RETROSPECT: RETROsynthesis via Sequential Prediction, and Chemically Transformed-ranking

Anonymous Authors[1]

Abstract

Single-step retrosynthesis needs both accurate first-ranked suggestions and candidate lists that are rich enough for downstream selection. We study this as a proposal-selection decomposition. Our system, RETROSPECT, combines a single Transformer proposal model, which we call the ChemAlign Transformer, with a LambdaMART reranker over structural, reaction-template, upstream-score, and optional DFT-derived descriptors. The generator is trained with hybrid root-aligned and random SMILES augmentation, Pre-LayerNorm, tied embeddings, exponential moving average weights, and a differentiable atom-balance auxiliary loss. On the full USPTO-50K test set of 5,007 reactions, the generator reaches 55.00% top-1 and 86.18% top-10 exact-match accuracy with 99.86% top-1 validity. On the merged candidate-pool benchmark used for reranking, which contains 5,007 test products and about 111 candidates per product, a LambdaMART model trained on the structural feature set reaches 59.4% top-1 with 0.7171 mean reciprocal rank. Feature ablations show that upstream proposal score and template-frequency statistics provide most of the reranking signal, while DFT and reaction-center DFT features provide smaller and less consistent gains. These results support a modular view of retrosynthesis: stronger single-model proposal and learned candidate selection are complementary, and the proposal model can serve as a drop-in component for ensemble systems such as RetroChimera (Maziarz et al., 2024).

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country.**AUTHORERR: Missing \icmlcorrespondingauthor.**

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

Retrosynthesis asks which precursor molecules can produce a target molecule. The problem is central to computer-aided synthesis planning, where a single-step model is repeatedly called inside a multi-step search procedure (Corey & Wipke, 1969; Coley et al., 2018; Segler et al., 2018). A useful single-step model must therefore satisfy two requirements. It must place a correct disconnection near the top of its ranked list, and it must preserve enough plausible alternatives for a planner or chemist to recover when the first suggestion is unavailable, unsafe, or strategically poor.

Many recent retrosynthesis systems fold proposal and ranking into one stage. Template-based methods classify or retrieve reaction templates (Dai et al., 2019; Somnath et al., 2021; Chen & Jung, 2021; Gaiński et al., 2024). Template-free methods generate precursor strings or graphs directly (Liu et al., 2017; Tetko et al., 2020; Wan et al., 2022; Maziarz et al., 2024; Igashov et al., 2024; Yadav et al., 2025; Han et al., 2024). Both families ultimately produce ranked candidates, but the mechanisms that enumerate plausible disconnections and the mechanisms that decide their order are not necessarily the same.

This paper focuses on that separation. We develop a stronger single proposal model, the ChemAlign Transformer, then study a learning-to-rank stage that reranks merged candidate pools produced by beam search and SMILES augmentation. The proposal model is intentionally single-model rather than ensemble-based. This lets us ask two cleaner questions. First, how much can a carefully trained Transformer proposal model achieve on its own? Second, once a proposal pool exists, which feature families actually improve its ordering?

Our contributions are:

1. We present RETROSPECT, a modular proposal-plus-reranking framework for single-step retrosynthesis that cleanly separates candidate generation from candidate selection.
2. We develop the ChemAlign Transformer, a stronger Augmented Transformer variant using hybrid root-aligned/random SMILES augmentation, Pre-

LayerNorm, exponential moving average weights, tied token embeddings, and a differentiable atom-balance auxiliary loss.

3. We implement a LambdaMART reranker over upstream score, structural descriptors, reaction-template descriptors, and optional DFT-derived features, and document the train-split freezing needed for frequency-style statistics.

4. We provide ablations that show where the reranking signal comes from: proposal score and template-derived features matter most in the current setup, while DFT and reaction-center DFT features are weaker and should be treated cautiously.

The resulting picture is deliberately conservative. Our verified numbers support a strong single-model proposal system and a useful reranking study, not a claim that every component uniformly surpasses the best published end-to-end baselines. Instead, we argue that the proposal model is competitive as a standalone system and attractive as a drop-in candidate source for ensemble frameworks such as RetroChimera (Maziarz et al., 2024).

2. Related work

Template-based retrosynthesis. Template-based systems classify or retrieve reaction templates, then apply them to a product molecule. GLN models product-template and reactant-template compatibility with graph logic (Dai et al., 2019). GraphRetro decomposes prediction into reaction-center identification and leaving-group completion (Somnath et al., 2021). LocalRetro predicts local reactivity with global attention and remains a strong high-k template-based baseline (Chen & Jung, 2021). RetroGFN uses generative flow networks to diversify template selections (Gaiński et al., 2024). These methods are interpretable and chemically constrained, but their proposal mechanism is still tied to a template inventory.

Template-free and semi-template methods. Sequence-to-sequence retrosynthesis treats product-to-reactant prediction as translation (Liu et al., 2017; Zheng et al., 2020; Tetko et al., 2020). Later systems add graph structure, edit operations, and reaction-center priors, including G2Gs (Shi et al., 2020), MEGAN (Sacha et al., 2021), GTA (Seo et al., 2021), Graph2SMILES (Tu & Coley, 2022), Retroformer (Wan et al., 2022), Graph2Edits (Zhong et al., 2023), G2Retro (Chen et al., 2023), and EditRetro (Han et al., 2024). Recent frontier systems include RetroChimera, which ensembles complementary proposal models with learned molecule-space ranking (Maziarz et al., 2024), and Retro SynFlow, which applies discrete flow matching from synthons to reactants (Yadav et al., 2025). Our work is closest in spirit to systems that explicitly acknowledge multiple proposal hypotheses, but we study the proposal model and reranker as separate units rather than introducing a new ensemble.

Aligned reaction representations. Root-aligned SMILES reduce source-target edit distance by starting product and reactant SMILES from corresponding atoms (Zhong et al., 2022). This representation is especially helpful for reactions because much of the molecular graph is preserved. Our generator uses a hybrid augmentation strategy, mixing root-aligned and random SMILES, so the model receives both tightly aligned supervision and traversal diversity.

Learning to rank for retrosynthesis. Ranking appears implicitly in beam search, template probabilities, and ensemble aggregation. RetroChimera uses learned ranking to combine predictions from several complementary models (Maziarz et al., 2024). Retro-Rank-In formulates inorganic retrosynthesis as ranking precursor candidates in a shared latent space (Prein et al., 2025). We focus on organic USPTO-50K retrosynthesis and study how a listwise LambdaMART reranker interacts with a single strong proposal model.

3. Method

Given target molecule T, RETROSPECT returns a ranked list of precursor sets $P_1, \ldots, P_K$. Figure 1 shows the pipeline.

3.1. Proposal model: the ChemAlign Transformer

The generator is an encoder-decoder Transformer over tokenized SMILES. It builds on the Augmented Transformer (Tetko et al., 2020), but we rename the model ChemAlign Transformer to emphasize reaction-aligned supervision, stronger optimization, and chemical regularization. The model uses six encoder layers, six decoder layers, hidden size 512, eight attention heads, feed-forward size 2048, dropout 0.1, sinusoidal positional encodings up to length 1000, and an 81-token vocabulary built from the training split only using a Molecular Transformer-style regex tokenizer (Schwaller et al., 2019). Source sequences append `<EOS>`; target sequences use `<BOS>` and `<EOS>` delimiters.

Training uses 20-fold offline augmentation over 40,008 training reactions, yielding 800,160 source-target pairs. Sixteen augmented pairs per reaction use root-aligned SMILES, which align product and reactant traversals around corresponding atoms (Zhong et al., 2022), and four use random SMILES to preserve traversal invariance. Randomly augmented multi-fragment reactants are canonically fragment-sorted to reduce output-order variance, while root-aligned reactants retain alignment order. Validation and test inputs

Figure 1. RETROSPECT separates candidate proposal from candidate selection. The generator produces candidates under multiple SMILES traversals, these candidates are merged and deduplicated, and a listwise reranker reorders the resulting proposal pool.

remain canonical. We use Pre-LayerNorm for stable optimization (Xiong et al., 2020), three-way weight tying among encoder embeddings, decoder embeddings, and output projection (Press & Wolf, 2017), Xavier-style initialization, and EMA weights with decay 0.999 for validation and inference.

Optimization details are summarized in Section B. We train with token-normalized cross-entropy, no label smoothing, Adam with $\beta_1 = 0.9$, $\beta_2 = 0.998$, and $\epsilon = 10^{-9}$, and a Noam schedule with factor 2.0 and 8,000 warmup steps. Batches are packed by target length up to 16,384 tokens per micro-batch, accumulated for two steps, and trained with AMP. Validation runs every 2,000 optimizer steps, and the best EMA checkpoint occurs at step 20,000.

The auxiliary atom-balance loss penalizes deviations between expected element counts under the decoder softmax and ground-truth element fractions. Let z_t be decoder logits at position t and let A map each token to element counts for 12 elements. We compute expected counts $\hat{a} = \sum_t \mathrm{softmax}(z_t)A$ and add an L1 penalty with coefficient 0.1. This loss is not a hard validity constraint, but it biases the model away from mass-balance violations while preserving differentiability.

3.2. LambdaMART reranking over merged candidate pools

The reranker is an XGBoost LambdaMART model trained with the listwise `rank:ndcg` objective (Chen & Guestrin, 2016; Burges, 2010). Candidate pools are produced by running the generator under multiple product traversals, then merging and deduplicating the resulting hypotheses. For each target-candidate pair, we compute four feature blocks: structural descriptors, optional DFT-derived descriptors, reaction-template descriptors, and the upstream proposal score. Structural descriptors include pharmacophore counts, functional-group indicators, atom-count deltas, and Morgan/MACCS fingerprint similarities (Rogers & Hahn, 2010). Reaction-template descriptors include whether a template can be extracted for the pair, hashed template identifiers at multiple radii, and train-frozen template-frequency statistics. DFT descriptors, when available, include reaction-level differences in HOMO, LUMO, gap, dipole, hardness, softness, and derived frontier-orbital features.

Before ranking, candidates are canonicalized to a shared precursor-set representation and merged by canonicalized precursor SMILES. If the same precursor set is proposed multiple times, we keep one merged row, retain source provenance, and carry forward the best available upstream score for that precursor set. LambdaMART is trained groupwise within each product, never across products. The training targets use graded relevance labels so exact precursor-set matches receive the highest gain while partial fragment overlap receives weaker positive labels. In the current tuned models, the upstream score is the single strongest feature, template-frequency features provide the next-largest gains, and DFT features remain secondary. Additional protocol details and the ranker hyperparameters are summarized in Sections D and E.

4. Experiments

Dataset. We evaluate on USPTO-50K, using the standard split of 40,008 training, 5,000 validation, and 5,007 test reactions (Schneider et al., 2016; Liu et al., 2017). We report exact-match top-k accuracy after canonicalizing predicted and ground-truth precursor sets. Unless noted otherwise, reaction class is unknown at test time.

Preprocessing and augmentation. Raw USPTO reactions follow the atom-mapped format popularized by Lowe (Lowe, 2012). We discard reagents, remove atom mapping, and canonicalize products and precursor sets with RDKit before training. The generator uses a double-canonicalization pass for stereochemical consistency, then applies 20 offline augmentations per training reaction, while validation and test examples remain canonical. Multi-fragment ordering is a real modeling issue in this benchmark: 70.7% of training reactions have two precursor fragments, 29.1% have one, and 0.2% have three. Randomly augmented reactant fragments are therefore canonically sorted during preprocessing, while R-SMILES variants preserve product-aligned fragment order.

Proposal and reranking protocols. The full-test proposal result uses canonical-input inference only and is the fairest single-model comparison to prior seq2seq systems. The reranking study uses larger merged candidate pools: train and validation pools are built from fewer beam-search augmentations per product than the test pool, while the test reranking benchmark contains 5,007 products with about 111 candidates per product on average. This mismatch makes the reranker closer to a robust reordering model than a memorizer of one fixed candidate width. We freeze all frequency-style statistics on the training split before applying them to validation and test examples.

Baselines. Our literature survey includes classical template-based baselines, modern template-free baselines, and recent strong systems including EditRetro (Han et al., 2024), RetroChimera (Maziarz et al., 2024), and Retro Syn-Flow (Yadav et al., 2025). We compare the ChemAlign Transformer directly against published single-step systems in the main table. We report the reranking ablations separately because they operate on a merged candidate-pool benchmark rather than the full 5,007-reaction end-to-end test set.

4.1. Single-model proposal results

Table 1 reports the generator alone. We position RETROSPECT as a single-model system rather than an ensemble, and this table should be read as the main evidence for that claim. The best full-test RETROSPECT generator checkpoint, which is also our best single-model configuration, reaches 55.00% top-1 and 86.18% top-10 with 99.86% top-1 validity. A 15K-reaction ablation shows that hybrid root-aligned SMILES is the dominant design choice, improving top-1 by 9.23 percentage points over the scaled baseline. Pre-LayerNorm, EMA, and atom-balance regularization add another 1.54 points. The cumulative ablation table appears in Section C.

4.2. Reranking on merged candidate pools

The full RETROSPECT row in Table 1 adds a LambdaMART reranker on the merged candidate pool of about 111 candidates per product. Reranking lifts top-1 from 55.00% to 59.4%, top-10 from 86.18% to 93.06%, and reaches 0.7171 mean reciprocal rank. Earlier feature ablations over a V1 proposal pool, reported in Section F, show that the choice of feature set shifts top-k by less than one percentage point and that adding DFT or reaction-center DFT does not clearly dominate simpler sets, so we run the V2 reranker on the structural feature set only. Full protocol and tuning details appear in Sections D and E.

5. Analysis

A stronger proposal model does not eliminate the value of selection. The generator is a competitive single model, but the reranking study shows that proposal and ordering still capture different signals. The proposal model determines whether plausible candidates enter the pool. The reranker exploits upstream score, template-derived priors, and structural compatibility to reorder those candidates once they exist.

Upstream score and template priors dominate the current reranker. The feature-importance profile and ablations indicate that the proposal score is the strongest single feature, while template-frequency and template-identity features add a large fraction of the remaining gain. This is useful scientifically because it narrows where future effort should go: better candidate scoring and better train-split template statistics are currently more valuable than larger DFT stacks.

DFT and reaction-center DFT are promising but not yet central. The DFT feature families can improve certain higher-k metrics or MRR, especially in reaction-center variants, but the gains are small and inconsistent across settings. In the current paper, they should therefore be treated as exploratory chemistry-aware features rather than the main explanation for performance.

The proposal model is modular enough for ensemble systems. RetroChimera already shows that complementary proposal mechanisms and learned ranking can outperform any one component (Maziarz et al., 2024). Our results suggest a practical use case: the ChemAlign Transformer is a stronger single-model proposal module than earlier Augmented Transformer baselines, so it is a natural component to test inside ensemble-and-ranking frameworks rather than only as a standalone decoder.

6. Limitations and broader impact

The main limitation is that results are reported on USPTO-50K, a widely used but small and biased benchmark extracted from patent reactions. The proposal result and the reranking result answer related but different questions: the full-test table measures standalone single-model exact-match accuracy, while the reranking study measures ordering quality on merged candidate pools with 5,007 products and about 111 candidates per product. If the correct precursor is absent from the proposal pool, no reranker can recover it. DFT-derived features and reaction-center features are also under-validated relative to the generator, and their benefits remain modest in the current ablations. Finally, exact-match accuracy rewards reproducing the recorded patent precursor set, even when other chemically valid disconnections exist.

The positive impact of improved retrosynthesis is faster synthesis planning for drug discovery, materials design, and chemical supply chains. Potential risks include accelerating access to harmful compounds or over-trusting algorithmic routes without expert review. We view RETROSPECT as a decision-support system for trained chemists rather than an autonomous synthesis planner. Practical deployment should include building-block availability checks, condition prediction, safety filters, and human review.

Table 1. Top-k exact-match accuracy on the USPTO-50K test set (5,007 reactions) with reaction class unknown. The ChemAlign Transformer row reports beam search alone; the RETROSPECT row adds a LambdaMART reranker trained on the structural feature set over the merged candidate pool of about 111 candidates per product. TB = template-based, ST = semi-template, TF = template-free.

Type	Method	Top-1	Top-3	Top-5	Top-10
TB	GLN (Dai et al., 2019)	52.5	74.7	81.2	87.9
TB	LocalRetro (Chen & Jung, 2021)	53.4	77.5	85.9	92.4
ST	GraphRetro (Somnath et al., 2021)	53.7	68.3	72.2	75.5
ST	G2Retro (Chen et al., 2023)	54.1	74.1	81.2	86.7
TF	Graph2Edits (Zhong et al., 2023)	55.1	77.3	83.4	89.4
TF	R-SMILES (Zhong et al., 2022)	56.3	79.2	86.2	91.0
TF	RetroChimera (Maziarz et al., 2024)	59.6	**82.8**	**89.2**	**94.2**
TF	Retro SynFlow (Yadav et al., 2025)	$60.0_{\pm 0.22}$	$77.9_{\pm 0.13}$	$82.7_{\pm 0.15}$	$85.3_{\pm 0.19}$
TF	EditRetro (Han et al., 2024)	**60.8**	80.6	86.0	90.3
Ours	ChemAlign Transformer only	55.00	76.13	81.33	86.18
Ours	RETROSPECT, structural	59.4	**82.02**	**87.51**	**93.06**

7. Conclusion

We presented RETROSPECT, a proposal-plus-reranking framework for single-step retrosynthesis. The ChemAlign Transformer provides a strong single-model proposal system through aligned SMILES supervision and improved optimization, while LambdaMART shows that learned candidate selection still adds useful signal once a rich proposal pool is available. The cleanest empirical conclusions are threefold. First, root-aligned augmentation is the dominant contributor to proposal quality. Second, reranking gains in the current setup come mainly from proposal score and template-derived priors. Third, DFT-based features remain exploratory. This decomposition makes the system scientifically easier to analyze and practically easier to reuse: the proposal model can stand alone, and it can also be inserted into ensemble systems such as RetroChimera for future work.

References

Burges, C. J. From RankNet to LambdaRank to LambdaMART: An overview. *Learning*, 11:23–581, 2010.

Chen, S. and Jung, Y. Deep retrosynthetic reaction prediction using local reactivity and global attention. *JACS Au*, 1(10):1612–1620, 2021.

Chen, T. and Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794. ACM, 2016. doi: 10.1145/2939672.2939785.

Chen, Z., Ayinde, O. R., Fuchs, J. R., Sun, X. N., and Ning, X. G2Retro as a two-step graph generative models for retrosynthesis prediction. *Communications Chemistry*, 6 (1):102, 2023. doi: 10.1038/s42004-023-00897-3.

Coley, C. W., Green, W. H., and Jensen, K. F. Machine learning in computer-aided synthesis planning. *Accounts of Chemical Research*, 51(5):1281–1289, 2018.

Corey, E. J. and Wipke, W. T. Computer-assisted design of complex organic syntheses. *Science*, 166(3902):178–192, 1969.

Dai, H., Li, C., Coley, C. W., Dai, B., and Song, L. Retrosynthesis prediction with conditional graph logic network. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Gaiński, P., Koziarski, M., Maziarz, K., Segler, M., Tabor, J., and Śmieja, M. Retrogfn: Diverse and feasible retrosynthesis using gflownets. *arXiv preprint arXiv:2406.18739*, 2024.

Han, Y., Xu, X., Hsieh, C.-Y., Ding, K., Xu, H., Xu, R., Hou, T., Zhang, Q., and Chen, H. Retrosynthesis prediction with an iterative string editing model. *Nature Communications*, 15(1):6404, 2024. doi: 10.1038/s41467-024-50617-1.

Igashov, I., Schneuing, A., Segler, M., Bronstein, M., and Correia, B. Retrobridge: Modeling retrosynthesis with markov bridges. In *International Conference on Learning Representations (ICLR)*, 2024.

Liu, B., Ramsundar, B., Kawthekar, P., Shi, J., Gomes, J., Nguyen, Q. L., Ho, S., Sloane, J., Wender, P., and Pande, V. Retrosynthetic reaction prediction using neural sequence-to-sequence models. *ACS Central Science*, 3 (10):1103–1113, 2017.

Lowe, D. M. *Extraction of Chemical Structures and Reactions from the Literature*. PhD thesis, University of Cambridge, 2012.

Maziarz, K., Liu, G., Misztela, H., Tripp, A., Li, J., Kornev, A., Gaiński, P., Hoefling, H., Fortunato, M., Gupta, R.,

and Segler, M. Chemist-aligned retrosynthesis by ensembling diverse inductive bias models. *arXiv preprint arXiv:2412.05269*, 2024.

Prein, T., Pan, E., Haddouti, S., Lorenz, M., Jehkul, J., Wilk, T., Moran, C., Fotiadis, M. P., Toshev, A. P., and Olivetti, E. Retro-rank-in: A ranking-based approach for inorganic materials synthesis planning. *arXiv preprint arXiv:2502.04289*, 2025.

Press, O. and Wolf, L. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 157–163, 2017.

Rogers, D. and Hahn, M. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010.

Sacha, M., Błaż, M., Byrski, P., Włodarczyk-Pruszyński, P., and Jastrzębski, S. Molecule edit graph attention network: Modeling chemical reactions as sequences of graph edits. *Journal of Chemical Information and Modeling*, 61(7): 3273–3284, 2021.

Schneider, N., Stiefl, N., and Landrum, G. A. What's what: The (nearly) definitive guide to reaction role assignment. *Journal of Chemical Information and Modeling*, 56(12): 2336–2346, 2016. doi: 10.1021/acs.jcim.6b00564.

Schwaller, P., Laino, T., Gaudin, T., Bolgar, P., Hunter, C. A., Bekas, C., and Lee, A. A. Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. *ACS Central Science*, 5(9):1572–1583, 2019.

Segler, M. H. S., Preuss, M., and Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature*, 555(7698):604–610, 2018.

Seo, S.-W., Song, Y. Y., Yang, J. Y., Bae, S., Lee, H., Shin, J., Hwang, S. J., and Yang, E. GTA: Graph truncated attention for retrosynthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 531–539, 2021. doi: 10.1609/aaai.v35i1.16131.

Shi, C., Xu, M., Guo, H., Zhang, M., and Tang, J. A graph to graphs framework for retrosynthesis prediction. In *International Conference on Machine Learning (ICML)*, pp. 8818–8827. PMLR, 2020.

Somnath, V. R., Bunne, C., Coley, C. W., Krause, A., and Barzilay, R. Learning graph models for retrosynthesis prediction. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

Tetko, I. V., Karpov, P., Van Deursen, R., and Godin, G. State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis. *Nature Communications*, 11(1):5575, 2020.

Tu, Z. and Coley, C. W. Permutation invariant graph-to-sequence model for template-free retrosynthesis and reaction prediction. *Journal of Chemical Information and Modeling*, 62(15):3503–3513, 2022.

Wan, Y., Hsieh, C.-Y., Liao, B., and Jia, S. Retroformer: Pushing the limits of end-to-end retrosynthesis transformer. In *International Conference on Machine Learning (ICML)*, pp. 22475–22490. PMLR, 2022.

Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., Zhang, H., Lan, Y., Wang, L., and Liu, T.-Y. On layer normalization in the transformer architecture. In *International Conference on Machine Learning (ICML)*, pp. 10524–10533. PMLR, 2020.

Yadav, R., Yan, Q., Wolf, G., and Bose, A. J. RETRO SynFlow: Discrete flow matching for accurate and diverse single-step retrosynthesis. *arXiv preprint arXiv:2506.04439*, 2025.

Zheng, S., Rao, J., Zhang, Z., Xu, J., and Yang, Y. Predicting retrosynthetic reactions using self-corrected transformer neural networks. *Journal of Chemical Information and Modeling*, 60(1):47–55, 2020.

Zhong, W., Yang, Z., and Chen, C. Y.-C. Retrosynthesis prediction using an end-to-end graph generative architecture for molecular graph editing. *Nature Communications*, 14 (1):3009, 2023. doi: 10.1038/s41467-023-38851-5.

Zhong, Z., Song, J., Feng, Z., Liu, T., Jia, L., Yao, S., Wu, M., Liu, T., and Song, M. Root-aligned SMILES: A tight representation for chemical reaction prediction. *Chemical Science*, 13(31):9023–9034, 2022. doi: 10.1039/D2SC02763A.

A. Dataset and preprocessing details

Table 2. Dataset and generator-preprocessing statistics used in the draft. Augmentation is applied only to the training split.

Statistic	Value	Note
Train / val / test reactions	40,008 / 5,000 / 5,007	Standard USPTO-50K split
Augmented training pairs	800,160	20× offline augmentation
R-SMILES / random ratio	16 / 4	Per training reaction
Vocabulary size	81	Train-split only
Mean source / target length	45.8 / 50.7	Sampled 5K training pairs
1 / 2 / 3 precursor fragments	29.1 / 70.7 / 0.2%	Training reactions

B. Generator training and inference configuration

Table 3. Generator configuration summary for the reproducible Baseline V2 generator used in this draft.

Parameter	Value	Note
Encoder / decoder layers	6 / 6	Transformer
d_{model} / heads / FFN	512 / 8 / 2048	Baseline V2
Dropout	0.1	Training and decoding stack
Optimizer	Adam	$\beta_1 = 0.9$, $\beta_2 = 0.998$, $\epsilon = 10^{-9}$
Learning-rate schedule	Noam	factor 2.0, warmup 8,000
Max tokens / accum steps	16,384 / 2	Dynamic token batching
Mixed precision / EMA	AMP fp16 / 0.999	EMA checkpoint used for inference
Best checkpoint step	20,000	Early-stopped training
Beam search	max length 200	Length norm $\alpha = 0.6$
Canonical test-time inference	Yes	Main table generator numbers
TTA extension	Available	Random product SMILES + aggregation

C. Generator ablation

Table 4. Cumulative generator ablation on a 15K-reaction subset with 3K test reactions.

Version	Added technique	Top-1	Delta
V1	Baseline, 5x random augmentation	34.97	–
V6	Larger Transformer	36.60	+1.63
V7	Hybrid R-SMILES	45.83	+9.23
V8	Pre-LN, EMA, atom-balance loss	47.37	+1.54
V8-SWA	Stochastic weight averaging	48.33	+0.96
V9	Label smoothing and weight decay	45.17	-3.16

D. Ranker protocol

Table 5. Compact summary of the LambdaMART reranking protocol.

Item	Setting
Objective	XGBoost LambdaMART with `rank:ndcg`, optimized for NDCG@10
Training groups	Candidates grouped by product, ranking never compares candidates across products
Labels	Graded relevance, exact precursor-set matches receive highest gain and partial fragment overlap receives weaker positive labels
Feature blocks	Upstream proposal score, structural descriptors, reaction-template descriptors, optional DFT descriptors
Template statistics	Frequency-style template features fit on the training split and frozen before validation/test scoring
Candidate pools	Train: 39,736 products, 29.9 candidates/product; val: 4,993 products, 30.1 candidates/product; test: 5,007 products, 111.4 candidates/product
Proposal coverage	Ground truth present in 96.3% of train pools, 91.0% of val pools, and 92.9% of test pools
Observed signal	Upstream score is dominant, template-frequency features are next most useful, DFT features are secondary in the current runs

E. Ranker tuning summary

Table 6. Summary of the tuned LambdaMART configuration and sweep outcome.

Parameter	Value	Note
Objective / eval metric	`rank:ndcg` / NDCG@10	Listwise ranking objective
Learning rate	0.01	Tuned from 0.05
Max depth	12	Tuned from 5
Min child weight	6	Tuned from 10
Subsample / colsample	0.8 / 0.4	Tree regularization
Gamma / max delta step	1.3 / 1	Tuned stabilizers
Tree method	`hist`	CPU-friendly training
Rounds / early stopping	5000 / 200	Best round at 1121
Sweep summary	100 trials	Best val NDCG@10 improved from 0.7822 to 0.7912

F. V1 ranker feature ablations

Table 7. Feature ablations on the merged candidate-pool benchmark using the V1 proposal pool. Upstream proposal score is included in every row; only the additional feature blocks are listed. The spread across feature sets is below one percentage point on top-1, which motivated training the V2 reranker on the structural feature set only.

Feature set	Top-1	Top-3	Top-5	Top-10	MRR
template	58.07	81.50	87.35	91.16	0.7079
structural + dft + template	58.58	79.73	83.46	87.48	0.7011
template + rc	58.33	81.97	87.99	92.21	0.7123
structural + dft + template + rc	58.56	82.10	88.83	92.04	**0.7168**
structural + DFT, no template	**58.76**	**82.34**	87.97	91.60	0.7155
A: structural + template	58.56	81.73	87.99	91.52	0.7123
B: structural + DFT + template	58.56	81.45	86.94	90.90	0.7105
C: DFT + template	58.50	81.73	87.11	90.32	0.7094