Supplementary Information for Retro-BLEU: Quantifying Chemical Plausibility of Retrosynthesis Routes through Reaction Template Sequence Analysis

Junren Li, Lei Fang*, Jian-Guang Lou

1 Reaction/Template n-gram overlap analysis under different partition settings

In addition to PaRoutes set-n1 & set-n5 test sets, we also proposed three settings to build known synthesis routes: 1) temporal partitioning, we use synthesis routes from patents published in and before 2013 as the known set, 2) random patent-based partitioning, we randomly partition the patents into 80% and 20%, routes from the 80% patents of patents are considered as the known set, and 3) random route-based partitioning, we randomly partition all the routes into 80% and 20%, the 80% split is considered as the known set.

Table 1: N-gram (reaction/templates) overlap ratio for patent routes under various partitioning settings.

n-grams		n=2			n=3			n=4	
ratio category	reaction	template	coverage	reaction	template	coverage	reaction	template	coverage
Temporal-based	70.0%	82.4%	100%	70.9%	73.4%	44.6%	70.6%	71.8%	19.6%
Random patent-based	80.8%	89.4%	100%	81.4%	83.8%	42.4%	82.6%	84.0%	18.4%
Random route-based	83.2%	94.2%	100%	83.8%	90.8%	42.1%	83.7%	89.8%	18.2%

Supporting Table 1 shows the overlap analysis on the three settings. All three partition methods exhibit a higher overlap ratio than the original partition, indicating that the recorded patent routes are highly conservative with known routes. Approximately 70% reaction n-grams appear in known synthesis routes when the value of n ranges from 2 to 4 under the temporal-based partition. The overlap ratio rises to above 80% for the other two partition settings.

2 Template n-gram overlap analysis under different radii

Table 2: N-gram (reaction/templates) overlap ratio for patent routes and model-generated routes on Retro*-190 under different radii.

n-grams		n=2			n=3			n=4	
template radius	r=0	r=1	r=2	r=0	r=1	r=2	r=0	r=1	r=2
Retro*-190[1]	61.2%	42.9%	34.8%	34.7%	28.9%	23.0%	24.5%	21.5%	16.2%
Retro*(165)[1]	53.0%	31.2%	20.6%	22.9%	16.9%	13.8%	17.2%	14.9%	11.4%
Retro*+(183)[2]	51.2%	29.5%	17.5%	20.0%	13.6%	9.7%	12.0%	10.2%	7.0%
EG-MCTS(183)[3]	33.8%	13.9%	8.0%	5.9%	5.2%	3.5%	2.9%	3.1%	2.2%
RetroGraph(189)[4]	42.0%	20.1%	12.3%	11.5%	7.6%	5.9%	5.6%	4.5%	2.9%

Here we present the results by using a radius of 0 and 2 to analyze the template n-gram overlap, as shown in Supporting Table 2. The template radius determines the size of the corresponding chemical environment surrounding the reaction center, which directly influences the sparsity of the chemical space. A template with a radius of 0 only represents the changing atoms in the reaction center, while a reaction template with an infinite radius essentially covers the entire reaction. In Table 2, it can be seen that the overlap ratio between patent and generated routes under a radius of 0 is less significant than a radius of 1 or 2. Moreover, the overlap template bigram ratio for patent routes at a radius of 2 is only 34.8%. This observation suggests that using a radius of 1 is an optimal choice for evaluating template sequences.

3 Top-k accuracy of Retro-BLEU and other baselines on set-n1

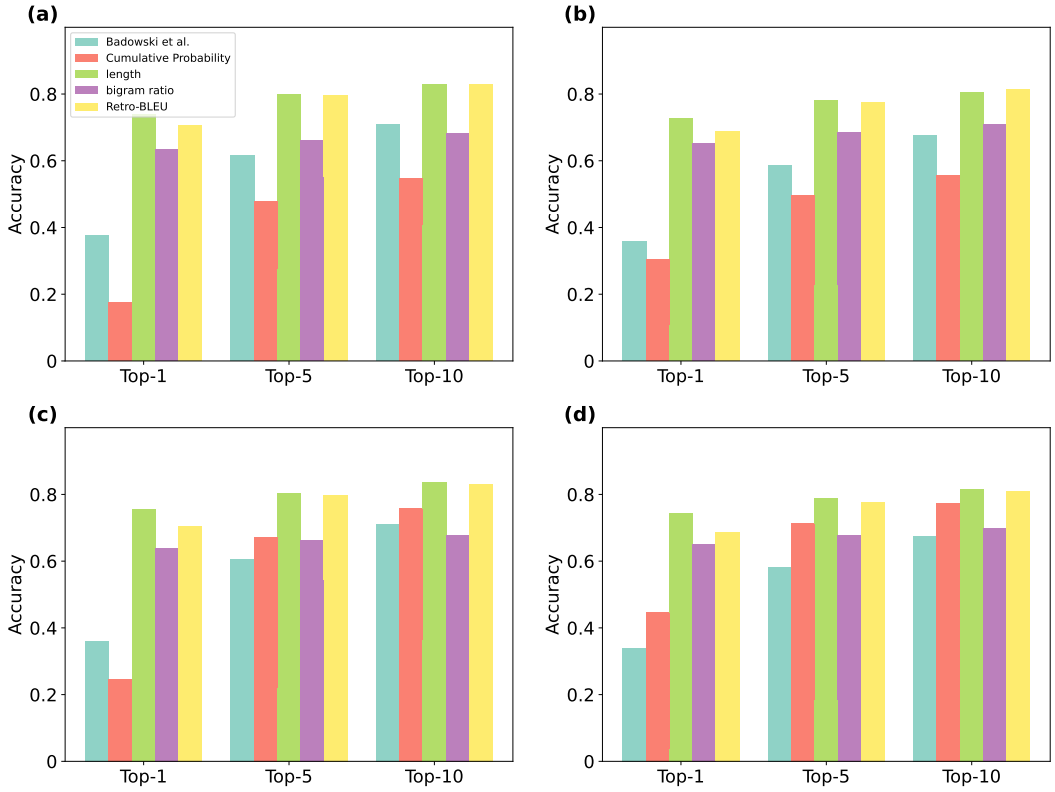

Figure 1: Top-k accuracies for Retro-BLEU and other metrics. The top and bottom of areas with the diagonal line markings represent the best-case and worst-case scenarios, respectively. (a) MCTS algorithm applied to set-n1, (b) Retro* algorithm applied to set-n1, (c) MCTS algorithm applied to set-n1 searchable routes. (d) Retro* algorithm applied to set-n1 searchable routes.

We also generated 2,896,036 routes using Monte Carlo Tree Search (MCTS)[5] and 3,008,148 routes using Retro* [1] for molecules in set-n1 with AiZynthFinder [6]. Herein, we provide the top-k accuracy of Retro-BLEU and other baselines on set-n1, as shown in Supporting Figure 1. The overall trends observed for set-n1 are consistent with those found for set-n5, except that the length metric slightly outperforms Retro-BLEU in the best scenarios. The reason is that routes in set-n1 (3.18) are shorter than in set-n5 (3.84), which means the patent routes are more likely to have the same length as shorter generated routes.

4 Exploring the relationship between Retro-BLEU and filtering strategies

Noticing the gap between computer-assisted designed routes and those that succeed in wet-lab experiments, Genheden *et al.* developed a filter policy, implemented as a neural network, to improve the performance and accuracy of MCTS in retrosynthetic route finding[7]. The filter policy evaluates the plausibility of reactions during the expansion phase, using a training dataset composed of artificially created unfeasible reactions. In our study, we adopted their setting with a threshold of 0.05, meaning reactions with a filter score below 0.05 are rejected, yielding a precision of 0.95 and a recall of 0.89 on their dataset[7].

In real-world applications, retrosynthesis planning programs produce multiple routes, which require chemists to employ their domain expertise to select chemically plausible routes among the available options. To facilitate this process, we classify the first ten generated routes for each target molecule in set-n1 using Retro-BLEU scores, dividing them into three categories: Retro-BLEU > 3.5, $3.5 <$ Retro-BLEU ≥ 4.5, and Retro-BLEU ≤ 3.5. Given that the initial routes lacked a filter policy[8], we employed the provided checkpoint[9] to assess their validity. After that, we calculated the plausibility rate for each category by determining the proportion of routes in which all the reactions passed the filter and present our findings in Supporting Table 3.

Table 3: Plausibility rate of routes generated for set-n1 target molecules categorized by Retro-BLEU score intervals

Algorithm	Retro-BLEU	[4.5, 2e]	[3.5, 4.5)	[2, 3.5)
MCTS	No. Routes	6,260	60,877	30,177
	Plausibility Rate	81.2%	72.9%	50.3%
Retro*	No. Routes	3,771	65,874	27,211
	Plausibility Rate	77.0%	66.8%	54.1%

Supporting Table 3 highlights a positive correlation between Retro-BLEU scores and the plausibility rate of retrosynthesis routes, as determined by the filter policy. This correlation suggests that as the Retro-BLEU score increases, the likelihood of a route being chemically feasible also increases. For instance, in the highest Retro-BLEU score category (4.5, 2e], both MCTS and Retro* algorithms exhibit higher plausibility rates (81.2% and 77.0%, respectively) compared to the lower Retro-BLEU score categories. This trend continues in the medium Retro-BLEU score category [3.5, 4.5), with plausibility rates of 72.9% and 66.8% for MCTS and Retro*, respectively.

The positive correlation between Retro-BLEU scores and plausibility rates supports that Retro-BLEU is a useful metric for evaluating retrosynthetic routes. As higher Retro-BLEU scores correspond to higher plausibility rates, it indicates that the metric effectively captures the chemical validity and practicality of the generated routes.

5 Sample routes suggested by Retro-BLEU

In addition to the example discussed in the main text, we present five additional cases randomly selected where the Retro-BLEU-suggested, model-generated routes score higher than their corresponding patent test routes. The suggested model-generated routes and corresponding patent routes are presented in Supplementary Figure 2 to Figure 6. These routes were generated using the Retro* algorithm, targeting set-n1 molecules. Interestingly, most routes selected by Retro-BLEU share similarities with the patent test routes, differing mainly in reagents, starting materials, or protection steps.

From these analyses, it can be concluded that while top-ranked routes generated by Retro-BLEU may not always surpass the patent routes, particularly due to the absence of specific reaction conditions or a bias towards common reactions, they represent meaningful alternatives to existing synthesis routes. This approach is substantially more effective than randomly selecting routes from retrosynthesis software.

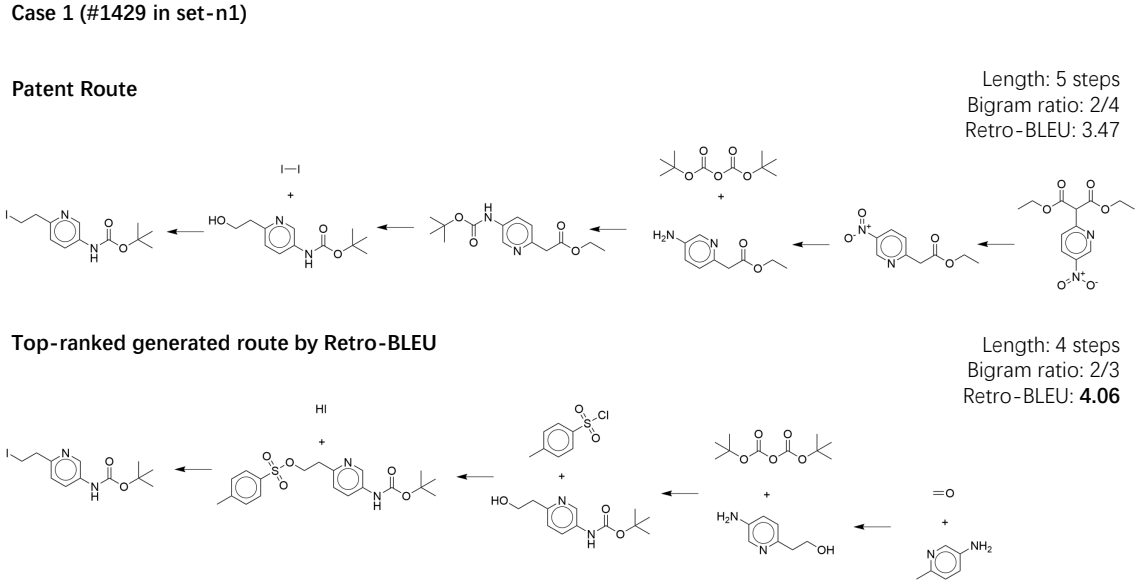

Figure 2: Randomly selected cases on set-n1 targets #1. The model-generated route utilizes a significantly simpler starting molecule, undergoing a similar acylation process to yield the hydroxyl precursor for subsequent iodization.

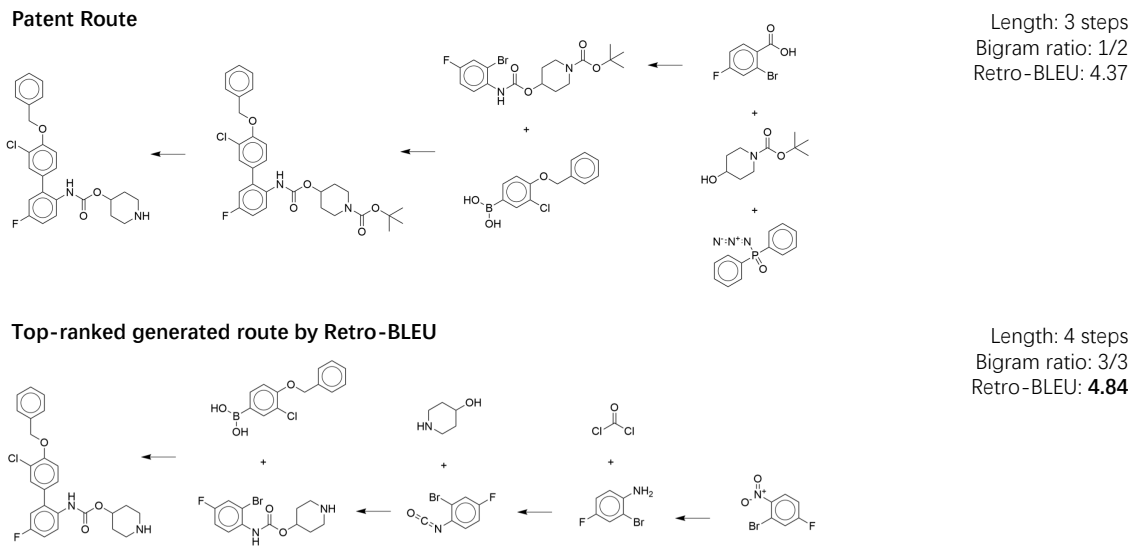

Figure 3: Randomly selected cases on set-n1 targets #2. In case 2, despite the model-generated route being one step longer, it starts from a simpler material compared to the patent. The two reactant parts are eventually coupled using a Suzuki-Miyaura reaction, mirroring the patent route's strategy.

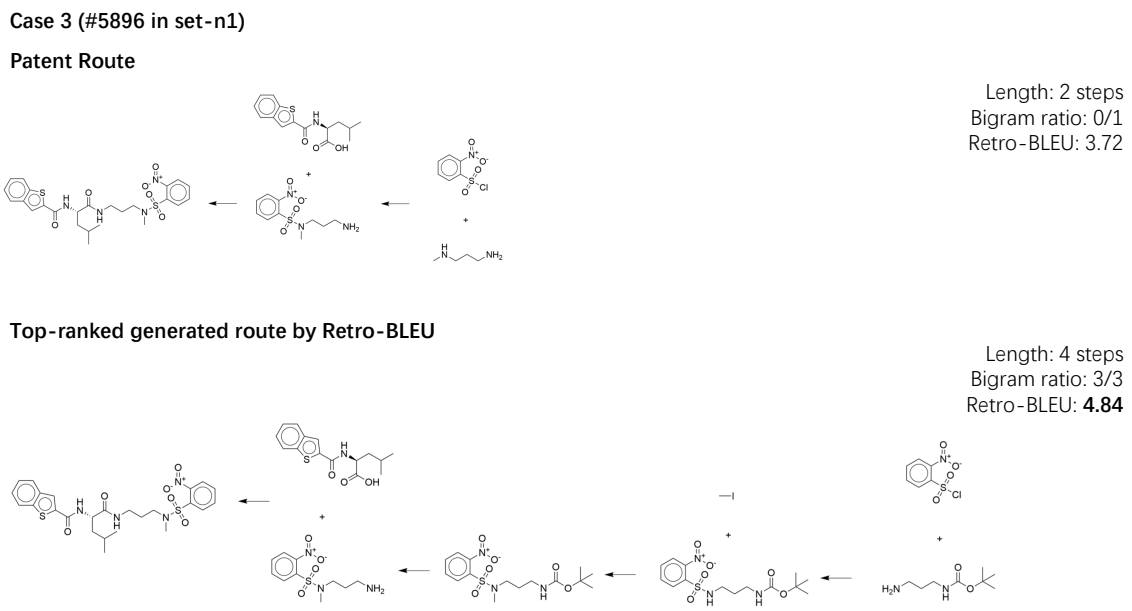

Figure 4: Randomly selected cases on set-n1 targets #3. The route in case 3 employs a protected amine as the starting material, leading to a lengthier synthesis due to the necessary deprotection process. However, all three bigrams in this route are recorded, resulting in a notably high Retro-BLEU score.

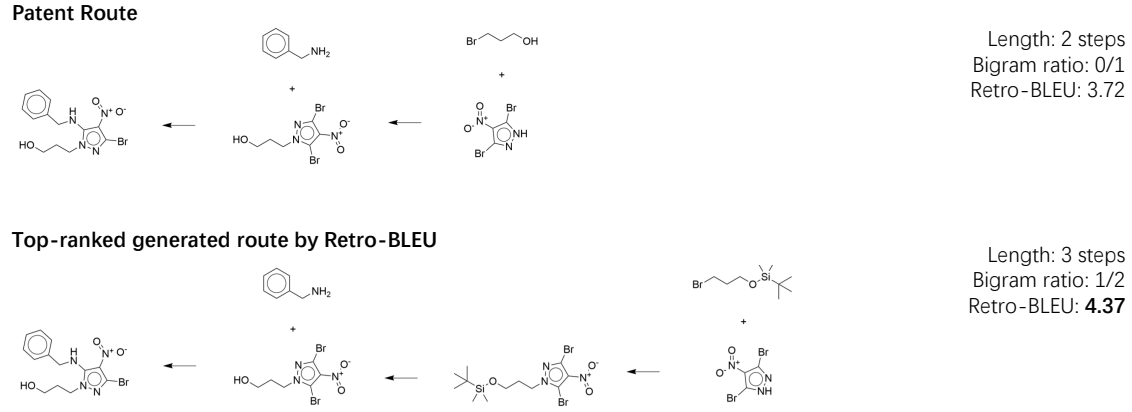

Figure 5: Randomly selected cases on set-n1 targets #4. The suggested model-generated route employs a silyl group to protect the unreacted hydroxyl group, a strategy reflected in the bigrams and contributing to its higher Retro-BLEU score.

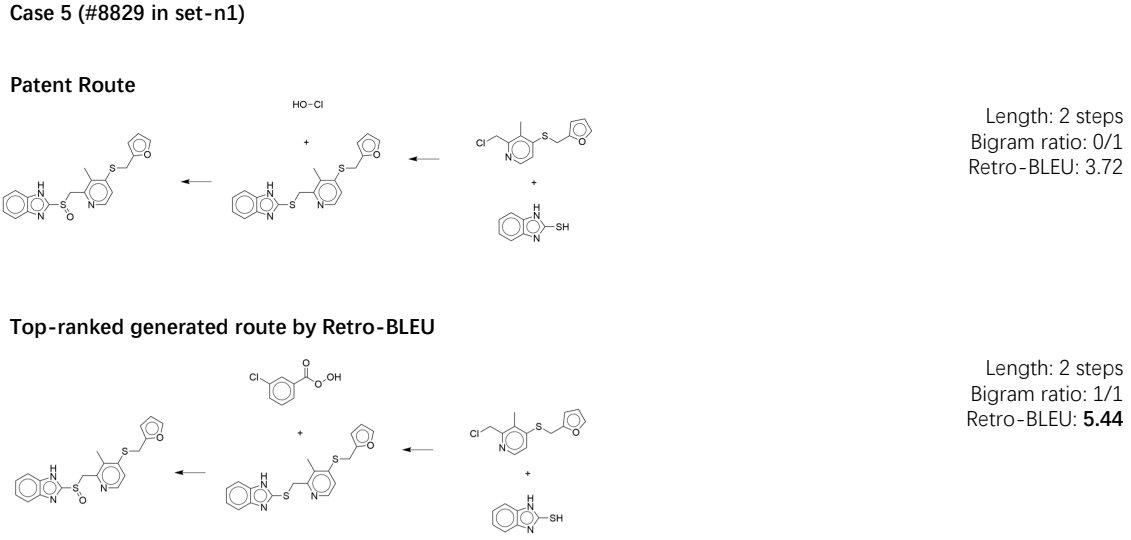

Figure 6: Randomly selected cases on set-n1 targets #5. The model-suggested route closely resembles the patent route, except for substituting the oxidation reagent with the more common mCPBA, indicating Retro-BLEU's preference for familiar synthesis reagents and strategies.

References

[1] Binghong Chen, Chengtao Li, Hanjun Dai, and Le Song. Retro*: learning retrosynthetic planning with neural guided a* search. In *International Conference on Machine Learning*, pages 1608–1616. PMLR, 2020.

[2] Junsu Kim, Sungsoo Ahn, Hankook Lee, and Jinwoo Shin. Self-improved retrosynthetic planning. In *International Conference on Machine Learning*, pages 5486–5495. PMLR, 2021.

[3] Siqi Hong, Hankz Hankui Zhuo, Kebing Jin, Guang Shao, and Zhanwen Zhou. Retrosynthetic planning with experience-guided monte carlo tree search. *Communications Chemistry*, 6(1):120, 2023.

[4] Shufang Xie, Rui Yan, Peng Han, Yingce Xia, Lijun Wu, Chenjuan Guo, Bin Yang, and Tao Qin. Retrograph: Retrosynthetic planning with graph search. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2120–2129, 2022.

[5] Marwin HS Segler, Mike Preuss, and Mark P Waller. Planning chemical syntheses with deep neural networks and symbolic ai. *Nature*, 555(7698):604–610, 2018.

[6] Samuel Genheden, Amol Thakkar, Veronika Chadimová, Jean-Louis Reymond, Ola Engkvist, and Esben Bjerrum. Aizynthfinder: a fast, robust and flexible open-source software for retrosynthetic planning. *Journal of cheminformatics*, 12(1):70, November 2020.

[7] Samuel Genheden, Ola Engkvist, and Esben Jannik Bjerrum. A quick policy to filter reactions based on feasibility in ai-guided retrosynthetic planning. 2020.

[8] Samuel Genheden and Esben Bjerrum. Paroutes: towards a framework for benchmarking retrosynthesis route predictions. *Digital Discovery*, 1(4):527–539, 2022.

[9] Samuel Genheden and Ola Engkvist. A quick policy to filter reactions based on feasibility in AI-guided retrosynthetic planning. 11 2020.