Retrosynthesis Prediction Revisited

Hongyu Tu
UMass Amherst

Shantam Shorewala*
Pinterest

Tengfei Ma
IBM Research

Veronika Thost*
MIT-IBM Watson AI Lab

{hongyutu, sshorewala}@umass.edu
{tengfei.ma1, veronika.thost}@ibm.com

Abstract

Retrosynthesis is an important problem in chemistry and represents an interesting challenge for AI since it involves predictions over sets of complex, molecular graph structures. Recently, a wealth of models ranging from language models to graph neural networks are being proposed. However, most studies evaluate over a single dataset and split only, focus on top-1 accuracy, and provide few insight into the actual capabilities of individual models. This prevents research from moving forward since issues to be addressed by future work are not identified. In this paper, we focus on the evaluation: we show that the currently used data does not fit to test generalization, one of the main goals stated in the literature; propose new splits of the USPTO reactions modeling various scenarios; study representatives of the main types of models over this data; and finally present the, to the best of our knowledge, first evaluation and comparison of these models in the multi-step scenario. Altogether, we show that the picture is more diverse than the results over the usually used USPTO-50k data suggest.

1 Introduction

Retrosynthesis is an important problem in chemistry (Robinson, 1917; Corey & Wipke, 1969): to synthesize a complex molecule in the lab based on purchasable molecules, a chemist needs to know the sequence of reactions to run. The problem represents an interesting challenge for AI since it involves predictions over molecular graph structures as well as a search component. In every step, a given *product molecule* (also, *target*) is split into a set of *reactant molecules*, the targets for the next steps.[2] Together, the steps form a *retrosynthetic route*, a tree structure.

The large number of models proposed for retrosynthesis prediction recently reflects the interest of the ML research community, for a good overview of the state of the art, we refer to the discussions in recent works (Tu & Coley, 2021; Zhong et al., 2022). The *single-step models* used for predicting the individual steps can be categorized into three types: (I) Template-based approaches (Segler & Waller, 2017) classify a target w.r.t. a set of reaction templates (i.e., rules how a target can be split if it meets certain conditions), which are usually mined from the training data or manually encoded, and assumed to encode valid reactions. (II) Template-free approaches (Liu et al., 2017) operate on the data alone and usually consider the molecules in text-based SMILES representation and the problem as a translation task (i.e., a sequence to sequence problem). (III) Semi-template based approaches (Yan et al., 2020) apply a given atom mapping (i.e., considered as a kind of template), relating the atoms in the targets to the corresponding ones in the reactants, during training. Current approaches in this category model a two-step procedure first identifying the

*The work was done while S.S. was at UMass Amherst and V.T. at IBM Research Zurich.

[2] In practice, there are other relevant factors, such as catalysts or reaction temperature, which are however usually ignored in current model development, amongst others, because the data is not available.

3rd AI for Science Workshop at NeurIPS 2022.

reaction center (i.e., where to split), and then completing the obtained molecule parts to reactants. There are a few open-source tools for multi-step synthesis planning (Genheden et al., 2020; ask; Chen et al., 2020). They apply a template-based MLP for the single steps by default, and resort to Retro* (Chen et al., 2020) (i.e., a variant of A* search) or MCTS for the multi-step planning.

Studies about single-step models primarily compare on the USPTO-50k dataset (Schneider et al., 2016), since many works report numbers only for that data, and a single (often random) split. This dataset was extracted from the United States patent data from 1976 to 2016 (Lowe, 2017). While the latter represents one of the few large open sources for reaction data, we note that most reactions (about 94%) in the USPTO-50k test set match templates one can extract from the training data and hence do not truly fit to test generalization. Yet many of the template-free proposals consider the ability to generalize as main advantage of their approaches (Liu et al., 2017; Somnath et al., 2021). Moreover, there is a strong focus on improving top-1 accuracy, which has been critisized as being insufficient in the chemistry community (Schwaller et al., 2019). Generally, few works provide insight into the actual capabilities of individual models and components, and concrete comparisons between models (i.e., in terms of predictions) are rare, possibly because many implementations are less accessible. The latter also often ignore practical aspects, such that applying the models out of the box would be hard. For example, most language-based models operate in batch mode reading and writing from files (Zheng et al., 2019; Chen et al., 2019), and others require involved preprocessing or training procedures (Dai et al., 2019; Yan et al., 2020). Altogether, this prevents research from moving forward since issues that are critical for applicability are not identified, and research is rarely translated into applications.

Related Work. While we above describe the general trend, there are certainly exceptions. In particular, several analyses in the chemistry domain provide more detail, in terms of all aspects, data, methodology, and also in measuring effectiveness in the multi-step scenario. **Data.** Some studies also evaluate over larger USPTO subsets (Dai et al., 2019) or use different splits (Seidl et al., 2022; Chen et al., 2019). **Methodology.** Recently, (Lin et al., 2022) proposed to re-rank predictions and, in the course of this, compared different single-step models and, specifically, also in terms of their actual predictions. Further, (Chen et al., 2019; Lin et al., 2022) distinguish the predictions in terms of reaction types. There have been further various proposals for metrics beyond top-k. (Schwaller et al., 2019, 2020) propose *round-trip accuracy*, the agreement between two models predicting the reaction in the forward and, respectively, backward direction; *coverage*, quantifying for how many of the products at least one valid suggestion of the set of reactants could be found; and *diversity*, the number of diverse, valid reactants after removing the buyable ones. Regarding diversity, they further consider a metric of statistical significance. (Tetko et al., 2020) propose to report *MaxFrag*, the recognition of the largest reactant, to estimate the ability of a model to deduce the correct reaction class; note the similarity to the coverage from (Schwaller et al., 2020). Most recently, (Lin et al., 2022) applied *MRR*, observing that it reflects a trend in between the ones of top-1 and top-3 accuracy, and *area under the top-k curve*, which parallels top-k accuracy; while both are similar to the latter in trend, especially MRR offers better explainability in that it summarizes the latter for different k. **Multi-step.** Lastly, some works apply the proposed single-step model in the muti-step planning setting by either chaining single-step predictions (Coley et al., 2017) or combined with a search algorithm (Ishida et al., 2022) but, to the best of our knowledge, there are no comparisons between models. Such a comparison has come closer, with the recently proposed PARoutes framework (Genheden & Bjerrum, 2022), which offers tools for extracting synthetic routes from the USPTO patent data, based on the work of (Mo et al., 2021), and for comparing predicted routes to the extracted ones. (Genheden & Bjerrum, 2022) also propose two route datasets, n1 and n5, and present an evaluation using template-based single-step models. However, this route data was designed for evaluating model capabilities under rather controlled conditions, it does not contain rare reactions and the training data is in a certain sense artificial (exactly three reaction examples per reaction template). Therefore, we used this framework for the creation of more general datasets, and in our evaluation of the multi-step setting. There are some other route datasets: (Chen et al., 2020) also extracted routes from the USPTO data that do not contain rare reactions, and (Ishida et al., 2022) manually curated routes. In summary, there are many reasonable proposals to evaluate single-step retrosynthesis models, yet these have not been picked up by the broader community and are not commonly used today. In the words of (Schwaller et al., 2019), *the evaluation of single-step retrosynthetic models is an overlooked research topic*.

Our study focuses on the evaluation of single-step models, applies some of these ideas, demonstrates their relevance for applications, and complements the above works by focusing on important aspects that have been less in focus so far, namely, prediction diversity, generalization, and performance in the multi-step scenario. Our contributions are the following.

- We propose new splits of the USPTO reactions. Since prediction diversity is important in view of the incomplete reaction data (e.g., missing conditions which may restrict applicability) and the models not being perfect, we collected products which can be synthesized from different reactant sets and thus have *multiple solutions* in the test set **USPTO-ms**. Moreover, we create datasets **USPTO-rt** and **USPTO-rd** containing (the reactions of) retrosynthetic *routes* such that we can compare the performance in the single and multi-step scenarios by either evaluating single-step models on the reactions with all reactions' targets as input, or by running them inside a synthesis planning tool with only the roots of the routes as input. This comparison shows that there is indeed a discrepancy between the performance in the two scenarios.
- In addition, we develop the *multi-step score (mss)* for estimating the multi-step performance based on the predictions on individual reactions, independently of a route search algorithm.
- We evaluate several representatives of the main models types over various datasets and present the, to the best of our knowledge, first multi-step evaluation and comparison of these models.
- To address issue of missing comparisons, we created all our test sets such that they do not overlap with the USPTO-50k train and validation data (based on the usually used split from (Coley et al., 2017; Dai et al., 2019)). Hence, researchers can use the existing checkpoints and evaluate easily over our new test sets. For the evaluation, we created a simple, abstract wrapper class for single-step retrosynthesis models; evaluation scripts based on this wrapper; and example wrappers for the models we evaluate in this work. Furthermore, we provide an extension of AIZynthFinder (Genheden et al., 2020), such that it can be used directly with our wrapper interface. All assets and code are provided at `https://github.com/vthost/retroeval`.

The fast adoption of Retro* (Genheden et al., 2020; Kim et al., 2021) has shown that the community is open to adopt latest research outcomes if they are accessible, efficient, and effective. The goal of our work is to ease the development of single-step models towards this direction.

2 Our Setting

Models considered in this paper.

- Template-based: **NeuralSym (NPP)** (Segler & Waller, 2017), a regular MLP; the **Conditional Graph Logic Network (GLN)** (Dai et al., 2019), which models template applicability using a conditional graphical model that is parameterized using a structure-based embedding; and the recently proposed **MHNreact (MHN)** (Seidl et al., 2022), which uses a modern Hopfield network (Widrich et al., 2020) to retrieve stored templates based on the input molecule embedding. Both GLN and MHN also encode the templates and hence can exploit similarities between them.
- Template-free: **Chemformer (CF)** and **Chemformer-Large (CFL)** (Irwin et al., 2022), regular BART language models (Lewis et al., 2020) that were pre-trained on a large molecule corpus and only differ in size; and **Graph2SMILES (G2S)** (Tu & Coley, 2021), combining a graph-based transformer encoder with a sequence-based decoder.
- Semi-template-based: **GraphRetro (GR)** (Somnath et al., 2021), modeling intermediate reaction steps with GNNs.

Note that we tried to set up other models, especially from the last category, but often failed. Further, we were able to run GLN only locally and therefore present less results for this model.

Metrics. *Top-k accuracy* (*top-k*) represents the fraction of test targets for which the correct reactant set is among the top-k ranked predicted reactant sets. Note that the former is usually used in a setting where a target with multiple correct reactant sets is considered as two different test points and the model is evaluated twice w.r.t. the two solutions, respectively. We further consider the *mean reciprocal rank* (*MRR*) and *MaxFrag* accuracy at k (*mf-k*), focusing on the recognition of the largest reactant. In the scenario where we focus on predicting multiple solutions, we consider *recall@k* (*r@k*), which represents the proportion of correct reactant sets among the top-k ranked predictions averaged over all targets. Throughout the paper, we report mf and our mss at 10 (dropping "-10" due to space) since we chose 10 as cutoff value in the multi-step planning; note that it provides a single-step model performance close to maximum and search remains feasible.

Table 1: Results over our new USPTOms test set.

Model	r@1	r@3	r@5	r@10	r@50
Data	48.4	98.9	99.9	100.0	100.0
Neuralsym++	11.8	19.1	22.1	25.0	27.9
GLN	12.6	20.8	23.3	25.9	27.7
MHNreact	12.0	20.5	23.4	25.8	28.2
Chemformer	12.9	16.3	16.7	17.0	17.2
Chemformer-Large	12.8	16.2	16.7	16.8	16.9
Graph2SMILES	12.2	19.6	22.1	24.4	26.8
GraphRetro	12.5	18.6	19.9	21.0	21.0

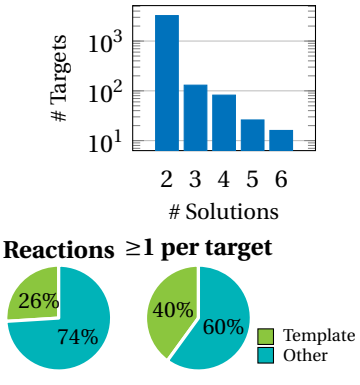

Figure 1: Details about USPTOms.

3 Revisiting USPTO-50k

In most experiments, we consider the models trained over USPTO-50k (or short, 50k). If available, we used the checkpoints from the original studies within our wrappers. Some of the *trends we observe turned out to be more general* (see also Sec. 6 and Appendix B):

- top-1: CF(L) is best, followed by (in no particular order) G2S and GR.
- MRR, top-5-50: Template-based models are best, followed by (in no particular order) G2S, GR.
- G2S performs more similar to GR than to CF(L). CF-L usually outperforms CL slightly.
- top-50: GLN/MHN basically reach maximal possible performance given the amount of test reactions matching templates (see Figure 4). In particular, they are already close to that at top-10. Note that, the template-based top-k predictions are selected from those templates that are actually applicable (i.e., their conditions are satisfied by the target) but our NeuralSym and MHN implementations pre-compute only the applicability of the first 200 templates; when considering all, MHN reaches 93.2% at top-50.
- MRR is between top-1 and top-5, and mf-10 is more similar to top-50 than to top-10.

4 Prediction Diversity: The Dataset USPTO-ms

USPTO-ms. For evaluating how well the models capture diversity, we collected reactions from the USPTO patent data (Lowe, 2017) where we have *multiple solutions* (i.e., multiple possible reactant sets) into the new test set USPTO-ms. Details about the creation process can be found in Appendix C. The data contains 3,501 different targets and 7,438 different reactions. Figure 1 shows further statistics. Note that we dropped reactions with more than six solutions since these were extremely rare. We see that only about a quarter of the reactions match templates from the USPTO-50k train/valid data, hence it is not only the diversity but also the novelty of the reactions which represents a challenge for the models. It is possible that the latter is the greater challenge.

Results, Table 1, Fig. 2. At first glance, the picture is very different from the one over USPTO-50k. In particular, we observe no considerable differences between the models at r@1, with the exception of NPP. Note that we include the scores on the test data to show the maximal r@k possible. Further, good scores at $k = 1..3$ are especially hard to obtain because most targets have two or three solutions. At higher k's it gets again easier for the models since they allow for mistakes. Here, the template-based models and G2S perform especially well. Since the table shows the average across targets, Figure 2, additionally shows how many solutions the models actually find per target among the top 5. We see that the r@k performance can be largely attributed to single solutions. In particular, CF(L) seems to lack diversity.

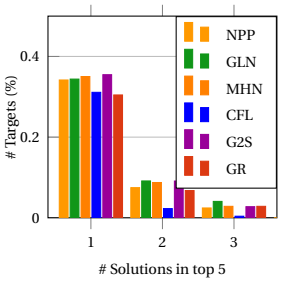

Figure 2: Solution counts.

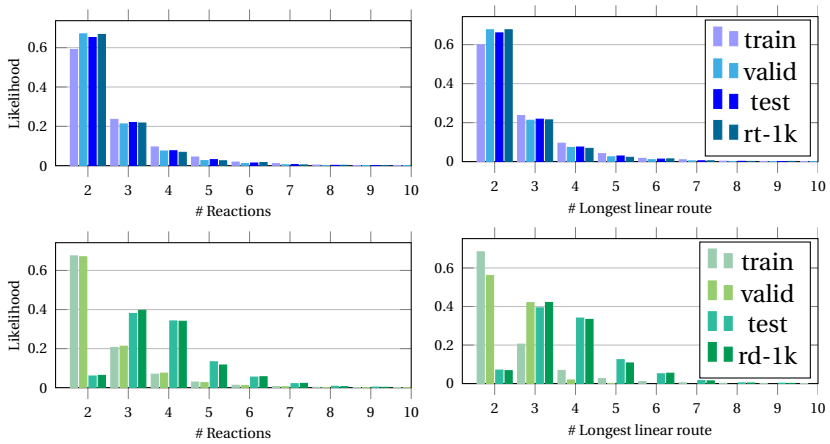

Figure 3: Overview of our rt (top) and rd datasets; we cut all horizontal values at 10 to ease comparison.

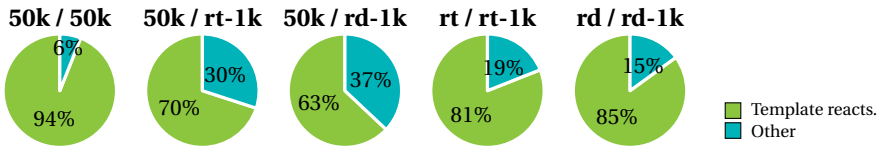

Figure 4: We trained over different datasets T and evaluated over various test sets E (abbrev. T / E) and show the shares of reactions in E that match one of our templates for the train/validation sets of T.

5 On Evaluating Single & Multi-Step Prediction

Prediction in Route Context, the Datasets rt & rd, Figure 3, Table 8. We followed (Mo et al., 2021; Genheden & Bjerrum, 2022) for extracting retrosynthesis routes from the USPTO data. However, we did not drop rare reactions and created two custom splits. In particular, we created a *time* split, the dataset **USPTO-rt** (or short, **rt**), and one where we extracted sets of very *diverse* routes as test and validation data and kept all remaining routes as train data, **USPTO-rd** (or short, **rd**). Further details about the creation are given in Appendix D. The figures show that the rt data contains train/valid/test distributions that are very similar, while there is great diversity in rd. The latter dataset contains more data overall because the test and validation routes are longer and we kept the data from all other patents for training. For rt we dropped all remaining data from patents at the test time points. We will use these full rt and rd datasets particularly to compare how the model performance changes when trained over data beyond USPTO-50k.

Test Subsets: rt-1k & rd-1k, Figure 7, Table 8. Since the models turned out to be more resource and especially time-intensive than the regularly used MLPs in the multi-step scenario, we extracted subsets of the original rt and rd test sets which will be in the focus of our study. Each contains (the reactions of) 1k routes. We additionally ensured that all reactions are unique, that we have maximally one route per patent, and that there is no overlap with the USPTO-50k train/valid data. The overview of our test sets compared to the ones from (Genheden & Bjerrum, 2022) in the figure shows that the statistics are similar. For comparison, we also evaluated over subsets of n1 and n5, but did not observe considerable differences; hence the latter lie primarily in the nature of the training data (i.e., the n1/n5 train data is more similar to the test data).

The Multi-Step Score (mss). Our multi-step score complements existing metrics by considering the route context. It is straightforward to calculate and interpret in that it captures the multi-step performance under the assumption that the model is combined with a perfect search algorithm. Formally, let m_k denote a single-step model which, given a target molecule, predicts a set of at most k sets of reactant molecules. For a given tree structure representing a retrosynthetic route, with nodes representing reactions, let S denote the set of all those reactions and S^i denote all reactions at level i (e.g., S^0 represents the singleton set containing the first reaction step at the

root node), such that $S = \bigcup_i S^i$. Then we define $\text{mss}_k(m, S^0, S^1, \ldots) := \frac{|T^\infty|}{|S|}$ with

$$T^0 := \emptyset$$

$$T^1 := \begin{cases} R, & \text{if } S^0 = \{\, R\text{»}P \,\},\ R \in m_k(P) \\ \emptyset, & \text{otherwise} \end{cases}$$

$$T^{i+1} := T^i \cup \begin{cases} \bigcup_{R\text{»}P \in S^i} R, & \text{if } \forall\ R\text{»}P \in S^i : R \in m_k(P) \text{ and } T^{i-1} \neq T^i \\ \emptyset, & \text{otherwise} \end{cases}$$

Table 2: Results over our 1k test sets.

	rt-1k						rd-1k					
Model	MRR	top-1	top-5	top-10	mf	mss	MRR	top-1	top-5	top-10	mf	mss
NPP	43.5	33.2	57.3	63.5	70.1	47.1	39.7	30.7	51.6	56.7	62.8	31.8
GLN	46.4	36.5	59.3	65.0	70.1	48.8	40.8	31.1	53.2	59.0	64.4	34.4
MHN	45.4	35.4	58.6	64.2	69.9	47.4	41.3	31.9	53.5	58.4	63.5	33.6
CF	41.6	38.3	45.7	46.1	52.0	27.6	36.8	33.3	41.0	41.3	47.8	18.3
CFL	41.8	38.8	45.7	45.8	51.5	28.1	38.0	34.9	41.8	42.1	48.2	18.2
G2S	42.2	35.2	51.0	54.5	62.0	35.1	39.5	33.1	47.6	50.8	57.4	25.7
GR	40.3	34.1	48.6	50.7	57.6	33.9	35.2	29.0	43.6	45.6	51.7	22.9

6 Evaluation in Route Context

Single-Step Results on 1k Test Sets, Tables 2, Figure 5. Our evaluation over all test sets generally confirms the trends outlined in Section 3 in the context of USPTO50k, including the additional results on n1 and n5 in Appendix F. Still, there are notable differences. The numbers are much lower, showing that the data from (Genheden & Bjerrum, 2022) as well as ours is more challenging. More interestingly, at top-1, the template-based model MHN is basically en par with G2S and GR. The Chemformers are clearly winning but considerably worse at higher k's. The results on all datasets are in similar ranges with the ones for rd-1k being lowest. In Figure 5, we can see that the performance can be attributed primarily to the targets matching templates seen during training. There are only few signals, although there are some, that the semi/non-template-based models are capable to generalize, especially G2S and GR. Our mss score differs considerably in magnitude. Interestingly, its trend is very similar to top-k for k > 1, although it is interpreted differently. Nevertheless note that it better reflects the route context, in that the numbers are lowest on the more complex routes (rd and n5) while, for example, top-5 on n5-1k is higher than on n1-1k; and the difference to mss on the routes in rt is larger than, for instance, the differences at MRR.

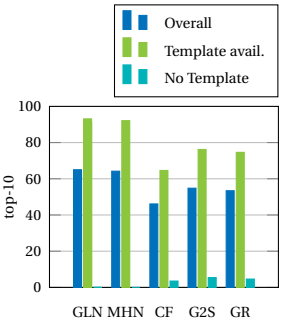

Figure 5: Splitting the rt-1k results into (non)-template-matching parts reveals that all models struggle on the new reactions.

Comparison of Predictions, Figure 6. We conclude the analysis of the single-step models trained on USPTO-50k by comparing the actual predictions, w.r.t. a top-5 threshold, in terms of the pairwise agreement between models on solutions, averaged over all six datasets, and the exact opposite, the percentage of solutions predicted by a single model only, w.r.t. the total number of those solutions per dataset. The former shows that there seems to be a correlation within a class of models. While the absolute results show greater similarity between G2S and GR (i.e., instead of G2S and CF(L)), here it is slightly larger between G2S and CFL, also between the template-based models. Note that (Lin et al., 2022) provide a similar but more fine-grained analysis comparing the actual ranks of predictions from different models and show that these may differ. The right picture shows that the usually best models, CFL and GLN w.r.t. top-1 and MRR/top-5, respectively, are interestingly indeed often the ones which predict solutions that are not predicted by other models.

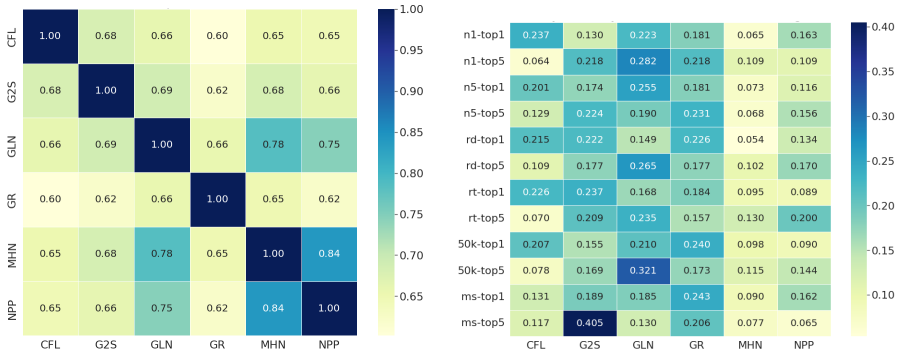

Figure 6: Comparison on actual solutions among the top-5 predictions, in terms of pairwise agreement between models (left) and solutions predicted by a single model only, in terms of relative percentage of those.

Scaling Up the Train Data, Tables 11, 12. On the 1k test sets we can compare to the models trained over USPTO-50k (see Table 2). The impact of retraining turns out to be much higher for rd-1k than for rt-1k; for example, we have a top-1 increase of about 10%, vs. 3-5% on rt-1k. This might be explained by the larger amount of training data in rd-1k and, in particular, by the greater similarity between train and test data. Recall that we have strict temporal split with rt. One very likely explanation, and consequence of the similarity, is the increase in test reactions matching templates seen during training. Figure 4, shows that the share for rd is now larger than the one for rt. Apparently, this impacts the performance of the template-free CF nearly as much as the template-based models. Finally, we note that there is no gap anymore between the latter; in fact, for larger k's, which we consider in the multi-step setting, NPP seems to provide better performance. The results on the full test sets are very similar and shown in Table 12.

The Multi-Step Scenario, Table 3. Our results give insights into the models' out-of-the-box usability. We show the top-5 accuracy w.r.t. finding the exact test route. Since this criterion is rather strong, we also report the maximal leaves overlap with this route within the top-k predicted routes and consider a subset of 100 routes where all steps match templates seen during training. Note that, interestingly, there is basically no difference between the top-5 and top-10 scores, so that we only show the former here. At first glance, we see that the performances are very different from the corresponding single-step experiments, and our selection of datasets reveals further differences. Over the manually picked routes where all reactions match templates seen during training, the template-based models show much better performance than all other models. The results for the other datasets are extremely low reflecting the struggle of the models with distributions beyond the train data and more complex routes (esp. rd). One main, additional problem we identified are the probabilities returned by the models, the search algorithms use those for cost estimates. In particular the template-free models based on common NLP frameworks return no normalized or regularized probabilities, assigning high ones to only to a few predictions, and hence prevent the search from exploring alternatives to their top suggestions altogether. This is problematic given that their top-1 performance is not perfect. We experimented with normalization and obtained best results (the ones shown) *for all models* by applying an additional softmax on the top-10 predictions given to AIZynthFinder. Note that (Tu & Coley, 2021) mention that additional engineering efforts are likely needed for G2S in the multi-step scenario, but the impact of such adaptations on both single and multi-step performance remains unclear. The results for G2S further reveal a shortcoming of the top-k evaluation in PARoutes: the routes are sorted including route length as one criterion but, in this way, very unlikely predictions may be taken as top ones, especially with the template-free models. We experimented with atom mapping or a forward-direction model as suggested in (Schwaller et al., 2020) to validate the outcomes of the template-free models, and hence filter out unlikely predictions, but this causes considerable overhead. In terms of metrics, we observe that the trends shown here are captured by most metrics with the exception of top-1 accuracy, and at different magnitudes. Our mss score is similar, but captures a different intent and some subtle difference (see previous subsection), in particular the large gap between the routes in rt and rd. Overall, NPP performs surprisingly well and is only outperformed by GLN; we could not find a specific reason for why it outperforms MHN but hypothesize that it is due to better regularization capability. Lastly, in comparison to the

Table 3: Model comparison in the multi-step scenario.

	rt-1k			rd-1k			rt-tpl-100		
	top-5	mlo-1	mlo-5	top-5	mlo-1	mlo-5	top-5	mlo-1	mlo-5
NPP	33.1	54.3	59.1	3.1	34.0	39.6	69.0	85.1	86.5
GLN	31.9	57.8	57.8	9.3	43.6	43.6	70.0	88.8	88.8
MHN	32.6	53.3	58.4	1.7	31.1	39.1	69.0	82.1	86.0
CF	15.8	42.7	44.4	4.0	35.4	37.8	33.0	60.9	62.6
CFL	16.0	43.5	45.0	3.9	37.3	39.4	34.0	62.1	64.2
G2S	12.7	28.5	49.3	0.9	22.4	38.8	29.0	38.5	68.1
GR	21.8	43.9	48.1	5.9	32.1	36.4	44.0	69.4	74.6

experiments with the template-based models in (Genheden & Bjerrum, 2022), our results are much lower - as expected -, but also show much greater variation between the datasets. This, again, seems to hint at a considerable impact of rare reactions, critical in the complex rd routes.

Discussion.

- Generally, the *identification of data aspects influencing predictions is critical for correctly interpreting the results*. We show that the templates from the train data reveal interesting insights.
- The variety of models does not seem to lead to a large variety of predictions; in our experiments, the models predicted few truly unique solutions (i.e., usually less than 1% of the test set).
- The main available *search algorithms for the multi-step setting expect scaled and regularized probabilities*, which are not given by all models. Hence it is not clear how those can be applied and how they perform. Similarly, we need *efficient techniques to estimate the validity of the predictions* of template-free models.
- *Currently, the template-based models are competitive* with other models and, in the multi-step setting, the most simple and easy to use MLP still seems to provide one of the best solutions since it is both efficient and effective. In our study, only GLN performs better.
- There is certainly major value in studying and comparing methods in a more theoretical context in the *single-step scenario* only, as it is usually done today. Our work supports such studies by showing the variation and commonalities in metrics beyond top-k accuracy. Moreover, we propose the mss score, as a first attempt to capture other aspects relevant in route context.
- The main open problem in terms of template-based models is the *improvement in ranking*. While the latter is in focus of current research, this improvement should not happen at the expense of top-k for higher k's given that multi-step prediction is the ultimate goal. Here, MRR offers a good overall estimate, and the template-based models represent strong baselines.
- *Generalization beyond reactions similar to the ones seen during training* represents the other major challenge and seems to be solved only partly to date. This leads to particularly low performance in the multi-step scenario, where a single new reaction leads to an overall failure.

Our work is intended to point out critical open issues rather than as a study of the SOTA. Considering recently proposed add ons (Fortunato et al., 2020; Sun et al., 2021; Lin et al., 2022), the actual SOTA is likely slightly better. We cannot fully generalize from our results because we focused on representative models and data only, and the research prototypes we probed have not been optimized for the multi-step setting. Note that template-free models have been successfully integrated in tools with careful engineering (Schwaller et al., 2020), and advanced search algorithms can overcome shortcomings of single-step models; e.g., (Chen et al., 2020; Kim et al., 2021) obtain close to maximal performance with NeuralSym combined with an optimized cost estimate.

7 Conclusions

We provide new datasets for retrosynthesis prediction which allow for evaluating prediction diversity, effectiveness in the multi-step context, and generalization, aspects that are usually not addressed when evaluating over USPTO-50k. We further provide evaluation infrastructure. Our evaluation of various single-step models suggests that the aspects our data addresses are critical for estimating multi-step performance. Altogether, our work shows that we need to evaluate more broadly to be able to address the challenges of retrosynthesis prediction in future research.

Acknowledgments

This work started within a course project at UMass Amherst. We thank Sahasra Iyer for initial discussions, and the tutors, especially Jay-Yoon Lee, and Andrew McCallum for all their support. We also thank Amol Thakkar for sharing his experience.

References

Askcos. *https://askcos.mit.edu/*.

Benson Chen, Tianxiao Shen, Tommi S. Jaakkola, and Regina Barzilay. Learning to make generalizable and diverse predictions for retrosynthesis. *CoRR*, abs/1910.09688, 2019. URL `http://arxiv.org/abs/1910.09688`.

Binghong Chen, Chengtao Li, Hanjun Dai, and Le Song. Retro*: Learning retrosynthetic planning with neural guided a* search. In *The 37th International Conference on Machine Learning (ICML 2020)*, 2020.

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.

Elias James 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.

Hanjun Dai, Chengtao Li, Connor Coley, Bo Dai, and Le Song. Retrosynthesis prediction with conditional graph logic network. *Advances in Neural Information Processing Systems*, 32, 2019.

Michael E Fortunato, Connor W Coley, Brian C Barnes, and Klavs F Jensen. Data augmentation and pretraining for template-based retrosynthetic prediction in computer-aided synthesis planning. *Journal of chemical information and modeling*, 60(7):3398–3407, 2020.

Samuel Genheden and Esben Bjerrum. Paroutes: a framework for benchmarking retrosynthesis route predictions. *ChemRxiv*, 2022. doi: 10.26434/chemrxiv-2022-wk8c3.

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):1–9, 2020.

Ross Irwin, Spyridon Dimitriadis, Jiazhen He, and Esben Jannik Bjerrum. Chemformer: a pretrained transformer for computational chemistry. *Machine Learning: Science and Technology*, 3 (1):015022, 2022.

Shoichi Ishida, Kei Terayama, Ryosuke Kojima, Kiyosei Takasu, and Yasushi Okuno. Ai-driven synthetic route design incorporated with retrosynthesis knowledge. *Journal of chemical information and modeling*, 62(6):1357–1367, 2022.

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

Alpha A. Lee, Qingyi Yang, Vishnu Sresht, Peter Bolgar, Xinjun Hou, Jacquelyn L. Klug-McLeod, and Christopher R. Butler. Molecular transformer unifies reaction prediction and retrosynthesis across pharma chemical space. *Chem. Commun.*, 55:12152–12155, 2019. doi: 10.1039/C9CC05122H. URL `http://dx.doi.org/10.1039/C9CC05122H`.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, 2020.

Min Htoo Lin, Zhengkai Tu, and Connor W Coley. Improving the performance of models for one-step retrosynthesis through re-ranking. *Journal of cheminformatics*, 14(1):1–13, 2022.

Bowen Liu, Bharath Ramsundar, Prasad Kawthekar, Jade Shi, Joseph Gomes, Quang Luu Nguyen, Stephen Ho, Jack Sloane, Paul Wender, and Vijay Pande. Retrosynthetic reaction prediction using neural sequence-to-sequence models. *ACS Central Science*, 3(10):1103–1113, 2017. doi: 10.1021/acscentsci.7b00303. URL https://doi.org/10.1021/acscentsci.7b00303. PMID: 29104927.

Daniel Lowe. Chemical reactions from US patents (1976-Sep2016). 6 2017. doi: 10.6084/m9.figshare.5104873.v1. URL https://figshare.com/articles/dataset/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873.

Yiming Mo, Yanfei Guan, Pritha Verma, Jiang Guo, Mike E Fortunato, Zhaohong Lu, Connor W Coley, and Klavs F Jensen. Evaluating and clustering retrosynthesis pathways with learned strategy. *Chemical science*, 12(4):1469–1478, 2021.

Robert Robinson. Lxiii.—a synthesis of tropinone. *Journal of the Chemical Society, Transactions*, 111:762–768, 1917.

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.

Philippe Schwaller, Vishnu H Nair, Riccardo Petraglia, and Teodoro Laino. Evaluation metrics for single-step retrosynthetic models. In *Second Workshop on Machine Learning and the Physical Sciences*. NeurIPS Vancouver, Canada, 2019.

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.

Marwin HS Segler and Mark P Waller. Neural-symbolic machine learning for retrosynthesis and reaction prediction. *Chemistry–A European Journal*, 23(25):5966–5971, 2017.

Philipp Seidl, Philipp Renz, Natalia Dyubankova, Paulo Neves, Jonas Verhoeven, Jorg K Wegner, Marwin Segler, Sepp Hochreiter, and Gunter Klambauer. Improving few-and zero-shot reaction template prediction using modern hopfield networks. *Journal of Chemical Information and Modeling*, 2022.

Vignesh Ram Somnath, Charlotte Bunne, Connor Coley, Andreas Krause, and Regina Barzilay. Learning graph models for retrosynthesis prediction. *Advances in Neural Information Processing Systems*, 34, 2021.

Ruoxi Sun, Hanjun Dai, Li Li, Steven Kearnes, and Bo Dai. Towards understanding retrosynthesis by energy-based models. *Advances in Neural Information Processing Systems*, 34:10186–10194, 2021.

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 (1):1–11, 2020.

Amol Thakkar, Thierry Kogej, Jean-Louis Reymond, Ola Engkvist, and Esben Jannik Bjerrum. Datasets and their influence on the development of computer assisted synthesis planning tools in the pharmaceutical domain. *Chemical science*, 11(1):154–168, 2020.

Zhengkai Tu and Connor W Coley. Permutation invariant graph-to-sequence model for template-free retrosynthesis and reaction prediction. *arXiv preprint arXiv:2110.09681*, 2021.

Michael Widrich, Bernhard Schäfl, Milena Pavlović, Hubert Ramsauer, Lukas Gruber, Markus Holzleitner, Johannes Brandstetter, Geir Kjetil Sandve, Victor Greiff, Sepp Hochreiter, et al. Modern hopfield networks and attention for immune repertoire classification. *Advances in Neural Information Processing Systems*, 33:18832–18845, 2020.

Chaochao Yan, Qianggang Ding, Peilin Zhao, Shuangjia Zheng, Jinyu Yang, Yang Yu, and Junzhou Huang. Retroxpert: Decompose retrosynthesis prediction like a chemist. *Advances in Neural Information Processing Systems*, 33:11248–11258, 2020.

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, 2019.

Zipeng Zhong, Jie Song, Zunlei Feng, Tiantao Liu, Lingxiang Jia, Shaolun Yao, Min Wu, Tingjun Hou, and Mingli Song. Root-aligned smiles: A tight representation for chemical reaction prediction. *Chem. Sci.*, 2022.

A Model Configuration and Experiment Notes

Single Step Models. For all models, we used the configurations suggested by the authors in the original works (see Section 2). The details can also be found in our repository.

NeuralSym(++). We ran hyperparameter tuning since the original paper (Segler & Waller, 2017) used different data; the final configuration(s) are detailed in the repository. We made some interesting observations in these experiments: (1) We also experimented with the more complex Highway network version described in (Segler & Waller, 2017), but the MLP provided better performance. (2) We needed to add batch normalization to reach a performance comparable to the one reported in (Dai et al., 2019); this design choice was later confirmed in communication with chemists who have experience with template-based models and noted the importance of regularization, in particular, for the muti-step experiments. (3) We also experimented with optimization, using the validation loss, top-1, top-5, and top-50 as criteria, respectively; interestingly, all these scenarios lead to basically the same hyperparameters (i.e., there was only some slight variation in the fingerprint dimension, 1028 vs. 2048, and layer number, 2 vs. 3). It is left open if this is due to the simplicity of the model or also holds for more complex ones. Lastly, note that for the experiments over the larger rt and rd data, we disregarded templates for which there is only one instance in the training data. This was shown to be beneficial in (Seidl et al., 2022), and we observed this as well.

Single Step Models - Influence of Beam Size. We observed some variation in performance with changed beam sizes. However, since we could not identify a specific best value for the models - and even not for individual models - we mostly resorted to the settings suggested by the authors; we made an exception with G2S setting n_best=topk*3 since we experimented with larger topk than (Tu & Coley, 2021), which used a fix value. However, we found that since we also evaluated top-50, while some of the works considered maximally top-10. We note that this also explains that there is no performance increase for GR in some cases from top-10 to top-50 (e.g., Table 5), but we obtained overall worse results for larger beam sizes and therefore used the ones originally proposed, keeping in mind that top-10 performance is likely more relevant in practice. In particular, note that our multi-step experiments focus on maximally 10 predictions, hence this setting is in line with the ones from the original studies.

Multi-Step Experiments. We used the default settings suggested in AIZynthFinder (see `https://molecularai.github.io/aizynthfinder/configuration.html`, accessed 08/23/2022) with the following three exceptions. We used Retro* as search algorithm since we obtained extremely bad results with MCTS, probably because it uses the probabilities returned by the models in a way that does not fit their current, mostly unregularized scale. The suggested maximal branching (cutoff_number) of 50 resulted in infeasible runtimes with most non-template-based models, hence we chose 10. We note that this was also due to our filtering of invalid molecule predictions. Since the numbers of the latter were extremely high with, for example, for CF, obtaining 50 valid reactant sets takes extremely long. Lastly, we increased the maximal number of reactions considered (max_transforms) to 10, which matches the maximal number in our data. These adaptations show the need for considering practical aspects in model design and evaluation.

Note that in our evaluation, for mss, we consider a form of micro average, considering the predictions for all routes in the test set at once, and compute roughly $\frac{\sum_T |T^\infty|}{\sum_S |S|}$.

B About USPTO-50k and USPTO-full

Datasets. The two most commonly used datasets are USPTO-50k (Schneider et al., 2016) and USPTO-full (Dai et al., 2019), both subsets of the US patent data (Lowe, 2017). As the names suggest, the latter, containing about 1M reactions, is much larger than the former; it is also used in much less works. It has been observed recently, after our experimentation, that there are a few non-sensical data points in USPTO-50k (Lin et al., 2022); however, the split we used is still useful in that it eases comparison to approaches we did not include in our study since it is one of ones most commonly used. We observed overlaps between the train and test data in USPTO-full and therefore did not experiment over this dataset.

Table 4: Results over USPTO-50k and USPTO-full reported in the literature, grouped by nature of approach (template/non-template/semi-template-based) and date of appearance. Note that (Seidl et al., 2022; Irwin et al., 2022) used random splits different from the one used in the other works.

	USPTO50k				USPTO-full		
Model	top-1	top-5	top-10	top-50	top-1	top-10	top-50
NPP	44.4	72.4	78.9	83.1	35.8	60.8	-
GLN	52.5	<u>75.6</u>	<u>83.7</u>	92.4	39.9	63.7	-
MHN	51.8	**81.2**	**88.1**	94.0	45.5	71.9	77.1
CF	53.6	61.1	61.7	-	-	-	-
CFL	**54.3**	62.3	63.0	-	-	-	-
G2S	52.9	70.0	72.9	-	45.7	63.4	-
GR	<u>53.7</u>	72.2	75.5	-	-	-	-

Table 5: Results we obtained over the USPTO-50K test set from (Coley et al., 2017; Dai et al., 2019).

Model	MRR	top-1	top-5	top-10	mf	top-50
Neuralsym++	61.3	48.5	78.0	84.7	88.3	88.7
GLN	65.0	52.4	81.2	87.9	90.2	92.4
MHNreact	63.6	50.8	79.9	86.6	89.2	91.5
Chemformer	61.0	56.4	66.5	66.9	71.2	67.0
Chemformer-Large	61.0	57.1	65.8	66.0	70.0	66.1
Graph2SMILES	60.4	52.2	70.7	74.9	80.0	78.7
GraphRetro	58.8	51.0	69.5	72.6	76.4	72.6

Results, Tables 4, 5. The table gives an overview of what is known about the SOTA. Our results over USPTO-50k (over a common split) are similar, see Table 5. Our additional analysis, presented in this paper, completes the picture in terms of scalability and other aspects, such as generalization.

Since not all models were originally evaluated over the same split (i.e., NPP, MHN, and CF(L)), we trained those. We re-evaluated all since this experiment served also to verify our wrapper implementations.

MHN vs. NeuralSym, Table 6. Not surprisingly, the more complex MHN outperforms NeuralSym. However, the comparison between these models is not entirely fair since the settings are quite different. While NeuralSym uses rather standard features, MHN applies a 30k-dimensional vector of different fingerprints and additionally uses the validation data during training. In order to obtain a better estimate of the actual impact of MHN's main feature, the template encoding, we therefore consider two other models, **NeuralSym+** (NeuralSym + MHN's initial features) and **NeuralSym++ (NPP)** (NeuralSym+ trained including the validation data). And we see that the gap indeed shrinks.

Table 6: Results we obtained over the USPTO-50K test set from (Coley et al., 2017; Dai et al., 2019).

Model	MRR	top-1	top-5	top-10	mf
NeuralSym	58.7	45.7	76.0	82.3	85.9
NeuralSym+	59.9	47.2	76.4	83.2	87.1
Neuralsym++	61.3	48.5	78.0	84.7	88.3

C Details about USPTO-ms

We used the raw data from (Lowe, 2017) and basically followed their data cleaning steps. In a nutshell, we then collected all products for which there are different reactions - after dropping reagents - , and dropped those reactions that are contained in the train/valid data of USPTO-50k (w.r.t. the spit of (Dai et al., 2019)). The exact steps are documented in our repository.

Table 7: Overview of the USPTO-based test sets which are subsets from the ones proposed by (Genheden & Bjerrum, 2022), compared to ours.

	n1-5k	n5-5k	n1-1k	n5-1k	rt-1k	rd-1k
# Routes	5,000	5,000	1,000	1,000	1,000	1,000
# Reactions	14,334	18,308	2,934	3,770	2,531	3,802
# Patents	5,000	4,482	1,000	976	1,000	1,000

D Details about rt & rd

Datasets, Table 8, Figures 7, 8 We basically followed the procedure proposed by (Mo et al., 2021; Genheden & Bjerrum, 2022) for extracting retrosynthesis routes from the USPTO data. The latter was obtained from (Thakkar et al., 2020), who provide a thoroughly cleaned dataset containing a reaction template with each reaction. We note that, in this way, a filtering was done already (i.e., all reactions where the template extraction failed), yet we considered the template availability as valuable enough, especially for later analysis. However, we did not drop rare reactions (e.g., whose template appears less often in the data) and adapted the last step, in which we created two custom splits. Our *time* split is based on the year data coming with the reactions. For the split containing especially *diverse* routes as test and validation data we proceeded as suggested in (Genheden & Bjerrum, 2022).

E Details about n1 & n5

Datasets, Table 7, Figures 7, 9. In order to see how general our data and results are, and to compare to the results from (Genheden & Bjerrum, 2022), we also experimented on their n1 and n5 datasets. Specifically, we extracted random test sets, of 1k and 5k routes for both n1 and n5, and evaluated our models trained over USPTO-50k over those. Note that we created these datasets in the same way as our rt and rd datasets, collecting all reactions contained in the considered routes. Table 7 shows the numbers of reactions, and Figure 9 shows the shares of reactions that match templates from the USPTO-50k training data. In this paper, we report the single-step results only for n1 and n5. We did not observe greater differences to the results over our datasets in the multi-step scenario and therefore did not complete the experiments, also due to the enormous resource requirements.

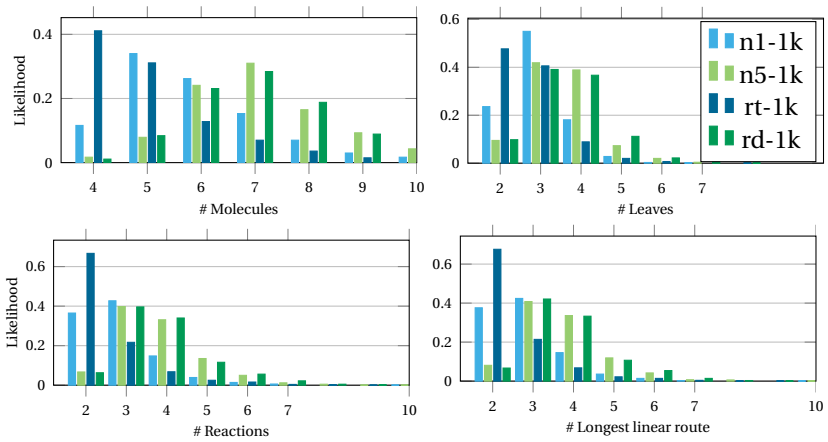

Figure 7: Overview of our test sets compared to subsets of n1 and n5 from (Genheden & Bjerrum, 2022).

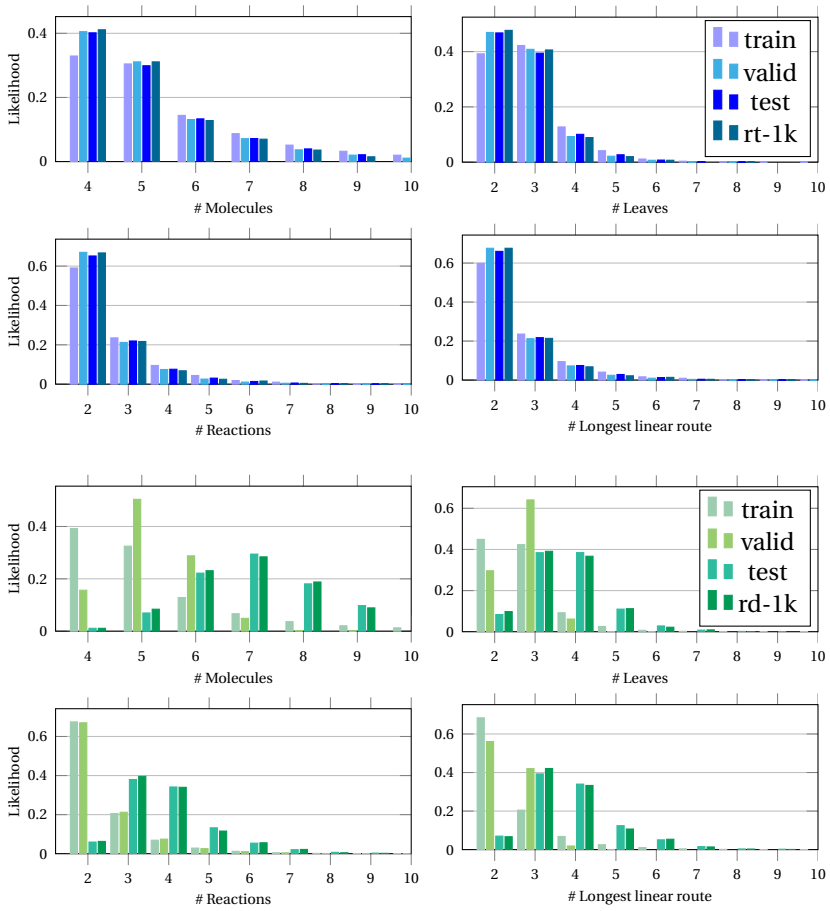

Figure 8: Overview of our rt (top) and rd datasets; we cut all horizontal values at 10 to ease comparison.

	rt			rd		
	train	valid	test	train	valid	test
# Routes	87,673	8,000	8,000	102,369	8,000	5,000
# Reactions	238,979	20,140	20,525	259,310	19,765	19,271
# Patents	18,379	8,000	8,000	20,925	8,000	5,000

Test set	# Reactions
rt-1k	2,531
rd-1k	3,802

Table 8: Overview of our USPTO-based datasets.

F Additional Results

Results, Tables 9, 10, 11, 12, Figure 10. We extracted random subsets of n1 and n5 of two sizes, 1k and 5k routes, to see how the results over the 1k sets generalize. The tables show that there are no considerable differences for all models. Figure 10 shows that all models struggle on the new reactions; in particular, there are no great differences between top-1 and top-10, the latter are only slightly larger. Tables 11 and 12 are discussed in the main paper. For GR, we obtained very disappointing numbers. For verification purposes, we therefore also ran the GR version trained over USPTO-50k and this returned much better results; a possible reason might be that the vocabularies constructed based on our larger training sets are too large and confuse the model (while the vocabulary size for USPTO-50k is 174, we have 289 for rt and 307 for rd), so it might be possible to obtain better results by dropping less common structures. Also note that we set a train time limit of 24 hours, as it was done in (Irwin et al., 2022). In particular NPP and CF might benefit from increasing that.

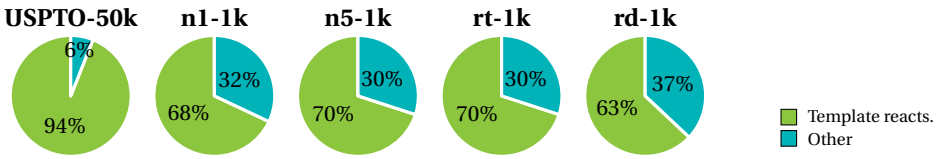

Figure 9: Share of reactions in the test sets that match templates from the USPTO-50k train/valid data.

Table 9: Results over the 1k test sets we extracted from n1 and n5.

	n1-1k						n5-1k					
Model	MRR	top-1	top-5	top-10	mf	mss	MRR	top-1	top-5	top-10	mf	mss
NPP	43.7	34.1	56.2	61.7	68.0	41.2	45.9	35.9	59.1	64.5	70.3	39.6
MHN	45.2	35.2	58.5	63.2	68.6	42.6	47.8	37.9	60.4	65.7	70.6	40.1
CF	39.7	35.8	44.4	44.8	51.3	24.2	42.7	38.8	47.6	48.1	53.7	22.9
CFL	40.5	37.3	44.6	44.9	51.0	24.0	43.8	40.5	47.9	48.2	53.5	22.4
G2S	41.3	33.9	50.5	54.3	61.7	33.9	44.8	37.7	53.5	56.9	63.4	29.8
GR	38.4	32.0	47.2	49.0	55.0	29.5	41.4	34.9	50.2	52.1	58.3	27.2

G Further Discussion

- We show that the templates from the train data reveal interesting insights and also observed a slight correlation between performance on reactions matching training templates and the number of template instances in the training data *for all models* (15-25% Pearson correlation, depending on the model). We also checked for a correlation between molecule complexity and prediction performance but did not observe any in our experiments. Reaction complexity might be a factor worth to study in future work.
- One problem with the non-template-based models is that they often produce invalid molecular structures or reactions. While the former is technically no problem because those can be easily identified with tools such as RDKit, such a filtering step may yield large runtime overhead. Moreover, there is few efficient technology for identifying invalid reactions. In order to get an idea of the validity of the predicted reactions, we briefly experimented with forward reaction prediction models as suggested in (Schwaller et al., 2020) and atom mapping tools for post processing in the multi-step experiments, but this was extremely time consuming as well. Given this trade-off, it is likely that non-template-based models will only be competitive if they either optimize for validity (amongst others) or include efficient post processing steps, to provide more efficient models; or if they show explicit benefits over template-based models, justifying the increase in runtime.
- While the trends in the paper tend to show that the template-based models are most effective to date, we emphasize that they are limited by nature and will never allow for generalization. Although the current models do not seem to solve this task satisfactory yet, we see some generalization capability. It will be interesting to see, how much generalization is possible with state-of-the-art ML technology once this aspect is considered explicitly in

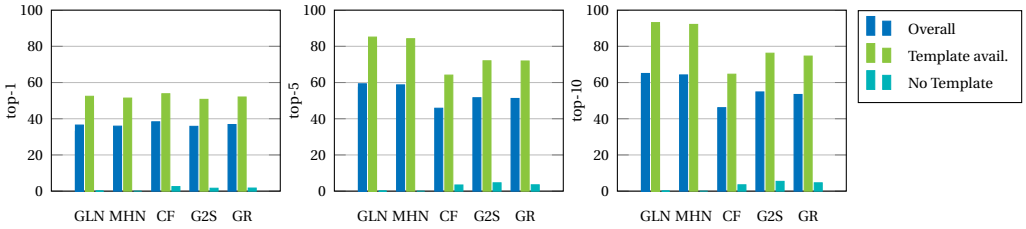

Figure 10: Splitting the rt-1k results into (non)-template-matching parts reveals that all models struggle on the new reactions; the results on the other test sets are similar.

Table 10: Results over the 5k test sets we extracted from n1 and n5.

	n1-5k						n5-5k					
Model	MRR	top-1	top-5	top-10	mf	mss	MRR	top-1	top-5	top-10	mf	mss
NPP	43.2	33.3	56.3	61.9	68.4	42.2	44.5	34.6	57.5	63.0	69.2	39.5
MHN	44.8	34.5	58.2	63.4	69.0	43.8	46.1	35.9	59.1	64.5	69.7	41.0
CF	40.5	36.9	45.1	45.5	51.8	25.0	41.6	38.0	46.1	46.5	52.4	22.1
CFL	40.9	37.6	45.1	45.3	51.4	24.8	42.0	38.8	46.0	46.2	52.0	21.9
G2S	41.5	34.4	50.2	54.0	61.6	33.3	42.8	35.7	51.7	55.1	61.9	29.4
GR	39.1	32.7	47.9	50.0	56.6	30.6	40.0	33.7	48.8	50.8	57.0	27.5

Table 11: Results over our 1k test sets when trained over the larger rt and rd datasets (see Table 8).

	rt-1k						rd-1k					
Model	MRR	top-1	top-5	top-10	mf	mss	MRR	top-1	top-5	top-10	mf	mss
NPP	49.6	37.4	66.4	73.7	78.7	57.3	53.1	41.3	68.3	76.1	80.6	51.3
MHN	50.6	38.6	66.7	73.4	77.0	55.9	52.9	40.8	69.1	75.7	79.2	50.1
CF	46.2	41.5	52.3	53.0	58.6	33.1	49.5	44.4	56.0	56.7	62.5	28.2

Table 12: Results for our new datasets.

	rt						rd					
Model	MRR	top-1	top-5	top-10	mf	mss	MRR	top-1	top-5	top-10	mf	mss
NPP	50.4	37.9	66.7	74.9	79.5	58.6	53.1	40.7	69.2	76.9	81.6	52.2
MHN	51.0	39.2	66.7	73.2	77.1	55.7	53.8	41.7	69.9	76.7	80.2	52.2
CF	47.5	42.4	53.9	54.7	60.6	34.6	50.2	45.0	56.8	57.6	63.4	28.9

future studies. We also note that there has been related work focusing on the generalization capability of the transformer across chemical space showing promising results (Lee et al., 2019); our study complements this type of investigation in pointing out reaction templates as efficient means to quantify generalization, even for non-template-based models.

H Limitations

Our study focuses on more general aspects of evaluating retrosynthesis prediction and, in the course of this, provides an exemplary evaluation. As it is in the nature of experimental studies, there are various other possible and interesting settings which can and sometimes definitely should be explored in future work. Below, we point out aspects which should be kept in mind when interpreting our results.

- It has been noted in the past that the USPTO data is lacking certain kinds of reactions and is biased towards certain kinds of reaction (Schneider et al., 2016), hence it may not provide sufficient samples to train ML models useful for practice. While this is likely the case, our study shows that the state of the art still struggles on the available data, in particular, in terms of generalization. Hence, the USPTO data seems to provide enough challenges for ML for the current moment.
- We considered several template-based models but only two different template-free ones, and GraphRetro, while there are many others for each model class. It will be interesting to see if our observations regarding these representatives are similar for other models in future studies.
- For most experiments, we used the models trained over USPTO-50k and the (hyper)parameter settings suggested in the original studies, since benchmarking the SOTA was out of the scope of this study. We experimented with some parameters, such as

beam sizes, and obtained slight variations in the results; however, for the parameters we considered, these variations did not lead to considerable changes in the overall trends.

- In the multi-step experiments we fixed the settings using mostly the default settings from AIZynthFinder and Retro* as search algorithm. We also resorted to creating the stocks representing the buyable molecules as suggested in (Genheden & Bjerrum, 2022). Certainly, variation in all these settings would lead to changes in the results. Nevertheless, our evaluation often shows some considerable and more general differences between the models, which will likely generalize to some extent.