

Reaction-Center-Guided Reranking of Transformer Product Predictions with a Product-Conditioned Graph Neural Network

Adham Norkobilov

adham@kstu.uz

Karshi State Technical University

Jaloliddin Eshbobaev

Tashkent Chemical-Technological Institute

Zafar Turakulov

Tashkent Chemical-Technological Institute

Sarvar Rejabov

Tashkent Chemical-Technological Institute

Utkir Safarov

Tashkent Chemical-Technological Institute

Islomova Farida

Tashkent Chemical-Technological Institute

Shuhrat Tuxtaev

Tashkent Chemical-Technological Institute

Khusan Niyozov

Tashkent Chemical-Technological Institute

Olimjon Maksudov

Tashkent State Technical University named after Islam Karimov

Research Article

Keywords: reaction prediction, reaction center, graph neural network, GINEConv, Transformer, candidate reranking, USPTO-50K, molecular graph

Posted Date: September 8th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10847205/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Forward reaction prediction is commonly formulated as sequence generation from reactant representations, but a purely sequence-based score does not explicitly assess whether a proposed product is locally consistent with the structural changes expected at a reaction center. We present a hybrid product-prediction framework that couples a compact SMILES Transformer with a product-conditioned graph neural network (GNN) and a learned candidate reranker. A high-confidence atom-mapped USPTO-50K subset containing 47,517 reactions was partitioned into record-disjoint training, validation, and held-out test sets (38,013/4,751/4,753). Reaction-center labels were enhanced to capture both bond changes and atom-state changes, eliminating 7,503 otherwise unlabeled state-only transformations. The GNN uses four GINEConv layers (hidden dimension 256) to score reactant atoms while conditioning on a product graph. In a diagnostic held-out reaction-center localization evaluation, where the recorded product graph defines the known transformation context, the GNN achieved ROC-AUC 0.9971, average precision 0.9556, F1 0.8911, and MCC 0.8860 using a validation-frozen threshold of 0.9574. For product prediction, the oracle product is not supplied: the frozen GNN is evaluated independently on each Transformer-generated candidate. A 5-beam Transformer baseline obtained 32.99% Top-1 and 45.34% Top-3 exact product accuracy. Combining Transformer ranking features with candidate-conditioned GNN reaction-center consistency descriptors increased held-out Top-1 accuracy to 35.37% and Top-3 accuracy to 46.52%. The Top-1 gain was 2.377 percentage points (7.21% relative), with 127 reactions recovered and only 14 harmed. The improvement was significant by exact McNemar testing ($p = 5.82 \times 10^{-24}$) and paired bootstrap analysis (95% CI for the gain, 1.894–2.882 percentage points). Complexity-stratified analysis showed significant gains across product-size and reactant-component strata, but no measurable advantage for reactions involving three or more reaction-center atoms. These results show that reaction-center consistency provides information complementary to sequence likelihood and can improve a compact Transformer without retraining the generator.

1. Introduction

Accurate forward reaction prediction is a core component of computer-assisted synthesis because it can be used to validate proposed synthetic steps, prioritize experiments, and constrain synthesis-aware molecular design. Early data-driven systems established two ideas that remain relevant: learned representations can capture reaction regularities, and candidate generation can be separated from candidate ranking [1,2]. A particularly important graph-based precedent is the Weisfeiler-Lehman Difference Network of Jin et al., which first localizes likely reaction centers, enumerates graph edits, and then scores candidate products [3]. The later Molecular Transformer reframed reaction prediction as SMILES-to-SMILES translation and demonstrated that attention-based sequence models can learn broad reaction statistics without hand-coded reaction rules [4]. Learned Transformer representations were subsequently shown to organize reaction classes and chemical reaction space [5].

Sequence models are attractive because they provide a simple end-to-end formulation, but SMILES strings do not encode molecular connectivity as directly or permutation-invariantly as graphs. This motivated hybrid and graph-native approaches. Graph2SMILES combines a molecular graph encoder with Transformer-style decoding and reduces dependence on input SMILES ordering [6], whereas Chemformer showed that Transformer pretraining can transfer across synthesis and other cheminformatics tasks [7]. In parallel, LocalRetro emphasized that many transformations are spatially local and modeled atom- and bond-centered reaction templates together with global attention [8]. These developments suggest that global sequence likelihood and local structural reactivity are related but non-identical signals, creating an opportunity for principled hybridization rather than replacement of one representation by the other.

Data quality is an equally important determinant of reported performance. Patent-derived reaction corpora contain atom-mapping errors, missing reaction context, noisy reagents, duplicated chemistry, and strong sampling biases.

Quantitative interpretation of the Molecular Transformer has shown that correct predictions can sometimes arise from dataset correlations that are not chemically causal, motivating stricter split design and interpretability analysis [9]. Reagent-aware curation can improve both data quality and downstream product prediction [10]. Atom-to-atom mapping is especially consequential for reaction-center learning because every structural edit depends on the correspondence between reactant and product atoms. LocalMapper addressed this issue through human-in-the-loop learning and reported highly reliable confident mappings on the widely used USPTO-50K benchmark [11]. The underlying USPTO-50K benchmark itself originates from a curated patent-reaction collection described by Schneider et al. [12].

Recent work has further questioned whether high in-distribution accuracy is sufficient evidence of chemical generalization. Bradshaw et al. showed substantial degradation under new-patent, temporal, and unseen-reaction-class evaluations [13]. Error-focused evaluation has likewise highlighted imperfections in common reaction datasets and the limitations of a single exact-match score [14], while chemically aware transfer learning demonstrates that mechanistically relevant data can matter more than raw pretraining scale in restricted domains [15]. Most recently, explicit reaction-center objectives have been incorporated into chemical reaction graph representation learning to better capture localized structural evolution [16]. A complementary 2025 direction is RXNGraphormer, which combines graph neural networks for intramolecular pattern recognition with Transformer-based intermolecular modeling in a unified reaction framework [17]. Reaction-center prediction has also been revisited directly through attention analysis of chemical language models, with supervised attention-based localization reported in 2026 [18]. Together, these studies shift the emphasis from headline benchmark accuracy toward chemically grounded representation, controlled evaluation, and transparent failure analysis.

A remaining methodological opportunity is to use reaction-center information not as a standalone product generator, but as a structural consistency signal for candidates already generated by a sequence model. This setting is attractive when a deployed or compact Transformer produces a useful Top-k candidate list: rather than retraining a substantially larger generator, a graph model can evaluate whether each proposed product is compatible with a focused transformation pattern in the reactants. Such a design recovers the classical generation-versus-ranking separation [2,3] while replacing manually engineered edit descriptors with learned, candidate-conditioned graph features. It also exposes an interpretable intermediate quantity atom-wise reaction-center probabilities that can be inspected independently of the final product rank.

In this work, we develop and evaluate a two-stage framework in which a compact SMILES Transformer generates five candidate products and a product-conditioned ReactionCenterGNN scores the corresponding reactant-side transformation pattern. Transformer-derived ranking descriptors and GNN-derived reaction-center consistency descriptors are combined in a regularized logistic reranker (Fig. 1). To minimize information leakage and model-selection bias, reaction records were uniquely reconciled and partitioned into record-disjoint training, validation, and held-out test sets. Reaction-center labels included both bond edits and atom-state changes, the GNN operating threshold was selected exclusively on validation data and frozen before testing, reranker selection used reaction-grouped out-of-fold validation, and the held-out test set was reserved for final evaluation. The study does not claim state-of-the-art absolute benchmark accuracy. Instead, it tests the narrower and reproducible hypothesis that graph-derived reaction-center consistency supplies statistically reliable information beyond Transformer sequence likelihood for reranking a fixed candidate set.

2. Materials and Methods

2.1. Study design and computational workflow

The workflow was organized to separate representation learning, model selection, and final testing. The master reaction table was first reconciled against the Transformer source/target files and the LocalMapper-remapped reaction records. The resulting 47,517 reactions were assigned to non-overlapping training, validation, and held-out test subsets. An enhanced reaction-center labeling procedure was then applied to the mapped reactions. The GNN was trained only on the training subset, and its decision threshold was selected on the validation subset. The Transformer generator was trained independently on the corresponding source-target split. Beam-search candidates were generated before reranker training. Candidate-level reranking models were selected using 5-fold GroupKFold by reaction on the validation set, after which the final rerankers were fitted to the full validation set and frozen. The test set was not used for model or threshold selection.

2.2. Reaction dataset, atom mapping, and data partitioning

The starting corpus was the public LocalMapper-remapped USPTO-50K release associated with Chen et al. [11], deposited on Figshare under DOI 10.6084/m9.figshare.25046471.v1. That release is based on the widely used USPTO-50K benchmark derived from the patent-reaction collection curated by Schneider et al. [12]. The study pipeline retained 47,517 reactions after its mapping-confidence, validity, and row-reconciliation checks. The same source/target partitioning was used consistently for both sequence and graph models, giving 38,013 training reactions (80.0%), 4,751 validation reactions (10.0%), and 4,753 held-out test reactions (10.0%). The split assignments were inherited from the fixed Transformer training, validation, and test source–target files and were propagated unchanged to the reconciled mapped master table and the GNN graph datasets; no additional resplitting was performed during GNN construction. Integrity checks identified 47,517 unique original rows, no duplicated original records, no missing mapped reactions, and no exact reaction-record overlap among the training, validation, and test partitions. During reconciliation of the Transformer and mapped test records, 4,703 of 4,753 reactions matched uniquely using canonical exact reaction keys; the remaining 50 were recovered uniquely with a stereochemistry-insensitive key and no ambiguous match remained. Importantly, this relaxed key was used only to reconcile row identity. Original stereochemical product representations were preserved for subsequent canonical exact-match evaluation.

The remapped master table used in this study did not contain an explicit reaction-class field. Reaction classes were therefore not reconstructed using heuristic templates solely for subgroup analysis. Instead, robustness was assessed using objective structural-complexity descriptors derived directly from the reaction graph and molecular sizes (Section 2.10). This choice avoids introducing a second, potentially noisy labeling layer into the evaluation.

2.3. Enhanced reaction-center label construction

Reaction centers were derived from atom-mapped reactant-product pairs. A mapped atom was labeled as a reaction-center atom if it participated in a bond formation, bond cleavage, or bond-order change, or if its local atomic state changed between reactants and product. Atom-state differences included hydrogen count, explicit hydrogen state, degree, hybridization, formal charge, and chirality. All comparisons were performed on sanitized atom-mapped molecules before atom-map identifiers were removed, and atom-state changes were evaluated only between atoms sharing the same map identifier on the reactant and product sides. The second criterion was essential: a bond-only definition left 7,503 reactions without any reaction-center atom, whereas diagnostic analysis showed that all 7,503 contained detectable atom-state changes. Incorporating these changes reduced the number of no-RC reactions to zero.

Across all 47,517 reactions, the enhanced labels contained a mean of 2.018 RC atoms per reaction and a median of two. The dominant class consisted of two-center transformations (35,397 reactions), while 7,024 reactions contained one RC atom. Larger multi-center transformations were progressively less frequent. A 1,000-reaction coverage audit confirmed that all enhanced RC map numbers were represented in the reactant graph. Figure 2b shows the final RC-size distribution.

2.4. Molecular graph representation

Each molecular component was converted to a graph with atoms as nodes and bonds as edges. Atom features were 39-dimensional and encoded element identity, formal charge, total hydrogen count, hybridization, aromaticity, ring membership, and degree. Edge features were six-dimensional indicators for single, double, triple, and aromatic bond type, conjugation, and ring membership. Reactant components were retained as separate molecular graphs during preprocessing and merged with index offsets for batched node-level prediction. The final v2 graph dataset was generated without preprocessing failures, RC coverage errors, or feature-dimension errors. The training split contained 1,633,839 reactant atoms, of which 76,685 (4.694%) were labeled as RC atoms, producing a class-imbalance ratio of 20.306 non-RC atoms per RC atom.

2.5. Product-conditioned ReactionCenterGNN

The ReactionCenterGNN was designed as a candidate-conditioned node classifier rather than a reactant-only reaction-center detector. A shared molecular encoder projected 39-dimensional node features and six-dimensional edge features into a 256-dimensional hidden space. Four GINEConv message-passing layers were then applied, each with a two-layer 256-512-256 multilayer perceptron, batch normalization, ReLU activation, 0.1 dropout, and residual connections. GIN provides an expressive neighborhood-aggregation framework [19], while the GINE formulation incorporates bond attributes directly into message passing [20]; this edge-aware variant was therefore used to retain chemically relevant bond information.

Reactant and candidate-product graphs were encoded by the shared graph encoder. Product node embeddings were aggregated by global mean pooling and projected to a 256-dimensional product-context vector. This vector was broadcast to every reactant node and concatenated with the corresponding 256-dimensional reactant-node embedding. A 512-256-128-1 classifier produced one logit per reactant atom. The architecture contained 1,559,045 trainable parameters. Because only 4.694% of training reactant atoms were labeled as reaction-center atoms, class imbalance was handled with BCEWithLogitsLoss using a training-only positive weight of 20.3058. The selected checkpoint occurred at epoch 25 and yielded validation AP 0.9545, ROC-AUC 0.9973, and loss 0.1314. All threshold selection was performed subsequently on validation predictions rather than on test labels.

The conditioning protocol differs deliberately between the auxiliary reaction-center task and the product-reranking task. During supervised GNN training and the diagnostic validation/test evaluation of reaction-center localization, the recorded product graph is used as the product context because the objective is to identify the reactant atoms associated with the known transformation. In the product-prediction experiment, by contrast, the recorded test product is never supplied to the reranker. The frozen GNN is evaluated separately for each Transformer beam candidate, and its atom-wise probabilities quantify how strongly that proposed candidate implies a localized transformation pattern in the reactants. Thus, the reaction-center metrics reported in Section 3.2 characterize the auxiliary localization task, whereas the final product-prediction metrics characterize candidate-conditioned inference without oracle-target leakage.

2.6. Validation-based threshold selection and GNN evaluation

Because RC atoms represented less than 5% of reactant atoms, both threshold-free and thresholded metrics were reported. Average precision (AP) and ROC-AUC quantified ranking performance independent of a fixed cutoff. Precision, recall, F1, Matthews correlation coefficient (MCC), confusion-matrix counts, and reaction-level exact RC accuracy were computed after thresholding. The operating threshold was selected exclusively from validation predictions by maximizing F1 and was then frozen at 0.957427 for every held-out analysis. Figure 3 visualizes the validation precision-recall-F1 trade-off across candidate thresholds. A conventional 0.5 cutoff was retained only as a reference and was never used to tune the final test result.

2.7. SMILES Transformer baseline and beam generation

The product generator was a compact encoder-decoder Transformer following the sequence-to-sequence formulation used in reaction prediction [4,21]. Reactants were represented as cleaned canonical SMILES with atom-map numbers removed, and products were represented analogously. A custom chemistry-aware tokenizer separated bracket atoms, multi-character elements such as Br and Cl, aromatic symbols, bond symbols, ring indices, and other SMILES syntax; the final vocabulary contained 201 tokens including padding, unknown, beginning-of-sequence, and end-of-sequence symbols. The model used $d_{\text{model}} = 256$, eight attention heads, four encoder layers, four decoder layers, and a feed-forward dimension of 512. Training proceeded for 60 epochs in two stages, and the final checkpoint had validation token accuracy 94.76% and validation loss 0.1617. This compact generator was treated as a fixed baseline for the reranking study rather than as a claim of state-of-the-art absolute USPTO-50K performance.

At inference, beam search with beam size 5 and length penalty 0.6 generated candidate products. Predictions and reference products were canonicalized with RDKit using stereochemistry-preserving canonical SMILES (`isomericSmiles=True`) before exact-match evaluation, avoiding penalties for equivalent SMILES serializations while retaining stereochemical distinctions. Candidate tables were deduplicated within each reaction on the basis of canonical product identity. As an implementation consistency check, Transformer scores recomputed by fixed-candidate teacher forcing agreed with the original beam-search scores to a maximum absolute difference of 5.21×10^{-7} , allowing saved beam candidates to be reused without introducing a scoring mismatch.

2.8. GNN consistency descriptors and learned reranking

For every beam candidate, the candidate product and the reactant graphs were passed through the frozen GNN. The atom-level probability vector was summarized into GNN consistency descriptors including the maximum RC probability, means over the two and three most probable atoms, total and mean probability, number of atoms above the frozen RC threshold, mean binary entropy, peak contrast, and deviation of the predicted RC count from the training-set mean. Absolute summaries were supplemented by within-reaction relative descriptors, yielding 15 GNN-derived features. Five Transformer-side descriptors captured the original beam/ranking information, candidate validity, normalized sequence score, score margin relative to the best sequence candidate, and within-reaction score normalization. The combined feature vector therefore contained 20 descriptors.

Three logistic rerankers were considered: Transformer-only, GNN-only, and combined Transformer + GNN. Continuous features were standardized and logistic regression used $C = 1.0$, `class_weight = 'balanced'`, a maximum of 3,000 iterations, and random state 42. Candidate labels indicated whether a canonical candidate matched the recorded product. Model comparison was performed with 5-fold GroupKFold in which all candidates from the same reaction were kept in the same fold. After out-of-fold validation, each reranker was fitted on all 4,751 validation reactions and frozen. No reranker parameter was adjusted after observing the held-out test predictions.

2.9. Product-prediction metrics and statistical testing

Product prediction was assessed by canonical exact-match Top-1 and Top-3 accuracy. Oracle Top-5 coverage was defined as the fraction of reactions for which the recorded product was present anywhere among the deduplicated five-beam candidates; it therefore represents an upper bound for any reranker restricted to the same candidate set. To quantify paired changes relative to the original Transformer beam ranking, each test reaction was assigned to one of four outcomes: both correct, both wrong, recovered (baseline wrong, hybrid correct), or harmed (baseline correct, hybrid wrong). Statistical significance of the paired Top-1 change was tested using an exact two-sided McNemar/binomial test on discordant reactions. A paired nonparametric bootstrap with 10,000 reaction-level resamples estimated the 95% confidence interval for the Top-1 accuracy difference.

2.10. Complexity-stratified robustness analysis

Because an explicit reaction-class field was absent from the remapped master table used in the final workflow, no post hoc class labels were invented. Instead, the held-out set was stratified using directly measurable complexity variables. Product size was divided into test-set tertiles (small, medium, large) based on heavy-atom count. Reactant mixtures were grouped as 1–2, 3–4, or at least 5 molecular components. Reactions were also grouped by enhanced RC size (1, 2, or at least 3 RC atoms). For every subgroup, baseline and hybrid Top-1 accuracy, recovered/harmed counts, exact McNemar p values, and paired bootstrap 95% confidence intervals were computed. Subgroup bootstrap intervals used 5,000 paired resamples.

2.11. Software and reproducibility

Chemical parsing, canonicalization, and molecular graph construction were performed in Python using RDKit; graph learning used PyTorch/PyTorch Geometric; and reranking, resampling, and statistical analysis used scikit-learn, NumPy, pandas, and SciPy-compatible numerical routines. The final archive accompanying submission should contain the exact environment specification, split indices, trained checkpoints, saved beam candidates, candidate-level GNN scores, and scripts used to reproduce every reported table and figure. Numerical values in the figures were generated directly from the saved evaluation outputs and were not manually altered during manuscript preparation. A consolidated summary of the final dataset, model configuration, training settings, and evaluation protocol is provided in Table 1.

Table 1
Final dataset and model configuration used for the reported experiments.

Component	Setting
Dataset	LocalMapper-remapped USPTO-50K; 47,517 mapping-confident reactions
Split	Split: Train 38,013 / Validation 4,751 / Test 4,753; no exact reaction-record overlap
RC labels	Bond edits + atom-state changes; mean 2.018 atoms/reaction; median 2
Graph features	39 node features; 6 edge features
ReactionCenterGNN	4 GINEConv layers; hidden 256; dropout 0.1; 1,559,045 parameters
GNN imbalance handling	BCEWithLogitsLoss; train-only pos_weight = 20.3058
Frozen RC threshold	0.957427, selected on validation by F1
Transformer	d_model 256; 8 heads; 4 encoder + 4 decoder layers; FFN 512; vocab 201
Generation	Beam size 5; length penalty 0.6; canonical deduplication
Reranker	StandardScaler + logistic regression; C = 1.0; balanced class weights
Reranker selection	5-fold GroupKFold by reaction on validation candidates
Held-out statistics	Exact McNemar test; paired bootstrap (10,000 resamples)
Public source	LocalMapper-remapped USPTO-50K release; Figshare DOI 10.6084/m9.figshare.25046471.v1; original USPTO-50K curation described by Schneider et al. [12]

3. Results

3.1. Dataset integrity and reaction-center label coverage

The final preprocessing workflow converted all 47,517 matched reactions to the v2 graph representation without failed reactions, coverage errors, or feature-dimension errors. Enhanced RC labeling was particularly consequential for state-only transformations: 7,503 reactions that had no bond-defined center all exhibited atom-state changes and therefore received valid RC labels under the enhanced definition. The resulting class balance was more challenging than in the earlier bond-centric preprocessing, with only 4.694% of training reactant atoms labeled as RC atoms. Nevertheless, the train, validation, and test partitions displayed very similar mean RC sizes (approximately two atoms per reaction), reducing the risk that the held-out set was trivially easier or harder because of label-count imbalance. Figure 2 shows the split sizes and overall RC-size distribution.

3.2. Held-out reaction-center localization

As an auxiliary diagnostic, the product-conditioned GNN was first evaluated for reaction-center localization using the recorded product graph to define the known transformation context. On the held-out test set, ROC-AUC was 0.9971 and AP was 0.9556 (Fig. 4), closely matching validation values of 0.9973 and 0.9545. At the validation-frozen threshold of 0.957427, held-out precision was 0.9084, recall 0.8745, F1 0.8911, and MCC 0.8860. The corresponding confusion matrix contained 8,421 true positives, 192,767 true negatives, 849 false positives, and 1,209 false negatives across 203,246 test reactant atoms. Reaction-level exact RC accuracy was 69.64%, with macro reaction precision 0.9096 and macro reaction recall 0.8891. These values assess the quality of the learned RC representation; they are not product-prediction accuracies and do not enter the final test reranking through the recorded product.

Threshold selection materially changed the precision-recall balance. At the default cutoff of 0.5, recall was very high (0.9746) but precision fell to 0.6924 and F1 to 0.8096. In contrast, the validation-frozen threshold emphasized high-confidence centers and produced a substantially better balanced F1 on both validation and test data. The close agreement between validation and test metrics supports use of the frozen validation operating point rather than post hoc test tuning (Fig. 3).

3.3. Transformer candidate generation and reranking ceiling

On the 4,753 held-out reactions, the original 5-beam Transformer achieved 32.99% Top-1 canonical exact accuracy and 45.34% Top-3 accuracy. The recorded product appeared somewhere within the five deduplicated candidates for 48.90% of test reactions, defining the Oracle Top-5 ceiling for any downstream reranker restricted to the same candidates. This gap between Top-1 (32.99%) and Oracle Top-5 (48.90%) indicates that a substantial portion of the remaining error was a ranking problem rather than a generation problem. It therefore provided a realistic target for a chemically informed reranker.

3.4. Hybrid reranking improves held-out product prediction

The ablation study demonstrates that Transformer and GNN information were complementary rather than interchangeable (Table 2; Fig. 5). A learned Transformer-only reranker increased Top-1 accuracy from 32.99% to 33.75% (+ 0.76 percentage points), recovering 36 reactions without harming any baseline-correct case. In contrast, the GNN-only reranker performed poorly as an independent ranking model (17.86% Top-1), recovering 295 baseline errors but displacing 1,014 correct predictions. This result is consistent with the intended role of the GNN: its consistency descriptors identify localized structural change but do not encode the full sequence-level prior required to choose among chemically similar candidate products.

Combining both feature families yielded the strongest performance. The Transformer + GNN reranker reached 35.37% Top-1 and 46.52% Top-3 exact accuracy, corresponding to an absolute Top-1 gain of 2.38 percentage points over the

original Transformer beam and 1.62 percentage points over the Transformer-only reranker. The number of correctly predicted test reactions increased from 1,568 to 1,681. Of the baseline errors, 127 were recovered, whereas only 14 baseline-correct reactions were harmed, producing a net gain of 113 reactions. Because the candidate set itself was unchanged, the improvement can be attributed to candidate ordering rather than expanded generation coverage.

Table 2

Held-out ablation study for candidate reranking. Recovered, harmed, and net-gain counts are calculated relative to the original Transformer beam ranking.

Configuration	Top-1 (%)	Top-3 (%)	Δ Top-1 vs Beam (pp)	Recovered vs Beam	Harmed vs Beam	Net gain vs Beam
Transformer Beam	32.99	45.34	0.00	0	0	0
Transformer-only reranker	33.75	45.97	+ 0.76	36	0	+ 36
GNN-only reranker	17.86	41.93	-15.13	295	1014	-719
Transformer + GNN reranker	35.37	46.52	+ 2.38	127	14	+ 113

3.5. Validation-to-test generalization and paired significance

The magnitude of the hybrid improvement was stable from validation to test. In 5-fold reaction-grouped out-of-fold validation, the original beam ranking achieved 33.28% Top-1 accuracy and the combined reranker achieved 35.57%, a gain of 2.29 percentage points. Held-out test performance was 32.99% and 35.37%, respectively (Fig. 6). The similarity of the validation and test deltas reduces concern that the reranker was tuned to idiosyncrasies of the validation candidate set.

Paired statistical testing strongly supported the Top-1 gain. Among the 4,753 held-out reactions, the hybrid model recovered 127 cases and harmed 14, giving 141 discordant pairs. The exact two-sided McNemar p value was 5.82×10^{-24} . A 10,000-resample paired bootstrap estimated a 95% confidence interval of 1.894 to 2.882 percentage points for the accuracy gain. The individual bootstrap intervals were 31.685–34.315% for the Transformer baseline and 34.042–36.714% for the combined model. Thus, the observed improvement was not only statistically detectable but also directionally consistent across paired resamples.

3.6. Error recovery is concentrated near the top of the beam

Error decomposition provides a more chemically useful view than aggregate accuracy alone. The baseline and hybrid reranker were both correct for 1,554 reactions and both wrong for 3,058 reactions. The hybrid model uniquely corrected 127 cases and uniquely damaged 14 (Fig. 7). The asymmetry between recovered and harmed cases explains the highly significant paired test and shows that the gain was not produced by indiscriminate reordering.

Most recovered products were already close to the original top prediction. Of the 127 recovered reactions, 95 (74.8%) had the correct product at original beam rank 2, 24 (18.9%) at rank 3, and four each at ranks 4 and 5. The mean original rank of recovered products was 2.35 and the median was 2 (Fig. 8). This pattern suggests that the GNN signal is most effective as a local correction to a reasonably calibrated sequence model rather than as a substitute for candidate generation.

3.7. Qualitative reaction-center interpretation

Representative examples illustrate the local structural signal learned by the GNN (Fig. 9). Panels 9a and 9b are baseline-correct cases, whereas panels 9c and 9d are recovered cases in which the recorded product originally

appeared at Transformer beam ranks 2 and 4, respectively. In each example, the highest predicted RC probabilities concentrate on atoms directly associated with the recorded structural transformation. For interpretability, these panels use the recorded product as the conditioning graph, matching the auxiliary RC-localization diagnostic described in Section 3.2. They are not reranking inputs. In the held-out product-prediction experiment, the frozen GNN is conditioned exclusively on Transformer-generated candidate products.

3.8. Complexity-stratified robustness

The hybrid gain remained positive and statistically significant across every product-size tertile. Small products improved by 2.41 percentage points (95% CI 1.56–3.31), medium products by 2.61 points (1.82–3.45), and large products by 2.08 points (1.32–2.92). Gains were also significant for reactions containing 1–2, 3–4, and at least 5 reactant components, ranging from 2.15 to 2.79 percentage points. Thus, the overall benefit was not confined to one molecular-size or mixture-complexity regime (Table 3; Fig. 10).

A different pattern emerged when reactions were stratified by reaction-center size. One-center reactions showed the largest improvement, + 4.82 percentage points (95% CI 3.12–6.52), while two-center reactions improved by + 2.25 points (1.71–2.80). In contrast, reactions with three or more RC atoms showed no net benefit (0.00 points; 95% CI -0.74 to 0.74; McNemar $p = 1.0$). This null subgroup is scientifically informative rather than a reason for exclusion: the current scalar consistency summaries appear well matched to sparse local changes but are insufficient for more complex multi-center transformations.

Table 3
Held-out complexity-stratified Top-1 performance and paired statistical evidence.

Stratification	Group	N	Transformer (%)	Hybrid (%)	Gain (pp)	95% CI (pp)	Recovered	Harmed	McNemar p
Product size	Small	1663	43.24	45.64	+ 2.41	[1.56, 3.31]	47	7	2.29×10 ⁻⁸
Product size	Medium	1650	31.52	34.12	+ 2.61	[1.82, 3.45]	47	4	2.42×10 ⁻¹⁰
Product size	Large	1440	22.85	24.93	+ 2.08	[1.32, 2.92]	33	3	2.27×10 ⁻⁷
Reactant components	1–2	712	30.90	33.15	+ 2.25	[1.12, 3.51]	17	1	1.45×10 ⁻⁴
Reactant components	3–4	1577	35.19	37.98	+ 2.79	[1.84, 3.74]	53	9	1.05×10 ⁻⁸
Reactant components	≥ 5	2464	32.18	34.33	+ 2.15	[1.58, 2.76]	57	4	4.85×10 ⁻¹³
Reaction-center size	1 RC atom	705	37.59	42.41	+ 4.82	[3.12, 6.52]	38	4	5.65×10 ⁻⁸
Reaction-center size	2 RC atoms	3506	34.71	36.97	+ 2.25	[1.71, 2.80]	87	8	6.74×10 ⁻¹⁸
Reaction-center size	≥ 3 RC atoms	542	15.87	15.87	0.00	[-0.74, 0.74]	2	2	1.00

4. Discussion

The central finding is that reaction-center consistency and sequence likelihood carry complementary information. The Transformer-only reranker produced a modest improvement over the raw beam order, indicating that some errors arise from calibration of sequence-side scores. The GNN-only reranker, by contrast, was not an effective replacement for the Transformer prior: although it recovered many baseline errors, it also displaced substantially more correct predictions. Once the two feature families were combined, the structural signal became useful in the intended role and yielded a reproducible + 2.38-percentage-point held-out Top-1 gain. The asymmetry is mechanistically sensible: localization of a plausible transformation is necessary for many reactions, but it is not sufficient to establish the overall probability of a complete product structure.

The observed behavior is consistent with the broader evolution of reaction-prediction methods. Reaction-center-first graph models demonstrated early that local edit structure can constrain forward product prediction [3], while sequence models such as the Molecular Transformer and Chemformer are particularly effective at learning global reaction statistics [4,7]. Graph-to-sequence and local-template approaches further show the value of explicit molecular connectivity and localized reactivity [6,8]. More recent reaction-representation work now introduces explicit reaction-center objectives within graph learning [16], while RXNGraphormer provides independent evidence that graph and Transformer representations can be complementary at larger scale [17]. Attention-based reaction-center localization offers another recent route to interpretable local reactivity signals [18]. The present candidate-conditioned formulation occupies a practical middle ground: it preserves the generative flexibility and prior probability supplied by the Transformer while adding an atom-resolved structural consistency signal only at reranking time.

The statistical analysis is important because the absolute gain is moderate. A 2.38-point improvement is meaningful only if it is stable and not driven by a small number of favorable cases. Here, the gain was nearly identical in reaction-grouped out-of-fold validation and held-out testing; the exact McNemar test strongly rejected symmetry of recovered and harmed cases; and the paired bootstrap confidence interval excluded zero by a wide margin. Complexity stratification further showed positive, statistically supported gains across all product-size and reactant-component bins. These checks support a genuine complementary signal while avoiding reliance on a single aggregate accuracy value.

The subgroup analysis also identifies a clear boundary of the present representation. Reactions with three or more enhanced RC atoms had low absolute accuracy and no reranking gain. Such transformations may contain multiple spatially separated probability peaks, coordinated bond rearrangements, or coupled atom-state changes that are poorly summarized by scalar descriptors such as maximum probability, entropy, and RC-count deviation. Future models should represent the predicted center as a structured subgraph, include explicit changed-bond features, or use cross-graph attention between reactant and candidate-product atoms rather than a single global mean-pooled product context.

The current absolute product-prediction accuracy should be interpreted in context. The compact Transformer was trained on the 47,517-reaction project split and was designed as a fixed generator for a paired reranking experiment, not as a direct reproduction of larger pretrained or heavily augmented benchmarks. Absolute Top-1 values are therefore not directly comparable with Molecular Transformer, Graph2SMILES, or Chemformer results obtained under different USPTO curation, reagent handling, augmentation, and split conventions [4,6,7]. The contribution tested here is the controlled improvement of the same fixed candidate set, for which every method sees identical generated alternatives and the Oracle Top-5 ceiling remains 48.90%.

Finally, data quality and generalization remain broader concerns. LocalMapper provides a substantially more reliable basis for mapping-dependent labels [11], and the present workflow eliminates exact cross-split duplicates, preserves stereochemistry in evaluation, and freezes all thresholds before testing. Nevertheless, the split is an in-distribution

reaction-level split rather than a patent-, time-, author-, or reaction-class-held-out challenge of the type advocated in recent generalization studies [13]. The observed stability should therefore be described as held-out generalization within this corpus, not evidence of extrapolation to novel chemistry. Reaction conditions are also incompletely represented in patent-derived reaction equations, so the task should be interpreted as recorded-product prediction from reaction strings rather than experimental feasibility prediction [9,10,13].

5. Limitations and Future Work

First, the reranker cannot recover a correct product that is absent from the Transformer Top-5 set; the 48.90% Oracle Top-5 coverage is therefore a hard ceiling for the current architecture. Improving candidate diversity and generator calibration is likely to provide the largest additional headroom. Second, the product-conditioned GNN summarizes the candidate product using global mean pooling. This is computationally efficient but discards explicit reactant-candidate atom correspondences. Cross-graph attention, learned atom matching, or candidate-specific changed-bond prediction may be especially important for multi-center reactions. Third, patent-derived reaction equations contain incomplete context. Solvent, temperature, concentration, scale, catalysts, and other procedural variables are not modeled explicitly. The reported task is therefore reaction-equation product prediction rather than a complete model of experimental feasibility. Fourth, the present split contains no exact reaction-record overlap across partitions but is not chronological, patent-held-out, or reaction-class-held-out. More stringent out-of-distribution evaluation is therefore required before claiming robustness to genuinely novel chemistry [13]. Fifth, the remapped master table used in the final workflow does not contain a trusted reaction-class field. We intentionally avoided reconstructing classes with heuristic templates solely to improve subgroup reporting and instead used objective complexity variables. Future work could evaluate class-held-out transfer using externally validated class annotations. Sixth, the paired bootstrap and McNemar analyses quantify uncertainty over held-out reactions, but the final GNN and Transformer results come from the selected training runs rather than a multi-seed ensemble. Independent retraining across several random seeds would strengthen estimates of optimization-induced variability without changing the paired test-set conclusions reported here.

Finally, the GNN is trained and diagnostically evaluated for reaction-center localization with the recorded product as context, whereas deployment-style reranking applies the same frozen model to generated candidate products. This creates a product-context distribution shift between the auxiliary supervised task and candidate scoring, even though the held-out reranking itself contains no oracle-target input. The significant candidate-level reranking gain shows that the learned signal transfers in the present setting, but future work should quantify this shift directly by comparing RC maps for correct candidates and chemically close near-miss candidates, particularly in recovered and harmed reactions.

6. Conclusions

A hybrid reaction-prediction framework was developed in which a compact SMILES Transformer generates candidate products and a product-conditioned GINE network supplies atom-resolved reaction-center consistency features for learned reranking. The study used a publicly traceable LocalMapper-remapped USPTO-50K source, explicit leakage checks, enhanced reaction-center labels, validation-frozen thresholding, reaction-grouped reranker selection, and paired statistical testing. The GNN achieved strong held-out RC discrimination (ROC-AUC 0.9971; AP 0.9556), while the combined reranker improved fixed-generator Top-1 product accuracy from 32.99% to 35.37%. The 2.377-percentage-point gain corresponded to 113 net additional correct reactions and was supported by exact McNemar testing and paired bootstrap confidence intervals. Benefits generalized across product-size and reactant-component strata, whereas the absence of gain for reactions with three or more RC atoms identifies a concrete target for future model development. Overall, the results support reaction-center consistency as a useful auxiliary signal for candidate ranking

and provide a practical route for adding graph-based structural awareness to an existing sequence generator without retraining the generator itself.

Declarations

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

AI-Assisted Language Editing

The authors confirm that artificial intelligence–based tools were used exclusively to improve the grammar, clarity, and stylistic consistency of the manuscript. The scientific content, interpretation of results, and conclusions were entirely developed and verified by the authors, who take full responsibility for the integrity of the work.

Funding:

This research received no external funding.

Author Contribution

A.N. conceived and supervised the study, developed the research methodology, interpreted the results, and drafted the manuscript. J.E. contributed to data preparation, model implementation, and computational experiments. Z.T. contributed to methodology development, model validation, and formal analysis. S.R. and U.S. contributed to software implementation, data processing, and visualization of the results. F.I. contributed to data curation, literature analysis, and manuscript preparation. S.T. and K.N. contributed to the chemical interpretation of the results, validation, and critical revision of the manuscript. O.M. contributed to computational resources, project administration, and critical review. All authors discussed the results, reviewed and approved the final version of the manuscript, and agreed to be accountable for the work.

Data Availability

The datasets generated and/or analysed during the current study are available from the corresponding author upon reasonable request.

References

1. Wei, J. N.; Duvenaud, D.; Aspuru-Guzik, A. Neural Networks for the Prediction of Organic Chemistry Reactions. *ACS Central Science* 2016, 2, 725–732. <https://doi.org/10.1021/acscentsci.6b00219>.
2. Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.; Jensen, K. F. Prediction of Organic Reaction Outcomes Using Machine Learning. *ACS Central Science* 2017, 3, 434–443. <https://doi.org/10.1021/acscentsci.7b00064>.
3. Jin, W.; Coley, C.; Barzilay, R.; Jaakkola, T. Predicting Organic Reaction Outcomes with Weisfeiler-Lehman Network. *Advances in Neural Information Processing Systems* 2017, 30.

4. Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Central Science* 2019, 5, 1572–1583. <https://doi.org/10.1021/acscentsci.9b00576>.
5. Schwaller, P.; Probst, D.; Vaucher, A. C.; Nair, V. H.; Kreutter, D.; Laino, T.; Reymond, J.-L. Mapping the Space of Chemical Reactions Using Attention-Based Neural Networks. *Nature Machine Intelligence* 2021, 3, 144–152. <https://doi.org/10.1038/s42256-020-00284-w>.
6. Tu, Z.; Coley, C. W. Permutation Invariant Graph-to-Sequence Model for Template-Free Retrosynthesis and Reaction Prediction. *Journal of Chemical Information and Modeling* 2022, 62, 3503–3513. <https://doi.org/10.1021/acs.jcim.2c00321>.
7. Irwin, R.; Dimitriadis, S.; He, J.; Bjerrum, E. J. Chemformer: A Pre-Trained Transformer for Computational Chemistry. *Machine Learning: Science and Technology* 2022, 3, 015022. <https://doi.org/10.1088/2632-2153/ac3ffb>.
8. Chen, S.; Jung, Y. Deep Retrosynthetic Reaction Prediction Using Local Reactivity and Global Attention. *JACS Au* 2021, 1, 1612–1620. <https://doi.org/10.1021/jacsau.1c00246>.
9. Kovács, D. P.; McCorkindale, W.; Lee, A. A. Quantitative Interpretation Explains Machine Learning Models for Chemical Reaction Prediction and Uncovers Bias. *Nature Communications* 2021, 12, 1695. <https://doi.org/10.1038/s41467-021-21895-w>.
10. Andronov, M.; Voinarovska, V.; Andronova, N.; Wand, M.; Clevert, D.-A.; Schmidhuber, J. Reagent Prediction with a Molecular Transformer Improves Reaction Data Quality. *Chemical Science* 2023, 14, 3235–3246. <https://doi.org/10.1039/D2SC06798F>.
11. Chen, S.; An, S.; Babazade, R.; et al. Precise Atom-to-Atom Mapping for Organic Reactions via Human-in-the-Loop Machine Learning. *Nature Communications* 2024, 15, 2250. <https://doi.org/10.1038/s41467-024-46364-y>.
12. Schneider, N.; Lowe, D. M.; Sayle, R. A.; Landrum, G. A. Development of a Novel Fingerprint for Chemical Reactions and Its Application to Large-Scale Reaction Classification and Similarity. *Journal of Chemical Information and Modeling* 2015, 55, 39–53. <https://doi.org/10.1021/ci5006614>.
13. Bradshaw, J.; Zhang, A.; Mahjour, B.; Graff, D. E.; Segler, M. H. S.; Coley, C. W. Challenging Reaction Prediction Models to Generalize to Novel Chemistry. *ACS Central Science* 2025, 11, 539–549. <https://doi.org/10.1021/acscentsci.5c00055>.
14. Tadanki, A. S.; Rao, H. S. P.; Priyakumar, U. D. Dissecting Errors in Machine Learning for Retrosynthesis: A Granular Metric Framework and a Transformer-Based Model for More Informative Predictions. *Digital Discovery* 2025, 4, 831–845. <https://doi.org/10.1039/D4DD00263F>.
15. Keto, A.; Guo, T.; Gönner, N.; Zhang, X.; Krenske, E. H.; Wiest, O. Improving Reaction Prediction through Chemically Aware Transfer Learning. *Digital Discovery* 2025, 4, 1232–1238. <https://doi.org/10.1039/D4DD00412D>.
16. Qiao, J.; Li, K.; Jin, J.; Wang, D.; Gao, W.; Zhang, S.; Liu, S.; Wei, L. Relation-Aware Pretraining and Reaction Center Modeling for Chemical Reaction Graph Representation Learning. *Journal of Chemical Theory and Computation* 2026. <https://doi.org/10.1021/acs.jctc.6c00815>.
17. Xu, L.-C.; Tang, M.-J.; An, J.; Cao, F.; Qi, Y. A Unified Pre-Trained Deep Learning Framework for Cross-Task Reaction Performance Prediction and Synthesis Planning. *Nature Machine Intelligence* 2025, 7, 1561–1571. <https://doi.org/10.1038/s42256-025-01098-4>.
18. Xiong, X.; Jia, R.; Tian, Y.; Chen, J.; Tian, B. Reaction Center Prediction by Analyzing Attention of a Chemical Language Model. *Digital Discovery* 2026, 5, 2458–2468. <https://doi.org/10.1039/D6DD00055J>.
19. Xu, K.; Hu, W.; Leskovec, J.; Jegelka, S. How Powerful Are Graph Neural Networks? *International Conference on Learning Representations (ICLR)*, 2019.

20. Hu, W.; Liu, B.; Gomes, J.; Zitnik, M.; Liang, P.; Pande, V.; Leskovec, J. Strategies for Pre-training Graph Neural Networks. International Conference on Learning Representations (ICLR), 2020. <https://arxiv.org/abs/1905.12265>.
21. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. Advances in Neural Information Processing Systems 2017, 30.

Figures

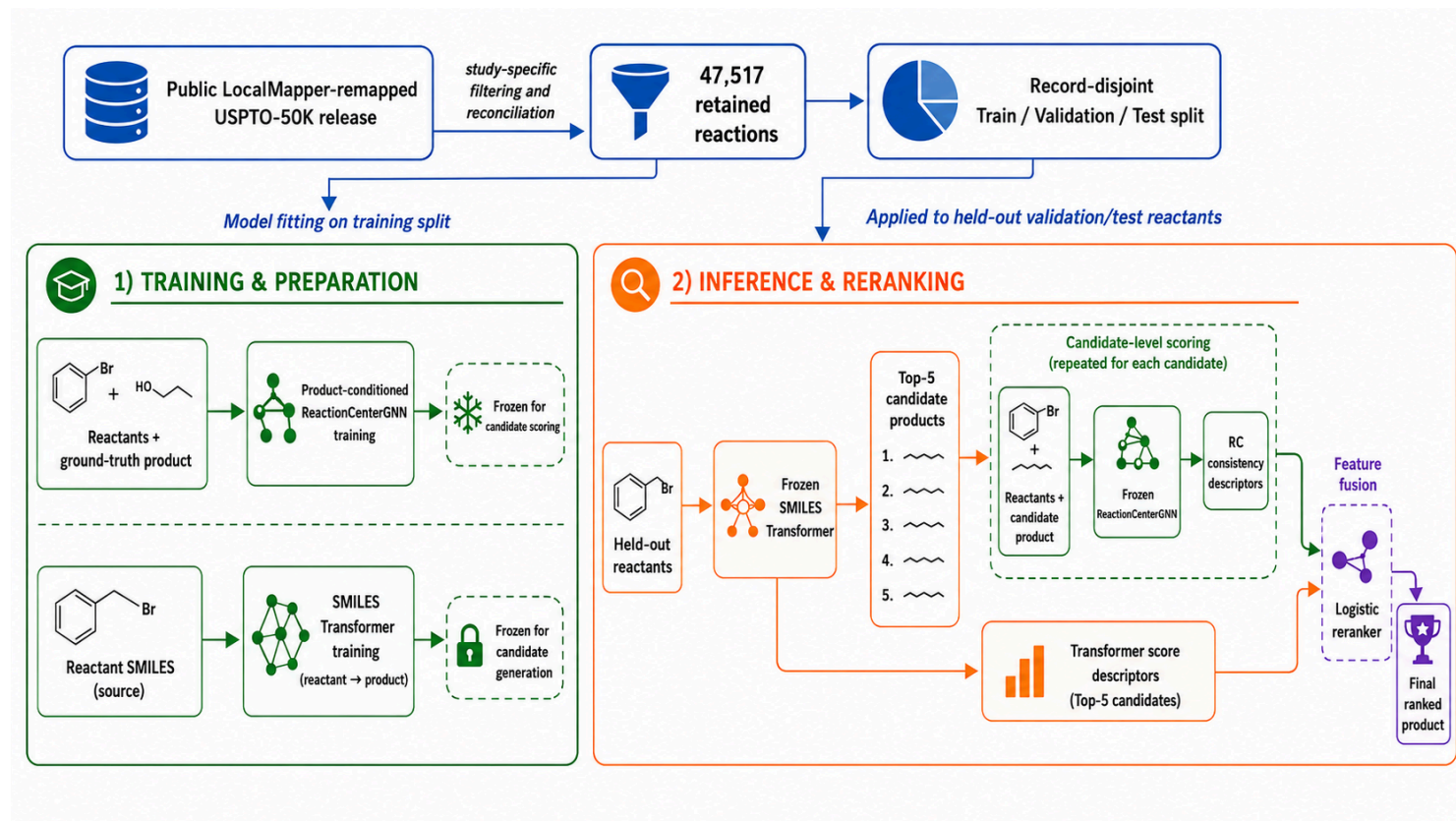


Figure 1

Proposed hybrid GNN–Transformer forward-reaction prediction framework. A public LocalMapper-remapped USPTO-50K release is partitioned into record-disjoint training, validation, and held-out test sets. During training and preparation, a product-conditioned ReactionCenterGNN is trained on reactants paired with the recorded product and later used as a frozen model, while a SMILES Transformer is trained on reactant-to-product pairs and later used as a fixed generator. During inference, the frozen SMILES Transformer generates Top-5 candidate products from the reactants. For each candidate, the frozen ReactionCenterGNN evaluates the reactants together with the candidate product to produce reaction-center (RC) consistency descriptors, while Transformer score descriptors are extracted from the Top-5 candidate set. These descriptor sets are fused and passed to a logistic reranker, which outputs the final top-ranked product

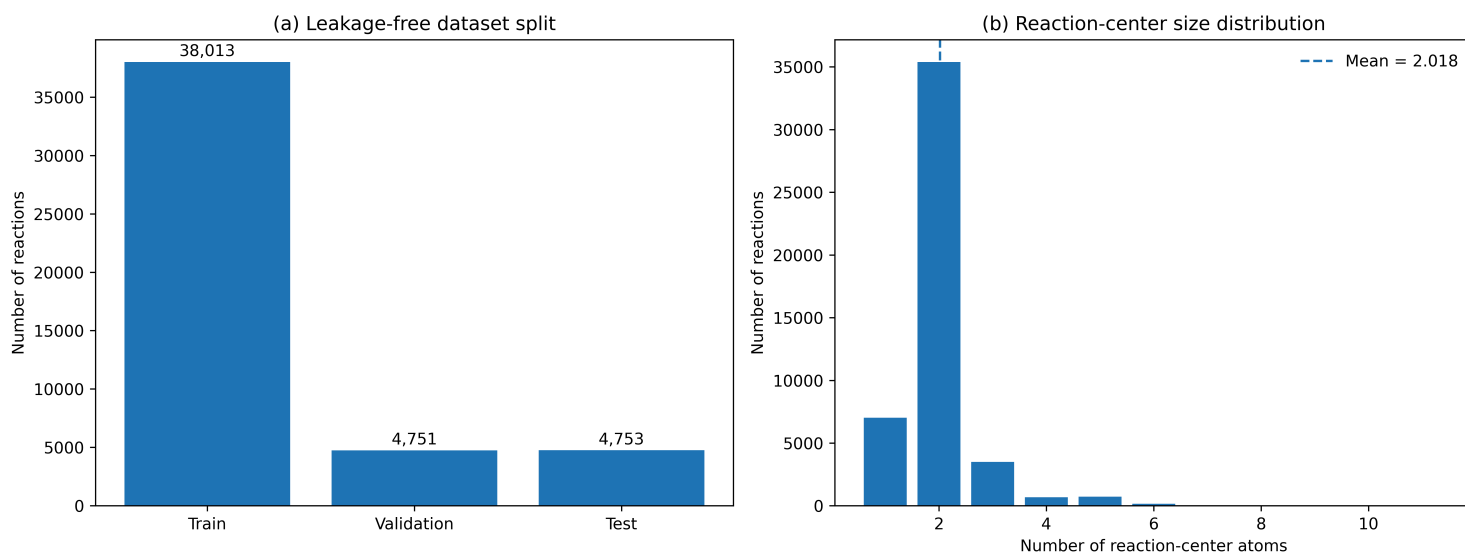


Figure 2

Dataset composition and enhanced reaction-center statistics. (left side) Record-disjoint training, validation, and held-out test partition of the 47,517 retained reactions derived from the public LocalMapper-remapped USPTO-50K release. (right side) Distribution of the number of enhanced reaction-center atoms per reaction; mean RC size = 2.018 atoms and median = 2.

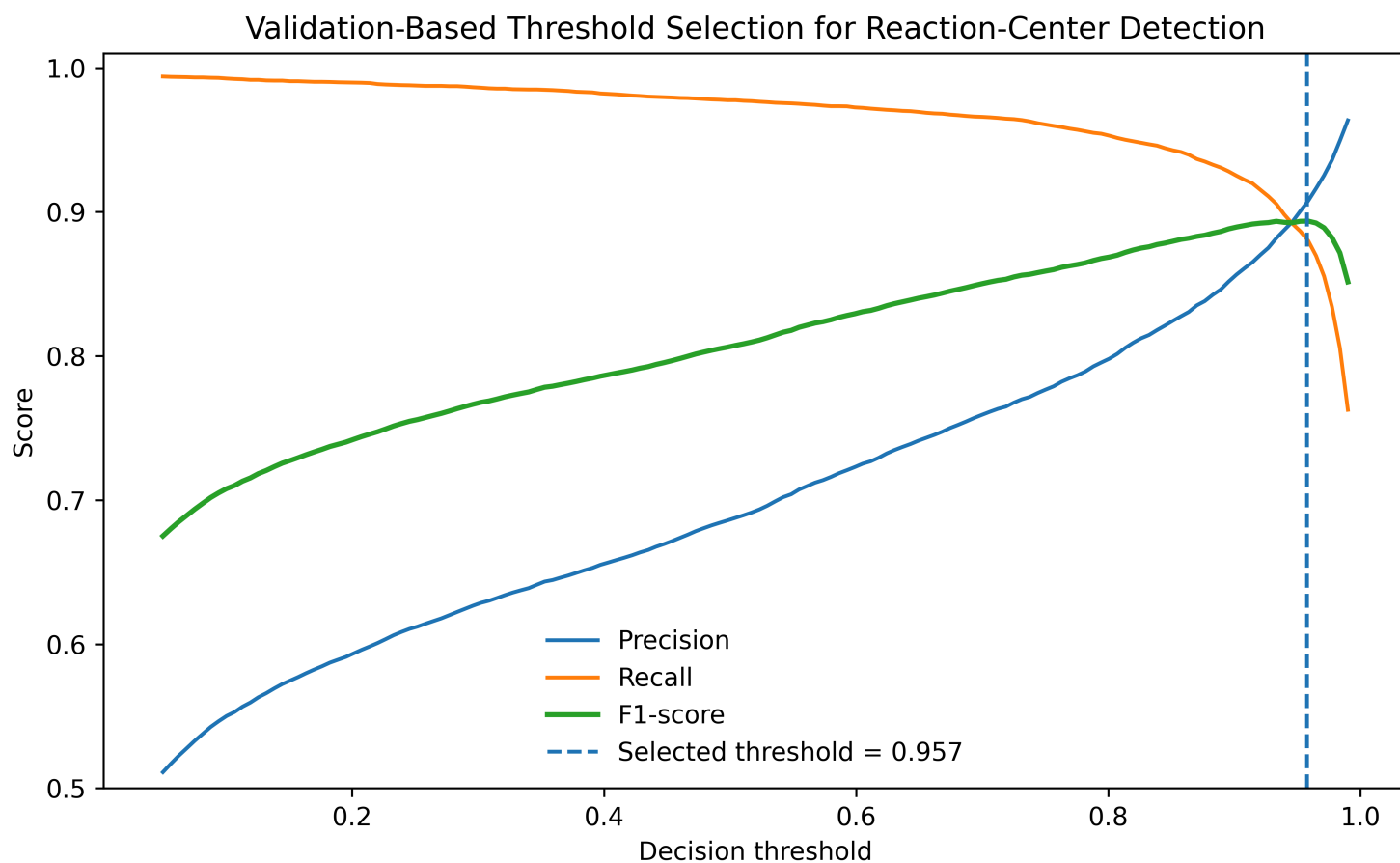


Figure 3

Validation-only threshold selection for reaction-center detection. Precision, recall, and F1 were evaluated across candidate thresholds; the selected operating point (0.9574) was frozen before held-out testing.

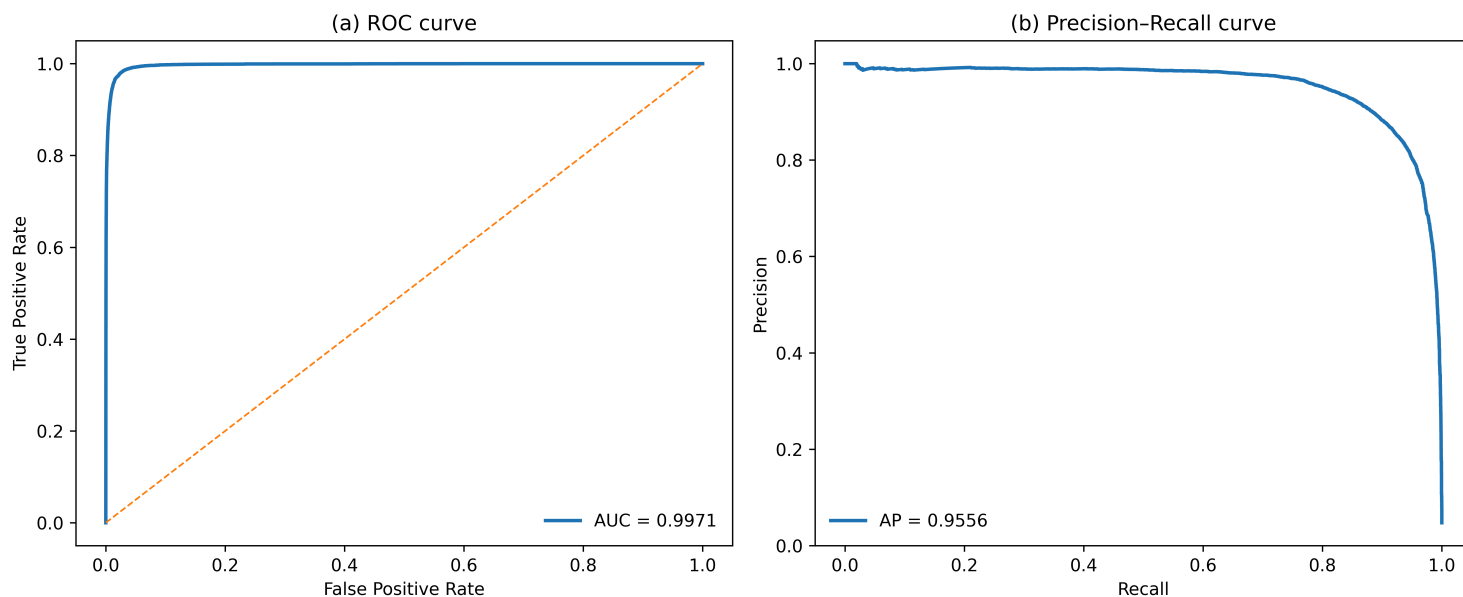


Figure 4

Held-out diagnostic reaction-center discrimination of the product-conditioned GNN when conditioned on the recorded product graph. (a) ROC curve (AUC = 0.9971). (b) Precision-recall curve (average precision = 0.9556). Product-prediction reranking instead conditions the same frozen GNN on Transformer-generated candidate products.

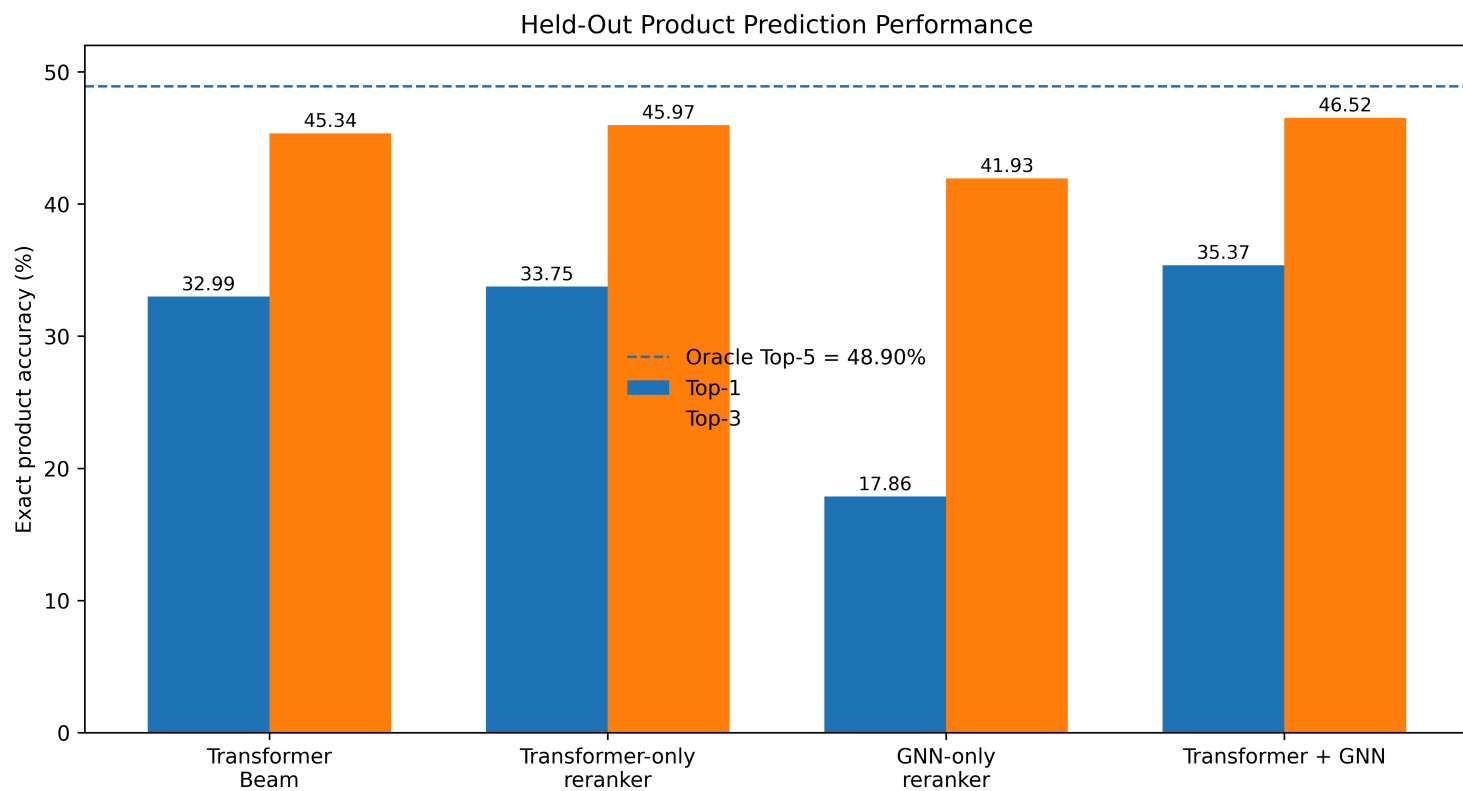


Figure 5

Held-out product-prediction performance of the original Transformer beam ranking and the learned rerankers. The dashed line denotes the Oracle Top-5 coverage (48.90%), which bounds reranking performance for the fixed candidate set.

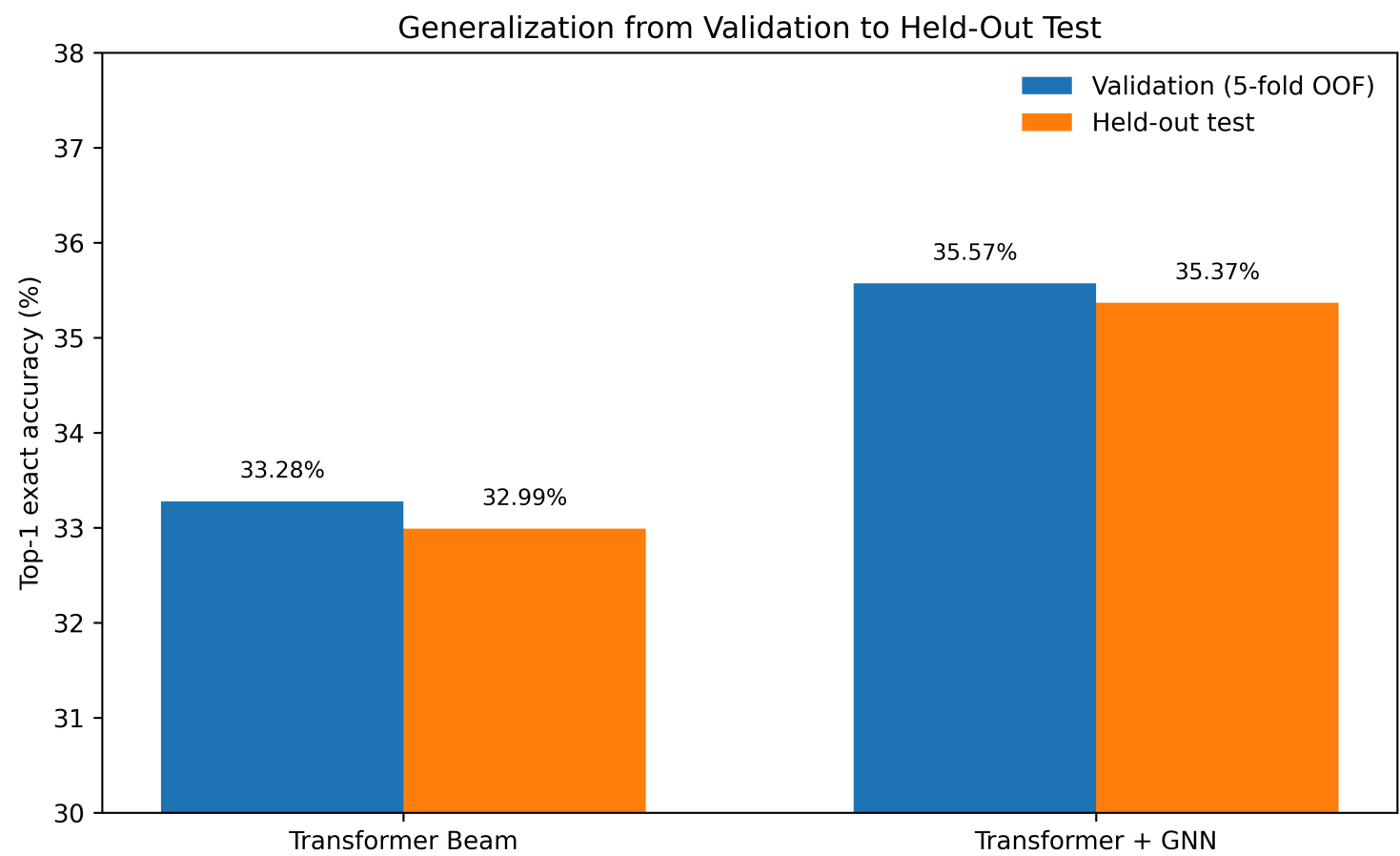


Figure 6

Generalization of the reranking gain from reaction-grouped out-of-fold validation to the held-out test set. The hybrid advantage is nearly unchanged between the two partitions.

Effect of GNN-Assisted Reranking on Transformer Predictions

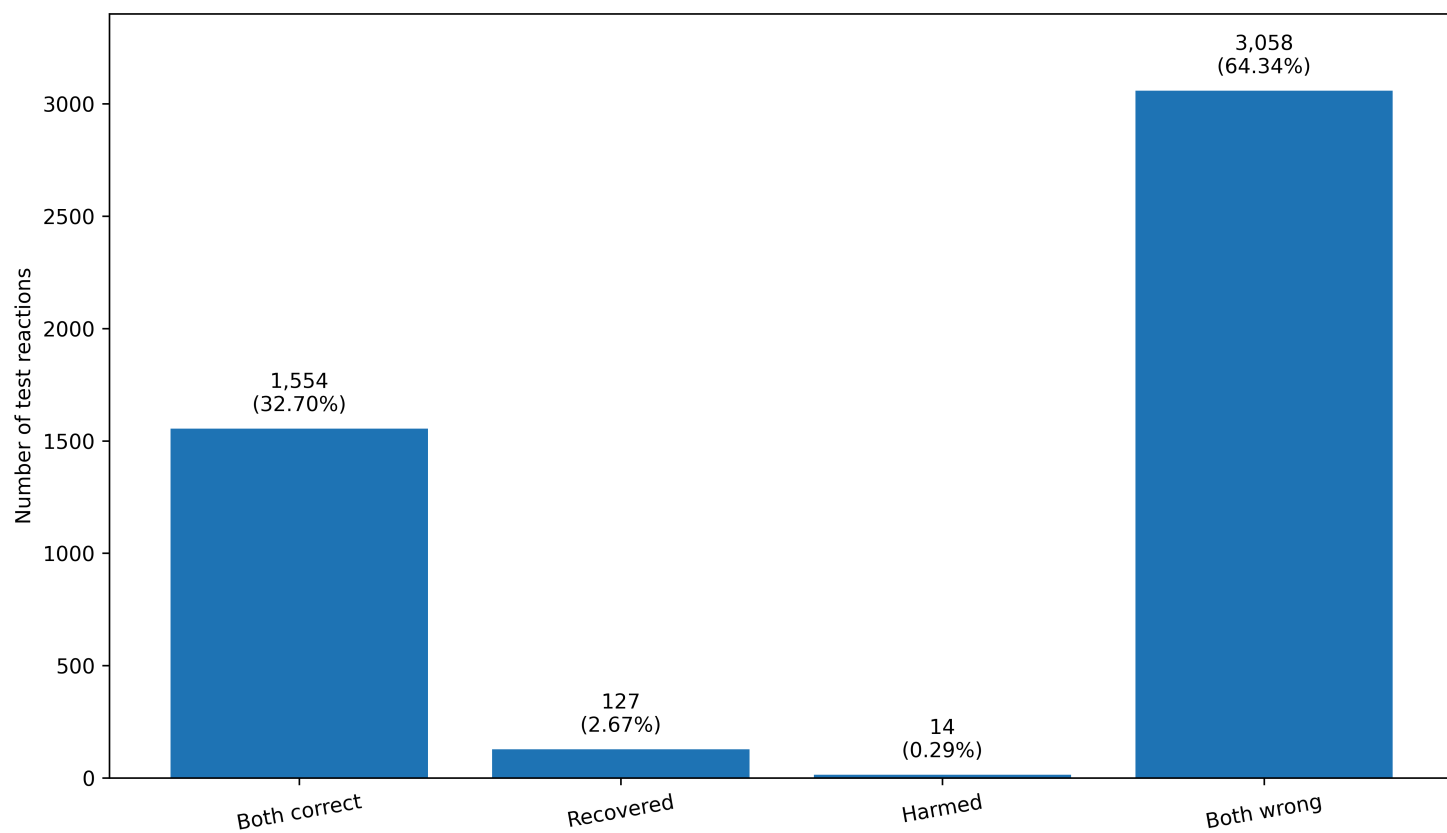


Figure 7

Reaction-level outcome categories for the original Transformer Top-1 prediction versus the final Transformer+GNN reranker on the held-out test set.

Original Beam Position of Products Recovered by GNN-Assisted Reranking

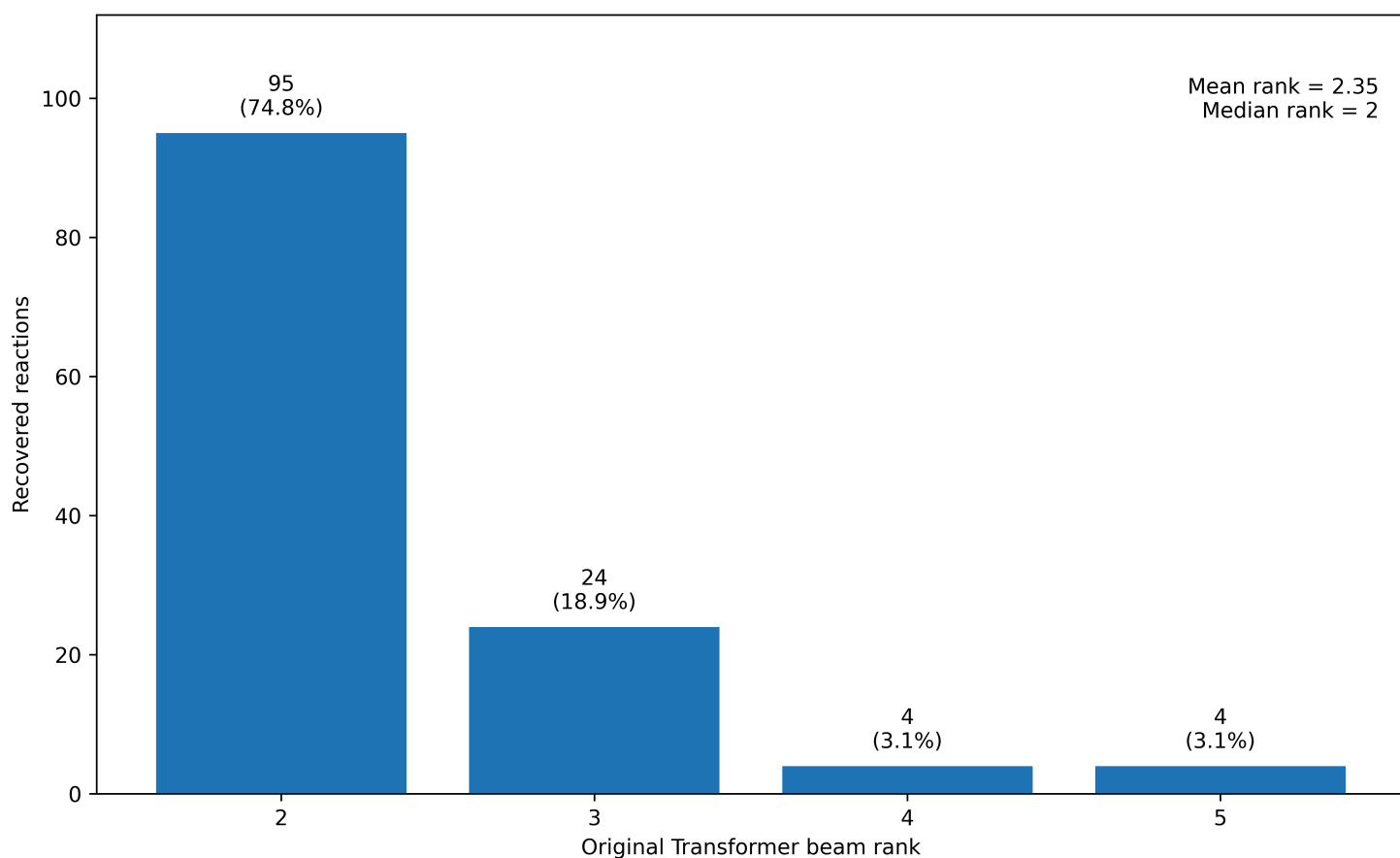


Figure 8

Original Transformer beam position of products recovered by the hybrid reranker. The correct product was at rank 2 in 95 of 127 recovered reactions; mean original rank = 2.35 and median = 2.

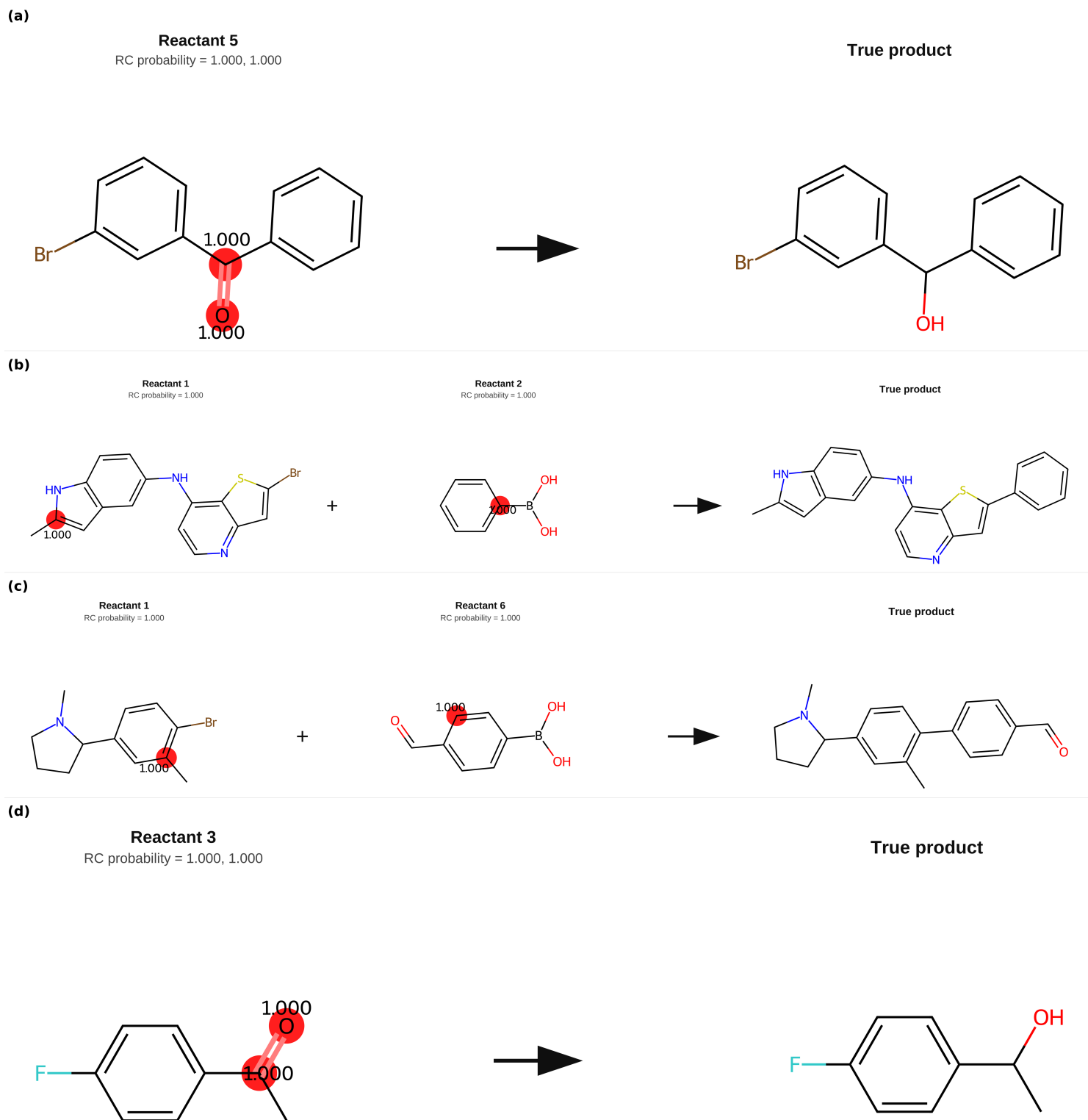


Figure 9

Representative recorded-product-conditioned reaction-center localizations used for qualitative interpretation. (a) Baseline-correct transformation, split index 1. (b) More complex baseline-correct two-reactant case, split index 304. (c) Hybrid recovery of the recorded product from original Transformer beam rank 2, split index 63. (d) Hybrid recovery from original beam rank 4, split index 3896. Red circles and numeric labels denote high-confidence reactant-side reaction-center probabilities predicted by the GNN. These visualizations inspect the auxiliary localization behavior; held-out product reranking is conditioned on Transformer-generated candidate products.

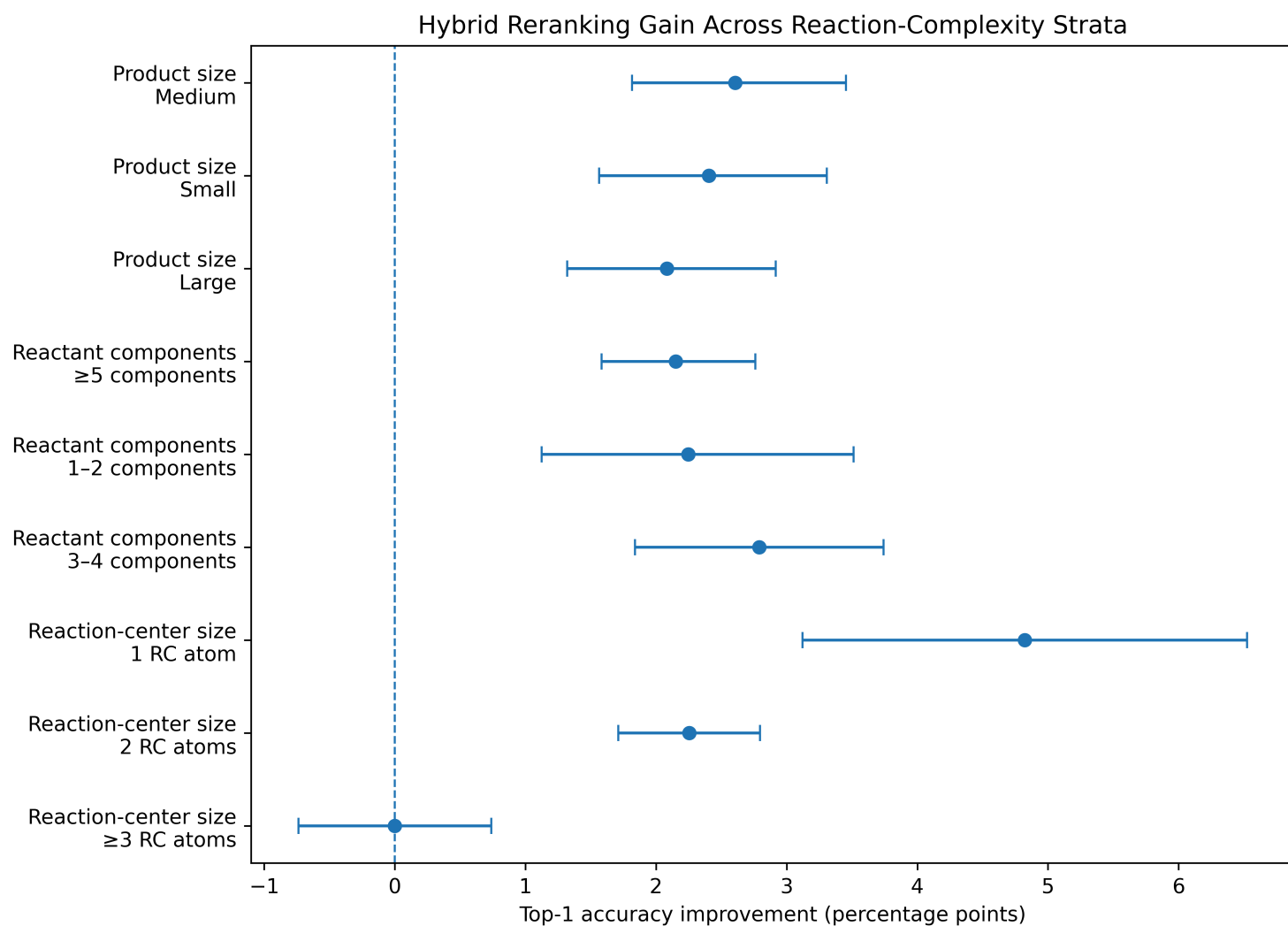


Figure 10

Top-1 accuracy gain of the hybrid reranker across objective reaction-complexity strata. Error bars show paired bootstrap 95% confidence intervals. The zero-gain result for reactions with at least three reaction-center atoms is retained as an explicit limitation.