the structural metrics defined in Section 3.4.

Interpreting the physical evidence. The random-100 audit characterizes one candidate per reaction, while the pilot resolves multiple conformationally distinct basin-to-basin connections in selected reactions. Its 11 canonical channels belong to distinct endpoint-basin pairs and share the mapped net bond-rearrangement topology within each reaction. These findings provide evidence for conformationally distinct connections, rather than different net mechanisms or multiple DFT-certified saddles for one canonical pair. The audited population is the frozen reference-construction subset; the large-scale benchmark scores structural recovery of 19,647 xTB proxies, without DFT certification of every reference or generated structure. Exhaustive physical ensemble discovery and quantitative rates or branching ratios are separate validation targets, since kinetic observables depend on the relative free-energy barriers and populations of the relevant reactive states rather than structural multiplicity alone (Eyring, 1935a;b).

F LIMITATIONS AND FUTURE DIRECTIONS

TREK evaluates structural recovery against a frozen xTB-derived reference inventory (Bannwarth et al., 2019), rather than a fully DFT-certified kinetic dataset. The ORCA/B97-3c (Neese, 2022b; Brandenburg et al., 2018b) and bidirectional IRC (Fukui, 1981b; Gonzalez & Schlegel, 1989a) audits in Appendix E provide subset-level physical evidence; they do not certify every reference or model-generated proposal. Extending these audits and releasing the resulting validation records are priorities for future benchmark versions. Structural agreement under the RMSD criteria does not by itself establish accurate barrier heights, rates, or branching ratios, which additionally depend on the underlying potential/free-energy landscape and transition-state thermodynamics (Eyring, 1935a; Truhlar et al., 1996; Bao & Truhlar, 2017; Hänggi et al., 1990). Moreover, complete recovery is defined relative to the retained reference inventory: finite conformer and path exploration need not enumerate every physically accessible TS, and unobserved endpoint combinations remain unlabeled. Finally, the shared proposal budget controls generation opportunities rather than wall-clock time or computational cost; the reported comparisons therefore characterize the evaluated checkpoints under the specified selector and allocator. These boundaries motivate broader physical validation and cost-aware evaluation while retaining a common, reproducible test of budgeted TS-set recovery.