Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Check for updates

Cite this: *Digital Discovery*, 2024, **3**, 482

Received 8th November 2023
Accepted 26th January 2024

DOI: 10.1039/d3dd00219e

rsc.li/digitaldiscovery

Retro-BLEU: quantifying chemical plausibility of retrosynthesis routes through reaction template sequence analysis†‡

Junren Li,[a] Lei Fang *[b] and Jian-Guang Lou [b]

Computer-assisted methods have emerged as valuable tools for retrosynthesis analysis. However, quantifying the plausibility of generated retrosynthesis routes remains a challenging task. We introduce Retro-BLEU, a statistical metric adapted from the well-established BLEU score in machine translation, to evaluate the plausibility of retrosynthesis routes based on reaction template sequences analysis. We demonstrate the effectiveness of Retro-BLEU by applying it to a diverse set of retrosynthesis routes generated by state-of-the-art algorithms and compare the performance with other evaluation metrics. The results show that Retro-BLEU is capable of differentiating between plausible and implausible routes. Furthermore, we provide insights into the strengths and weaknesses of Retro-BLEU, paving the way for future developments and improvements in this field.

1 Introduction

Retrosynthesis analysis plays a crucial role in the design and discovery of new chemical compounds.[1] A retrosynthesis route usually consists of multiple chemical reactions, decomposing the target molecule into commercially available starting materials in a step-by-step manner.[2] Recently, deep learning-based approaches have substantially expedited the process of retrosynthesis planning, known as Computer-Assisted Synthesis Planning (CASP).[3,4] For example, ASKCOS,[5] an open-source platform, can easily generate hundreds of retrosynthesis routes for a given target molecule. However, it is important to note that not every generated route is guaranteed to be feasible, as the predicted reactants may not yield the expected product in actual lab scenarios.[6] Therefore, it is crucial to develop metrics to assess the validity and plausibility of these model-generated routes.

Existing metrics to evaluate the retrosynthesis routes can be broadly classified into two primary categories:

- Metrics based on intrinsic properties of generated routes, *e.g.*, route length, reactants price,[7] or the coverage of the starting materials in recorded routes.[8] While these metrics provide valuable information, they cannot capture the chemical plausibility or practicality of a given route.[2,7] For example, protection and deprotection steps are essential for obtaining the target product by preventing undesired reactions, which increases the route length.
- Metrics based on trained models, *e.g.*, reaction cost,[9] which calculates a route-level probability score by multiplying the probabilities of each reaction step. A typical planning system generally consists of a single-step retrosynthesis model[10–12] and a multi-step searching algorithm.[9,13,14] The probabilities generated by single-step models represent the model's confidence derived from the underlying training data,[15] they do not correspond to actual reaction probabilities, which are influenced by various factors such as reaction kinetics and the presence of catalysts. Moreover, the model's performance degrades when the size of the template library is increased, resulting in less reliable probabilities,[16] and the metric is also affected by the route length, because a route comprising more steps typically exhibits a lower cumulative probability.

The goal of retrosynthesis planning is to provide valid routes for synthesis design. A valid route indicates that all reactions of the route can be performed in the real-world lab scenario, instead of simply applying reaction templates to arbitrary chemical environments. Nonetheless, the metrics mentioned above cannot determine the route validity, which leaves a gap between current CASP programs and actual laboratory experiments. In order to accurately determine if a reaction can take place, a theoretical evaluation or wet-lab experiment is indispensable. Such assessments necessitate substantial computational resources (starting from first principles, which are typically challenging to compute precisely) or involve considerable labor costs. These challenges motivate us to approach the

[a] College of Chemistry and Molecular Engineering, Peking University, No. 5 Yiheyuan Road, Beijing, China

[b] Microsoft Corporation, Building 2, No. 5 Dan Ling Street, Beijing, China. E-mail: leifa@microsoft.com

† This work was done when Junren Li was an intern at Microsoft.

‡ Electronic supplementary information (ESI) available: We provide the reaction/template *n*-gram overlap analysis under different partition settings, template *n*-gram overlap analysis under different radii, and the relationship between Retro-BLEU and filtering strategies in the ESI. See DOI: **https://doi.org/10.1039/d3dd00219e**

problem from a statistical perspective, seeking statistical measures correlated with chemical plausibility, and thus enabling us to quantify the plausibility of chemical reaction routes.

In natural language processing (NLP), widely accepted evaluation metrics for tasks such as machine translation or text generation/summarization include Bilingual Evaluation Understudy (BLEU)[17] and Recall-Oriented Understudy for Gisting Evaluation (ROUGE),[18] and they focus on precision and recall when evaluating with human translation, respectively. Both BLEU and ROUGE rely on the concept of *n*-grams to compute the overlap between generated text and the reference text. *n*-grams are sequences of "*n*" consecutive words. For example, unigrams represent single words, bigrams represent two consecutive words, and so on. Drawing an analogy to NLP, retrosynthetic routes (typically represented as trees) can also be considered collections of reaction sequences, as we demonstrated in Fig. 1. Each sequence corresponds to a specific reaction pathway connecting the target product to leaf nodes, which represent a set of individual starting materials. Similar to how consecutive words in a sentence often exhibit semantic correlations, consecutive reactions in validated synthesis routes also demonstrate interrelated synthetic strategies, reflecting the underlying logic and coherence in chemical transformations.[19,20] For instance, the nitro group can be easily introduced into an aromatic ring, then reduced to an amine, followed by other substitution reactions, as a simple example of sequential reactions. Since there is no absolute best route for retrosynthesis planning, the precision of sequential reactions is more important than recall. This motivates us to modify the BLEU score for the scenario of retrosynthesis, resulting in Retro-BLEU. The key difference between the basic BLEU score and Retro-BLEU lies in the data being analyzed. While BLEU deals with text, Retro-BLEU is designed for reaction sequences, which can be obtained from retrosynthetic routes. In this context, *n*-grams represent sequences of "*n*" consecutive reactions (or consecutive reaction templates, as we will discuss later) instead of words. This adaptation allows us to apply the concept of *n*-grams from natural language processing to the domain of retrosynthetic routes, enabling a more relevant and meaningful comparison between generated routes and known synthesis routes. By calculating the precision of matching reaction *n*-grams between generated routes and known synthesis routes, Retro-BLEU offers a quantifiable approach to assess the quality of generated synthetic pathways.

2 Dataset and methods

2.1 *n*-Gram overlap analysis

We first study the *n*-gram overlap of the reaction sequences in retrosynthesis routes, which will be further used in Retro-BLEU. In NLP, the overlap is calculated over *n*-grams in the reference text to measure the semantic correlations between the text generated by a model and the reference text. While in retrosynthesis, our goal is to measure the plausibility of a synthesis route, the *n*-gram overlap is determined over *n*-grams in all known experimentally validated synthesis routes. The rationale behind this is that if a proposed route shares a significant number of *n*-grams with known and successful synthesis routes, it is more likely to be chemically plausible and experimentally viable.

We employ two datasets, the PaRoutes[21] dataset and the Retro*-190 (ref. 9) dataset, to determine if there is a noticeable relationship between the *n*-gram overlap found in model-generated routes and patent test routes.

- PaRoutes: following PaRoutes,[21] we collected 457 447 experimentally validated routes from the US Patent and Trademark Office (USPTO) dataset.[22] PaRoutes also provides two sets of 10 000 diverse, non-overlapping routes with a depth of at most 10 reactions: set-n1 and set-n5. The difference is the number of routes extracted from each patent before checking for overlapping routes: one route for set-n1 and five routes for set-n5, please refer to PaRoutes[21] for details. Due to space limitations, we mainly report the results on set-n5 because the results on set-n1 are similar. We constructed the known *n*-grams from the patent dataset, excluding those patents containing the corresponding 10 000 target instances. As a result, the remaining patents are denoted as the corresponding "known routes". This approach was taken to mimic the scenario when evaluating retrosynthesis routes for new targets, ensuring a fair and unbiased comparison. In addition, we generated 2 958 811 routes for set-n5 molecules using Monte Carlo Tree Search (MCTS)[13] and 2 799 023 routes for set-n5 molecules using Retro* (ref. 9) with AiZynthFinder,[23] *i.e.*, for each target molecule, we generated approximately 300 routes. We use the default parameter settings in AiZynthFinder and employ the top-50 predictions from the single-step model in each step.

- Retro*-190: Retro*-190 (ref. 9) is a collection of 190 challenging target molecules specifically designed to test the performance of retrosynthesis search algorithms. Retro*[9] provided the shortest route for each target by concatenating reactions from various patents until the starting materials are available in *eMolecules*§, which are considered as patent test routes. It should be noted that these routes are pseudo-routes because their corresponding reaction sequences may not be chemically logical. We employ the results of several state-of-the-art search algorithms as model-generated routes to compare the *n*-gram overlap with known routes. The 299 902 training routes from Retro* are considered as known routes to build the known *n*-grams.

On each dataset, we collected *n*-consecutive reactions (with *n* ranging from 2 to 4) from the set of corresponding known synthesis routes to construct the known reaction *n*-gram sequences. We utilized the SMILES (Simplified Molecular-Input Line-Entry System) representation for these reactions, as the canonical SMILES of each molecule is unique, allowing for efficient identity checking between tuples. For example, if a route is 6 steps long, we would take the first four reactions, the middle four reactions, and the last four reactions as three 4-grams. For each of the tested route, which includes both patent test routes and model-generated routes, we extracted

§ https://www.emolecules.com/products/building-blocks

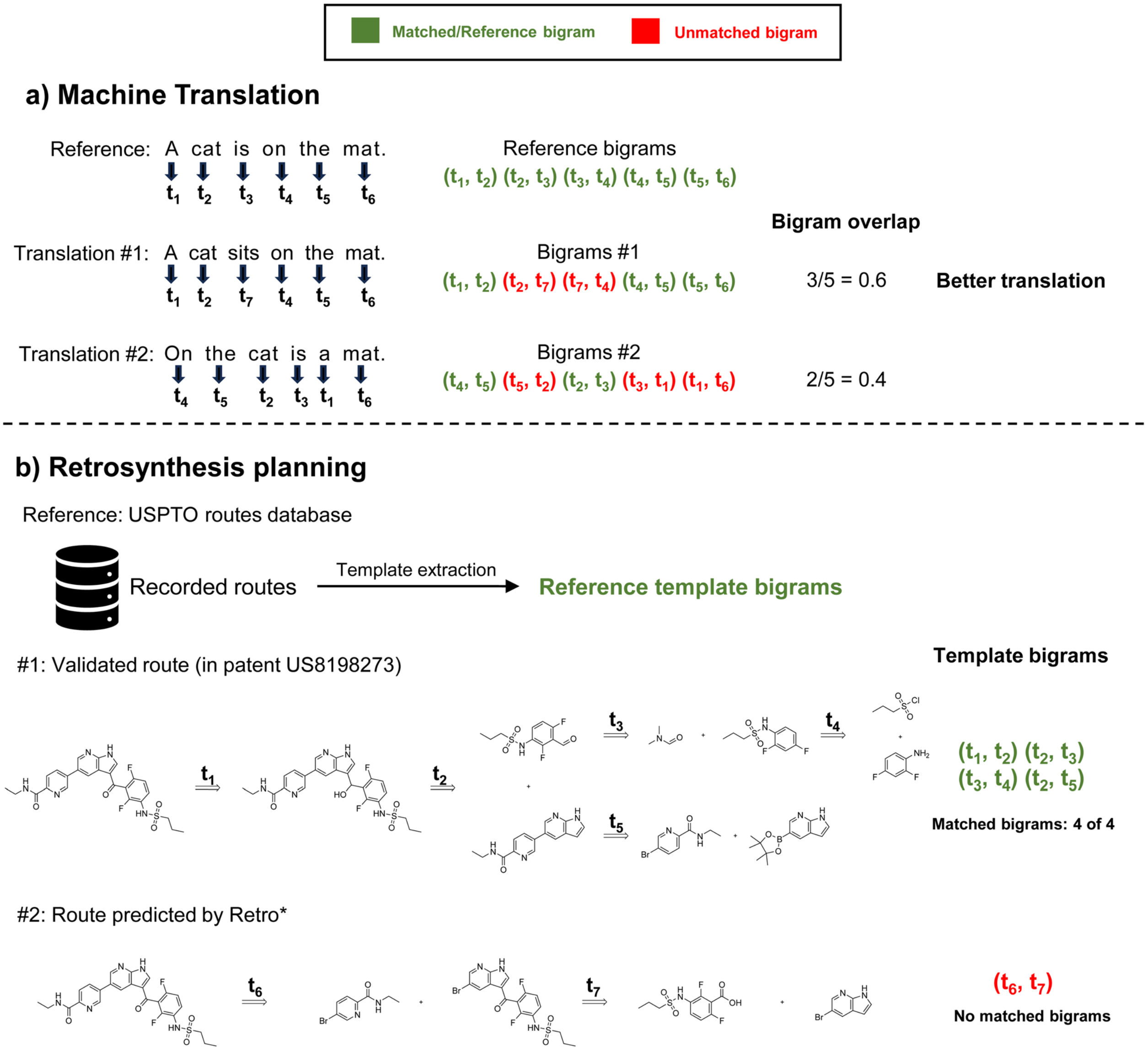
🟩 Matched/Reference bigram
🟥 Unmatched bigram
a) Machine Translation
Reference: A cat is on the mat.
t_1 t_2 t_3 t_4 t_5 t_6
Reference bigrams
(t_1, t_2) (t_2, t_3) (t_3, t_4) (t_4, t_5) (t_5, t_6)
Bigram overlap
Translation #1: A cat sits on the mat.
t_1 t_2 t_7 t_4 t_5 t_6
Bigrams #1
(t_1, t_2) (t_2, t_7) (t_7, t_4) (t_4, t_5) (t_5, t_6)
3/5 = 0.6
Better translation
Translation #2: On the cat is a mat.
t_4 t_5 t_2 t_3 t_1 t_6
Bigrams #2
(t_4, t_5) (t_5, t_2) (t_2, t_3) (t_3, t_1) (t_1, t_6)
2/5 = 0.4
b) Retrosynthesis planning
Reference: USPTO routes database
Recorded routes
Template extraction
Reference template bigrams
#1: Validated route (in patent US8198273)
Template bigrams
t_1
t_2
t_3
t_4
t_5
(t_1, t_2) (t_2, t_3)
(t_3, t_4) (t_2, t_5)
Matched bigrams: 4 of 4
#2: Route predicted by Retro*
t_6
t_7
(t_6, t_7)
No matched bigrams

Fig. 1 An comparative view of evaluation in machine translation and retrosynthesis planning using bigram overlap: (a) in machine translation, the BLEU-2 score, which can be considered as bigram overlap in this case, can be used to select high-quality translation (b) in retrosynthesis planning, template bigrams overlap can be used to select chemically plausible routes.

n-consecutive reactions and computed their overlap ratio with the known reaction *n*-grams (routes from the same patents are excluded when constructing dataset), then we averaged the ratio across all routes to obtain the overall overlap ratio. To be specific, the fraction of *n*-gram overlap is calculated as follow:

$$f_n(r) = \frac{N_{\text{known}}(r)}{N_{\text{total}}(r)} \tag{1}$$

where r denotes the route, N_{known} is the number of known *n*-grams in the route r, and N_{total} is the total number of *n*-grams in the route r.

We also calculated the coverage, which is the average ratio of routes having *n*-grams, *e.g.*, routes shorter than 3 steps do not contain any trigram reaction sequences.

As shown in Table 1, on PaRoutes set-n5, nearly half of the reaction *n*-grams are recorded/known when evaluating the patent test routes. This observation suggests that a significant portion of the reaction sequences in set-n5 patent test routes overlap with those found in known synthesis routes, indicating that chemists often rely on familiar and well-understood reaction sequences when designing new synthesis strategies. However, the overlap ratio declined to less than 10% on generated routes for both MCTS and Retro*. On Retro*-190, which is a quite challenging dataset, the overlap ratio of the pseudo-routes decreases to approximately 10% at the bigram level, because these routes contain many unobserved reactions from the training data. This decrease can be attributed to the sparsity of reaction sequences.[27] When considering reactions as individual tokens, the space formed by continuous *n*-grams is

Table 1 *n*-Gram (reaction/templates) overlap ratio for patent test routes and model-generated routes

	$n = 2$			$n = 3$			$n = 4$			
n-Grams ratio category	Reaction	Template	Coverage	Reaction	Template	Coverage	Reaction	Template	Coverage	Avg. length
PaRoutes set-n5	49.0%	70.0%	100%	47.6%	52.3%	92.9%	46.4%	48.7%	51.3%	3.84
MCTS-1[a]	7.9%	24.0%	94.3%	7.3%	5.5%	56.7%	7.4%	1.6%	24.8%	3.75
MCTS-10	3.9%	21.2%	99.0%	2.2%	2.9%	76.7%	1.0%	0.5%	41.2%	4.50
MCTS-all	1.4%	29.1%	100%	0.3%	3.4%	98.6%	0.1%	0.3%	92.6%	8.65
Retro*-1	4.5%	14.9%	94.3%	3.4%	2.1%	57.5%	2.4%	0.7%	22.4%	3.23
Retro*-10	3.1%	16.1%	99.0%	1.7%	1.6%	75.0%	1.0%	0.4%	35.2%	3.79
Retro*-all	2.3%	25.1%	100%	0.6%	2.4%	98.0%	0.1%	0.2%	84.6%	5.38
Retro*-190 (ref. 9)[b]	10.1%	42.9%	100%	4.6%	28.9%	90.5%	1.5%	21.5%	77.9%	6.67
Retro* (165)[9]	6.0%	31.2%	100%	3.4%	16.9%	88.5%	2.1%	14.9%	75.2%	6.35
Retro*+ (183)[24c]	3.6%	29.5%	100%	1.6%	13.6%	90.7%	0.9%	10.2%	79.8%	6.82
EG-MCTS (183)[25]	1.2%	13.9%	100%	0.5%	5.2%	90.1%	0.1%	3.1%	72.7%	5.69
RetroGraph (189)[26]	2.1%	20.1%	100%	0.9%	7.6%	90.5%	0.3%	4.5%	75.1%	6.40

[a] The numbers represent how many routes are used in the evaluation, *i.e.*, top-1 predicted routes, top-10 predicted routes, and all predicted routes (approximately 300 for each target). [b] The number in the parentheses denotes the solved routes among the 190 targets. [c] We use the variant of Retro*+ without value functions.

extremely sparse, because encountering unseen reactions is inevitable during the synthesis of novel molecules. Nonetheless, this does not imply that we should regard unseen reactions as invalid choices.

The sparsity of the reaction space encourages us to develop a more flexible evaluation of generated routes, emphasizing the underlying chemical transformations. In the context of chemical reactions, templates can be considered as an induction and generalization form of reactions. Therefore, we conducted a similar analysis on template sequences, using the same approach as in analyzing the overlap of reaction sequences. Atom-mapping information is a prerequisite for extracting templates. The patent routes have atom-mapping information within, for the test routes on Retro*-190, we employed the commonly used tool RXNMapper[28] to map the atom numbers. Afterwards, the reaction templates are extracted with the rxnutils[29] package and we use SMARTS (SMILES Arbitrary Target Specification) strings to demonstrate these templates.

We tested the reaction templates with radii ranging from 0 to 2. The chosen radius for a template determines the extent of the chemical environment encapsulated around the reaction center, which in turn influences the sparsity of the chemical space formed by the bigrams. A template with a radius *r* encompasses the surrounding *r* atoms, specifically, a template with a radius of 0 focuses only on the atoms undergoing change at the reaction center. For example, a radius of 0 proves to be insufficiently representative, as a single template might correspond to multiple reactions. However, selecting a large radius can lead to overly restrictive template coverage. At a radius of 2, the overlap template bigram ratio for patent routes drops to a mere 34.8%, resulting in bigrams too sparse for effective evaluation. Therefore, we set the radius to 1 when evaluating template sequences, offering a meaningful compromise between specificity and coverage.

Herein, we present the results for a radius of 1, results for other radii can be found in ESI Table 2,‡ indicating that using a radius of 1 is an optimal choice for evaluating template sequences. As shown in Table 1, the patent-extracted routes on PaRoutes set-n5 have a significant portion of known consecutive template sequences, much higher than using reaction sequences. Meanwhile, the overall template sequence overlap ratio is considerably higher than the reaction sequence. Similarly, the test routes on Retro*-190 have 42.9% of recorded template bigrams, while model-generated routes exhibit lower overlaps.

It is important to note that coverage is closely related to the average route length. When more generated routes are examined for each target, the average length increases, resulting in higher coverage. However, only the bigram coverage consistently remains near 100%. Taking the coverage into account, we propose that the bigram overlap ratio should be considered when assessing the chemical plausibility. Furthermore, it should be noted that the template bigram overlap ratio increases when the average route length increases. This might be due to randomly paired sequences as the route extends, which may contain unproductive steps, such as performing unnecessary protection before converting functional groups. This observation implies that route length should also be considered when evaluating the plausibility of generated routes.

2.2 Retro-BLEU metric

Based on *n*-gram overlap analysis, we propose Retro-BLEU as a method for evaluating the plausibility of retrosynthetic routes from a statistical perspective, although we acknowledge that the ultimate confirmation of a reaction's plausibility relies on wet lab experiments. The *n*-gram overlap with known routes between validated routes and generated routes demonstrates a noticeable difference. Considering that the overlap ratio could be affected by route length, we integrated both the route length and overlap ratio of template *n*-grams with known routes in Retro-BLEU score:

$$\mathrm{Score}_{\text{Retro-BLEU}}(r) = \exp\frac{L}{\max(L, \mathrm{len}(r))} + \exp f_n(r) \qquad (2)$$

where L is a hyperparameter, len(r) is the number of reaction steps in route r, and $f_n(r)$ is the n-gram overlap with known routes. We use the bigram overlap ratio, *i.e.*, we set n to 2. Retro-BLEU penalizes routes with lengths exceeding L. Given that the average length of known routes is 2.79, we set L to 3. In practice, we might have an estimation of the number of required reactions by averaging the length of the model-generated reaction routes or leveraging other length-prediction tools like DFRscore.[30]

We compare Retro-BLEU with four other baselines:

- The route score by Badowski *et al.*[7] This score takes into account route length and convergence. However, due to insufficient experimental data, the cost of each reaction and the yields can only be set using heuristics. We adapted the original implementation from PaRoutes in our comparisons:[21]

$$\mathrm{Score}_{\mathrm{Badowski}}(r) = \min_{x\in \mathrm{pred}(r)} \mathrm{cost}(x) \tag{3}$$

where pred(r) denotes all the child nodes for the route r, *i.e.*, the preceding reactions. The cost of a reaction x is defined as:

$$\mathrm{cost}(x) = \varepsilon(x) + \sum_{r\in \mathrm{pred}(x)} \frac{\mathrm{cost}(x)}{\mathrm{yield}(x)} \tag{4}$$

where $\varepsilon(x)$ is the fixed reaction cost of performing the reaction. Since both the reaction cost $\varepsilon(x)$ and the reaction yield (denoted as yield(x)) are unknown, the authors heuristically set their values to 1 and 80%, respectively.

- Cumulative probability: we recursively add the logarithmic probability obtained from the single-step retrosynthesis model NeuralSym[31] for each reaction in the route. Note that for reactions in patent test routes that cannot be predicted by the single-step model, we set its probability to 1×10^{-10} when calculating the cumulative probability.

$$\mathrm{Score}_{\mathrm{cum}}(r) = \sum_{x} \log p(x) \tag{5}$$

where x denotes each reaction in route r.

- Length: we use the number of reactions in the route as a metric, with shorter routes being preferable.

$$\mathrm{Score}_{\mathrm{length}}(r) = N_x(r) \tag{6}$$

where N_x denotes the number of reactions in route r.

- Bigram ratio: we rank the routes based on the bigram overlap ratio. As we discussed earlier, a higher bigram ratio suggests that the route more closely resembles known successful routes, and is therefore considered better.

For each set of routes, we compute the route score using the aforementioned baselines and Retro-BLEU score. Then, we calculate the rank of the patent-recorded route among all the tested routes for the same target, leading to our top-k metric. Since multiple routes may share the same scores (*e.g.*, the same length under the route length metric), we assess the routes in terms of both best-case and worst-case scenarios. These scenarios represent instances where the patent route is identified either first or last among routes with the same score, respectively.

3 Results and discussions

3.1 Differentiating patent-extracted routes

With Retro-BLEU, we can differentiate between plausible and implausible routes. We report the results on the PaRoutes dataset. We first merge patent test routes with model-generated routes and assess their rankings using Retro-BLEU and other metrics in terms of top-k accuracy. This evaluation method takes into account the ranking position of patent test routes among the top-k. We believe that validated routes derived from patents are chemically feasible¶ and, as such, should receive higher scores (rank higher). Although it is possible that model-generated routes can also be effective, experimentally evaluating numerous routes is impractical. Thus, we evaluate the ranking of patent test routes as their plausibility is experimentally verified.[21] In addition, we refined a subset of set-n5 such that every reaction within each patent route in these subsets could be predicted among the top-50 predictions, we named these routes "searchable routes", resulting in 6345 routes. The reaction space in these routes more closely resembles that of the generated routes, and we consider this as a fairer comparison for utilizing the cumulative probability baseline.

Fig. 2 shows the results on set-n5 for MCTS and Retro*, and the results on set-n1 can be found in ESI Fig. 1,‡ demonstrating a similar outcome to the one discussed here. In Fig. 2, the gap between the best- and worst-case scenarios is marked with diagonal lines. Retro-BLEU achieves the best overall ranking accuracy with a relatively small gap between the best- and worst-case when compared with other evaluation metrics on both MCTS and Retro* generated routes.

3.2 Bigram examples

Analyzing the extracted bigrams can help reveal the underlying correlations between bigrams and validated routes. We analyzed some common template bigrams from known routes, referred to as positive bigrams, and some common template bigrams extracted from generated routes that were not present in known routes, referred to as negative bigrams. We present three positive bigrams and three negative bigrams from the most frequently occurring bigrams when planning routes for set-n5 molecules using MCTS in Fig. 3. Please note that the template bigrams presented here follow the retrosynthesis order, and we visualized them from the original auto-extracted SMARTS strings. This means that the template of the product is positioned to the left of the reaction sequence, and it is iteratively decomposed into the templates of the reactants on the right.

The first positive bigram illustrates a common process for generating an amide, which involves initially hydrolyzing an ester and then coupling it with another amine. Considering the difficulty of amidation reactions and the low reactivity of esters, hydrolyzing esters into more reactive carboxylic acids is often necessary. Following this step, amidation can be completed with the help of condensing agents. Similarly, in the second

¶ Except for the errors during text extraction.

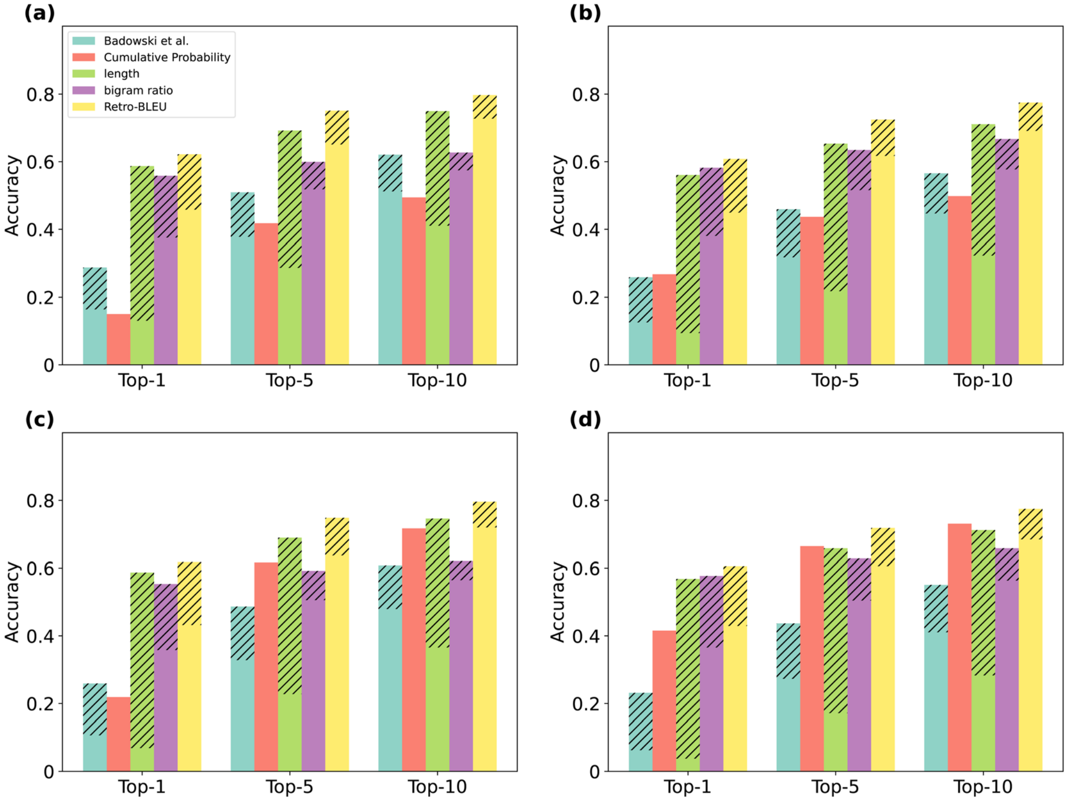

Fig. 2 Top-k accuracies for Retro-BLEU and other metrics. The top and bottom of areas with the diagonal line markings represent the best-case and worst-case scenarios, respectively. (a) MCTS algorithm applied to set-n5, (b) Retro* algorithm applied to set-n5, (c) MCTS algorithm applied to set-n5 searchable routes. (d) Retro* algorithm applied to set-n5 searchable routes.

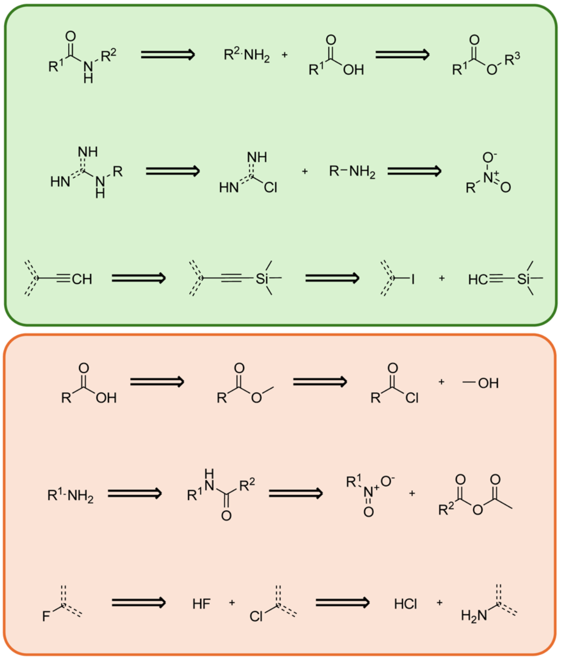

Fig. 3 Most frequent positive (highlighted in green) and negative (highlighted in red) template bigrams.

positive bigram, the product is deconstructed into a primary amine and a nitrogen-substituted heterocycle. The primary amine is derived from a nitro group through a reduction process. This bigram demonstrates an excellent strategy for linking two molecular fragments together, which is commonly employed in drug-like molecule synthesis. The third positive bigram comprises a Sonogashira coupling reaction followed by deprotection to form an exocyclic triple bond. Trimethylsilyl-based protection prevents the formation of side products from excessive coupling. The subsequent deprotection process provides an opportunity for coupling on the other side of the triple bond. These positive bigrams represent well-established reaction strategies, whereas negative bigrams often contain redundant reactions that are not practical in synthesis applications.

In the first negative bigram, the overall reaction involves the hydrolysis of acyl chloride. However, the negative template bigram uses two steps to complete the entire process: an alcoholysis and a hydrolysis on the ester intermediate. In common practice, this reaction can be simply executed by adding acyl chloride into water. When performing the initial alcoholysis step, the search algorithm is unaware that the molecule will ultimately be converted into a carboxylic acid, leading to a redundant step. The second negative bigram aims to convert the R-substituted nitro compound into a primary amine, which could be achieved by directly using reductants to reduce the nitro group. However, the template bigram incorporates extra reagents, resulting in an unnecessarily extended reaction sequence. Similarly, the third negative bigram, which involves converting fluorobenzene to the more easily accessible aminobenzene, can be accomplished in a single step using the Schiemann reaction. These negative bigrams reveal an inherent limitation in the current retrosynthesis planning approaches, in which consecutive reactions were not considered, consequently resulting in the potential generation of redundant steps. We can potentially build these negative bigrams using various data mining techniques, which can help in early stopping unnecessary searches during route finding.

3.3 Synthesis route examples

We show two cases to further verify the correlation between Retro-BLEU and chemical plausibility in Fig. 4. The first case in Fig. 4 compares a relatively long route (5 steps) from the patent database with the shortest one generated by AiZynthFinder using the MCTS algorithm. The patent route synthesized the target molecule in 5 steps, starting from simple molecules, and featured a convergent route. It incorporated several synthesis strategies such as linking two acyl amines together step-by-step using a 2,2,2-trichloroethyl chloroformate. All four template bigrams have been previously documented, resulting in a Retro-BLEU score of 4.54. The first generated route shown in Fig. 4 illustrates an unfeasible approach due to potential selectivity issues in the second step. This route attempts to combine two components using bis(trichloromethyl) carbonate in a single step. The presence of two identical reactive trichloromethyl groups in bis(trichloromethyl) carbonate may cause the molecule to couple twice, resulting in overreacted by-products. The template for this three-component reaction is quite rare and is

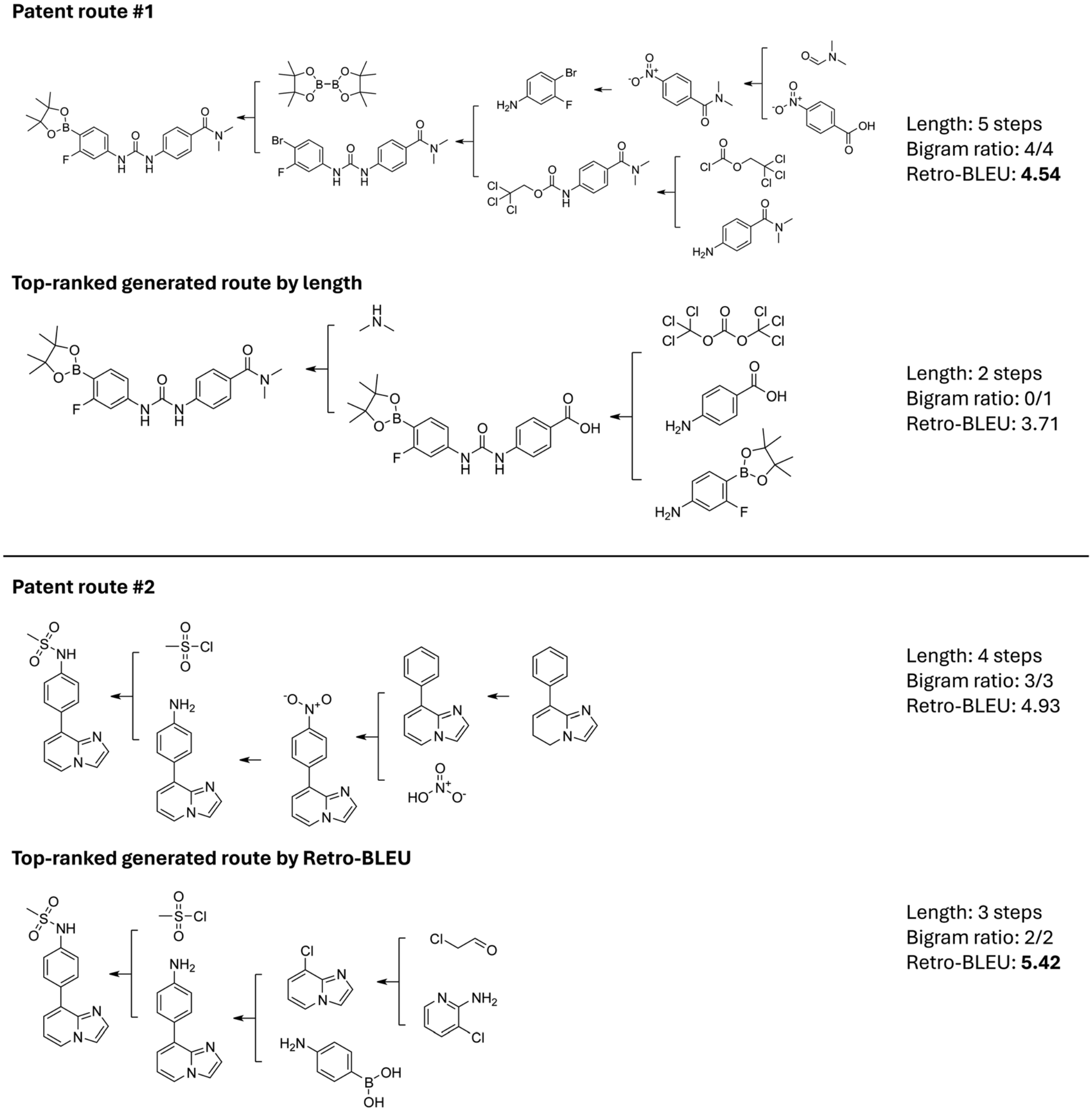

Fig. 4 Examples of using Retro-BLEU to select feasible retrosynthesis routes. Top: the recorded route in patent WO2012/58 671 and the shortest generated route. Bottom: the recorded route in patent US4596872 and the top-ranked generated route ranked by Retro-BLEU. The search algorithm employed in these examples is MCTS.

likely only applicable to specific reactants or accompanied by subsequent deprotection reactions. However, in this generated route, the next step involves the coupling of carboxylic acid and secondary amine, a reaction sequence that has not been recorded. In this case, the shortest path score is only 3.71, which is significantly lower than the corresponding patent route. While many searching algorithms aim to finding shorter routes, the length can be deceptive and leading to unhelpful routes for synthesis scientists. Although this generated route consists of only two steps, it cannot synthesize the expected molecule in an actual laboratory environment. The difference in Retro-BLEU scores demonstrates that our metric is capable of identifying potential invalid template combination patterns. This enables preliminary screening of synthetic routes and assists scientists in avoiding unfeasible approaches.

It is worth mentioning that for approximately 20% of target molecules, the patent-extracted routes are not ranked in the top-1 positions. We selected another example where the Retro-BLEU score of the patent route is lower than some of the generated routes, as shown in the second case in Fig. 4. The patent route synthesizes the target molecule within four steps, primarily modifying the substituent on the benzene ring; however, the starting material remains relatively expensive[‖]. In our comparison, we selected the generated route with the highest Retro-BLEU score (5.42), which surpasses the patent route's score of 4.84. Notably, this route was originally ranked in the 25th position by AiZynthFinder's searching process. The template bigrams in the generated route have all been recorded, and it can be considered an alternative synthesis route by first synthesizing the two aromatic systems and then coupling them together using a Suzuki–Miyaura coupling reaction. The generated route is shorter than the one previously reported in the patent, and the starting materials are also simplified. We also include 5 more randomly selected targets where the top-ranked generated route has a higher Retro-BLEU score than the patent test route in the ESI Section 5,[‡] providing a more comprehensive view of the Retro-BLEU suggested routes. This comparison indicates that model-generated routes can also serve as valuable supplements to existing routes if carefully selected based on Retro-BLEU scores. Therefore, we believe that Retro-BLEU can serve as a valuable metric to distinguish plausible routes from a vast number of model-generated routes, ultimately enhancing the efficiency and effectiveness of synthetic route selection.

4 Conclusion

Retro-BLEU, as a statistical metric, has its limitations. Although our study confirms a correlation between Retro-BLEU and chemical plausibility, it should be noted that correlation does not imply causation. Because Retro-BLEU does not explicitly encode chemical knowledge, it might face difficulties in evaluating routes involving rare or novel reactions due to their limited occurrences in the database, potentially underestimating innovative routes. Additional factors, such as production rate and reaction costs, among others, should also be considered when evaluating a synthesis route, given that sufficient data is available. The limitation can be mitigated in the future by incorporating more data in constructing the template sequence database to ensure a comprehensive and diverse representation of reactions. It should also be noted that the threshold determining the viability of a retrosynthetic route is adaptable upon the length of the generated routes, as the absolute Retro-BLEU score may vary with route length. To demonstrate this approach, when assessing synthetic routes of a specific length, we can examine the Retro-BLEU score distribution for all patent routes of the same length. Furthermore, it is crucial to regularly update the template sequence database with the latest research findings and innovative reactions. This ensures that the metric remains relevant and effective in evaluating state-of-the-art retrosynthetic routes, thereby maintaining its ability to identify and assess novel synthesis pathways in the rapidly evolving field of chemistry.

In conclusion, we introduce Retro-BLEU as a metric to evaluate and rank retrosynthetic routes generated by computer-aided synthesis planning tools. Retro-BLEU offers a statistical approach to assess chemical plausibility by analyzing reaction template sequences, and it significantly outperforms other baselines in selecting experimentally validated patent test routes. This accelerates the utilization of retrosynthesis planning tools and enables researchers to identify feasible routes more efficiently. We encourage further research for evaluating model-generated retrosynthetic routes, which will support synthetic chemistry progress and facilitate the discovery and synthesis of novel molecules, benefiting the broader scientific community.

Data availability

We primarily utilized data from PaRoutes to conduct our research. PaRoutes is an open-source benchmark that can be accessed at **https://github.com/MolecularAI/PaRoutes**. We used the 2.0 version of the benchmark dataset (which is the default setting) provided in the GitHub repository. Additionally, we employed AiZynthFinder v3.7.0 and Retro* for our retrosynthesis planning, which are other common open-source tools in this field. They can be accessed at **https://github.com/MolecularAI/aizynthfinder** and **https://github.com/binghong-ml/retro_star**, respectively. The code for Retro-BLEU is publicly available at **https://github.com/catalystforyou/Retro-BLEU**.

Author contributions

Junren Li: investigation, methodology, data curation, software, formal analysis, writing – original draft. Lei Fang: conceptualization, methodology, resources, writing – review & editing, project administration. Jian-Guang Lou: supervision.

Conflicts of interest

There are no conflicts of interest to declare.

[‖] The price was $280 per g on **https://www.biosynth.com/** accessed in July, 2023.

Acknowledgements

The authors thank Dr Kangjie Lin for helpful discussions.

Notes and references

1 E. J. Corey, *Chem. Soc. Rev.*, 1988, **17**, 111–133.

2 Y. Jiang, Y. Yu, M. Kong, Y. Mei, L. Yuan, Z. Huang, K. Kuang, Z. Wang, H. Yao, J. Zou, C. W. Coley and Y. Wei, *Engineering*, 2023, **25**, 32–50.

3 P. Schwaller, A. C. Vaucher, R. Laplaza, C. Bunne, A. Krause, C. Corminboeuf and T. Laino, *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 2022, **12**, e1604.

4 Z. Zhong, J. Song, Z. Feng, T. Liu, L. Jia, S. Yao, T. Hou and M. Song, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2024, **14**(1), e1694.

5 C. W. Coley, D. A. Thomas III, J. A. Lummiss, J. N. Jaworski, C. P. Breen, V. Schultz, T. Hart, J. S. Fishman, L. Rogers, H. Gao, *et al.*, *Science*, 2019, **365**, eaax1566.

6 E. Kim, D. Lee, Y. Kwon, M. S. Park and Y.-S. Choi, *J. Chem. Inf. Model.*, 2021, **61**, 123–133.

7 T. Badowski, K. Molga and B. A. Grzybowski, *Chem. Sci.*, 2019, **10**, 4640–4651.

8 S. Liu, Z. Tu, M. Xu, Z. Zhang, L. Lin, R. Ying, J. Tang, P. Zhao and D. Wu, *Int. Conf. Mach. Learn.*, 2023, 22028–22041.

9 B. Chen, C. Li, H. Dai and L. Song, *International Conference on Machine Learning*, 2020, pp. 1608–1616.

10 H. Dai, C. Li, C. Coley, B. Dai and L. Song, *Advances in Neural Information Processing Systems*, 2019.

11 P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas and A. A. Lee, *ACS Cent. Sci.*, 2019, **5**, 1572–1583.

12 L. Fang, J. Li, M. Zhao, L. Tan and J.-G. Lou, *Nat. Commun.*, 2023, **14**, 2446.

13 M. H. Segler, M. Preuss and M. P. Waller, *Nature*, 2018, **555**, 604–610.

14 K. Lin, Y. Xu, J. Pei and L. Lai, *Chem. Sci.*, 2020, **11**, 3355–3364.

15 J. Li, L. Fang and J.-G. Lou, *J. Cheminf.*, 2023, **15**, 58.

16 A. Thakkar, T. Kogej, J.-L. Reymond, O. Engkvist and E. J. Bjerrum, *Chem. Sci.*, 2020, **11**, 154–168.

17 K. Papineni, S. Roukos, T. Ward and W.-J. Zhu, *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

18 C.-Y. Lin, *Text summarization branches out*, 2004, pp. 74–81.

19 K. Molga, S. Szymkuć, P. Gołębiowska, O. Popik, P. Dittwald, M. Moskal, R. Roszak, J. Mlynarski and B. A. Grzybowski, *Nat., Synth.*, 2022, **1**, 49–58.

20 E. P. Gajewska, S. Szymkuć, P. Dittwald, M. Startek, O. Popik, J. Mlynarski and B. A. Grzybowski, *Chem*, 2020, **6**, 280–293.

21 S. Genheden and E. Bjerrum, *Digital Discovery*, 2022, **1**, 527–539.

22 D. M. Lowe, *PhD thesis*, University of Cambridge, 2012.

23 S. Genheden, A. Thakkar, V. Chadimová, J.-L. Reymond, O. Engkvist and E. Bjerrum, *J. Cheminf.*, 2020, **12**, 70.

24 J. Kim, S. Ahn, H. Lee and J. Shin, *Int. Conf. Mach. Learn.*, 2021, 5486–5495.

25 S. Hong, H. H. Zhuo, K. Jin, G. Shao and Z. Zhou, *Commun. Chem.*, 2023, **6**, 120.

26 S. Xie, R. Yan, P. Han, Y. Xia, L. Wu, C. Guo, B. Yang and T. Qin, *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022, pp. 2120–2129.

27 S. Stocker, G. Csányi, K. Reuter and J. T. Margraf, *Nat. Commun.*, 2020, **11**, 5505.

28 P. Schwaller, B. Hoover, J.-L. Reymond, H. Strobelt and T. Laino, *Sci. Adv.*, 2021, **7**, eabe4166.

29 C. Kannas and S. Genheden, *ChemRxiv*, 2022, preprint, DOI: **10.26434/chemrxiv-2022-wt440-v2**.

30 H. Kim, K. Lee, C. Kim, J. Lim and W. Y. Kim, *J. Chem. Inf. Model.*, 2023, DOI: **10.1021/acs.jcim.3c01134**.

31 M. H. S. Segler and M. P. Waller, *Chem.–Eur. J.*, 2017, **23**, 5966–5971.