RETROVIRTSYN: ADAPTIVE MULTI-CENTER VIRTUAL SYNTHONS FOR SINGLE-STEP RETROSYNTHESIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Two-stage single-step retrosynthesis is susceptible to error propagation because reactant generation depends critically on reaction-center predictions made in the preceding stage. Conventional semi-template methods typically retain a single reaction-center candidate and construct synthons through hard bond disconnection based on that prediction. When the predicted center is incorrect, synthon construction converts prediction errors into structural deviations, amplifying their impact on downstream reactant generation. To address this limitation, we propose RetroVirtSyn, a retrosynthesis framework based on adaptive multi-center virtual synthons. RetroVirtSyn avoids hard bond disconnection; instead, it adaptively retains multiple candidate reaction centers based on prediction confidence and encodes them as confidence-weighted probabilistic conditions to construct virtual synthons directly on the intact product graph. Empirically, RetroVirtSyn achieves 63.2% Top-1 accuracy on USPTO-50K, outperforming the previous state of the art by 3.2 percentage points, while also achieving the best Top-3, Top-5, and Top-10 accuracies among the compared methods. These results demonstrate that virtual synthons provide an effective interface between reaction-center prediction and reactant generation, mitigating error propagation and thereby improving retrosynthesis accuracy.

1 INTRODUCTION

Recent advances in artificial intelligence have substantially accelerated molecular discovery by enabling the generation and optimization of novel compounds with desired biological and chemical properties (Gómez-Bombarelli et al., 2018; Zhavoronkov et al., 2019). However, the practical application of AI-generated molecules remains constrained by their synthetic accessibility, as identifying feasible synthetic routes for newly designed compounds is often a labor-intensive process requiring extensive expert knowledge (Gao & Coley, 2020). Retrosynthesis planning provides a computational solution to this challenge by predicting precursor molecules or synthetic transformations from a target product, serving as a fundamental component of automated synthesis design and AI-driven drug discovery (Corey & Wipke, 1969; Segler et al., 2018).

Among various retrosynthesis tasks, single-step retrosynthesis aims to predict plausible reactants for a given product and acts as a core module of multi-step synthesis planning (Segler et al., 2018; Schwaller et al., 2020). Recent advances in machine learning have greatly promoted the development of single-step retrosynthesis, yielding diverse modeling strategies for product-to-reactant transformation (Dai et al., 2019; Shi et al., 2020; Yan et al., 2020; Wan et al., 2022). The most fundamental distinction among these strategies lies in how chemical and structural priors are incorporated into the prediction process, and these different modeling approaches naturally form three major paradigms: template-based, template-free, and semi-template methods.

Template-based methods match target products against reaction templates extracted from known reactions and rank valid candidate templates via learned models (Coley et al., 2017; Segler et al., 2018; Dai et al., 2019). Restricted by the coverage and granularity of template libraries, such methods struggle to generalize to unseen reaction patterns. Template-free methods directly learn the product-to-reactant mapping and typically formulate retrosynthesis as an end-to-end sequence or

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.

graph generation task (Karpov et al., 2019; Schwaller et al., 2020; Wan et al., 2022). However, most such end-to-end approaches treat retrosynthesis as a black-box transformation and fail to provide explicit intermediate reasoning for molecular structure modifications.

To introduce more explicit structural priors, semi-template methods decompose retrosynthesis into two stages: they first identify reaction centers and derive intermediate synthons, and then generate corresponding reactants conditioned on these synthons (Shi et al., 2020; Yan et al., 2020; Somnath et al., 2021). However, conventional semi-template methods typically construct synthons via rigid bond disconnections based on predicted reaction centers. If the reaction center is misidentified, the resulting erroneous synthon will serve as the condition for subsequent reactant generation. Consequently, errors introduced in the first stage are encoded into the intermediate representation and propagated downstream. When predicted synthons participate in model training, such errors contaminate the training distribution; during inference, these errors directly affect subsequent reactant generation (Yan et al., 2020; Somnath et al., 2021). Another challenge arises from the one-to-many nature of retrosynthesis: a single product molecule may possess multiple chemically plausible reaction centers, corresponding to various feasible synthetic routes (Schwaller et al., 2020; Chen et al., 2023). Restricting the intermediate representation to a single deterministic transformation limits the model's ability to capture diverse retrosynthetic options. Therefore, an effective retrosynthesis framework should retain informative reaction-center guidance, avoid irreversible structural corruption induced by incorrect predictions, and preserve sufficient flexibility to represent multiple plausible retrosynthetic transformations.

Motivated by these limitations, we propose RetroVirtSyn, an adaptive multi-center virtual synthon framework for single-step retrosynthesis. Rather than committing to a single predicted reaction center through hard bond disconnection, RetroVirtSyn preserves the intact product topology and constructs virtual synthons by encoding candidate reaction centers as confidence-weighted probabilistic structural conditions. This design transforms reaction-center information from a hard structural operation into a flexible guidance signal, allowing the model to retain useful chemical priors while reducing the negative impact of incorrect predictions. Furthermore, RetroVirtSyn introduces an adaptive multi-center strategy that dynamically determines the number of retained reaction centers according to prediction confidence. High-confidence cases focus on the most reliable center, whereas low-confidence cases preserve multiple plausible centers, enabling the model to represent a broader range of possible retrosynthetic transformations.

We evaluate RetroVirtSyn on the widely used USPTO-50K benchmark for single-step retrosynthesis (Schneider et al., 2016). RetroVirtSyn achieves Top-1, Top-3, Top-5, and Top-10 accuracies of 63.2%, 81.3%, 84.1%, and 86.5%, respectively. Compared with the strongest baseline at each cutoff, these results correspond to absolute improvements of 3.2, 2.2, 1.4, and 0.4 percentage points, respectively. These results demonstrate that RetroVirtSyn consistently improves retrosynthesis accuracy across multiple Top-k criteria while alleviating error propagation between reaction-center prediction and reactant generation.

To summarise, the main contributions of this work are the following:

- We introduce a topology-preserving virtual-synthon representation to mitigate error propagation caused by hard synthon fragmentation. Instead of committing to a single disconnected structure, it encodes and fuses multiple possible reaction-center hypotheses on the intact product graph, preserving structural context while guiding downstream reactant generation.
- We introduce a confidence-aware dynamic Top-m strategy for adaptive candidate reaction-center selection. This adaptive selection avoids unnecessary multi-center noise while reducing premature commitment to an uncertain Top-1 prediction.
- We formulate reactant generation as a discrete graph flow conditioned on virtual synthons. The virtual-synthon representation provides generation conditions that encode candidate reaction-center information while retaining the intact product graph as structural context.

2 Related work

Retrosynthesis Modeling. Single-step retrosynthesis follows three paradigms. Template-based methods match products against templates extracted from known reactions, offering interpretability but limited template coverage (Coley et al., 2017; Segler & Waller, 2017; Dai et al., 2019). Template-

free methods directly generate reactants through sequence or graph models, offering flexibility while leaving reaction-relevant transformations largely implicit (Sacha et al., 2021; Tu & Coley, 2022; Wan et al., 2022; Igashov et al., 2024). Semi-template methods provide explicit structural guidance through reaction-center identification followed by synthon-based reactant reconstruction (Shi et al., 2020; Yan et al., 2020; Somnath et al., 2021). Our method belongs to the semi-template paradigm, as it follows the two-stage formulation, but departs from conventional hard-disconnection-based approaches by introducing virtual synthons as topology-preserving probabilistic intermediate representations.

Reaction-Center Possibilities and Synthon Representations. Reaction-center predictions guide reactant generation, but a product may admit multiple plausible transformations. Semi-template methods differ mainly in synthon completion. G2Gs sequentially adds atoms and bonds through variational graph generation (Shi et al., 2020). RetroXpert generates reactant SMILES conditioned on products and synthons (Yan et al., 2020). GraphRetro attaches leaving groups selected from a precomputed vocabulary (Somnath et al., 2021). To account for multiple possible reaction centers, GraphRetro retains alternative edits through beam search, and G^{2}Retro considers multiple reaction centers and synthon-completion paths to generate diverse reactant candidates (Somnath et al., 2021; Chen et al., 2023). Both instantiate separate, discretely edited synthons. RetroVirtSyn instead uses prediction confidence to adaptively retain and fuse candidate centers into virtual synthons, providing multi-center structural guidance without explicit bond disconnection.

Error Propagation in Two-Stage Retrosynthesis. Two-stage retrosynthesis creates dependencies between reaction-center prediction and reactant generation: Incorrect bond disconnection produces erroneous synthons and further impairs reactant generation. GraphRetro explicitly notes that its completion module cannot recover true reactants from incorrectly edited synthons (Somnath et al., 2021). Existing methods address this issue through different strategies. RetroXpert augments reactant-generation training with incorrectly predicted synthons (Yan et al., 2020). RetroDiff (Wang et al., 2025) reverses the generation order, sampling external groups conditioned on the product before predicting their interconnections. These strategies highlight the importance of intermediate predictions passed between stages. We therefore argue that reaction-center error propagation fundamentally reflects deficiencies in the two-stage information interface. RetroVirtSyn uses an inter-stage representation to mitigate these errors, aiming to preserve useful center guidance and maintain accurate reactant generation despite incorrect center predictions.

3 METHOD

3.1 OVERVIEW AND PROBLEM FORMULATION

Given a product molecule, single-step retrosynthesis aims to generate its corresponding reactant molecules. We represent the product as a molecular graph $G_p = (V, E, X, B)$, where V is the atom set, E is the bond set, and X and B contain discrete atom and bond features, respectively. The target reactants are jointly represented as a graph G_R, which may contain multiple disconnected molecular components.

RetroVirtSyn connects reaction-center prediction and reactant generation through an adaptive virtual-synthon representation. Figure 1 illustrates the overall framework. Candidate reaction centers are selected according to their prediction confidence using a dynamic Top-m strategy, which retains one or three hypotheses. Each selected hypothesis is encoded on the intact product graph, and the resulting views are fused into a unified virtual synthon. Finally, a conditional discrete graph flow-matching model generates reactants using $C = [G_s, G_p, C_{\text{center}}]$, where C is the complete generation condition, G_s is the fused virtual synthon, C_{center} records the selected reaction centers and their confidence scores, and brackets denote the collection of conditioning inputs.

The virtual synthon provides local reaction-center guidance, while the product graph supplies the complete molecular context. This design allows intermediate reaction-center predictions to guide reactant generation without explicitly disconnecting bonds during synthon construction.

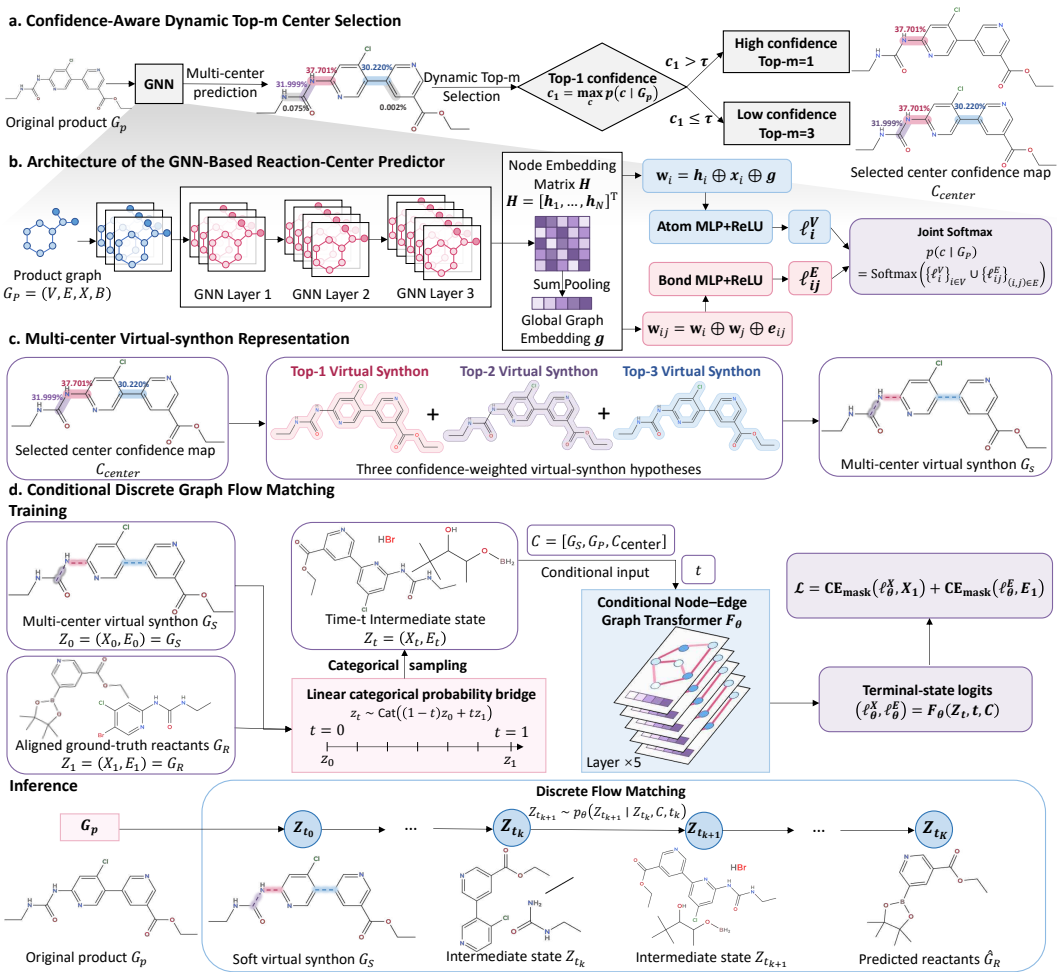


Figure 1: **RetroVirtSyn framework.** (a) Confidence-aware dynamic top-m center selection. (b) Architecture of the GNN-based reaction-center predictor. (c) Multi-center virtual-synthon representation without explicit fragment construction. (d) Conditional discrete graph flow matching for reactant generation.

3.2 Adaptive Reaction-Center Selection

We use a GNN predictor to score candidate reaction centers and adaptively retain hypotheses according to prediction confidence. When the leading prediction is confident, this procedure provides focused reaction-center guidance; otherwise, it preserves additional reaction-center possibilities.

Reaction-center scoring. As illustrated in Figure 1(b), a three-layer message-passing GNN initializes $h_i^{(0)} = x_i$ and encodes the product graph as

$$h_i^{(l+1)} = \text{GNN}^{(l)}\left(h_i^{(l)}, \left\{(h_j^{(l)}, e_{ij}) : j \in \mathcal{N}(i)\right\}\right), \quad l = 0, 1, 2, \tag{1}$$

where x_i is the input feature vector of atom i, $h_i^{(l)}$ is its representation at layer l, e_{ij} is the feature vector of bond (i, j), $\mathcal{N}(i)$ is the neighborhood of atom i, and $\text{GNN}^{(l)}$ is the corresponding message-passing layer.

The final atom embeddings are collected into a matrix $H = [h_1, \ldots, h_N]^\top$ and sum-pooled as $g = \sum_{i \in V} h_i$ to obtain a global graph representation. Atom and bond scoring features are then constructed as $w_i = h_i \oplus x_i \oplus g$ and $w_{ij} = w_i \oplus w_j \oplus e_{ij}$, where $h_i = h_i^{(3)}$ is the final atom embedding, $N = |V|$ is the number of product atoms, H is the node-embedding matrix, g is the

global graph embedding, w_i and w_{ij} are the atom and bond scoring representations, respectively, and $\oplus$ denotes concatenation.

Two MLP heads with ReLU activations produce atom logits $\ell_i^V = \mathrm{MLP}_V(w_i)$ and bond logits $\ell_{ij}^E = \mathrm{MLP}_E(w_{ij})$, which are jointly normalized:

$$p_c = p(c \mid G_p) = \mathrm{Softmax}\big(\{\ell_i^V\}_{i\in V} \cup \{\ell_{ij}^E\}_{(i,j)\in E}\big)_c, \tag{2}$$

where MLP_V and MLP_E are the atom and bond prediction heads, ℓ_i^V and ℓ_{ij}^E are their scalar logits, and p_c is the predicted score for candidate center c. The candidate set $\mathcal{K}_P$ contains all scored reaction-center candidates, and the subscript c selects the corresponding Softmax output. Joint normalization makes the reaction-center candidates directly comparable.

Adaptive selection. The reaction-center predictor produces multiple possible centers. To retain additional candidates when the leading prediction is less confident, we define

$$q = \max_{c\in\mathcal{K}_P} p_c, \qquad m(q) = \begin{cases} 1, & q > \tau, \\ m_{\max}, & q \le \tau, \end{cases} \tag{3}$$

where q is the highest candidate score, τ is the selection threshold, and $m_{\max}$ is the expanded rank cutoff. The adaptive configuration uses $\tau = 0.5$ and $m_{\max} = 3$, whereas the fixed configuration uses $m = 1$ for every product. Experiments supporting the threshold choice are provided in the appendix.

We select candidates according to their recorded ranks and preserve their original score magnitudes using $S^{(m)} = \{c \in \mathcal{K}_P : r(c) \le m\}$ and $s_c = p_c\,\mathbf{1}[c \in S^{(m)}]$, where $r(c)$ is the candidate rank recorded by the selection procedure, $S^{(m)}$ is the retained candidate set, s_c is the score retained after selection, and $\mathbf{1}[\cdot]$ is the indicator function. Here, $m = m(q)$ specifies the rank cutoff rather than the total number of distinct retained candidates.

Finally, the retained scores are mapped to graph-aligned node and edge tensors as $a_i = s_{c_i^V}$ and $b_{ij} = s_{c_{ij}^E}$, where c_i^V is the atom-center candidate associated with atom i, c_{ij}^E is the bond-center candidate associated with bond (i, j), and a_i and b_{ij} are the corresponding entries of the node confidence tensor a and edge confidence tensor b, respectively.

These hypotheses represent alternative reaction centers and do not imply their simultaneous participation in a reaction. Their confidence features are subsequently used to construct the virtual synthon.

3.3 Virtual Synthon Construction

We construct virtual synthons by encoding the selected candidate reaction centers as confidence-valued features on the intact product graph. As illustrated in Figure 1(c), each selected center defines a separate soft view, and these views are fused into a unified multi-center virtual synthon.

Let $R = |S^{(m)}|$ denote the number of retained candidates, enumerated as $c_{(1)}, \ldots, c_{(R)}$. For each candidate, we construct node and edge confidence features:

$$a_i^{(r)} = s_{c_{(r)}}\,\mathbf{1}[c_{(r)} = c_i^V], \qquad b_{ij}^{(r)} = s_{c_{(r)}}\,\mathbf{1}[c_{(r)} = c_{ij}^E], \tag{4}$$

where $a_i^{(r)}$ and $b_{ij}^{(r)}$ are the confidence features at atom i and bond (i, j) in view r, respectively; c_i^V and c_{ij}^E identify the corresponding atom-center and bond-center candidates; and $\mathbf{1}[\cdot]$ is the indicator function.

The resulting center-specific virtual synthon is $G_s^{(r)} = (V, E, X, B;\, a^{(r)}, b^{(r)})$, where $G_s^{(r)}$ is the soft view associated with candidate $c_{(r)}$, and $a^{(r)}$ and $b^{(r)}$ collect its node and edge confidence features, respectively. The semicolon separates the original molecular graph from the additional confidence features.

The selected views are then fused on their shared product graph:

$$G_s = \mathrm{Fuse}\Big(G_s^{(1)}, \ldots, G_s^{(R)}\Big) = (V, E, X, B;\, a, b), \tag{5}$$

where Fuse integrates the center-specific annotations and G_s is the unified virtual synthon.

The original confidence magnitudes are preserved without renormalization. Together with G_p and C_{center}, G_s provides the generator with local reaction-center guidance and complete product context, without requiring hard bond disconnection during synthon construction.

3.4 Conditional discrete graph flow matching

Training. Following discrete flow matching (Gat et al., 2024), we connect the virtual-synthon state $Z_0 = (X_0, E_0)$ and aligned reactant state $Z_1 = (X_1, E_1)$ through a linear probability path. Here, X and E are one-hot atom and bond categories, with the initial categories extracted from G_s. For $t \sim \mathcal{U}(0, 1)$, we sample $Z_t = (X_t, E_t)$ as

$$X_t \sim \mathrm{Cat}((1-t)X_0 + tX_1), \qquad E_t \sim \mathrm{Cat}((1-t)E_0 + tE_1), \tag{6}$$

where Cat samples categories at each position.

A five-layer node–edge Graph Transformer with eight attention heads predicts terminal atom and bond logits as $(\ell_\theta^X, \ell_\theta^E) = F_\theta(Z_t, t, C)$, using the previously defined condition $C = [G_s, G_p, C_{\text{center}}]$. Training minimizes masked terminal-state cross-entropy:

$$\mathcal{L} = \mathrm{CE}_{\text{mask}}\left(\ell_\theta^X, X_1\right) + \mathrm{CE}_{\text{mask}}\left(\ell_\theta^E, E_1\right). \tag{7}$$

The mask excludes padded nodes and invalid edges, and edge logits are symmetrized to preserve the undirected graph structure.

Inference. Starting from $Z_{t_0} = Z_0$, we discretize time as $0 = t_0 < t_1 < \cdots < t_K = 1$, with $K = 50$. At each step, the Transformer predicts terminal logits for the discrete transition:

$$(\ell_\theta^X, \ell_\theta^E) = F_\theta(Z_{t_k}, t_k, C), \qquad Z_{t_{k+1}} \sim p_\theta(Z_{t_{k+1}} \mid Z_{t_k}, C, t_k), \tag{8}$$

where p_θ is the transition distribution from t_k to t_{k+1}. After each update, we symmetrize the bond-category tensor and restore the states of non-dummy product atoms designated as fixed. The condition C remains fixed throughout sampling, and the final state $Z_{t_K} = (\widehat{X}_R, \widehat{E}_R)$ is decoded into the reactant graph $\widehat{G}_R$.

4 Experiments

4.1 Experimental Setup

Dataset. We train and evaluate RetroVirtSyn on USPTO-50K (Schneider et al., 2016), a standard benchmark for retrosynthesis modeling containing approximately 50,000 atom-mapped reactions extracted from US patents. We use the same training/validation/test split as RetroBridge (Igashov et al., 2024), GLN (Dai et al., 2019), and RetroSynFlow (Yadav et al., 2025). The dataset covers 10 reaction classes, and all experiments are conducted under the reaction-class-unknown setting. The low-confidence center-prediction subset comprises samples from the overall evaluation set with reaction-center prediction confidence $q \leq 0.5$.

Baselines. We compare RetroVirtSyn with the baselines in Table 1, grouped into three categories: (1) template-based methods: RetroSim (Coley et al., 2017), NeuralSym (Segler & Waller, 2017), and GLN (Dai et al., 2019); (2) semi-template-based methods: GraphRetro (Somnath et al., 2021), G2G (Shi et al., 2020), RetroXpert (Yan et al., 2020), RetroPrime (Wang et al., 2021), SemiRetro (Gao et al., 2022), RetroDiff (Wang et al., 2025), and GDiffRetro (Sun et al., 2025); and (3) template-free methods: Transformer and Retroformer$_{\text{aug}}$ (Wan et al., 2022), SCROP (Zheng et al., 2020), Augmented Transformer (Tetko et al., 2020), DualTF$_{\text{aug}}$ (Sun et al., 2021), MEGAN (Sacha et al., 2021), Tied Transformer (Kim et al., 2021), GTA$_{\text{aug}}$ (Seo et al., 2021), RetroDCVAE (He et al., 2022), Graph2SMILES (Tu & Coley, 2022), Chemformer (Irwin et al., 2022), RetroBridge (Igashov et al., 2024), DiffAlign (Laabid et al., 2025), RARB (Qiao et al., 2025), and RetroProdFlow and RetroSynFlow (Yadav et al., 2025). Table 1 assembles these literature baselines.

Evaluation. We report Top-k exact-match accuracy for $k \in \{1, 3, 5, 10\}$, defined as the proportion of products whose recorded reactant set appears among the top-k predictions after SMILES canonicalization. To evaluate forward consistency beyond exact matching, we compute round-trip

Table 1: Top-k accuracy (exact match) on the USPTO-50k test dataset.

	Model	Top-1	Top-3	Top-5	Top-10
Template-based	RetroSim (Coley et al., 2017)	37.3	54.7	63.6	74.1
	NeuralSym (Segler & Waller, 2017)	44.4	65.3	72.4	78.9
	GLN (Dai et al., 2019)	52.5	69.0	75.6	83.7
Semi-template-based	GraphRetro (Somnath et al., 2021)	53.7	68.3	72.2	75.5
	G2G (Shi et al., 2020)	48.9	67.6	72.5	75.5
	RetroXpert (Yan et al., 2020)	50.4	61.1	62.3	63.4
	RetroPrime (Wang et al., 2021)	51.4	70.8	74.0	76.1
	SemiRetro (Gao et al., 2022)	54.9	75.3	80.4	84.1
	RetroDiff (Wang et al., 2025)	52.6	71.2	81.0	85.3
	GDiffRetro (Sun et al., 2025)	58.9	79.1	81.9	–
Template-free	Transformer (Wan et al., 2022)	42.4	58.6	63.8	67.7
	Retroformer$_{\mathrm{aug}}$ (Wan et al., 2022)	52.9	68.2	72.5	76.4
	SCROP (Zheng et al., 2020)	43.7	60.0	65.2	68.7
	Aug. Transformer (Tetko et al., 2020)	48.3	–	73.4	77.4
	DualTF$_{\mathrm{aug}}$ (Sun et al., 2021)	53.6	70.7	74.6	77.0
	MEGAN (Sacha et al., 2021)	48.1	70.7	78.4	<u>86.1</u>
	Tied Transformer (Kim et al., 2021)	47.1	67.1	73.1	76.3
	GTA$_{\mathrm{aug}}$ (Seo et al., 2021)	51.1	67.6	74.8	81.6
	RetroDCVAE (He et al., 2022)	53.1	68.1	71.6	74.3
	Graph2SMILES (Tu & Coley, 2022)	52.9	66.5	70.0	72.9
	Chemformer (Irwin et al., 2022)	54.3	–	62.3	63.0
	RetroBridge (Igashov et al., 2024)	50.8	74.1	80.6	85.6
	DiffAlign (Laabid et al., 2025)	54.7	73.3	77.8	81.8
	RARB (Qiao et al., 2025)	58.3	75.2	78.8	81.5
	RetroProdFlow (Yadav et al., 2025)	50.0	74.3	81.2	85.8
	RetroSynFlow (Yadav et al., 2025)	60.0	77.9	82.7	85.3
Ours	Ours (Top-$m = 1$)	**63.5**	<u>79.9</u>	<u>83.4</u>	85.8
	Ours (dynamic Top-m)	<u>63.2</u>	**81.3**	**84.1**	**86.5**

accuracy and coverage using a pretrained Molecular Transformer (Schwaller et al., 2019), following prior evaluation protocols (Igashov et al., 2024; Yadav et al., 2025). A candidate reactant set passes round-trip validation if the model's top-1 product prediction matches the target product. Round-trip accuracy measures the fraction of the top-k candidates passing this validation, averaged across products, whereas round-trip coverage measures the proportion of products with at least one passing candidate among the top-k predictions. All metrics are reported as percentages.

4.2 Main Results

Top-k exact-match performance. Table 1 summarizes the results on USPTO-50K. Both RetroVirtSyn configurations achieve higher Top-1 accuracy than the reported baseline results: the fixed $m = 1$ strategy reaches 63.5%, and dynamic Top-m achieves 63.2%, compared with 60.0% for RetroSynFlow (Yadav et al., 2025). This improvement supports the effectiveness of virtual synthons in improving retrosynthesis accuracy compared with methods that construct synthons through hard bond disconnections (Somnath et al., 2021; Sun et al., 2025; Yadav et al., 2025). The dynamic configuration achieves Top-3, Top-5, and Top-10 accuracies of 81.3%, 84.1%, and 86.5%, respectively, the highest values in the table. These exceed the strongest baseline at each corresponding rank by 2.2, 1.4, and 0.4 percentage points (Sun et al., 2025; Yadav et al., 2025; Sacha et al., 2021). Compared with fixed $m = 1$, dynamic Top-m improves all three metrics with a 0.3-percentage-point decrease in Top-1 accuracy. These results show that adaptively retaining multiple candidate reaction centers in virtual synthons largely preserves the accuracy of the first prediction while increasing the likelihood that the predicted candidates include the correct reactants.

Table 2: Round-trip coverage and round-trip accuracy (%) on USPTO-50K.

		Round-Trip Coverage				Round-Trip Accuracy			
	Model	$k=1$	$k=3$	$k=5$	$k=10$	$k=1$	$k=3$	$k=5$	$k=10$
Template-based	GLN	82.5	92.0	94.0	–	82.5	71.0	66.2	–
	LocalRetro	82.1	92.3	94.7	–	82.1	71.0	66.7	–
	RetroGFN	–	–	–	–	76.7	69.1	65.5	60.8
Template-free	MEGAN	78.1	88.6	91.3	–	78.1	67.3	61.7	–
	Graph2SMILES	–	–	–	–	76.7	56.0	46.4	–
	$\text{Retroformer}_{\text{aug}}$	–	–	–	–	78.6	71.8	67.1	–
	RetroBridge	85.1	95.7	97.1	97.7	85.1	73.6	67.8	56.3
	RetroSynFlow (w/ RS)	88.9	97.3	98.4	98.9	88.9	73.9	69.1	61.8
	RetroProdFlow (w/ RS)	91.4	97.6	98.7	99.3	91.4	84.1	80.1	75.1
	RetroProdFlow (w/o RS)	84.4	95.3	96.9	97.7	84.4	72.8	66.8	57.6
Virtual synthons	Ours (w/o RS)	89.2	96.4	97.5	98.3	89.2	74.9	68.2	62.0
	Ours (w/ RS)	**92.1**	**97.9**	**99.0**	**99.6**	**92.1**	**85.5**	**80.7**	**77.0**

Table 3: Ablation study of virtual synthon conditioning and center-selection strategies on USPTO-50K (%).

Configuration	Top-1	Top-3	Top-5	Top-10
Hard-disconnected synthons	60.3	76.8	80.7	83.7
w/o virtual-synthon conditioning	59.7	77.8	82.9	85.3
w/ virtual synthons (Top-$m=1$)	**63.5**	79.9	83.4	85.8
w/ virtual synthons (Top-$m=3$)	62.7	79.4	82.4	84.1
w/ virtual synthons (Top-$m=10$)	63.0	79.2	82.3	84.0
w/ virtual synthons (dynamic $m \in \{1,3\}$)	63.2	**81.3**	**84.1**	**86.5**

Round-trip evaluation. Exact-match accuracy does not account for the fact that multiple reactant sets, including those absent from the dataset, may yield the same product (Igashov et al., 2024; Yadav et al., 2025). We therefore evaluate RetroVirtSyn using round-trip coverage and accuracy, adopting the same reward-based steering (RS) strategy as RetroSynFlow (Yadav et al., 2025). As shown in Table 2, RetroVirtSyn with RS outperforms all compared models on both metrics at every evaluated k, achieving 99.6% coverage and 77.0% accuracy at $k=10$. Without RS, RetroVirtSyn also outperforms RetroProdFlow without RS across all reported metrics, with round-trip accuracy improving by 4.8 and 4.4 percentage points at $k=1$ and $k=10$, respectively. These results show that RetroVirtSyn's advantage in forward consistency is not solely attributable to RS, as it persists without this strategy.

4.3 Ablation Studies

We conduct ablation studies (Table 3) to evaluate the effects of virtual-synthon conditioning and adaptive reaction-center selection. With fixed $m=1$, virtual synthons outperform both hard-disconnected synthons and the baseline without virtual-synthon conditioning across all four metrics. Top-1 accuracy reaches 63.5%, exceeding these two configurations by 3.2 and 3.8 percentage points, respectively, supporting the effectiveness of virtual-synthon conditioning. Among the fixed configurations, $m=1$ consistently outperforms $m=3$ and $m=10$, indicating that retaining more candidate reaction centers alone does not guarantee better predictions. Adaptive selection between $m=1$ and $m=3$ achieves the highest Top-3, Top-5, and Top-10 accuracies among all evaluated configurations. Compared with fixed $m=1$, it improves these metrics by 1.4, 0.7, and 0.7 percentage points, respectively, with a 0.3-percentage-point decrease in Top-1 accuracy. These results show that confidence-guided reaction-center selection improves Top-3, Top-5, and Top-10 accuracy while largely preserving Top-1 performance.

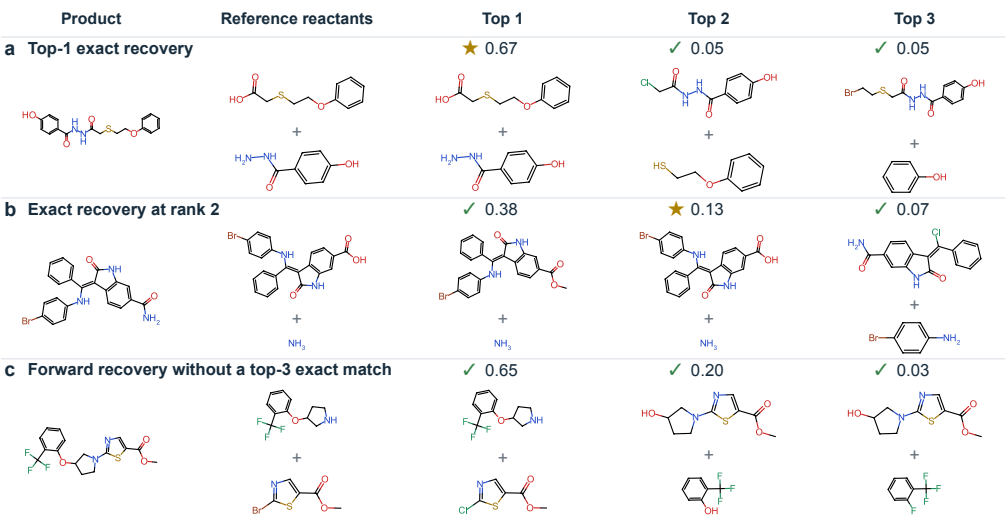


Figure 2: Qualitative examples of RetroVirtSyn predictions. A star indicates an exact match, a checkmark indicates a round-trip match but not an exact match.

4.4 Examples

Figure 2 presents three examples from the USPTO-50K test set, showing the top three predictions from RetroVirtSyn and their corresponding scores. In case (a), the model ranks the reference reactants first with a score of 0.67, well above the second-ranked prediction at 0.05. The other two candidates also pass forward-model validation. In case (b), the reference reactants appear at rank 2 with a score of 0.13, while the first-ranked candidate scores 0.38. Although this higher-scoring candidate differs from the reference, it passes forward-model validation. In case (c), none of the top three predictions matches the reference reactants, yet all three pass forward-model validation. The first-ranked candidate, scoring 0.65, uses a chlorinated counterpart of the brominated reference reactant. These examples illustrate that RetroVirtSyn can recover recorded reactants and generate alternative candidates consistent with the target product under the forward model.

5 Conclusion

In this work, we introduce RetroVirtSyn, a single-step retrosynthesis framework based on adaptive multi-center virtual synthons and conditional discrete graph flow matching. By adaptively selecting and integrating candidate reaction centers, the method incorporates multiple reaction-center possibilities into reactant generation. We demonstrate that virtual synthons provide a more informative intermediate representation for generating reactants. On USPTO-50K, the adaptive configuration achieves 63.2% Top-1 accuracy, exceeding the strongest reported baseline by 3.2 percentage points, and obtains the highest Top-3, Top-5, and Top-10 accuracies among the compared methods, surpassing previous state-of-the-art results. Round-trip evaluation shows improved forward consistency both with and without reward-based steering. Ablation studies further support the effectiveness of virtual-synthon conditioning and confidence-based reaction-center selection. These experiments show that choosing an appropriate reaction-center representation—preserving product connectivity and conveying reaction-center information through virtual synthons—can improve retrosynthesis performance.

These findings highlight the importance of how reaction-center predictions are represented and passed to downstream generation. Future work will incorporate reaction conditions into RetroVirtSyn to provide more practical synthesis guidance and integrate it with existing multi-step planning algorithms to evaluate its performance in multi-step retrosynthesis.

AI USE STATEMENT

In this work, we used generative AI tools for manuscript writing and polishing, as well as literature retrieval and discovery of related work. We have not used generative AI tools for generating research hypotheses, designing model architectures, conducting model training, or producing experimental data, and other required-disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for textual refinement and rephrasing. We have reviewed all AI-assisted work: all AI-polished texts and retrieved literature candidates were manually examined and validated by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work performs computational analysis on publicly available chemical reaction datasets and focuses on the retrosynthesis prediction task. This study does not involve human subjects or animal experiments. There are no privacy, security, or conflict-of-interest issues in this research.

REPRODUCIBILITY STATEMENT

The dataset used in this work is the publicly available USPTO-50k retrosynthesis benchmark. Detailed model architecture, hyperparameter settings, training pipeline, and evaluation metrics are provided in the appendix.

REFERENCES

Ziqi Chen, Oluwatosin R. Ayinde, James R. Fuchs, Huan Sun, and Xia Ning. G2Retro as a two-step graph generative models for retrosynthesis prediction. *Communications Chemistry*, 6(1):102, 2023. doi: 10.1038/s42004-023-00897-3. URL https://www.nature.com/articles/s42004-023-00897-3.

Connor W. Coley, Luke Rogers, William H. Green, and Klavs F. Jensen. Computer-assisted retrosynthesis based on molecular similarity. *ACS Central Science*, 3(12):1237–1245, 2017. doi: 10.1021/acscentsci.7b00355.

Elias J. Corey and W. Todd Wipke. Computer-assisted design of complex organic syntheses: Pathways for molecular synthesis can be devised with a computer and equipment for graphical communication. *Science*, 166(3902):178–192, 1969. doi: 10.1126/science.166.3902.178.

Hanjun Dai, Chengtao Li, Connor W. Coley, Bo Dai, and Le Song. Retrosynthesis prediction with conditional graph logic network. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/0d2b2061826a5df3221116a5085a6052-Abstract.html.

Wenhao Gao and Connor W. Coley. The synthesizability of molecules proposed by generative models. *Journal of Chemical Information and Modeling*, 60(12):5714–5723, 2020. doi: 10.1021/acs.jcim.0c00174.

Zhangyang Gao, Cheng Tan, Lirong Wu, and Stan Z. Li. SemiRetro: Semi-template framework boosts deep retrosynthesis prediction, 2022. URL https://arxiv.org/abs/2202.08205.

Itai Gat, Tal Remez, Neta Shaul, Felix Kreuk, Ricky T. Q. Chen, Gabriel Synnaeve, Yossi Adi, and Yaron Lipman. Discrete flow matching. In *Advances in Neural Information Processing Systems*, volume 37, pp. 133345–133385, 2024. doi: 10.52202/079017-4239. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f0d629a734b56a642701bba7bc8bb3ed-Abstract-Conference.html.

Rafael Gómez-Bombarelli, Jennifer N. Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D. Hirzel, Ryan P. Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, 4(2):268–276, 2018. doi: 10.1021/acscentsci.7b00572.

Hua-Rui He, Jie Wang, Yunfei Liu, and Feng Wu. Modeling diverse chemical reactions for single-step retrosynthesis via discrete latent variables. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pp. 717–726, 2022. doi: 10.1145/3511808.3557397.

Ilia Igashov, Arne Schneuing, Marwin H. S. Segler, Michael M. Bronstein, and Bruno E. Correia. RetroBridge: Modeling retrosynthesis with markov bridges. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/acb94e709f02895fd98b5867f0b184f3-Paper-Conference.pdf.

Ross Irwin, Spyridon Dimitriadis, Jiazhen He, and Esben Jannik Bjerrum. Chemformer: A pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology*, 3 (1):015022, 2022. doi: 10.1088/2632-2153/ac3ffb.

Pavel Karpov, Guillaume Godin, and Igor V. Tetko. A transformer model for retrosynthesis. In *Artificial Neural Networks and Machine Learning – ICANN 2019: Workshop and Special Sessions*, volume 11731 of *Lecture Notes in Computer Science*, pp. 817–830. Springer, 2019. doi: 10.1007/978-3-030-30493-5_78. URL https://link.springer.com/chapter/10.1007/978-3-030-30493-5_78.

Eunji Kim, Dongseon Lee, Youngchun Kwon, Min Sik Park, and Youn-Suk Choi. Valid, plausible, and diverse retrosynthesis using tied two-way transformers with latent variables. *Journal of Chemical Information and Modeling*, 61(1):123–133, 2021. doi: 10.1021/acs.jcim.0c01074.

Najwa Laabid, Severi Rissanen, Markus Heinonen, Arno Solin, and Vikas Garg. Equivariant denoisers cannot copy graphs: Align your graph diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/f8f78f8043f35890181a824e53a57134-Abstract-Conference.html.

Anjie Qiao, Zhen Wang, Jiahua Rao, Yuedong Yang, and Zhewei Wei. Advancing retrosynthesis with retrieval-augmented graph generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(19):20004–20013, 2025. doi: 10.1609/aaai.v39i19.34203.

Mikołaj Sacha, Mikołaj Błaż, Piotr Byrski, Paweł Dąbrowski-Tumański, Mikołaj Chromiński, Rafał Loska, Paweł Włodarczyk-Pruszyński, and Stanisław Jastrzębski. Molecule edit graph attention network: Modeling chemical reactions as sequences of graph edits. *Journal of Chemical Information and Modeling*, 61(7):3273–3284, 2021. doi: 10.1021/acs.jcim.1c00537.

Nadine Schneider, Nikolaus Stiefl, and Gregory A. Landrum. 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.

Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A. Hunter, Costas Bekas, and Alpha A. Lee. Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. *ACS Central Science*, 5(9):1572–1583, 2019. doi: 10.1021/acscentsci.9b00576.

Philippe Schwaller, Riccardo Petraglia, Valerio Zullo, Vishnu H. Nair, Rico Andreas Haeuselmann, Riccardo Pisoni, Costas Bekas, Anna Iuliano, and Teodoro Laino. Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chemical Science*, 11(12):3316–3325, 2020. doi: 10.1039/C9SC05704H.

Marwin H. S. Segler and Mark P. Waller. Neural-symbolic machine learning for retrosynthesis and reaction prediction. *Chemistry – A European Journal*, 23(25):5966–5971, 2017. doi: 10.1002/chem.201605499.

Marwin H. S. Segler, Mike Preuss, and Mark P. Waller. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature*, 555:604–610, 2018. doi: 10.1038/nature25978.

Seung-Woo Seo, You Young Song, June Yong Yang, Seohui Bae, Hankook Lee, Jinwoo Shin, Sung Ju Hwang, and Eunho Yang. GTA: Graph truncated attention for retrosynthesis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(1):531–539, 2021. doi: 10.1609/aaai.v35i1.16131.

Chence Shi, Minkai Xu, Hongyu Guo, Ming Zhang, and Jian Tang. A graph to graphs framework for retrosynthesis prediction. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8818–8827. PMLR, 2020. URL https://proceedings.mlr.press/v119/shi20d.html.

Vignesh Ram Somnath, Charlotte Bunne, Connor W. Coley, Andreas Krause, and Regina Barzilay. Learning graph models for retrosynthesis prediction. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9405–9415, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/4e2a6330465c8ffcaa696a5a16639176-Abstract.html.

Ruoxi Sun, Hanjun Dai, Li Li, Steven Kearnes, and Bo Dai. Towards understanding retrosynthesis by energy-based models. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/5470abe68052c72afb19be45bb418d02-Abstract.html.

Shengyin Sun, Wenhao Yu, Yuxiang Ren, Weitao Du, Liwei Liu, Xuecang Zhang, Ying Hu, and Chen Ma. GDiffRetro: Retrosynthesis prediction with dual graph enhanced molecular representation and diffusion generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(12): 12595–12603, 2025. doi: 10.1609/aaai.v39i12.33373.

Igor V. Tetko, Pavel Karpov, Ruud Van Deursen, and Guillaume Godin. State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis. *Nature Communications*, 11: 5575, 2020. doi: 10.1038/s41467-020-19266-y.

Zhengkai Tu and Connor W. Coley. Permutation invariant graph-to-sequence model for template-free retrosynthesis and reaction prediction. *Journal of Chemical Information and Modeling*, 62(15): 3503–3513, 2022. doi: 10.1021/acs.jcim.2c00321.

Yue Wan, Chang-Yu Hsieh, Ben Liao, and Shengyu Zhang. Retroformer: Pushing the limits of end-to-end retrosynthesis transformer. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 22475–22490. PMLR, 2022. URL https://proceedings.mlr.press/v162/wan22a.html.

Xiaorui Wang, Yuquan Li, Jiezhong Qiu, Guangyong Chen, Huanxiang Liu, Benben Liao, Chang-Yu Hsieh, and Xiaojun Yao. RetroPrime: A diverse, plausible and transformer-based method for single-step retrosynthesis predictions. *Chemical Engineering Journal*, 420:129845, 2021. doi: 10.1016/j.cej.2021.129845.

Yiming Wang, Yuxuan Song, Yiqun Wang, Minkai Xu, Rui Wang, Hao Zhou, and Wei-Ying Ma. RetroDiff: Retrosynthesis as multi-stage distribution interpolation. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 1918–1926. PMLR, 2025. URL https://proceedings.mlr.press/v258/wang25e.html.

Robin Yadav, Qi Yan, Guy Wolf, Joey Bose, and Renjie Liao. RETRO SYNFLOW: Discrete flow-matching for accurate and diverse single-step retrosynthesis. In *Advances in Neural Information Processing Systems*, volume 38, pp. 61492–61521, 2025. doi: 10.52202/085713-1846. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/4f86833d5cc98ec32e470ef1c8cb82e3-Abstract-Conference.html.

Chaochao Yan, Qianggang Ding, Peilin Zhao, Shuangjia Zheng, Jinyu Yang, Yang Yu, and Junzhou Huang. RetroXpert: Decompose retrosynthesis prediction like a chemist. In *Advances in Neural Information Processing Systems*, volume 33, pp. 11248–11258, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/819f46e52c25763a55cc642422644317-Abstract.html.

Alex Zhavoronkov, Yan A. Ivanenkov, Alex Aliper, Mark S. Veselov, Vladimir A. Aladinskiy, Anastasiya V. Aladinskaya, Victor A. Terentiev, Daniil A. Polykovskiy, Maksim D. Kuznetsov, Arip Asadulaev, Yury Volkov, Artem Zholus, Rim R. Shayakhmetov, Alexander Zhebrak, Lidiya I. Minaeva, Bogdan A. Zagribelnyy, Lennart H. Lee, Richard Soll, David Madge, Li Xing, Tao Guo, and Alán Aspuru-Guzik. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nature Biotechnology*, 37(9):1038–1040, 2019. doi: 10.1038/s41587-019-0224-x.

Shuangjia Zheng, Jiahua Rao, Zhongyue Zhang, Jun Xu, and Yuedong Yang. Predicting retrosynthetic reactions using self-corrected transformer neural networks. *Journal of Chemical Information and Modeling*, 60(1):47–55, 2020. doi: 10.1021/acs.jcim.9b00949.

A Model and implementation details

A.1 Architecture and optimization

The reactant generator uses a Graph Transformer to predict categorical atom and bond distributions from intermediate molecular graphs. Its backbone follows the conditional graph flow formulation of Yadav et al. (2025), with separate node, edge, and graph-level representations. Table 4 summarizes the network dimensions and optimization settings.

The conditioning input combines the virtual-synthon source, the original product graph, and a graph encoding the selected reaction-center scores. These conditions are aligned to the product atom mapping and concatenated with the current graph features. The model also receives the current flow time and graph-level features to predict the target reactant graph.

Training uses AdamW with AMSGrad for 500 epochs at a constant learning rate. The categorical cross-entropy objective operates on valid node and edge entries, excluding all-zero masked targets. Product atom identities are preserved by replacing their prediction logits with the corresponding source entries before loss evaluation. These entries contribute no gradient to the node predictor; the model learns the identities of additional atoms and the bond categories.

Table 4: Architecture and optimization settings for the virtual-synthon generator.

Setting	Value
Graph Transformer layers / attention heads	5 / 8
Node / edge / global hidden dimensions	256 / 64 / 64
Node / edge / global feed-forward dimensions	256 / 128 / 128
Input MLP dimensions	256 / 128 / 128
Training batch size	32
Optimizer	AdamW, AMSGrad enabled
Learning rate / schedule	2×10^{-4} / constant
Adam (β_1, β_2) / weight decay	$(0.9, 0.999)$ / 10^{-12}
Training seed	42

A.2 Sampling

To compare reactant recovery by hard synthons and dynamic virtual synthons under different inference budgets, we evaluated both models using 5, 10, 20, 50, and 100 inference steps while keeping their parameters fixed. We examined Top-k exact-match accuracy and reaction-center coverage under both limited and extended inference budgets. The results are presented in Table 5 and visualized in Figure 3.

For both methods, overall performance improves with increasing inference steps and gradually approaches a plateau, indicating that additional inference steps facilitate correct reactant recovery. Across all evaluated step counts, however, dynamic virtual synthons consistently achieve higher exact-match accuracy and reaction-center coverage than hard synthons. These results support their more robust recovery performance across inference budgets.

Under limited inference budgets of 5–20 steps, hard synthons exhibit substantially lower performance. Dynamic virtual synthons achieve clear advantages across Top-1 through Top-50 accuracy, suggesting that adaptive candidate selection helps mitigate the effects of initial reaction-center prediction errors even with few inference steps. At 50–100 steps, the performance of both models begins to saturate, yet dynamic virtual synthons retain a clear advantage. This persistent gap suggests that their benefits extend beyond settings with limited inference steps and are not eliminated by increasing the inference budget.

Overall, dynamic virtual synthons provide better reactant recovery under both rapid inference with few steps and extended inference with more steps. Their consistent advantages over hard synthons are compatible with greater tolerance to reaction-center prediction errors across the evaluated inference settings.

Table 5: **Reactant generation performance under different sampling budgets.** Top-k exact-match accuracy and coverage are reported as percentages. Model parameters are fixed across inference-step settings.

Steps K	Condition	Top-1	Top-3	Top-5	Top-10	Top-50	Coverage (%)
5	Hard-synthon	39.4	48.6	55.4	60.6	65.2	68.0
	Dynamic virtual	48.4	55.0	60.6	66.0	68.4	71.2
10	Hard-synthon	51.8	64.0	66.4	69.6	73.2	76.4
	Dynamic virtual	56.4	69.2	74.4	76.4	78.2	80.6
20	Hard-synthon	54.6	69.6	72.6	74.6	79.6	81.2
	Dynamic virtual	61.6	76.2	77.8	80.4	81.8	83.8
50	Hard-synthon	57.0	74.4	76.4	79.6	82.6	83.2
	Dynamic virtual	63.2	81.3	84.1	86.5	89.1	89.3
100	Hard-synthon	56.2	74.2	79.0	80.8	83.2	83.2
	Dynamic virtual	60.8	80.0	80.6	83.2	85.4	85.4

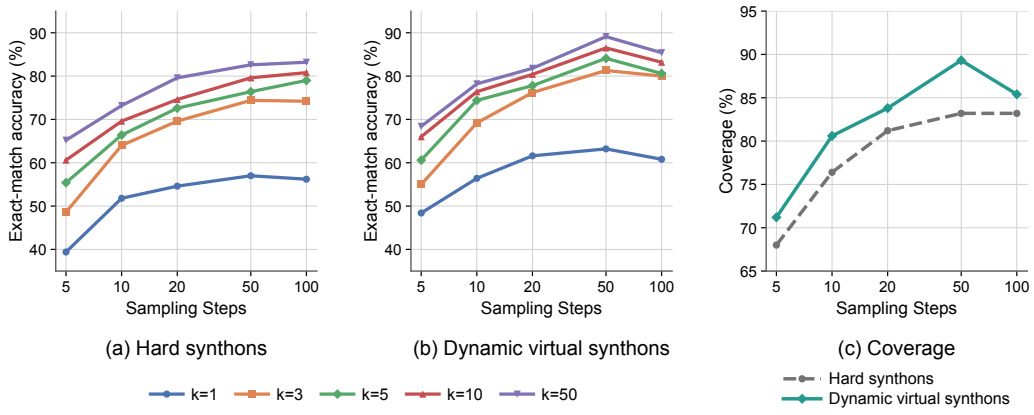


Figure 3: **Effect of sampling budget on reactant generation performance.** Top-k exact-match accuracy for (a) hard synthons and (b) dynamic virtual synthons; (c) coverage for both methods. Points show the values reported in Table 5, with model parameters fixed across sampling budgets. The sampling-step axis uses a logarithmic scale.

A.3 TRAINING PSEUDOCODE

Algorithm 1 Virtual-synthon-conditioned training of RetroVirtSyn

Input: Product/reactant pairs $\mathcal{D} = \{(G_p, G_R)\}$, predicted center scores $\{p_c\}$, generator F_θ, selection threshold $\tau = 0.5$, expanded rank cutoff $m_{\max} = 3$.

for each training minibatch $\mathcal{B} \subset \mathcal{D}$ **do**
  Align product/reactant graphs and reserve the fixed dummy atom slots.
  $q \leftarrow \max_{c \in \mathcal{K}_P} p_c, \quad m \leftarrow m(q)$ ▷ *Eq. 3*
  $S^{(m)} \leftarrow \{c \in \mathcal{K}_P : r(c) \le m\}$ ▷ *Select candidate centers*
  $s_c \leftarrow p_c \, \mathbf{1}[c \in S^{(m)}]$ ▷ *Preserve confidence magnitudes*
  Build center-specific views $G_s^{(1)}, \ldots, G_s^{(R)}$, $R = |S^{(m)}|$.
  $G_s \leftarrow \mathrm{Fuse}(G_s^{(1)}, \ldots, G_s^{(R)})$ ▷ *Eq. 5*
  $Z_0 = (X_0, E_0) \leftarrow G_s, \quad Z_1 = (X_1, E_1) \leftarrow G_R$
  $C \leftarrow [G_s, G_p, C_{\mathrm{center}}]$ ▷ *Assemble conditioning inputs*
  $t \sim \mathcal{U}(0, 1)$ ▷ *Sample flow time*
  $X_t \sim \mathrm{Cat}((1-t)X_0 + tX_1), \quad E_t \sim \mathrm{Cat}((1-t)E_0 + tE_1)$
  $Z_t \leftarrow (X_t, E_t), \quad (\ell_\theta^X, \ell_\theta^E) \leftarrow F_\theta(Z_t, t, C)$
  Symmetrize edge logits and restore fixed product-atom entries.
  $\mathcal{L} \leftarrow \mathrm{CE}_{\mathrm{mask}}(\ell_\theta^X, X_1) + \mathrm{CE}_{\mathrm{mask}}(\ell_\theta^E, E_1)$
  Update θ using AdamW with the minibatch loss $\mathcal{L}$.
end for
Return trained generator F_θ

A.4 SAMPLING PSEUDOCODE

Algorithm 2 Virtual-synthon-conditioned sampling with RetroVirtSyn

Input: Product graph G_p, predicted center scores $\{p_c\}$, trained generator F_θ, sampling steps $K = 50$, number of draws M, $\tau = 0.5$, $m_{\max} = 3$.

$q \leftarrow \max_{c \in \mathcal{K}_P} p_c, \quad m \leftarrow m(q)$ ▷ *Eq. 3*
$S^{(m)} \leftarrow \{c \in \mathcal{K}_P : r(c) \le m\}, \quad s_c \leftarrow p_c \, \mathbf{1}[c \in S^{(m)}]$
Build and fuse selected views into G_s using Eq. 5; reserve dummy atom slots.
Extract one-hot categories (X_0, E_0) from G_s; fix $C = [G_s, G_p, C_{\mathrm{center}}]$.
Set valid-node mask $\mathcal{M}$ and fixed non-dummy product-atom indices $\mathcal{I}_{\mathrm{fix}}$.
$t_k \leftarrow k/K$ for $k = 0, \ldots, K$; $\mathcal{R} \leftarrow [\,]$ ▷ $t_0 = 0$, $t_K = 1$
for $j = 1, \ldots, M$ **do**
  $X \leftarrow X_0, \quad E \leftarrow E_0$ ▷ *Initialize* $Z_{t_0} = Z_0$
  for $k = 0, \ldots, K-1$ **do**
    $\delta_k \leftarrow t_{k+1} - t_k, \quad (\ell_\theta^X, \ell_\theta^E) \leftarrow F_\theta((X, E), t_k, C)$
    $\widehat{p}^X \leftarrow \mathrm{Softmax}(\ell_\theta^X), \quad \widehat{p}^E \leftarrow \mathrm{Softmax}(\ell_\theta^E)$
    $u^X \leftarrow (\widehat{p}^X - X)/(1 - t_k), \quad u^E \leftarrow (\widehat{p}^E - E)/(1 - t_k)$
    $Q^X \leftarrow \mathrm{Normalize}(X + \delta_k u^X)$
    $Q^E \leftarrow \mathrm{Normalize}(E + \delta_k u^E)$
    $X'_i \sim \mathrm{Cat}(Q_i^X)$ for each valid node i
    $E'_{ab} \sim \mathrm{Cat}(Q_{ab}^E)$ for valid $a < b$; $E'_{ba} \leftarrow E'_{ab}$
    Zero diagonal edge entries and apply mask $\mathcal{M}$.
    $X'_i \leftarrow X_{0,i}$ for $i \in \mathcal{I}_{\mathrm{fix}}$ ▷ *Restore product atoms*
    $X \leftarrow X', \quad E \leftarrow E'$ ▷ *Advance to* $Z_{t_{k+1}}$
  end for
  $\widehat{G}_R^{(j)} \leftarrow \mathrm{Decode}(X, E, \mathcal{M})$ ▷ *Decode final state* Z_{t_K}
  Append $\widehat{G}_R^{(j)}$ to $\mathcal{R}$.
end for
Return generated reactant graphs $\mathcal{R}$

B SELECTION OF THE CANDIDATE RANKING CUTOFF

This experiment examined the choice of the candidate reaction-center cutoff m, balancing the number of retained candidates against reaction-center coverage. Specifically, we evaluated the rationale for retaining one candidate by default, adaptively expanding to three candidates, and setting the maximum cutoff at $m = 3$.

We first evaluated reaction-center coverage across candidate ranks on the overall evaluation set. Coverage increased from 86.21% at Top-1 to 95.14% at Top-2 and 97.43% at Top-3. Further expansion to Top-5 and Top-10 yielded only modest increases, reaching 98.77% and 99.57%, respectively (Figure 4a). Most of the coverage gain was therefore achieved by $m = 3$, with diminishing returns from retaining additional candidates. These results support setting the maximum cutoff at $m = 3$ to avoid unnecessary computational overhead.

To examine how the benefit of candidate expansion varied with prediction confidence, we divided samples into high-confidence ($q > \tau$) and low-confidence ($q \leq \tau$) groups. We used $\tau = 0.5$ as an operational threshold for determining whether additional candidate reaction centers were needed. Predictions above this threshold were treated as relatively reliable, whereas those at or below it were assigned an expanded candidate set. For high-confidence samples, reaction-center coverage reached 91.52% at $m = 1$, and expansion to $m = 3$ provided a comparatively smaller gain of 7.50 percentage points. By contrast, coverage on the low-confidence center-prediction subset increased from 50.67% at $m = 1$ to 88.00% at $m = 3$, a gain of 37.33 percentage points (Figure 4b).

These results indicate that candidate expansion provides a substantially greater benefit for low-confidence samples than for high-confidence samples. This difference supports retaining a single candidate reaction center ($m = 1$) by default and expanding the candidate set to $m = 3$ only for low-confidence predictions.

As a complementary analysis, we evaluated the retained probability mass on the overall evaluation set, defined as

$$M_m = \sum_{c \in S^{(m)}} p_c, \tag{9}$$

where $S^{(m)}$ denotes the set of the top m candidate reaction centers and p_c is the predicted probability assigned to center c. The mean retained probability mass was 94.97% for the high-confidence branch ($m = 1$) and 93.11% for the low-confidence center-prediction subset ($m = 3$). Thus, both branches retained most of the predicted probability mass on average, leaving a small proportion outside the candidate set. This strategy preserves candidates accounting for most of the probability mass while limiting the inclusion of low-probability reaction centers.

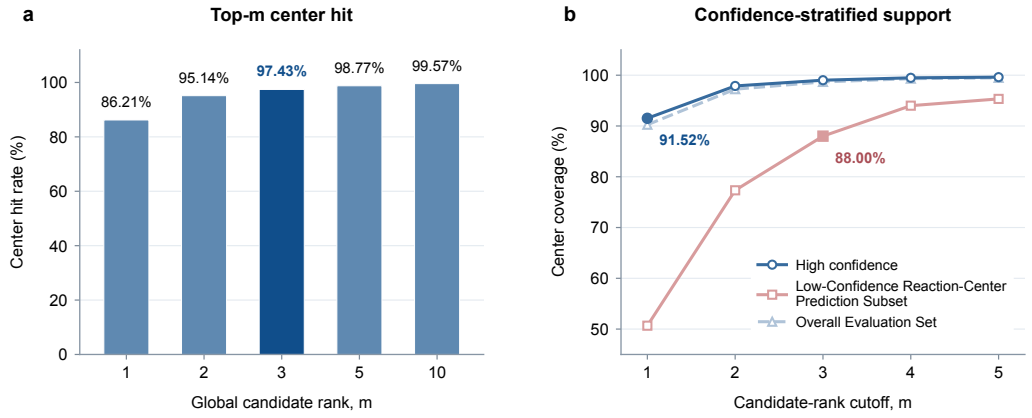


Figure 4: **Reaction-center support as a function of rank.** (a) Top-m center hit rates from the recorded predictor evaluation. (b) Coverage of the annotated center by $S^{(m)}$, using the recorded candidate ranks, on the Overall Evaluation Set restricted to products with available center labels. Confidence groups use the threshold $\tau = 0.5$; the gray curve combines both groups.

C SELECTION OF THE CONFIDENCE THRESHOLD

C.1 SELECTION OF THE CONFIDENCE THRESHOLD DURING TRAINING

To determine the confidence threshold τ used in the adaptive branching strategy, we conducted a set of exploratory experiments. For each candidate threshold, we independently trained a model on the same training subset of 10,000 products and evaluated it on a fixed validation subset of 512 products. During inference, five candidates were used in their original sampling order. The random seeds for constructing the training and validation subsets were set to 2026 and 2027, respectively.

For each threshold, we recorded the number of samples assigned to the high-confidence branch, n_1 ($m = 1$), and the low-confidence branch, n_3 (adaptive expansion to $m = 3$). We also measured the corresponding Top-1, Top-3, and Top-5 accuracies, along with their mean. The results are presented in Table 6.

Table 6: **Effect of the training confidence threshold on reactant generation performance.**

τ	n_1	n_3	Top-1	Top-3	Top-5	Mean
0.5	490	22	**45.90**	**58.79**	**62.50**	**55.73**
0.6	473	39	45.70	56.45	59.96	54.04
0.7	453	59	45.51	57.62	60.35	54.49
0.8	430	82	43.36	56.25	60.74	53.45
0.9	390	122	44.92	57.23	59.96	54.04

As the threshold τ increases, more samples are assigned to the low-confidence branch, resulting in a steady increase in n_3. Consequently, more samples require candidate expansion to $m = 3$, increasing computational overhead. Among the evaluated thresholds, $\tau = 0.5$ achieves the highest Top-1, Top-3, and Top-5 accuracies, as well as the highest mean accuracy. At this threshold, n_3 is only 22, meaning that few samples require candidate expansion, yielding the lowest overall computational cost. Considering both prediction accuracy and the inference budget, we therefore select $\tau = 0.5$ as the threshold for distinguishing high-confidence from low-confidence reaction-center predictions.

C.2 SELECTION OF THE CONFIDENCE THRESHOLD DURING INFERENCE

During inference, the threshold τ determines the candidate ranking cutoff m for the virtual source: $m = 1$ when $q > \tau$, and $m = 3$ otherwise. Table 7 summarizes the threshold sweep results supporting the choice of $\tau = 0.5$.

Evaluation was performed on a fixed, randomly sampled subset of 512 validation products, using a subset sampling seed of 2026. We varied only the inference confidence threshold, considering values in $\{0.5, 0.6, 0.7, 0.8, 0.9\}$. Five candidates were generated per product and evaluated in their original sampling order, with a sampling seed of 42 and a batch size of 32. Threshold selection prioritized Top-1 accuracy, with ties resolved by comparing Top-3 accuracy and then Top-5 accuracy.

Table 7: **Effect of the inference confidence threshold on reactant generation performance.**

τ	n_1	n_3	Top-1	Top-3	Top-5	Mean
0.5	492	20	**51.37**	**65.43**	**69.14**	**61.98**
0.6	468	44	51.17	65.23	68.75	61.72
0.7	439	73	50.98	65.23	68.95	61.72
0.8	411	101	50.98	64.84	68.36	61.39
0.9	379	133	50.00	64.84	68.95	61.26

The table reports the numbers of samples assigned to the high-confidence branch, n_1 ($m = 1$), and the low-confidence branch, n_3 ($m = 3$), together with the corresponding accuracies. Among the

evaluated thresholds, $\tau = 0.5$ achieved the highest Top-1 accuracy of 51.37%, alongside the highest Top-3 and Top-5 accuracies of 65.43% and 69.14%, respectively. We therefore selected $\tau = 0.5$ as the confidence threshold for inference.

C.3 Effectiveness of Confidence-Based Sample Selection for Candidate Set Expansion

We compared confidence-based and random selection to evaluate whether prediction confidence identifies products that benefit from candidate set expansion. Both strategies expanded the same number of products. Within the Overall Evaluation Set, we used a confidence threshold of $\tau = 0.5$: products with $q > \tau$ formed the High confidence group, whereas those with $q \leq \tau$ formed the Low-Confidence Reaction-Center Prediction Subset. The dynamic Top-m strategy retained $m = 1$ for the former and expanded to $m = 3$ for the latter.

For each random control, we sampled products without replacement from the Overall Evaluation Set, matching the size of the Low-Confidence Reaction-Center Prediction Subset. Selected products were expanded to $m = 3$, while all remaining products retained $m = 1$. Random selection used neither prediction confidence nor reaction labels or generated outputs, so the selected subsets could contain both high- and low-confidence products. We constructed five random subsets using seeds 2026–2030. All strategies used the same dynamic Top-m model, with 100 generated candidates per product, 50 nominal sampling steps, temperature 1, and chunk size 10. Each product used the sampling seed $(42 + \text{dataset position}) \bmod 2^{32}$. Overlapping products reused identical paired outputs under $m = 1$ and $m = 3$, and unselected products retained their original $m = 1$ outputs. This design compared selection strategies while holding the number of expanded products fixed.

Table 8 reports performance changes within each selected subset after expansion from $m = 1$ to $m = 3$. Confidence-based selection increased reaction-center coverage by 37.33 percentage points, compared with 7.73 ± 2.77 percentage points across the five random selections. Although random expansion improved center coverage in every trial, its effects on downstream exact-match accuracy varied in direction. By contrast, confidence-based expansion yielded positive gains across all reported Top-k metrics. Top-1 and Top-50 accuracy increased by 3.33 and 5.33 percentage points, respectively, compared with mean random-selection gains of only 0.27 and 0.13 percentage points. The mean and standard deviation describe variation across the five random selections.

Table 8: **Confidence-based versus random support expansion.** Changes are percentage points relative to fused $m = 1$, evaluated within each strategy's selected subset after expansion to $m = 3$. Δ Center Coverage measures annotated-center coverage; ΔTop-k measures exact-match accuracy. Random mean $\pm$ SD uses the five selection seeds.

Selection strategy	Δ Center Coverage	ΔTop-1	ΔTop-3	ΔTop-5	ΔTop-10	ΔTop-50
Random seed 1	+10.00	+2.67	+1.33	+2.67	+2.67	+2.00
Random seed 2	+9.33	−0.67	−0.67	−0.67	0.00	0.00
Random seed 3	+6.67	−2.67	−3.33	−2.00	−1.33	−2.00
Random seed 4	+9.33	+1.33	+0.67	+0.67	+0.67	+0.67
Random seed 5	+3.33	+0.67	+1.33	+1.33	0.00	0.00
Random mean $\pm$ SD	$+7.73 \pm 2.77$	$+0.27 \pm 2.03$	-0.13 ± 1.97	$+0.40 \pm 1.80$	$+0.40 \pm 1.46$	$+0.13 \pm 1.45$
Dynamic Top-m	+37.33	+3.33	+1.33	+4.00	+3.33	+5.33

Figure 5 shows the corresponding net gains. Confidence-based expansion covered the annotated reaction center for 56 additional products and produced a net increase of five products with a Top-1 exact match. With the same number of expanded products, confidence-based selection yielded greater coverage gains and larger average reactant-recovery gains than random selection. Under the evaluated inference settings, these results support using $\tau = 0.5$ to identify the Low-Confidence Reaction-Center Prediction Subset for expansion. Prioritizing this subset directs the expansion budget towards products that benefit more from $m = 3$, supporting the confidence-based selection mechanism in the dynamic Top-m strategy.

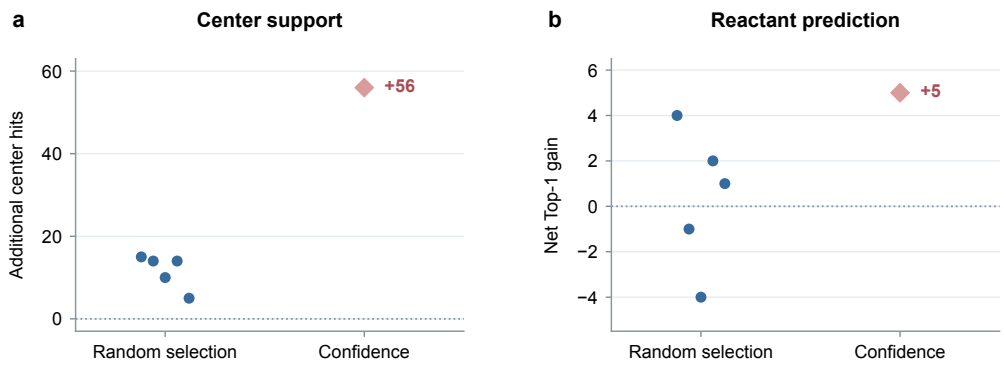


Figure 5: **Equal expansion budget, different downstream returns.** Each policy expands the same number of products from fused $m = 1$ to $m = 3$. Panel (a) shows additional annotated centers covered; panel (b) shows net additional Top-1 exact matches. Each blue point is one of five random-selection seeds; the muted-red diamond denotes selection of the low-confidence center-prediction subset. The same model, 100 draws, and per-product random seeds are used. Random groups can overlap, so these are selection replicates rather than independent training runs.

C.4 Effectiveness of the Confidence-Based Selection Mechanism

To evaluate the confidence-adaptive switching mechanism in our dynamic Top-m strategy, we conducted comparative experiments on the low-confidence center-prediction subset ($q \leq \tau$, with $\tau = 0.5$). We compared fixed multi-candidate inference strategies with confidence-driven dynamic Top-m inference to assess the benefits of selectively expanding candidate sets according to prediction confidence. All strategies were evaluated under the same experimental settings, with 100 samples generated per product. The results are presented in Table 9.

Table 9: **Fixed and dynamic Top-m strategies on the Low-Confidence Reaction-Center Prediction Subset ($q \leq \tau$, $\tau = 0.5$).** Values are Top-k exact-match accuracies (%). Each strategy uses 100 generated draws per product.

Strategy	Top-1	Top-3	Top-5	Top-10	Top-50
Top-$m = 1$	23.33	44.67	50.00	55.33	58.00
Top-$m = 3$	26.67	46.00	54.00	58.67	63.33
Top-$m = 5$	**27.33**	48.00	55.33	60.67	64.67
Dynamic Top-m	**27.33**	**52.00**	**56.00**	**62.67**	**67.33**

The fixed Top-m strategies applied a uniform candidate cutoff to all samples, using $m = 1$, 3, or 5. Increasing the fixed candidate count modestly improved reactant recovery for low-confidence samples. As m increased from 1 to 5, Top-1 accuracy rose from 23.33% to 27.33%, while Top-50 accuracy increased from 58.00% to 64.67%. However, this global expansion strategy does not account for prediction confidence and introduces additional computation for high-confidence samples that may not require candidate expansion.

By contrast, dynamic Top-m inference uses a confidence threshold to adaptively select the candidate count. It retains $m = 1$ for high-confidence predictions and expands the candidate set to $m = 3$ only for low-confidence predictions. The dynamic strategy achieved a Top-1 accuracy of 27.33%, matching the best fixed strategy ($m = 5$). It also outperformed all evaluated fixed strategies in Top-3, Top-5, Top-10, and Top-50 accuracy, reaching 52.00%, 56.00%, 62.67%, and 67.33%, respectively.

These results suggest that confidence-based candidate expansion can improve downstream reactant recovery for samples with low-confidence reaction-center predictions. By selectively expanding

candidate sets, the dynamic strategy aims to mitigate the effects of reaction-center prediction errors while avoiding additional expansion for high-confidence samples. The observed performance supports the benefits of the proposed dynamic strategy for reactant recovery on the low-confidence subset.

D Reaction-Center Errors and Downstream Recovery

D.1 Tolerance to and Recovery from Reaction-Center Prediction Errors through Soft Bond Disconnection

This subsection investigates whether downstream retrosynthesis generation can recover the correct target reactants when the highest-confidence (Top-1) reaction-center prediction is incorrect.

Table 10 compares two model configurations: a hard-synthon model and a virtual-synthon model with fixed $m = 1$. For each product, both models use only the single predicted Top-1 reaction center as reaction-center input and generate 100 samples conditioned on that center. Samples are divided into two groups according to whether the Top-1 prediction matches the ground-truth reaction center. The **Covered** group (90.25%) comprises samples with a correct Top-1 reaction-center prediction, whereas the **Uncovered** group (9.75%) comprises samples with an incorrect Top-1 reaction-center prediction.

Table 10: **Reactant accuracy stratified by reaction-center prediction status.** Values are within-group Top-k exact-match accuracies (%). Covered and Uncovered denote correct and incorrect single Top-1 reaction-center predictions, respectively. The two models use the same Overall Evaluation Set. Fraction is the percentage of the Overall Evaluation Set in each group.

Coverage	Fraction (%)	Method	Top-1	Top-3	Top-5	Top-10	Top-50
Covered	90.25	Hard	66.29	83.09	86.46	89.08	90.73
		Virtual, $m = 1$	**69.48**	**84.63**	**87.47**	**89.40**	**92.06**
Uncovered	9.75	Hard	5.52	19.32	27.60	33.97	40.13
		Virtual, $m = 1$	**8.49**	**36.52**	**46.07**	**53.08**	**64.12**
Overall	100.00	Hard	60.34	76.84	80.69	83.67	85.76
		Virtual, $m = 1$	**63.51**	**79.90**	**83.40**	**85.82**	**89.30**

The Covered group accounts for 90.25% of the overall evaluation set and comprises samples with correct Top-1 reaction-center predictions. Our analysis focuses on the Uncovered group, which accounts for the remaining 9.75%. Even when provided with only a single incorrect reaction-center prediction, the virtual-synthon model achieves higher reactant recovery accuracy. Its Top-1 accuracy reaches 8.49%, compared with 5.52% for the hard-synthon model, and it also outperforms the hard-synthon model in Top-3, Top-5, Top-10, and Top-50 accuracy. These results indicate that the soft bond disconnection mechanism provides greater tolerance to reaction-center prediction errors, allowing correct reactants to be recovered despite incorrect reaction-center guidance.

On the overall evaluation set, the virtual-synthon model achieves a Top-1 accuracy of 63.51%, compared with 60.34% for the hard-synthon model. Together, these findings support the ability of soft bond disconnection to tolerate incorrect reaction-center inputs and improve downstream reactant recovery. The virtual-synthon model retains an advantage even on challenging samples where the sole predicted reaction center is incorrect and no alternative centers are available.

D.2 Dynamic Top-m Strategy: Tolerance to Reaction-Center Prediction Errors and Reactant Recovery

To assess whether dynamic virtual synthons recover correct reactants despite reaction-center prediction errors, we grouped test samples by Top-1 prediction correctness and candidate-set coverage. Group A comprises samples with correct Top-1 reaction-center predictions. In Group B, the Top-1 prediction is incorrect, but the retained candidate set includes the ground-truth reaction center. In Group C, the Top-1 prediction is incorrect, and the retained candidate set excludes the ground-truth

reaction center. Table 11 reports the results. Because candidate sets vary across methods, Groups B and C contain different samples and occupy different fractions of the Overall Evaluation Set.

Table 11: **Reactant recovery by reaction-center prediction status for fixed and dynamic Top-m strategies.** Values are within-group Top-k exact-match accuracies (%). Fraction is the percentage of the Overall Evaluation Set. Group A has correct Top-1 center predictions; Groups B and C have incorrect Top-1 predictions with and without ground-truth center coverage, respectively. Coverage is evaluated using each method's candidate set, so Groups B and C differ across methods. Bold values indicate the highest reported accuracy within each group and metric.

Center group	Method	Fraction (%)	Top-1	Top-3	Top-5	Top-10	Top-50
A: Top-1 correct	Top-$m = 1$	86.58	**71.87**	86.80	89.48	91.22	93.52
	Top-$m = 3$	86.58	70.44	87.01	89.72	90.94	92.44
	Top-$m = 10$	86.58	70.56	86.61	89.24	90.65	91.82
	Dynamic virtual	86.58	71.30	**88.54**	**90.82**	**92.61**	**94.24**
B: Top-1 incorrect; candidate set covered	Top-$m = 1$	10.66	11.46	41.75	51.65	58.25	68.74
	Top-$m = 3$	12.09	14.73	33.90	39.38	44.18	47.43
	Top-$m = 10$	13.36	14.73	32.09	38.29	41.55	44.65
	Dynamic virtual	9.57	**15.37**	**45.89**	**53.68**	**61.26**	**70.13**
C: Top-1 incorrect; candidate set uncovered	Top-$m = 1$	2.75	**3.01**	**12.03**	**16.54**	24.06	**37.59**
	Top-$m = 3$	1.33	0.00	1.56	1.56	1.56	1.56
	Top-$m = 10$	0.06	0.00	0.00	0.00	**33.33**	33.33
	Dynamic virtual	3.85	1.61	6.99	10.22	13.44	22.04
Overall	Top-$m = 1$	100.00	**63.51**	79.90	83.40	85.82	**89.30**
	Top-$m = 3$	100.00	62.74	79.42	82.43	84.06	85.76
	Top-$m = 10$	100.00	63.03	79.24	82.34	84.02	85.45
	Dynamic virtual	100.00	63.24	**81.29**	**84.12**	**86.52**	89.11

When the Top-1 reaction-center prediction is correct, all methods achieve high Top-k accuracy. Dynamic virtual synthons achieve the highest Top-3, Top-5, Top-10, and Top-50 accuracies in Group A. Their Top-1 accuracy is 71.30%, slightly below the 71.87% achieved by fixed $m = 1$. Thus, the dynamic strategy maintains strong recovery on samples with correct initial predictions.

The central comparison concerns Group B, where an incorrect Top-1 prediction coexists with ground-truth center coverage in the candidate set. The fixed $m = 1$ baseline achieves 11.46% Top-1 accuracy in its Group B. Fixed $m = 3$ and $m = 10$ achieve higher Top-1 accuracy but lower Top-3 and higher-rank accuracies in their respective Group B subsets. Dynamic virtual synthons record the highest accuracies across all reported metrics in Group B, reaching 15.37% Top-1, 61.26% Top-10, and 70.13% Top-50 accuracy. These results show substantial recovery despite incorrect initial predictions when alternative candidate centers are retained. However, the differing subsets prevent attributing the performance differences solely to sampling allocation.

When the ground-truth reaction center is absent from the candidate set, Top-1 recovery remains low across all methods. In Group C, dynamic virtual synthons achieve 1.61% Top-1 accuracy, compared with 3.01% for fixed $m = 1$. The dynamic strategy does not consistently outperform the fixed strategies in this group. The results are consistent with limited recovery when the correct center is unavailable, although the differing group sizes restrict direct comparisons.

On the Overall Evaluation Set, dynamic virtual synthons achieve the highest Top-3, Top-5, and Top-10 accuracies, reaching 81.29%, 84.12%, and 86.52%, respectively. Fixed $m = 1$ retains the highest Top-1 and Top-50 accuracies, at 63.51% and 89.30%, compared with 63.24% and 89.11% for dynamic virtual synthons. These findings support improved recovery at several candidate ranks, together with strong performance on the dynamic strategy's covered center-error subset.

E Propagation of Reaction-Center Prediction Errors and Error Mitigation by Virtual Synthons

E.1 Impact of Reaction-Center Error Propagation

To quantify the downstream performance loss caused by reaction-center prediction errors, we fixed the hard-synthon model weights and varied only the reaction-center input during inference. Specifically, we compared Top-k exact-match accuracy using ground-truth reaction centers (Oracle) versus model-predicted reaction centers. This controlled comparison isolates the performance degradation attributable to reaction-center prediction errors. Table 12 reports results on the Overall Evaluation Set using the same hard-synthon baseline as Table 13.

Table 12: **Impact of reaction-center prediction errors on hard-synthon reactant recovery.** Oracle and Predicted denote Top-k exact-match accuracies (%) using ground-truth and model-predicted reaction centers, respectively. Accuracy drop is Oracle minus Predicted, in percentage points. Model weights are fixed across both conditions.

Metric	Oracle (%)	Predicted (%)	Accuracy drop (pp)
Top-1	65.58	60.34	5.24
Top-3	82.22	76.84	5.38
Top-5	85.66	80.69	4.97
Top-10	87.73	83.67	4.06

Replacing ground-truth reaction centers with model-predicted centers reduces accuracy across all evaluated Top-k metrics. The accuracy drops are 5.24, 5.38, 4.97, and 4.06 percentage points at Top-1, Top-3, Top-5, and Top-10, respectively. The gap is largest at Top-3 and narrows at higher ranks, but remains 4.06 percentage points at Top-10. These results indicate that reaction-center prediction errors propagate to downstream reactant generation.

E.2 Mitigation of Reaction-Center Prediction Errors by Virtual Synthons

We further conducted reaction-center intervention experiments comparing hard synthons, fixed-fusion virtual synthons (Fused $m = 1$), and dynamic virtual synthons (Dynamic virtual). These experiments assessed the robustness of each framework to reaction-center prediction errors. We define Δ_{all} as the Oracle-minus-Predicted accuracy difference on the Overall Evaluation Set. We define Δ_{wrong} as the corresponding difference on the subset with incorrect reaction-center predictions. A smaller Δ indicates a smaller accuracy gap between the two input conditions, but must be interpreted alongside absolute accuracy. The results are presented in Table 13.

Table 13: **Reaction-center intervention results for hard and virtual synthons.** Predicted and Oracle denote Top-k exact-match accuracies (%) under predicted and ground-truth reaction-center inputs, respectively. Δ_{all} and Δ_{wrong} are Oracle-minus-Predicted differences in percentage points on the Overall Evaluation Set and the center-error subset. Bold values indicate the smallest gap for each metric; lower is better.

		Overall Evaluation Set			Center-error subset	
Metric	Condition	Predicted	Oracle	$\Delta_{\text{all}}\downarrow$	Predicted	$\Delta_{\text{wrong}}\downarrow$
Top-1	Hard-synthon	60.34	65.58	5.24	14.66	32.25
	Fused $m = 1$	63.51	67.13	**3.62**	9.72	32.10
	Dynamic virtual	63.24	67.13	3.89	11.42	**30.40**
Top-10	Hard-synthon	83.67	87.73	4.06	45.06	27.31
	Fused $m = 1$	85.82	87.41	1.59	51.23	**15.59**
	Dynamic virtual	86.52	87.56	**1.04**	47.53	20.83

On the Overall Evaluation Set, both virtual-synthon approaches exhibit smaller accuracy gaps, Δ_{all}, than the hard-synthon baseline at Top-1 and Top-10. For Top-10 accuracy, Δ_{all} is 4.06 percentage points for hard synthons but only 1.04 percentage points for dynamic virtual synthons. Dynamic virtual synthons also achieve higher predicted-center accuracy, at 86.52% versus 83.67%. Together, these results support reduced downstream performance loss under reaction-center prediction errors in the evaluated configurations.

F Controlled Experiments on Training–Inference Distribution Shift

A straightforward approach to retrosynthesis modeling is to construct synthons using ground-truth reaction centers during training, thereby avoiding training-time errors from inaccurate first-stage reaction-center predictions. However, this approach introduces a training–inference distribution shift. The model learns from synthons constructed using ground-truth centers during training, whereas inference relies on synthons constructed using noisy predicted centers. We compared three experimental settings to evaluate our virtual-synthon approach against this alternative, which eliminates reaction-center prediction errors during training but introduces a distribution mismatch at inference.

The three experimental settings are as follows: **A:** Hard synthons constructed using ground-truth reaction centers during training and predicted reaction centers during inference; **B:** Hard synthons constructed using predicted reaction centers during both training and inference; and **C:** Virtual synthons constructed using predicted reaction centers during both training and inference. Table 14 reports the results.

Table 14: **Reactant generation performance under different training-center sources and synthon representations.** Values are Top-k exact-match accuracies (%). GT denotes ground-truth centers and Pred. predicted centers. All three settings use predicted centers during inference. Bold values indicate the best result in each accuracy column.

Setting	Synthon	Training center	Inference center	Top-1	Top-3	Top-5	Top-10	Top-50
A	Hard	GT	Pred.	62.26	76.80	78.82	79.78	80.75
B	Hard	Pred.	Pred.	60.34	76.84	80.69	83.67	85.76
C	Virtual	Pred.	Pred.	**63.51**	**79.90**	**83.40**	**85.82**	**89.30**

A versus B. Both settings use predicted reaction centers during inference, while their training reaction centers come from different sources. Training with ground-truth centers (A) yields higher Top-1 accuracy than training with predicted centers (B). However, A performs worse than B in Top-5, Top-10, and Top-50 accuracy. This pattern is consistent with a training–inference distribution mismatch. A model trained exclusively on synthons constructed using ground-truth centers may generalize less effectively to noisy predicted centers, reducing recovery performance when multiple candidates are considered.

C versus B. Both settings use predicted reaction centers during training and inference, with hard synthons in B and virtual synthons in C. Virtual synthons (C) outperform the hard-synthon baseline (B) across all evaluated Top-k metrics. These results support improved overall reactant recovery for the virtual-synthon configuration under predicted-center inputs.

C versus A. C outperforms A across all evaluated Top-k metrics. Setting A eliminates reaction-center prediction errors during training but introduces a mismatch between training and inference conditions. By contrast, C uses predicted reaction centers during both training and inference, maintaining consistency in how reaction-center inputs are obtained. Without using ground-truth centers to construct training synthons, C achieves higher Top-1 and higher-rank Top-k accuracies. Under the evaluated conditions, these results favor the virtual-synthon configuration over training hard synthons exclusively with ground-truth centers. The pattern is consistent with greater robustness to predicted-center inputs.

G Structural Heterogeneity and Similarity to the Training Set

Stratified analyses reveal that the predictive gains from dynamic virtual synthons vary with product structural features and similarity to the training set. Dynamic virtual synthons improve Top-1 accuracy across most evaluated structural subgroups, with identifiable differences between subgroups (Figure 6 and Table 15). For medium-sized products containing 21–30 heavy atoms, dynamic virtual synthons outperform hard synthons by 4.25 percentage points in Top-1 accuracy, demonstrating an advantage for products of moderate size. For larger products with ≥ 41 heavy atoms, the two models have equal Top-1 accuracy, with no performance degradation from the dynamic strategy at this rank.

For products containing 2–3 rings or ≥ 4 rings, dynamic virtual synthons consistently improve Top-1 accuracy. Hard synthons show a slight advantage only for products with 0–1 rings. When stratified by the number of bond edits, dynamic virtual synthons show a clear advantage for reactions involving a single bond edit. Hard synthons achieve slightly higher Top-1 accuracy only on the subset requiring ≥ 3 bond edits. Reaction-class results under class-unknown inference are reported in Table 16.

Further stratification by each product's maximum Tanimoto similarity to the training set (Table 17) shows that dynamic virtual synthons outperform hard synthons in Top-1 accuracy across all similarity ranges. This indicates improved first-ranked reactant predictions for products with both high and low similarity to the training set. Top-10 accuracy also improves across all evaluated similarity ranges.

Table 15: **Structural stratification of reactant recovery.** Accuracies and group proportions are percentages. Each descriptor partitions the Overall Evaluation Set.

			Top-1		Top-10	
Descriptor	Group	Proportion (%)	Hard	Dynamic	Hard	Dynamic
Heavy atoms	≤ 20	32.15	58.34	61.11	84.22	85.96
	21–30	36.53	58.92	63.17	81.98	85.84
	31–40	25.03	64.02	65.84	85.53	88.17
	≥ 41	6.29	64.14	64.14	83.22	86.84
Rings	0–1	16.21	57.98	57.73	82.63	84.80
	2–3	46.80	58.82	63.47	83.72	86.20
	≥ 4	36.99	63.29	65.36	84.05	87.69
Heteroatoms	0–3	12.92	51.44	56.25	80.13	82.05
	4–6	39.00	58.39	61.09	83.49	86.15
	≥ 7	48.09	64.31	66.85	84.76	88.03
Bond edits	1	23.78	43.95	49.61	79.63	81.11
	2	67.29	68.69	71.21	87.67	90.96
	≥ 3	8.92	41.07	39.44	64.27	67.52

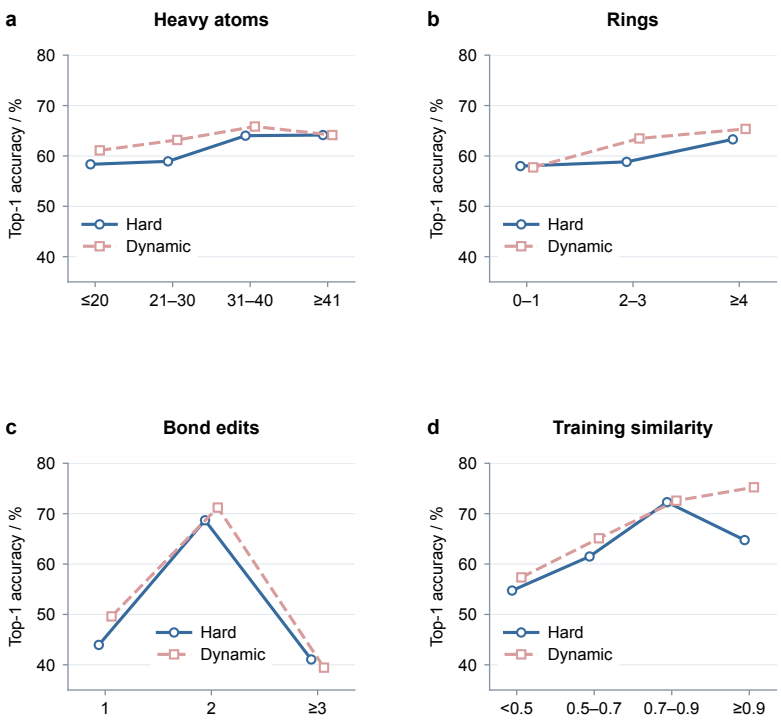


Figure 6: **Top-1 reactant recovery across structural and training-similarity groups.** Empirical accuracies are stratified by (a) product heavy-atom count, (b) ring count, (c) bond-edit count, and (d) maximum Tanimoto similarity to training products. Group proportions are reported in Tables 15 and Table 17. Similarity bins are $[0, 0.5)$, $[0.5, 0.7)$, $[0.7, 0.9)$, and $[0.9, 1]$. Lines connect discrete groups to guide the eye. Points are fixed-cohort proportions, not averages over training seeds.

Table 16: **Reaction-class stratification under class-unknown inference.** Proportions are percentages of the Overall Evaluation Set. Class labels are used only for this post-hoc analysis. Very small classes, such as class 4 (0.12%), cannot support broad comparative conclusions.

		Top-1		Top-10	
Class	Proportion (%)	Hard	Dynamic	Hard	Dynamic
1	31.38	63.92	67.55	83.77	89.25
2	24.61	75.11	77.21	91.42	93.36
3	10.31	47.39	45.58	69.88	73.49
4	0.12	66.67	66.67	66.67	83.33
5	1.26	96.72	96.72	96.72	96.72
6	16.97	46.71	53.17	84.39	85.37
7	9.36	53.10	55.09	84.96	83.63
8	1.70	46.34	52.44	78.05	76.83
9	3.81	40.22	41.30	60.87	67.93
10	0.48	82.61	82.61	95.65	91.30

Table 17: **Maximum training-product similarity and exact-match accuracy (%).** Proportions are percentages of the Overall Evaluation Set.

		Top-1		Top-10	
Similarity	Proportion (%)	Hard	Dynamic	Hard	Dynamic
< 0.50	40.39	54.74	57.36	80.37	82.98
$[0.50, 0.70)$	43.55	61.50	65.11	84.89	87.88
$[0.70, 0.90)$	13.89	72.28	72.58	88.82	91.65
≥ 0.90	2.17	64.76	75.24	87.62	92.38

H Association between Reaction-Center Confidence and Downstream Prediction Performance

This section examines the association between reaction-center confidence and prediction performance through confidence-stratified accuracy analysis (Table 18) and reaction-center confidence calibration analysis (Table 19), complemented by the confidence distributions in Figure 7.

Table 19 groups samples into bins according to their predicted reaction-center confidence and compares the mean confidence with the empirical reaction-center accuracy within each bin. Reaction-center accuracy generally increases with predicted confidence, indicating a strong association between the two. The expected calibration error (ECE) is 4.78%, and the Brier score is 0.08921. These results suggest that reaction-center confidence is reasonably well calibrated and provides a useful indication of the likelihood that a reaction-center prediction is correct.

Table 18 groups samples by reaction-center confidence and reports reaction-center identification accuracy alongside downstream Top-1 and Top-10 accuracies for the Hard and Dynamic models. Downstream Top-1 reactant prediction accuracy generally increases with reaction-center confidence, indicating an association between confidence and downstream retrosynthesis performance. In the highest-confidence group, reaction-center identification accuracy reaches 98.89%, while the Dynamic model achieves a downstream Top-1 accuracy of 72.13%. These results suggest that reaction-center confidence provides a useful signal for anticipating downstream retrosynthesis performance and could support confidence-based prediction and abstention strategies.

Table 18: **Confidence-stratified reaction-center and downstream prediction accuracy.** Values are percentages. Proportions are calculated within the center-labeled evaluation set. Center accuracy uses labeled reactions; downstream Top-k accuracy uses all evaluated reactions in each bin.

			Top-1		Top-10	
Confidence	Proportion (%)	Center accuracy	Hard	Dynamic	Hard	Dynamic
≤ 0.50	3.11	36.67	32.00	29.33	57.33	60.67
$(0.50, 0.80]$	13.61	55.71	40.03	40.18	68.95	76.26
$(0.80, 0.95]$	15.49	78.88	51.60	56.02	79.14	83.29
$(0.95, 0.99]$	15.74	90.00	62.16	67.54	85.41	84.23
> 0.99	52.06	98.89	69.38	72.13	89.90	92.41

Figure 7 shows that correct reaction-center predictions have confidence concentrated near one, whereas incorrect predictions span a broader range. For both Hard and Dynamic, correct Top-1 reactant predictions tend to have higher reaction-center confidence. However, substantial overlap remains, with many incorrect reactant predictions also receiving high center confidence.

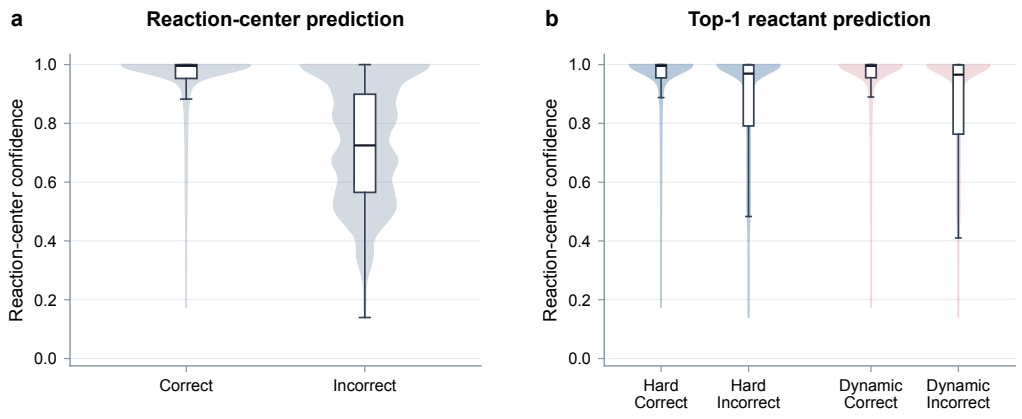


Figure 7: **Reaction-center confidence distributions by prediction outcome.** (a) Correct and incorrect reaction-center predictions, excluding records without center labels. (b) Correct and incorrect Top-1 reactant predictions from Hard and Dynamic on the same evaluation reactions; correctness denotes exact match. Violins use a common Gaussian bandwidth with reflection at zero and one, are restricted to observed ranges, and have equal maximum widths. Boxes show medians and interquartile ranges; whiskers extend to the most extreme observations within 1.5 interquartile ranges. Outlier markers are omitted, but all observations enter the densities.

Table 19: **Reaction-center confidence calibration.** Ten equal-width bins were used; only nonempty bins are shown. Intervals are left-closed and right-open, with 1 included in the final bin. Proportions are calculated within the center-labeled evaluation set. Mean confidence and accuracy are percentages.

Bin	Proportion (%)	Mean confidence	Center accuracy
0.1–0.2	0.06	16.77	33.33
0.2–0.3	0.14	26.24	28.57
0.3–0.4	0.81	35.48	28.21
0.4–0.5	2.09	46.09	40.59
0.5–0.6	4.18	54.85	49.01
0.6–0.7	4.51	65.29	50.92
0.7–0.8	4.91	75.45	65.82
0.8–0.9	7.89	85.51	72.97
0.9–1.0	75.40	98.60	95.63