Do Chemformers dream of organic matter? Evaluating a transformer model for multi-step retrosynthesis

Annie M. Westerlund[a*], Siva Manohar Koki[a,b], Supriya Kancharla[a,b], Alessandro Tibo[a], Lakshidaa Saigiridharan[a], Rocío Mercado[b], Samuel Genheden[a]

[a] Department of Molecular AI, Discovery Sciences, R&D, AstraZeneca, Gothenburg, Sweden

[b] Department of Computer Science and Engineering, Chalmers University of Technology, 412 96 Gothenburg, Sweden

Abstract

Synthesis planning of new pharmaceutical compounds is a well-known bottleneck in modern drug design. Template-free methods, such as transformers, have recently been proposed as an alternative to template-based methods for single-step retrosynthetic predictions. Here, we trained and evaluated a transformer model, called Chemformer, for retrosynthesis predictions within drug discovery. The proprietary dataset used for training comprised ~18M reactions from literature, patents, and electronic lab notebooks. Chemformer was evaluated for the purpose of both single-step and multi-step retrosynthesis. We found that the single-step performance of Chemformer was especially good on reaction classes common in drug discovery, with most reaction classes showing a top-10 round-trip accuracy above 0.97. Moreover, Chemformer reached a higher round-trip accuracy compared to a template-based model. By analyzing multi-step retrosynthesis experiments, we observed that Chemformer found synthetic routes leading to commercial starting materials for 95% of the target compounds, an increase by more than 20% compared to the template-based model. In addition to this, we discovered that Chemformer suggested novel disconnections corresponding to reaction templates which are not included in the template-based model. The conclusions drawn from this work allow for designing a synthesis planning tool where

template-based and template-free models work in harmony to optimize retrosynthetic recommendations.

Introduction

Synthesis planning of new pharmaceutical compounds is a well-known bottleneck in modern drug design [1]. The *retrosynthesis* approach was developed already in 1969 [2] and is now considered standard practice. The goal of retrosynthesis is to find disconnections that break down a target molecule into smaller building blocks, referred to as *precursors* or *reactants*. This step is repeated until all precursors are readily available for purchase. Typically, retrosynthesis is divided into single-step retrosynthesis which only deals with predicting precursors and multi-step retrosynthesis which describes the iterative application. Recent developments have resulted in numerous open-source software tools which help chemists with the various aspects of synthesis planning, including retrosynthesis [3]. We recently developed AiZynthFinder [4], an open-source software which is used daily by chemists. It performs multi-step retrosynthesis by coupling a single-step neural network with a Monte Carlo Tree Search (MCTS) [5]. The single-step neural network model used in the AiZynthFinder platform is trained to recommend reaction templates given a molecular fingerprint of the target molecule (a *product* molecule) as input. Such a model is called a *template-based* model.

Template-based models are dependent on the data processing needed to extract and curate reaction templates, including atom-mapping, reaction center identification, and subsequent filtering based on e.g., template abundance [6]. Moreover, because the model uses pre-specified templates as classes, it is not able to recommend novel disconnections [3], [7], [8]. In contrast, template-free models, such as Chemformer [9], do not require templates and can be trained directly on the Simplified-Molecular-Input-Line-Entry-System (SMILES) [10]

representation of the reactions. These models are therefore not restricted by specific reaction templates and can in theory extrapolate beyond the training set [8].

Chemformer is a sequence-to-sequence transformer model based on the BART language model [11]. It was pre-trained using SMILES to represent the “language” of chemical structures [9]. Research investigating the use of similar transformer models for single-step retrosynthesis prediction [6], [9], [12]–[21], as well as forward prediction [8], [9], [12], [17], [20], [22]–[24], and reagent prediction [25] have demonstrated a superior performance of these models compared to more classic approaches such as template-based models or graph neural networks [19]. Since the first implementations of a retrosynthesis transformer model [6], [19], several enhancements and developments have been presented, including diversity-enhanced transformer [13], triple-transformer validation loop [12], and disconnection-prompted transformer [15]. Although earlier studies show great promise for retrosynthesis transformer models, there are still questions to address before introducing a transformer model like Chemformer into the AiZynthFinder production platform. First, most studies consider models trained on United States Patent and Trademark Office (USPTO) data. While USPTO is a good initial benchmarking dataset, it does not span the chemical space relevant for drug discovery [26]. Second, a recent study showed that the top-N accuracy of single-step predictions does not necessarily correlate with how well a model performs in a multi-step setting [21]. Third, although the overall performance of template-based and template-free models has been examined, a study which thoroughly investigates the generated chemistry and individual reaction classes is still missing. All these issues hint that a complete evaluation of Chemformer for both single-step and multi-step retrosynthesis in the scope of drug discovery is needed before integrating Chemformer in the AiZynthFinder platform.

Here, we fine-tune the pre-trained Chemformer [9] on a large set of proprietary data from AstraZeneca [26] for the task of retrosynthesis prediction. By carrying out thorough data analysis, we dissect the predictive performance and behavior of the model. To investigate

performance relating to different reaction classes common in drug discovery, we first evaluate the model on the test set split of the proprietary dataset. We then compare its single-step performance with the AiZynthFinder template-based model on a shared test set from PaRoutes-USPTO [27]. Finally, we apply multi-step retrosynthesis to 5,000 AstraZeneca target compounds and analyze the proposed routes and reactions from the two models. The full set of analyses yields a deeper understanding of the strengths and flaws of Chemformer and the more commonly used template-based model. The conclusions drawn from this work allow for designing a synthesis planning tool where template-based and template-free models work in harmony to optimize retrosynthetic recommendations.

Methods

Reaction and compound datasets

The Chemformer was trained on a proprietary dataset from AstraZeneca consisting of 18.7 million public and proprietary reactions. This is the same dataset used for training the template-based model in AiZynthFinder [28]. The dataset was randomly divided into training, validation, and test sets (percentages - 98:1:1). To compare Chemformer to the AiZynthFinder template-based model on the single-step retrosynthesis task, we used an external test set from PaRoutes [27]. The test set is a subset (1,025 reactions) which shares no overlap with the proprietary training set. Because the PaRoutes dataset is on the same format as the proprietary dataset, there was no additional processing step needed. To evaluate the models on multi-step retrosynthesis, we carried out experiments on 5,000 AstraZeneca target compounds, referred to collectively as the *AZ design* dataset. This dataset represents modern drug design goals and typical target molecules for the AiZynthFinder application.

Model fine-tuning

We fine-tuned the pre-trained Chemformer model [9] for backward (retrosynthesis) prediction, and forward (product) prediction. The forward Chemformer model was trained

with reactants as input and was used for the evaluation of retrosynthesis predictions. A forward model usually takes both reagents and reactants as input since it yields a more accurate model [20]. We therefore validated our forward model with one trained using both reactants and reagents as input, Figure S1. The reactants-only model achieved almost as high top-N accuracy as the forward model trained with reactants as well as reagents. We therefore trust the model trained without reagents for round-trip prediction, which aims to check the agreement between a backwards prediction on a target molecule, and a forward prediction on the precursors to the target molecule. Hereafter, the backward Chemformer will be referred to as *Chemformer*, while the forward model trained without reagents will be referred to as *forward Chemformer*.

The fine-tuning was done for 20 epochs with the Adam optimizer [29], using a learning rate of 0.001, batch size of 32, and augmentation probability of 0.5. We refer to the original Chemformer paper for more information about model architecture and training procedure [9]. Model inference was done with a GPU-optimized beam search, using a beam size of 10 and batch size of 128. To obtain statistics, we evaluated each batch individually and aggregated evaluation metrics by taking the average across each batch. PyTorch 1.8.0 [30] and PyTorch lightning 1.2.3 [31] were used for model setup and training, while RDKit 2020.09.1 [32] was used for manipulating SMILES strings.

Single-step model evaluation

We evaluated the Chemformer model using exact matching accuracy and round-trip accuracy [6]. In exact matching, the backward model-predicted reactants were directly compared to the ground-truth reactants in the dataset. Such a comparison naturally does not consider the one-to-many nature of retrosynthesis prediction; one product can be broken down in many ways. Round-trip accuracy, on the other hand, takes this into account. Round-trip accuracy was computed by first carrying out forward Chemformer inference using beam search [33] with beam size 1 (corresponding to greedy search) and using the predicted reactants as input. If the forward-predicted product was the same as the backward model's

input product, the prediction was marked as correct. For the multi-step retrosynthesis analysis, we used a beam size of 10 in the round-trip forward Chemformer inference. Recent work has shown that round-trip accuracy may agree more with chemists' judgements of correct reaction suggestions than exact matching accuracy [20]. However, because the round-trip accuracy depends on the accuracy of the forward model, and does not entail exact accuracy, these two metrics can be considered complementary. We therefore used both metrics to evaluate the Chemformer performance, in particular accuracy and precision on individual reaction classes. For the precision analysis, the reaction classes of the top-1 predictions were determined using NextMove's NameRxn software [34] and the obtained classes were used to group the data. We specifically evaluated reaction classes typically used in drug design [35]. These include N-acylation to amide, reductive amination, Suzuki coupling, protections, deprotections, N-acylation to urea, N-substitution with alkyl-X, N-sulfonylation, Buchwald-Hartwig and Ullman couplings, N-arylation with Ar-X, and Sonogashira reaction. In addition to this, we also analyzed rearrangements, ring-forming reactions, and Ugi reactions, which are reaction classes known to be challenging for retrosynthesis models. The top-1 predictions were furthermore evaluated by computing the Tanimoto similarity to the ground truth using Morgan fingerprints with radius 2.

Multi-step retrosynthesis experiments

We performed multi-step retrosynthesis experiments with the AiZynthFinder software [4]. To allow a more efficient utilization of the computationally expensive Chemformer, we extended the MCTS algorithm with three features. First, we implemented immediate instantiation of child nodes in the tree search upon expansion of the parent node. For template-based models it is advantageous to delay this process due to the expensive operation of template application, but for a template-free model, such as Chemformer, there is no efficiency to be gained by delaying child node instantiation. Second, whenever a child node is expanded, i.e., when the product SMILES is given as input to the Chemformer model, we also supply the SMILES of sibling nodes. This exploits the efficient batch prediction of Chemformer and

yields a pre-cursor prediction for more than one product at a time. Third, to ensure that each product is given to the Chemformer only once, we cache the Chemformer predictions during the search.

The multi-step retrosynthesis was performed using 100 iterations and a maximum depth of 7. The maximum search time was set to 120 s and 300 s for the template-based and Chemformer models, respectively. For the Chemformer predictions, we used a beam size of 10, whereas for the template-based model we added the top-50 predictions to the search tree. We used an internal AstraZeneca stock collection as stop-criteria in the search. Moreover, a quick filter policy was used to remove unfeasible reactions from the search tree.

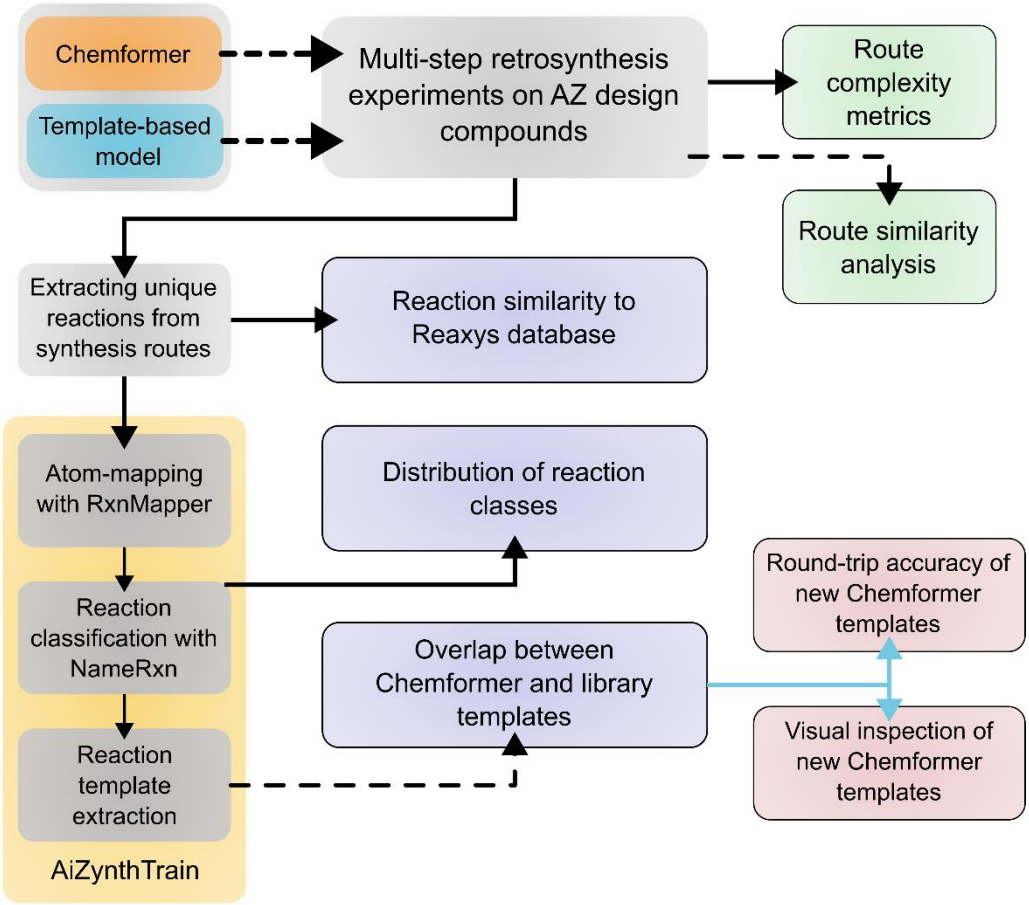

Figure 1 *Illustration of the analysis workflow used for multi-step model evaluation. Green boxes represent analysis carried out the obtained synthesis routes, violet boxes represent analyses carried out on extracted reactions, and red boxes correspond to analyses on the new Chemformer templates identified here. The steps carried out with the AiZynthTrain pipeline are highlighted by a yellow box.*

Multi-step model evaluation

Figure 1 illustrates the analysis workflow used to carry out multi-step model evaluation. Following the multi-step retrosynthesis experiments with AiZynthFinder, we first extracted basic route-complexity metrics, including for example number of solved routes and average depth. To evaluate similarity of generated routes from Chemformer and the template-based model, respectively, we computed 1) the fraction of routes in common across the top-5 ranked routes, and 2) cluster homogeneity. Cluster homogeneity was calculated by first combining all routes for one target molecule from the template-based and Chemformer models, and computing the approximate distance between all pairs of routes using the efficient LSTM proxy for tree-edit distance calculator [36], [37]. The distances obtained were then used to cluster the routes with HDBSCAN using 5 nearest neighbors [38]. Next, we labeled each route according to which the model it was obtained with and computed the cluster homogeneity as the average largest-model-fraction across all clusters. The cluster homogeneity is a score between 0.5 and 1, where 0.5 corresponds to an equal number of routes from both models in every cluster (routes were similar) and 1 corresponds to the routes of the two models being completely separated in different clusters (routes were different).

The ability of Chemformer to extrapolate beyond the training data during multi-step retrosynthesis was evaluated by performing atom-mapping with RxnMapper [39], reaction classification with NameRxn, and reaction template extraction on the predicted reactions. These steps were done using the AiZynthTrain pipeline tool [28]. To identify novel templates among the extracted Chemformer templates, we first quantified the overlap between the Chemformer templates and three sets of templates: 1) the set of templates which had been suggested by the template-based model, 2) the set of templates available in the template-based model (modeling templates), and 3) the full set of templates extracted from the training data prior to filtering (library templates). Reaction similarities of the extracted reactions compared to the Reaxys database were then computed. For each query reaction,

the database was first filtered based on the reactive functions: only reactions with the same reactive functions as the query reaction were kept. The reaction similarities of the query reaction and the selected database subset were calculated using Tanimoto similarity of reaction fingerprints. Each reaction fingerprint was calculated by subtracting reactant molecular fingerprints from the product molecular fingerprint [40]. Here, each molecular fingerprint was constructed by concatenating the Morgan fingerprint (radius 3 and length 512) together with the RDKit Daylight fingerprint (length 512).

Results and discussion

Evaluating single-step retrosynthesis performance

Accuracy and precision of common reaction classes in drug discovery. The performance of a model for the purpose of multi-step retrosynthesis is in part dependent on its performance on single-step predictions. To get a full picture of the Chemformer performance relating to different reaction classes important in drug discovery, we computed top-N exact matching and round-trip accuracy for specific subsets of the test set, Figure 2. As can be seen, the exact matching accuracy of the overall test set was modest (Top-1: 0.34, Top-10: 0.61 - Table S1). However, the round-trip accuracy was higher (Top-1: 0.72, Top-10: 0.92 - Table S2), indicating that the overall predictions were valid from a “chemical” point-of-view.

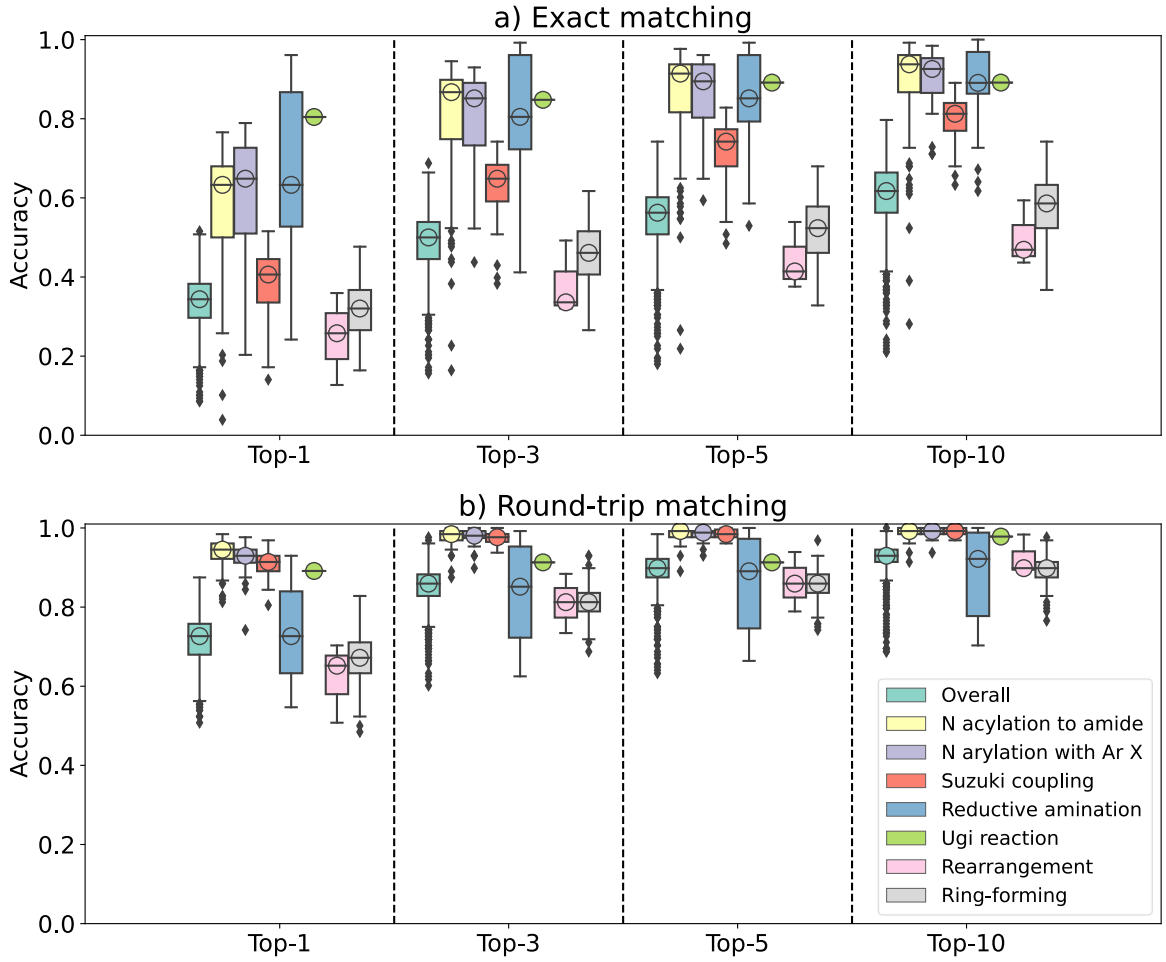

Figure 2 *Top-N a) exact matching and b) round-trip matching accuracies of selected reaction classes, evaluated on the test set split of the proprietary data. The distributions were obtained over inference batches.*

Surprisingly, we observed an especially high exact matching accuracy for Ugi reactions (Figure 2a; Table S1 – Top-1: 0.80, Top-10: 0.98). This four-component reaction has relatively few data points (4,508 data points: ~0.024% of the entire dataset) but nevertheless the Chemformer model performs well. A contributing factor to this is that the reaction space of Ugi reactions is rather limited. In contrast to the situation with Ugi reactions, Suzuki cross-couplings displayed quite low exact matching accuracy (Top-1: 0.38, Top-10: 0.79 - Table S1). However, the Suzuki cross-couplings instead had among the highest round-trip accuracies, Figure 2b (Top-1: 0.91, Top-10: 0.99 - Table S2), highlighting the weaknesses of only using exact matching accuracy. For Suzuki reactions there are several possible

reactants with different boron groups. Although the different possible reactants would be equally interesting from a retrosynthesis perspective, they are not the “ground truth” and are therefore penalized by the exact matching protocol. We further noticed a large spread in accuracy within the reductive amination class for both exact and round-trip matching. In general, however, the round-trip accuracy showed that Chemformer indeed performed well on the most common reaction classes in drug discovery [35], Figure 2b and Table S2. Apart from the ring-forming and reductive amination reaction classes, the top-10 round-trip accuracy of the various reaction classes was higher than 0.9; if we also disregard rearrangements, the round-trip accuracy is higher than 0.97, Table S2.

A model can in theory have a high accuracy but low precision of a specific reaction class. Such a model would consistently make specific predictions, even when those predictions are not applicable. Therefore, we determined reaction classes with NextMove’s *NameRxn* tool using the input product and predicted reactants, and thereafter calculated the precision (top-N exact matching and round-trip accuracy) of subsets defined by the identified reaction classes, Figure S2. The results agreed well with the accuracies reported in Figure 2. Specifically, the high exact matching accuracy of Ugi reactions was reflected in a high precision of predicted Ugi reactions (Top-1: 0.76), indicating that once an Ugi reaction was predicted, it was usually correct. As expected, predicted Suzuki reactions had a low exact matching precision but a high round-trip precision, Figure S2. Reductive amination had a similar average precision in both exact and round-trip matching precision (Figure S2), as was the case for accuracy, Figure 2.

To characterize the predictions further, we calculated the top-1 Tanimoto similarity, the fraction of valid sampled SMILES, and the fraction of uniquely top-N sampled SMILES, Figure S3. The Tanimoto similarity demonstrated that even if the top-1 predictions were not always an exact match, they were often similar to the ground truth. The top-1 exact matching accuracy of Suzuki cross-couplings, for example, was low, while the Tanimoto similarity was on average 0.76 (Figure 2a and Figure S3a). Apart from a few outlier batches, the fraction of

valid sampled SMILES was close to 1 (median 0.99; Figure S3a). Moreover, the fraction of uniquely sampled SMILES was 0.77 for the overall test set. An exception was found among Ugi reactions where the fraction of uniquely sampled SMILES was 0.43. Recalling the high exact matching accuracy of Ugi (Figure 2a), the high precision (Figure S2a) and the high Tanimoto similarity (0.93; Figure S3a), we hypothesized that the model was certain about when to apply Ugi reactions and therefore did not generate a diverse set of suggestions in those cases. Finally, we noticed that the fraction of uniquely sampled SMILES was lower for the forward model than the backward model, Figure S3 (0.38 vs. 0.77), which supported the rather high observed round-trip accuracies.

Comparing Chemformer single-step performance with the template-based model. The results on the test set split of the proprietary data showed promise for applying Chemformer to retrosynthesis prediction but do not address how the model generally performs compared to the template-based model. Figure 3 compares the performance of Chemformer and the template-based model on the shared PaRoutes-USPTO [27] test set. Chemformer consistently reached a higher top-N accuracy compared to the template-based model (mean 7.19 and 2.85 percentage difference in top-N exact and round-trip accuracy, respectively), which further validated the single-step prediction ability of Chemformer.

Taken together, the results from the single-step retrosynthesis analysis demonstrated that although the exact matching accuracy was modest on the proprietary test set, the predictions could be validated with the forward Chemformer model via round-trip matching. Moreover, the performance of Chemformer was especially good on reaction classes common in drug discovery, with most reaction classes showing a top-10 round-trip accuracy above 0.97. The Chemformer model also outperformed the template-based model on the PaRoutes-USPTO test set. Confident that Chemformer performed well on single-step predictions, we moved on to evaluate its performance in a multi-step setting.

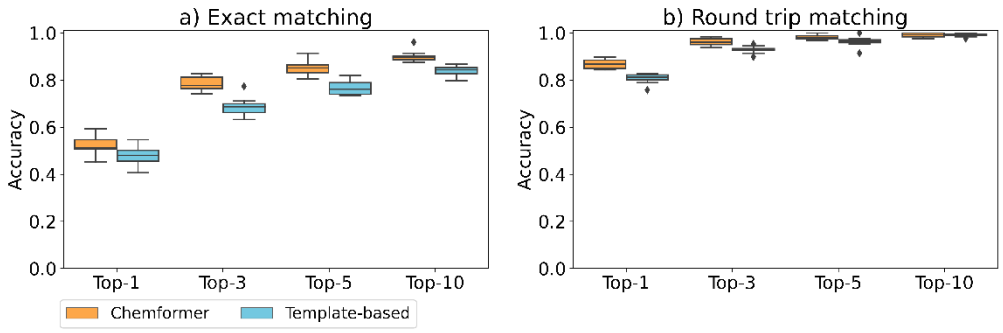

Figure 3 *Top-N a) exact matching and b) round-trip matching overall accuracies of Chemformer (orange) and the AiZynthFinder template-based model (blue) evaluated on the PaRoutes-USPTO test set. The distributions were obtained over inference batches.*

Evaluating multi-step retrosynthesis performance

General multi-step metrics and route similarity. To study the two models for the purpose of multi-step retrosynthesis, we employed AiZynthFinder with each single-step model as the expansion policy together with Monte Carlo tree search (MCTS). For the Chemformer experiments, we used our modified MCTS algorithm with immediate child node instantiation and cached batch predictions. AiZynthFinder experiments were then applied to the AZ design target dataset. Table 1 and Table 2 list the resulting performance metrics and the percentage of solved routes depending on solvability case, respectively. Chemformer obtained at least one solved route for 95% of the targets, while the template-based model solved routes for only 72% of the targets. The template-based model, on the other hand, was one order of magnitude faster than Chemformer to find solutions (15.8 vs. 158.0 seconds median search time), Table 1. Table 2 shows that 23% of the targets were solved by Chemformer but not the template-based model, while a mere 0.7% of the targets were solved only by the template-based model. Roughly 4% of the targets remained unsolved by either models.

Model	Median search time (s)	Percentage solved (%)
Chemformer	158.04	95.06
Template-based	15.82	72.38

Table 1 *General multi-step retrosynthesis metrics for Chemformer and the template-based model coupled with MCTS applied to the AZ design dataset.*

Solvability case	Percentage (%)
Solved by both	71.64
Only solved by Chemformer	23.42
Only solved by template-based	0.74
Solved by none	4.20
Solved by any	95.80

Table 2 *Percentage of solved targets in the AZ design dataset given different solvability cases.*

Because most of the targets solved by the template-based model were also solved by Chemformer, it is of interest to examine the routes for targets solved by both and see whether those were identical, or at least similar. Noteworthy, a recent multi-step analysis including the Chemformer and the template-based model showed that while the Chemformer had a higher route success rate (i.e., a greater percentage of compounds where at least one solved synthesis route was produced), the template-based model achieved the highest top-N route accuracy and building block accuracy [21]. These results, which were obtained on the PaRoutes benchmark set [27], indicate that the two models indeed generated different routes. To investigate this on the AZ design dataset, we compared the top-5 ranked routes and calculated the fraction of identical routes, Figure 4a. Moreover, we clustered all the extracted routes and computed cluster homogeneity as a metric for route similarity, Figure 4b. In short, a cluster homogeneity closer to 0.5 means that an equal amount of Chemformer and template-based routes had been grouped together, implying that the routes were very similar. A cluster homogeneity closer to 1 instead means that the routes generated by the two models were not grouped together, implying that the routes were very different. The top-5 overlap- and similarity-analyses showed that for targets which were not solved by both models, the routes were typically not identical in top-5, and not very similar according to

cluster homogeneity, Figure 4. In contrast, routes solved by both were more similar and had a higher fraction of identical top-5 routes. Still, the results suggest that the proposed routes were different between the two models.

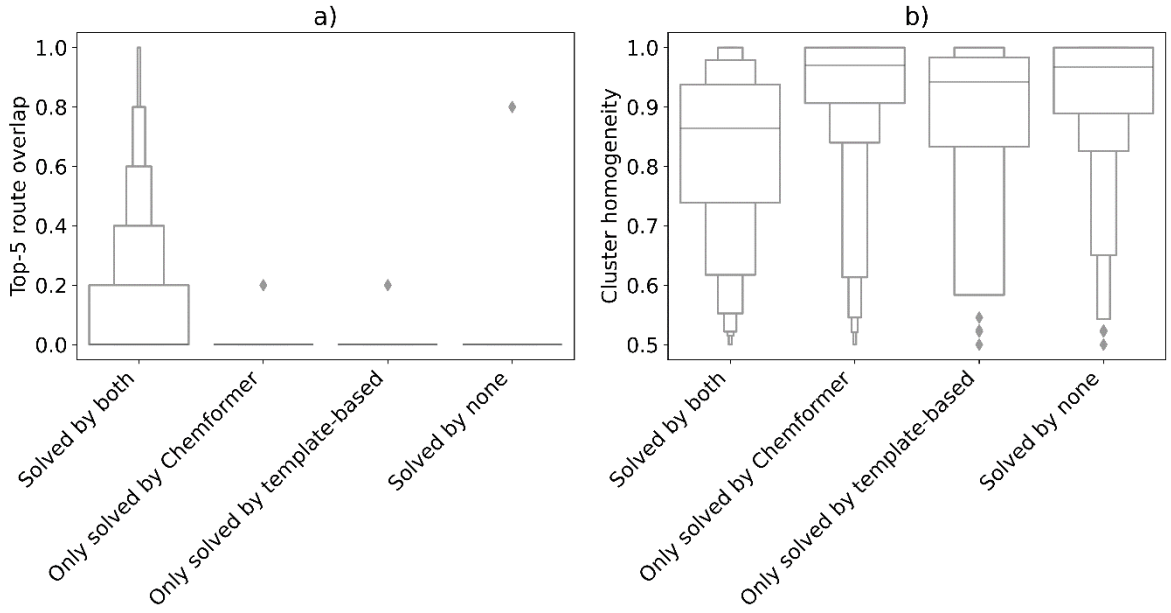

Figure 4 *Route identity and similarity as functions of solvability case: a) fraction of top-5 route overlap of each target, and b) cluster homogeneity of clusters generated for each target in the AZ design dataset. The data distribution is represented by boxen-plots. The centerline corresponds to the median, each box following the centerline encapsulates half of the residual data (50%, 25%, etc.).*

An initial quantification of the difference between routes was done by extracting route complexity metrics, Table 3. Chemformer on average produced more reactants and fewer single-reactant solutions (linear transforms) than the template-based model. The Chemformer routes were also typically shorter than the routes suggested by the template-based model. Moreover, Chemformer appeared to generate more trees than the template-based model, agreeing with recent findings [21]. Altogether, the initial route analyses showed that the solutions produced by Chemformer and the template-based model were different. Specifically, Chemformer solved routes for more target compounds than the template-based model, but with an order of magnitude higher median search time. Moreover, the suggested routes of targets solved by both models were typically not identical

between the two models, and the route complexity differed depending on which model had been used in the tree search. This naturally raised the question of how the solutions were different in terms of the chemistry suggested by each model.

Model	# Reactants / reaction	# Building blocks	Linear transforms (%)	Ave. depth	# Routes	# Solved routes
Chemformer	1.83	4.48	21	3.43	2625.51	368.74
Template-based	1.73	4.96	28	4.49	141.12	37.70

Table 3 *Complexity metrics of routes extracted with Chemformer and the template-based model as expansion policies, averaged over AZ design targets.*

Differences in Chemformer and template-based predicted reactions. To investigate how the routes generated by Chemformer and the template-based model were chemically different, we extracted the unique reactions from all routes and performed a number of analyses of those reactions. Figure 5 shows the round-trip accuracy (top-10 of forward predictions) of each extracted reaction related to targets solved by a) no model and b) both models. The round-trip accuracies of targets solved by no model were comparable between the two models, although the accuracies are slightly lower than the round-trip accuracy from the single-step model analysis (Figure 2). For targets solved by both, we noticed a slightly higher round-trip accuracy of the template-based model, although the difference is likely not significant. Notably, the round-trip accuracy was similar regardless of the solvability case. The inability to find solved routes for certain target compounds was therefore likely not due to the models suggesting invalid reactions. Based on these results, we conclude that both methods suggested feasible reactions as assessed by the forward model.

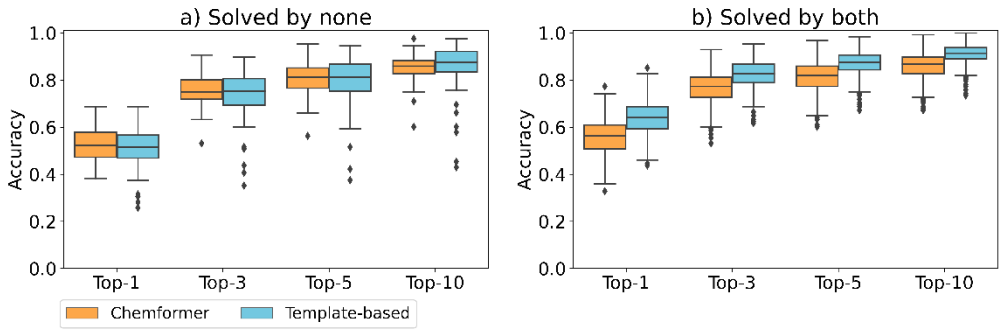

Figure 5 *Round-trip accuracy of (unique) extracted reactions from routes generated using Chemformer (orange) and template-based model (blue) as expansion policy. The subplots show the results of reactions extracted from routes of AZ design targets which were a) solved by none and b) solved by both models. Each data point represents one inference batch.*

An advantage with the template-based model is that we can directly map back the predictions to experiments. In contrast, Chemformer operates more like a black box. To further investigate whether the Chemformer predictions represented reasonably feasible chemistry, we analyzed if the predictions represented known experimental reaction data by computing the reaction similarity to the Reaxys reaction database, Figure 6. The results showed that while the Chemformer and template-based model predictions had comparable reaction similarity distributions across all solvability cases, the reaction similarity of Chemformer reactions was slightly lower. In line with this, we observed that Chemformer in general produced more reactions which the NextMove reaction classifier was unable to recognize, compared to the template-based model, Figure S4. However, the distribution of the reaction similarity was rather wide, hinting that both models predicted reactions which cannot be found in Reaxys, Figure 6.

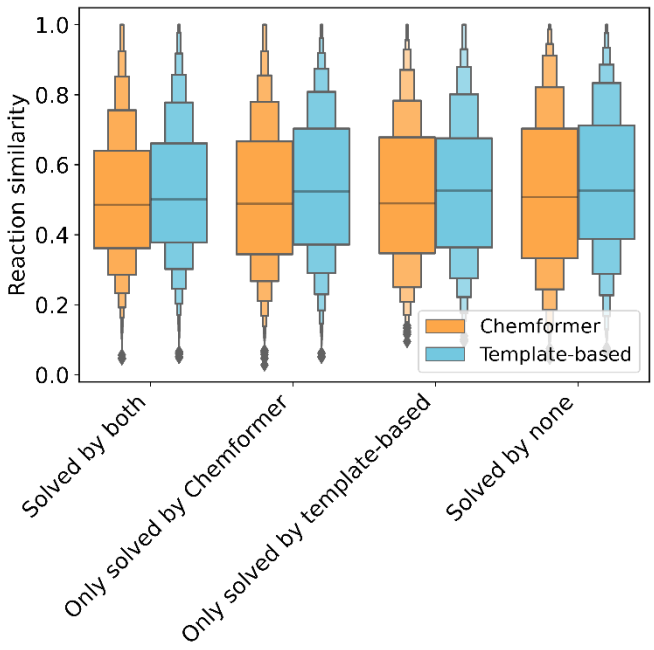

Figure 6 *Reaction similarity to reactions in Reaxys database. Comparing Chemformer (orange) and the template-based model (blue) across different solvability cases. Each data point represents a reaction from routes generated on AZ design targets. The data distribution is represented by boxen-plots. The centerline corresponds to the median, each box following the centerline encapsulates half of the residual data (50%, 25%, etc.).*

Next, we extracted reaction templates for all reactions predicted by the Chemformer model using the same approach as was used to extract templates for the template-based model. The aim was to investigate to what extent Chemformer suggested reactions corresponding to previously unseen templates, and how the relative amount of such novel templates differed between the four solvability cases. For targets solved by both models, it was clear that the models did not use the same set of templates, Figure 7a. This again reflected the route differences observed by the cluster analysis and Figure 3 and the complexity metrics in Table 3. Of the ~21.9k Chemformer templates, only ~3.5k were the same as those used by the Template-based model. The other ~18.4k Chemformer templates to some extent belonged to the set of templates making up the template-based model (modeling template library: 7.6k) and the full set of templates (full template library: 4228), Figure 7b. However, a large portion of the templates extracted from the Chemformer predictions was neither in the modeling template library nor in the full template library (Figure 7b: 10k). These results applied regardless of solvability case, Figures S5-S7. In other words, Chemformer invented

novel templates, unseen in the training data. These findings demonstrate that the Chemformer model is capable of extrapolating outside the known template space. Nonetheless, it is unclear what fraction of extrapolated reactions may simply be Chemformer hallucinations; a well-known characteristic of language-models. The identified novel templates require further validation and experimental characterization.

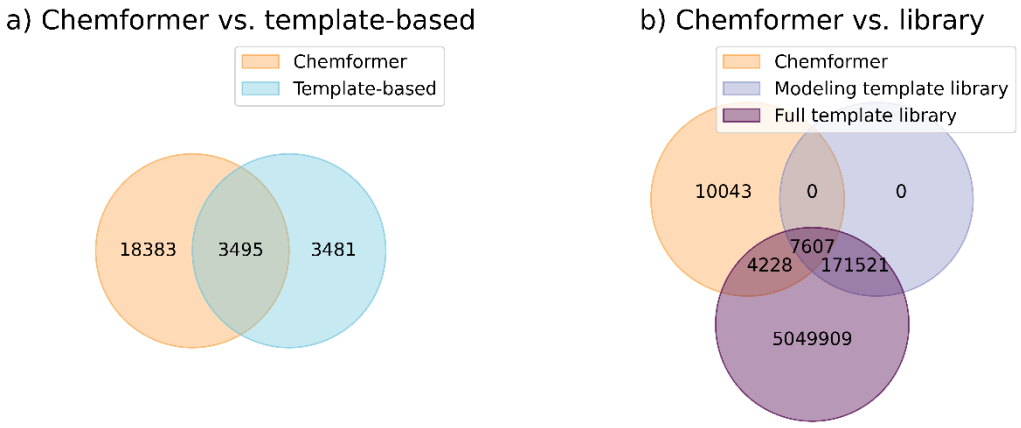

Figure 7 *Venn diagrams showing the overlap of exploited templates used in multi-step retrosynthesis on AZ design targets solved by both models. a) Comparing templates used by Chemformer and the template-based model, b) comparing templates used by Chemformer to the full template library of the proprietary dataset as well as the modeling template library used to train the template-based model.*

Characterizing Chemformer novel templates. As a first evaluation of the new templates, we computed the reaction similarity to the Reaxys database, Figure S8. As expected, the new templates also displayed lower reaction similarity to the reaction database compared to the library templates. When investigating where these new templates were generated, we found that they were more often generated at lower top-N rankings compared to library templates, Figure S9a,d. Moreover, for unsolved routes, the new templates appeared later in the synthesis sequence than library templates, while the opposite was true for solved routes, Figure S9b,e. The distributions of number of reacting atoms were similar between the two sets of reactions, Figure S9c,f. Interestingly, we found that new templates were involved in solved routes for roughly 82% of the targets.

To see if we could quantitatively validate the new templates, we computed the round-trip accuracy of the corresponding predicted reactions, Figure 8. The round-trip accuracy was significantly higher for library templates than new templates, which can explain the slight discrepancy observed in round-trip accuracy when comparing the extracted reactions of Chemformer and template-based models (Figure 5). Furthermore, we noticed that the round-trip accuracy of the Chemformer library templates were slightly higher than that of the template-based predictions, especially in the case of targets for which none of the models could find a solved route. These results, however, cannot validate the new chemistry of the Chemformer. However, a lower round-trip accuracy does not directly imply invalid reactions: the reason for the lower accuracy could simply be because the forward model was not trained on such reactions. Although outside the scope of this work, future efforts will be aimed at realizing some of the identified novel reactions in the wet-lab.

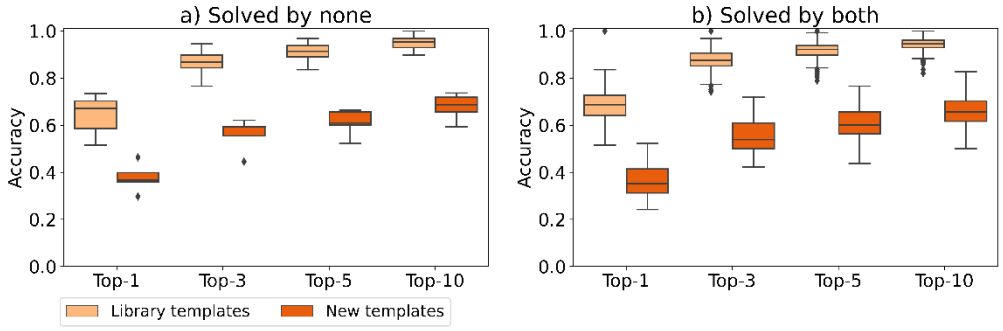

Figure 8 *Round-trip accuracy of (unique) extracted reactions from routes generated using Chemformer as expansion policy where library templates (light orange) or new templates generated by Chemformer (dark orange) were used. The subplots correspond to AZ design targets a) solved by none and b) solved by both. Each data point represents one inference batch.*

Because we could not quantitatively validate the new template reactions, we turned to visual inspection of a randomly selected subset of reactions. Figure S10-S14 show a short list of such reactions where the product was found in the public PubChem [41] database. One trend was particularly interesting: we noticed that although the suggested reactions appeared to correspond to reasonable chemistry, they oftentimes included multiple steps. Among the templates extracted from the routes in which both models found a synthetic route, 24% have two distinct, non-overlapping reaction sites. Figure 9 shows an example

reaction where Chemformer suggested multiple reaction steps in one retrosynthesis step. Predictions such as this one are likely the reason for the surplus of Chemformer reactions being classified as “Unrecognized”, Figure S4. Moreover, these multi-step predictions naturally yield and can explain the previously observed shorter retrosynthesis routes compared to the longer routes obtained with the template-based model, Table 3. To validate the two separate reactions, we modified our template-extraction routine to extract two templates for these reactions. This is a non-trivial problem, but we could nevertheless confirm that for 64% of the two-step reactions we could find at least one of the separate templates in the “full template library”-set (55% if we compare to the smaller “modelling template library"-set), at least partially validating the predictions made by the Chemformer. These results show that many of the novel Chemformer templates in fact do not correspond to novel chemistry, but the combination of existing transformations. Two-set transformations are in fact common tools used in organic synthesis. Even so, they tend to be problematic as single reaction templates since they constitute two reaction centers, which would typically be difficult to apply to a novel target compound. Consequently, the relative abundance of such templates is low, and therefore most of these templates would not be included in a template-based model. As a result, one of the shortcomings of using the template-based model is that the tree search may find routes with consecutive template-applications which could have been combined in a single retrosynthetic step. Predicted two-set transformations could yield a more efficient tree search, and the ability of Chemformer to suggest such transformations

may thus be an advantage compared to a template-based model.

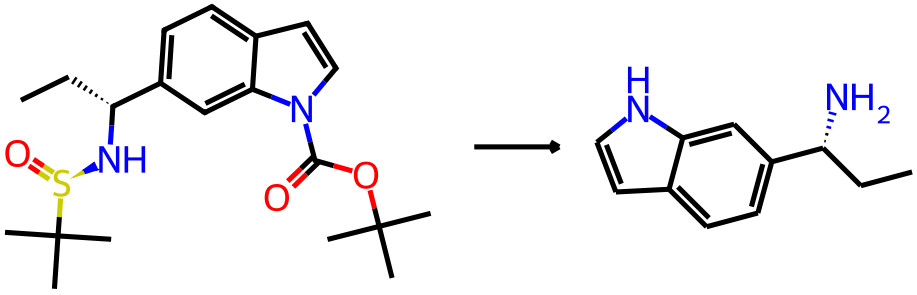

Figure 9 *Example of a Chemformer-generated reaction including multiple steps. The specific reaction was selected for display because the product molecule is present in the public PubChem [41] database.*

In summary, Chemformer generated potentially new chemistry in the form of previously unseen templates and these templates could not be fully validated by the forward Chemformer. However, through visual inspection of specific examples, we could qualitatively validate some of the new template reactions.

Conclusions

In this paper, we evaluated a transformer model, *Chemformer*, for the task of single-step and multi-step retrosynthesis in drug discovery. The model was obtained by fine-tuning a pre-trained transformer model [9] on a proprietary dataset from AstraZeneca for the task of retrosynthesis predictions. The main results and findings are summarized in Table 4 to enable a quick overview of the study. Although a lower single-step prediction accuracy was observed for rearrangements and ring-forming reactions, Chemformer achieved high accuracy for reaction classes especially relevant for drug discovery. Moreover, the fact that Chemformer reached higher accuracy than the template-based model on the shared test set showed promise for multi-step retrosynthesis.

Single-step retrosynthesis	Multi-step retrosynthesis
Modest overall exact matching accuracy (Top-1: 0.34, Top-10: 0.61), but high round-trip accuracy (Top-1: 0.72, Top-10: 0.92).	Chemformer solved routes for more target compounds than the template-based model, but with an order of magnitude higher median search time.
Ugi reactions obtained a high exact matching accuracy (0.80, Top-10: 0.98).	The suggested routes of targets solved by both models were typically not identical between the two single-step models.
Suzuki couplings had low exact matching accuracy (Top-1: 0.38, Top-10: 0.79), but among the highest round-trip accuracies (Top-1: 0.91, Top-10: 0.99).	Chemformer generated more and shorter routes compared to the template-based model.
Especially high round-trip accuracy for reaction classes common in drug discovery (top-10 > 0.97 for most classes).	The round-trip accuracy of reactions from targets solved by both models or no model was similar between the two models.
Chemformer outperformed the template-based model on the PaRoutes-USPTO test set (percentage difference 7.19 and 2.85 for exact and round-trip accuracy, respectively).	Reaction similarity to Reaxys database was comparable between Chemformer and the template-based model.
	Chemformer predicted reactions corresponding to reaction templates not present in the template library.
	Multi-step reactions were identified in 24% of the novel templates.
	Multi-step reaction predictions yield shorter routes and are in line with how chemists carry out experiments.
	The novel Chemformer templates could not be validated with round-trip accuracy, but some of the new templates were validated through visual inspection.

Table 4 *Summary of the main results from the work presented here.*

In multi-step retrosynthesis, Chemformer solved routes for more targets than the template-based model but took much longer time to generate those. The generated Chemformer routes were shorter than the template-based routes and included “new chemistry” in the form of previously unseen templates. Although the reactions behind the new templates had lower round-trip accuracy than those of library templates, we were able to qualitatively validate them. We found that Chemformer often proposed multi-step reactions which yielded shorter

routes compared to the template-based model. The shorter routes may be considered a feature but could potentially introduce additional work for the chemist's postprocessing in some cases. Given the results obtained here, we suggest a mixed-policy strategy for the industrial production platform. Specifically, we suggest applying MCTS with the template-based model to each target to get solved routes for most targets fast. After this, MCTS with Chemformer could be applied to the unsolved targets to generate solved routes for almost all targets. This way, the two models work together to optimize synthesis planning output in a resource-efficient manner.

Conflicts of interest

The authors would like to declare no conflicts of interest.

Acknowledgements

The authors would like to thank Mikhail Kabeshov and Thierry Kogej for insightful discussions about reaction validity.

References

[1] J. C. A. Oliveira *et al.*, "When machine learning meets molecular synthesis," *Trends Chem.*, vol. 4, no. 10, pp. 863–885, Oct. 2022, doi: 10.1016/j.trechm.2022.07.005.

[2] E. J. Corey and W. T. Wipke, "Computer-Assisted Design of Complex Organic Syntheses," *Science*, vol. 166, no. 3902, pp. 178–192, Oct. 1969, doi: 10.1126/science.166.3902.178.

[3] C. W. Coley, W. H. Green, and K. F. Jensen, "Machine Learning in Computer-Aided Synthesis Planning," *Acc. Chem. Res.*, vol. 51, no. 5, pp. 1281–1289, May 2018, doi: 10.1021/acs.accounts.8b00087.

[4] S. Genheden, A. Thakkar, V. Chadimová, J.-L. Reymond, O. Engkvist, and E. Bjerrum, "AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning," *J. Cheminformatics*, vol. 12, no. 1, p. 70, Nov. 2020, doi: 10.1186/s13321-020-00472-1.

[5] M. Świechowski, K. Godlewski, B. Sawicki, and J. Mańdziuk, "Monte Carlo Tree Search: A Review of Recent Modifications and Applications," *Artif. Intell. Rev.*, vol. 56, no. 3, pp. 2497–2562, Mar. 2023, doi: 10.1007/s10462-022-10228-y.

[6] P. Schwaller *et al.*, "Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy," *Chem. Sci.*, vol. 11, no. 12, pp. 3316–3325, Mar. 2020, doi: 10.1039/C9SC05704H.

[7] V. R. Somnath, C. Bunne, C. W. Coley, A. Krause, and R. Barzilay, "Learning Graph Models for Retrosynthesis Prediction." arXiv, Jun. 04, 2021. doi: 10.48550/arXiv.2006.07038.

[8] P. Schwaller, T. Gaudin, D. Lányi, C. Bekas, and T. Laino, "'Found in Translation': predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models," *Chem. Sci.*, vol. 9, no. 28, pp. 6091–6098, Jul. 2018, doi: 10.1039/C8SC02339E.

[9] R. Irwin, S. Dimitriadis, J. He, and E. J. Bjerrum, “Chemformer: a pre-trained transformer for computational chemistry,” *Mach. Learn. Sci. Technol.*, vol. 3, no. 1, p. 015022, Jan. 2022, doi: 10.1088/2632-2153/ac3ffb.
[10] D. Weininger, “SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules,” *J. Chem. Inf. Comput. Sci.*, vol. 28, no. 1, pp. 31–36, Feb. 1988, doi: 10.1021/ci00057a005.
[11] M. Lewis *et al.*, “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*, Online: Association for Computational Linguistics, Jul. 2020, pp. 7871–7880. doi: 10.18653/v1/2020.acl-main.703.
[12] D. Kreutter and J.-L. Reymond, “Multistep retrosynthesis combining a disconnection aware triple transformer loop with a route penalty score guided tree search,” *Chem. Sci.*, Sep. 2023, doi: 10.1039/D3SC01604H.
[13] A. Toniato, A. C. Vaucher, P. Schwaller, and T. Laino, “Enhancing diversity in language based models for single-step retrosynthesis,” *Digit. Discov.*, vol. 2, no. 2, pp. 489–501, Apr. 2023, doi: 10.1039/D2DD00110A.
[14] Y. Jiang, Y. Wei, F. Wu, Z. Huang, K. Kuang, and Z. Wang, “Learning Chemical Rules of Retrosynthesis with Pre-training,” *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 4, Art. no. 4, Jun. 2023, doi: 10.1609/aaai.v37i4.25640.
[15] A. Thakkar, A. Vaucher, A. Byekwaso, P. Schwaller, A. Toniato, and T. Laino, “Unbiasing Retrosynthesis Language Models with Disconnection Prompts.” ChemRxiv, Sep. 20, 2022. doi: 10.26434/chemrxiv-2022-gx9gb.
[16] D. Kreutter and J.-L. Reymond, “Multistep retrosynthesis combining a disconnection aware triple transformer loop with a route penalty score guided tree search.” ChemRxiv, Mar. 29, 2023. doi: 10.26434/chemrxiv-2022-8khth-v2.
[17] B. Zhang, J. Lin, L. Du, and L. Zhang, “Harnessing Data Augmentation and Normalization Preprocessing to Improve the Performance of Chemical Reaction Predictions of Data-Driven Model,” *Polymers*, vol. 15, no. 9, Art. no. 9, Jan. 2023, doi: 10.3390/polym15092224.
[18] D. Probst, M. Manica, Y. G. Nana Teukam, A. Castrogiovanni, F. Paratore, and T. Laino, “Biocatalysed synthesis planning using data-driven learning,” *Nat. Commun.*, vol. 13, no. 1, Art. no. 1, Feb. 2022, doi: 10.1038/s41467-022-28536-w.
[19] S. Zheng, J. Rao, Z. Zhang, J. Xu, and Y. Yang, “Predicting Retrosynthetic Reactions Using Self-Corrected Transformer Neural Networks,” *J. Chem. Inf. Model.*, vol. 60, no. 1, pp. 47–55, Jan. 2020, doi: 10.1021/acs.jcim.9b00949.
[20] F. Jaume-Santero *et al.*, “Transformer Performance for Chemical Reactions: Analysis of Different Predictive and Evaluation Scenarios,” *J. Chem. Inf. Model.*, vol. 63, no. 7, pp. 1914–1924, Apr. 2023, doi: 10.1021/acs.jcim.2c01407.
[21] P. Torren-Peraire *et al.*, “Models Matter: The Impact of Single-Step Retrosynthesis on Synthesis Planning.” arXiv, Aug. 10, 2023. doi: 10.48550/arXiv.2308.05522.
[22] P. Schwaller *et al.*, “Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction,” *ACS Cent. Sci.*, vol. 5, no. 9, pp. 1572–1583, Sep. 2019, doi: 10.1021/acscentsci.9b00576.
[23] G. Pesciullesi, P. Schwaller, T. Laino, and J.-L. Reymond, “Transfer learning enables the molecular transformer to predict regio- and stereoselective reactions on carbohydrates,” *Nat. Commun.*, vol. 11, no. 1, Art. no. 1, Sep. 2020, doi: 10.1038/s41467-020-18671-7.
[24] D. P. Kovács, W. McCorkindale, and A. A. Lee, “Quantitative interpretation explains machine learning models for chemical reaction prediction and uncovers bias,” *Nat. Commun.*, vol. 12, no. 1, Art. no. 1, Mar. 2021, doi: 10.1038/s41467-021-21895-w.
[25] M. Andronov, V. Voinarovska, N. Andronova, M. Wand, D.-A. Clevert, and J. Schmidhuber, “Reagent prediction with a molecular transformer improves reaction data quality,” *Chem. Sci.*, vol. 14, no. 12, pp. 3235–3246, Mar. 2023, doi: 10.1039/D2SC06798F.

[26] A. Thakkar, T. Kogej, J.-L. Reymond, O. Engkvist, and E. J. Bjerrum, “Datasets and their influence on the development of computer assisted synthesis planning tools in the pharmaceutical domain,” *Chem. Sci.*, vol. 11, no. 1, pp. 154–168, Dec. 2019, doi: 10.1039/C9SC04944D.
[27] S. Genheden and E. Bjerrum, “PaRoutes: towards a framework for benchmarking retrosynthesis route predictions,” *Digit. Discov.*, vol. 1, no. 4, pp. 527–539, Aug. 2022, doi: 10.1039/D2DD00015F.
[28] S. Genheden, P.-O. Norrby, and O. Engkvist, “AiZynthTrain: Robust, Reproducible, and Extensible Pipelines for Training Synthesis Prediction Models,” *J. Chem. Inf. Model.*, vol. 63, no. 7, pp. 1841–1846, Apr. 2023, doi: 10.1021/acs.jcim.2c01486.
[29] D. P. Kingma and L. J. Ba, “Adam: A Method for Stochastic Optimization,” presented at the International Conference on Learning Representations (ICLR), Ithaca, NYarXiv.org, 2015. doi: https://doi.org/10.48550/arXiv.1412.6980.
[30] A. Paszke *et al.*, “PyTorch: an imperative style, high-performance deep learning library,” in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA: Curran Associates Inc., 2019, pp. 8026–8037.
[31] F. Wa, “Pytorch lightning,” *GitHub*, vol. 3, 2019, Accessed: Jul. 21, 2023. [Online]. Available: https://cir.nii.ac.jp/crid/1370013168774120069
[32] G. Landrum *et al.*, “rdkit/rdkit: 2020_09_1 (Q3 2020) Release.” Zenodo, Oct. 20, 2020. doi: 10.5281/zenodo.4107869.
[33] A. Viterbi, “Error bounds for convolutional codes and an asymptotically optimum decoding algorithm,” *IEEE Trans. Inf. Theory*, vol. 13, no. 2, pp. 260–269, Apr. 1967, doi: 10.1109/TIT.1967.1054010.
[34] “NextMove Software.” https://www.nextmovesoftware.com/ (accessed Jul. 21, 2023).
[35] M. Hartenfeller *et al.*, “A Collection of Robust Organic Synthesis Reactions for In Silico Molecule Design,” *J. Chem. Inf. Model.*, vol. 51, no. 12, pp. 3093–3098, Dec. 2011, doi: 10.1021/ci200379p.
[36] S. Genheden, O. Engkvist, and E. Bjerrum, “Clustering of Synthetic Routes Using Tree Edit Distance,” *J. Chem. Inf. Model.*, vol. 61, no. 8, pp. 3899–3907, Aug. 2021, doi: 10.1021/acs.jcim.1c00232.
[37] S. Genheden, O. Engkvist, and E. Bjerrum, “Fast prediction of distances between synthetic routes with deep learning,” *Mach. Learn. Sci. Technol.*, vol. 3, no. 1, p. 015018, Jan. 2022, doi: 10.1088/2632-2153/ac4a91.
[38] L. McInnes, J. Healy, and S. Astels, “hdbscan: Hierarchical density based clustering,” *J. Open Source Softw.*, vol. 2, no. 11, p. 205, Mar. 2017, doi: 10.21105/joss.00205.
[39] P. Schwaller, B. Hoover, J.-L. Reymond, H. Strobelt, and T. Laino, “Extraction of organic chemistry grammar from unsupervised learning of chemical reactions,” *Sci. Adv.*, vol. 7, no. 15, p. eabe4166, Apr. 2021, doi: 10.1126/sciadv.abe4166.
[40] N. Schneider, D. M. Lowe, R. A. Sayle, and G. A. Landrum, “Development of a Novel Fingerprint for Chemical Reactions and Its Application to Large-Scale Reaction Classification and Similarity,” *J. Chem. Inf. Model.*, vol. 55, no. 1, pp. 39–53, Jan. 2015, doi: 10.1021/ci5006614.
[41] S. Kim *et al.*, “PubChem 2023 update,” *Nucleic Acids Res.*, vol. 51, no. D1, pp. D1373–D1380, Jan. 2023, doi: 10.1093/nar/gkac956.

Supplementary materials

Supplementary figures

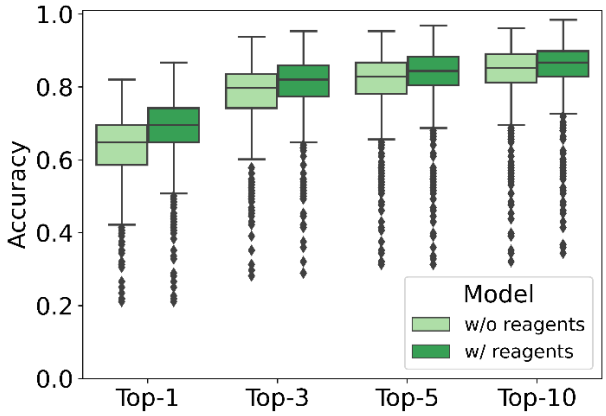

Figure S1 *Forward Chemformer model top-N accuracy, comparing a model trained without reagents (light green) to a model trained with reagents (dark green).*

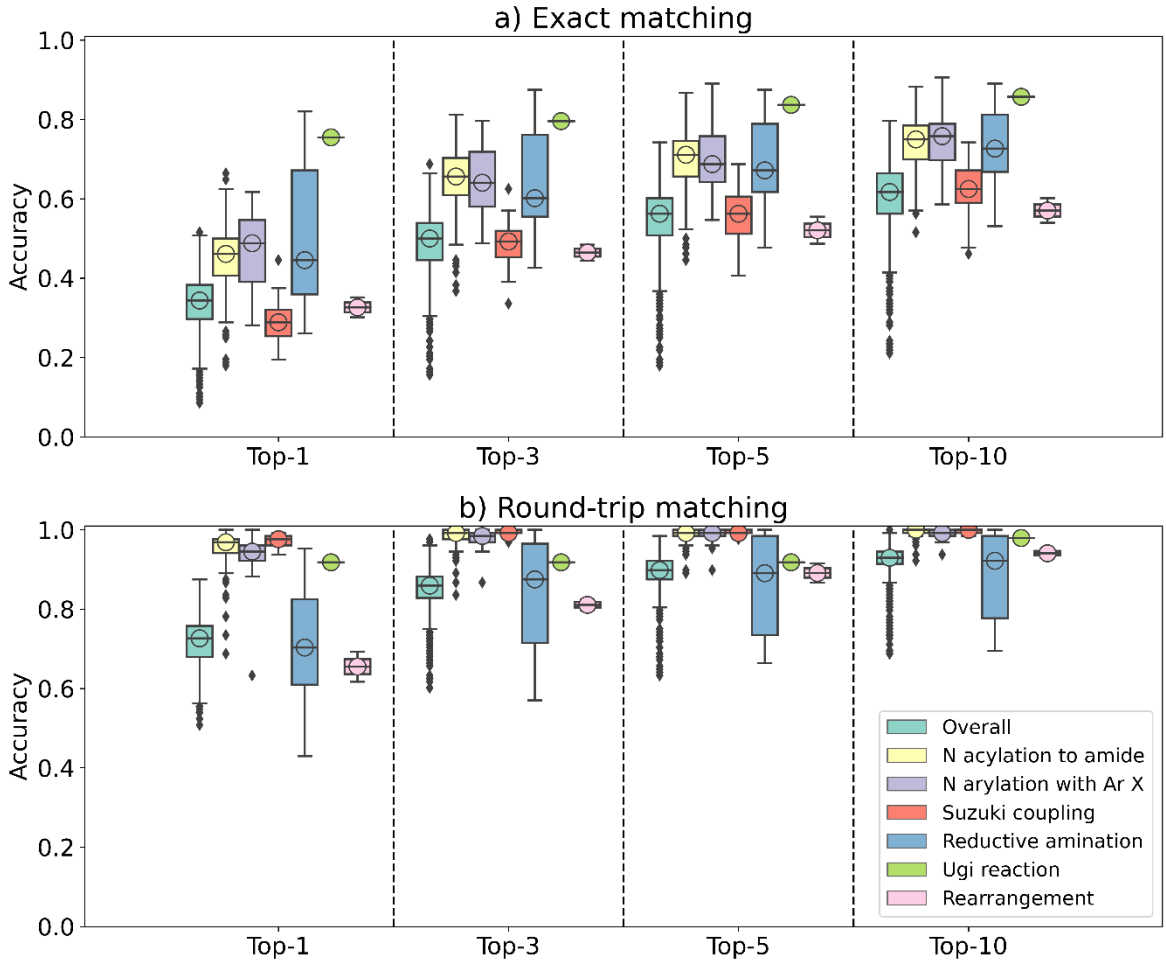

Figure S2 *Top-N a) exact matching and b) round-trip matching precisions of selected reaction classes, evaluated on the test set split of the proprietary data. The distributions were obtained over inference batches.*

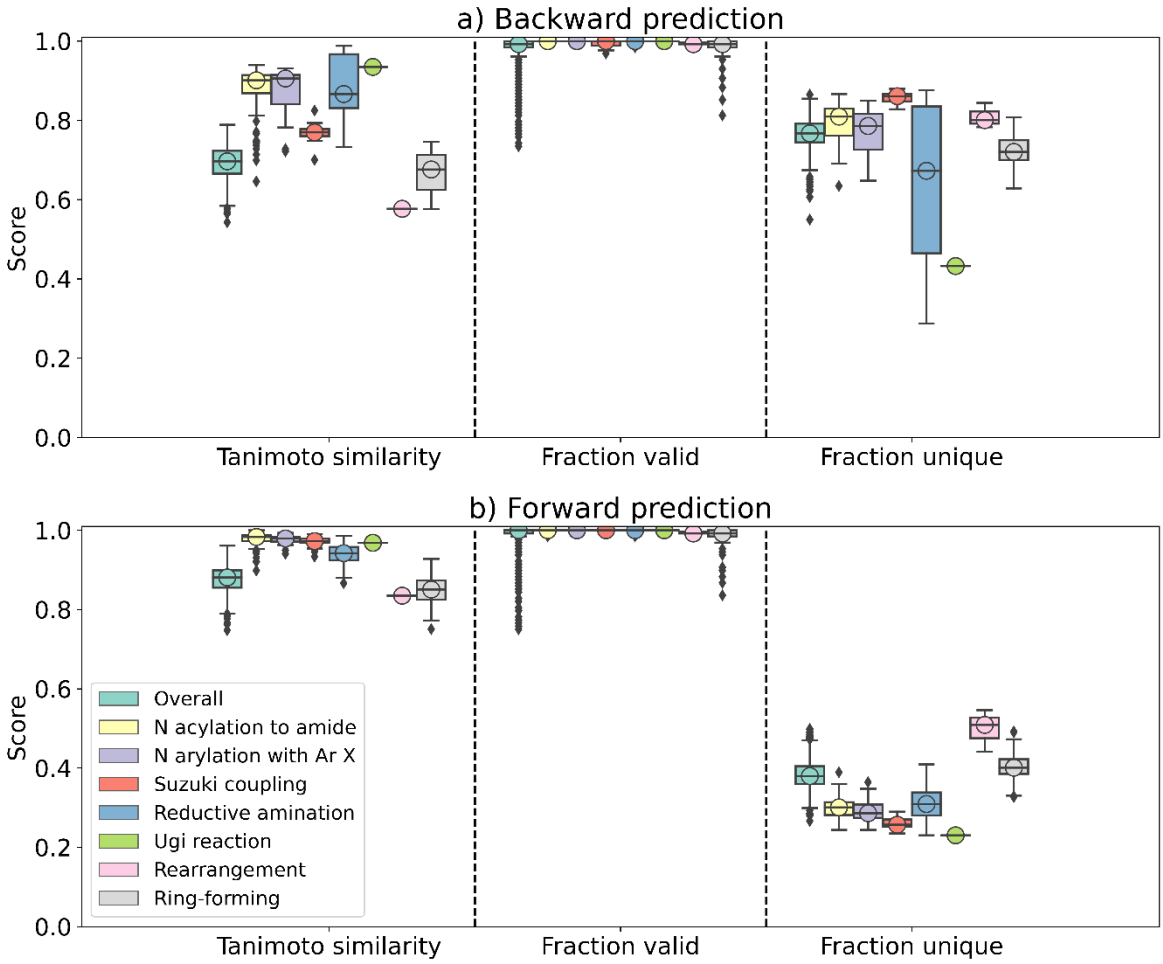

Figure S3 *Metrics characterizing predictions made by Chemformer a) backward and b) forward models. The distributions were obtained over inference batches.*

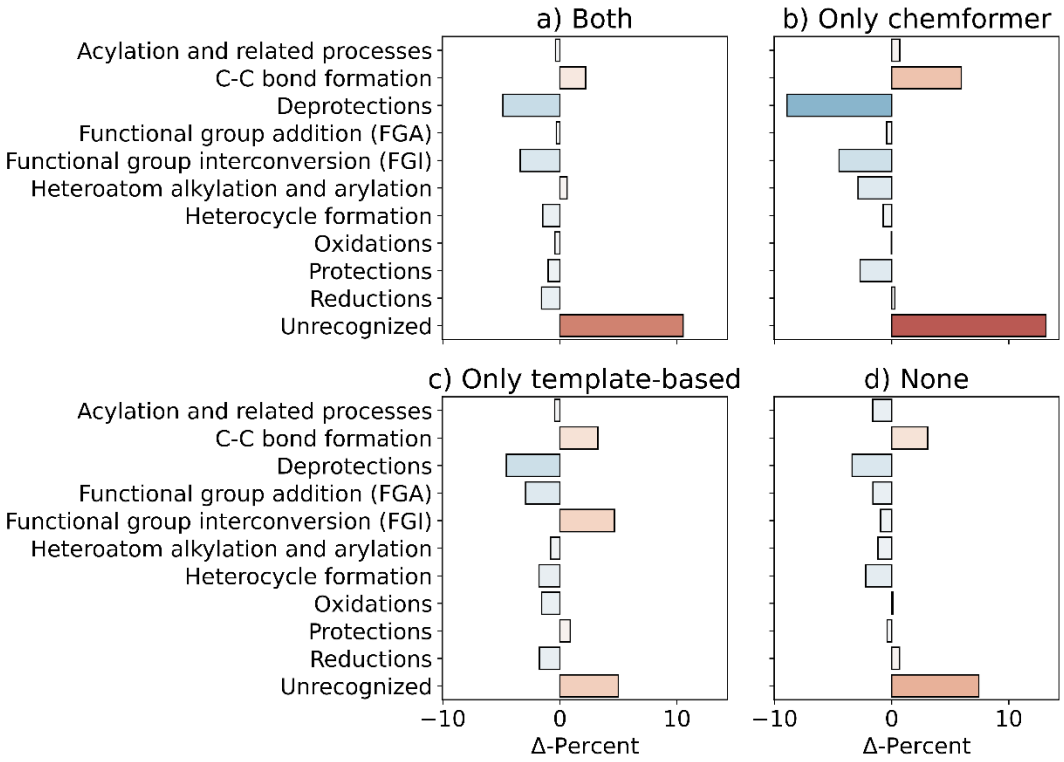

Figure S4 *Difference in reaction class distributions of Chemformer and template-based generated reactions on the AZ design targets. A positive Δ-Percent denotes a positive surplus for Chemformer.*

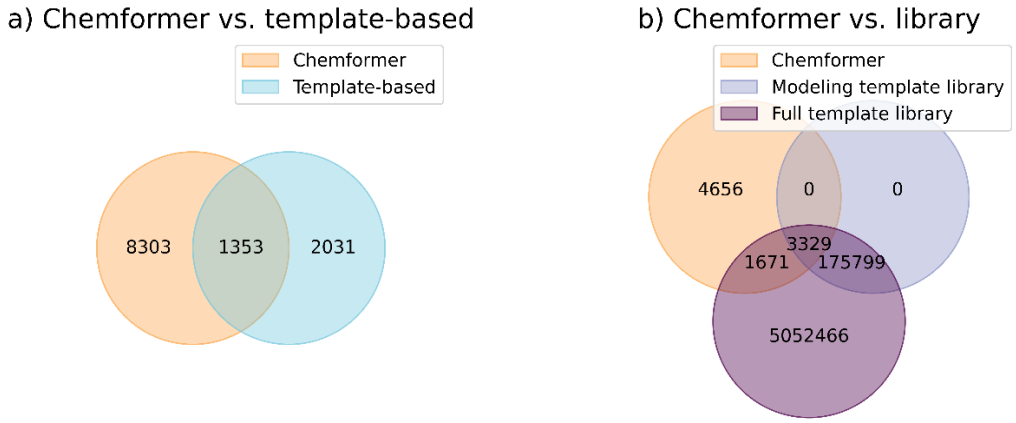

Figure S5 *Venn diagrams showing the overlap of exploited templates used in multi-step retrosynthesis on AZ design targets solved only by Chemformer. a) Comparing templates used by Chemformer and the template-based model, b) comparing templates used by Chemformer to the full template library of the proprietary dataset as well as the modeling template library used to train the template-based model.*

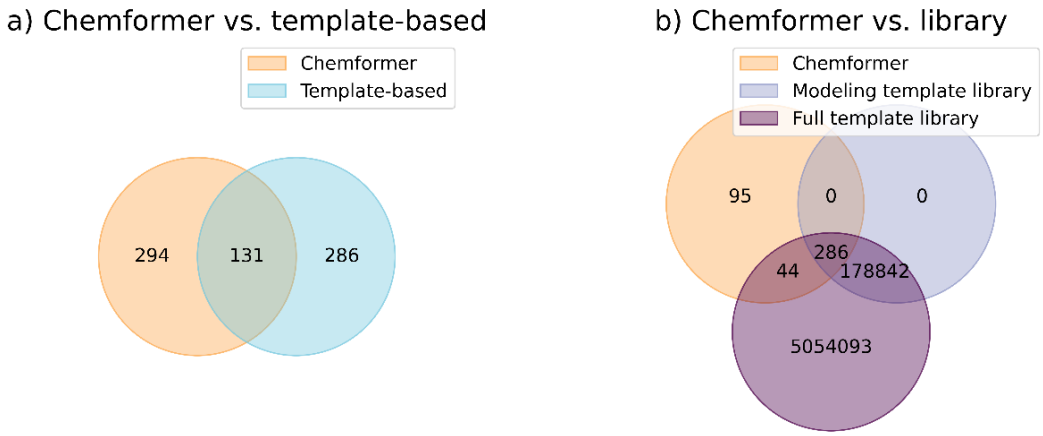

Figure S6 *Venn diagrams showing the overlap of exploited templates used in multi-step retrosynthesis on AZ design targets solved only by the template-based model. a) Comparing templates used by Chemformer and the template-based model, b) comparing templates used by Chemformer to the full template library of the proprietary dataset as well as the modeling template library used to train the template-based model.*

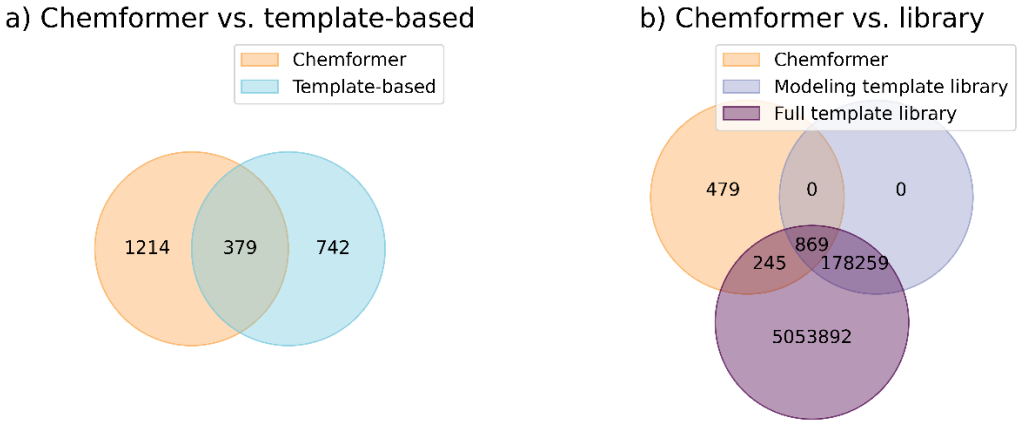

Figure S7 *Venn diagrams showing the overlap of exploited templates used in multi-step retrosynthesis on AZ design targets solved by none of the two models. a) Comparing templates used by Chemformer and the template-based model, b) comparing templates used by Chemformer to the full template library of the proprietary dataset as well as the modeling template library used to train the template-based model.*

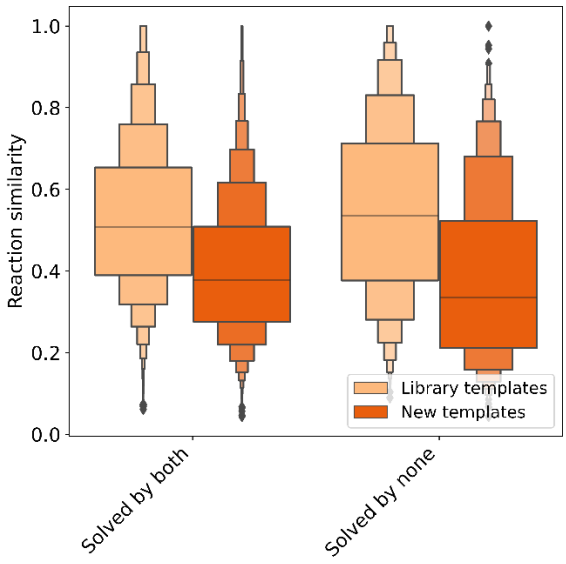

Figure S8 *Reaction similarity to a database of Chemformer generated library templates (light orange) and new templates (dark orange) as a function of solvability case. AZ design targets. Solved by both models, solved by no model.*

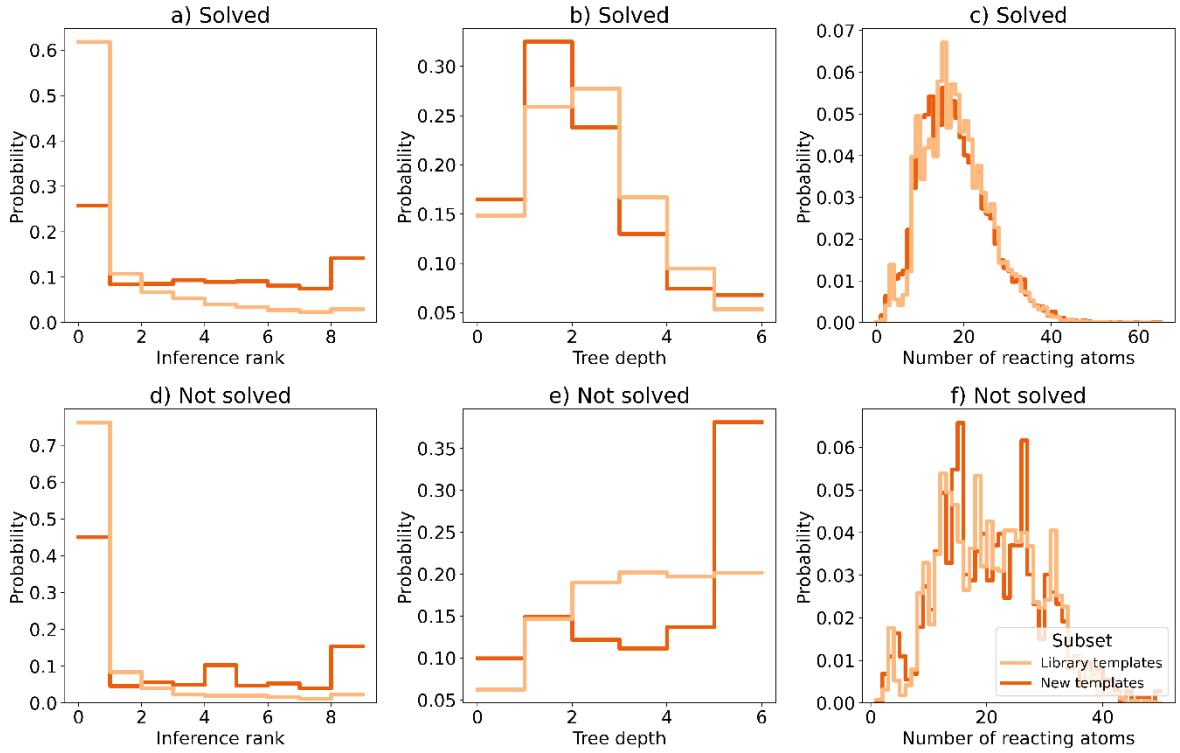

Figure S9 *Histograms over inference rank (a,d), tree depth (b, e) and number of reacting atoms (c, f) on AZ design targets. Chemformer-generated library templates are shown as light orange and new templates as dark orange. Subplots a-c) visualize statistics for solved routes, while d-f) visualize statistics for unsolved routes.*

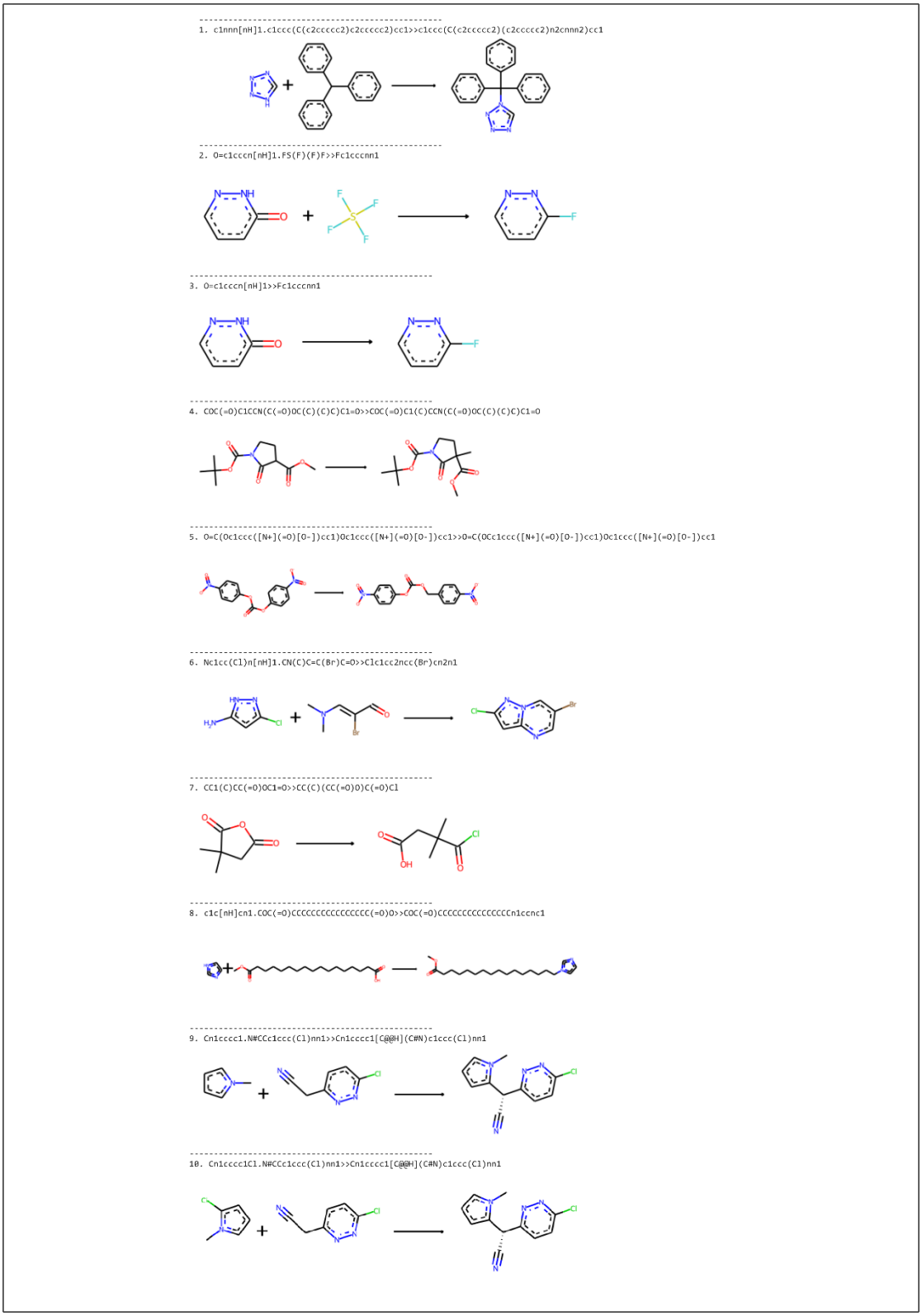

Figure S10 *Example reactions generated by Chemformer where the product molecule was found in the public Pubchem database.*

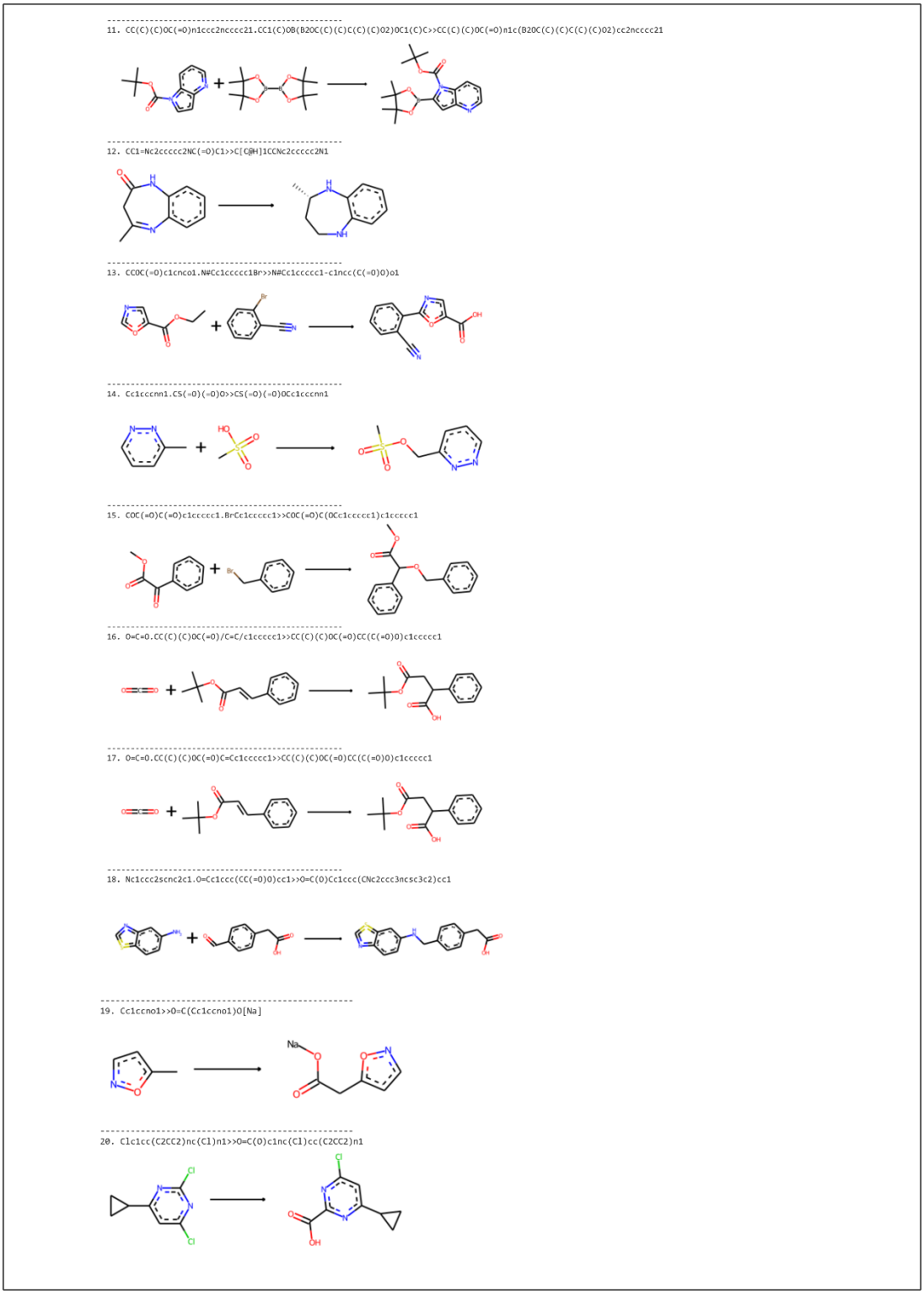

Figure S11 *Example reactions generated by Chemformer where the product molecule was found in the public Pubchem database.*

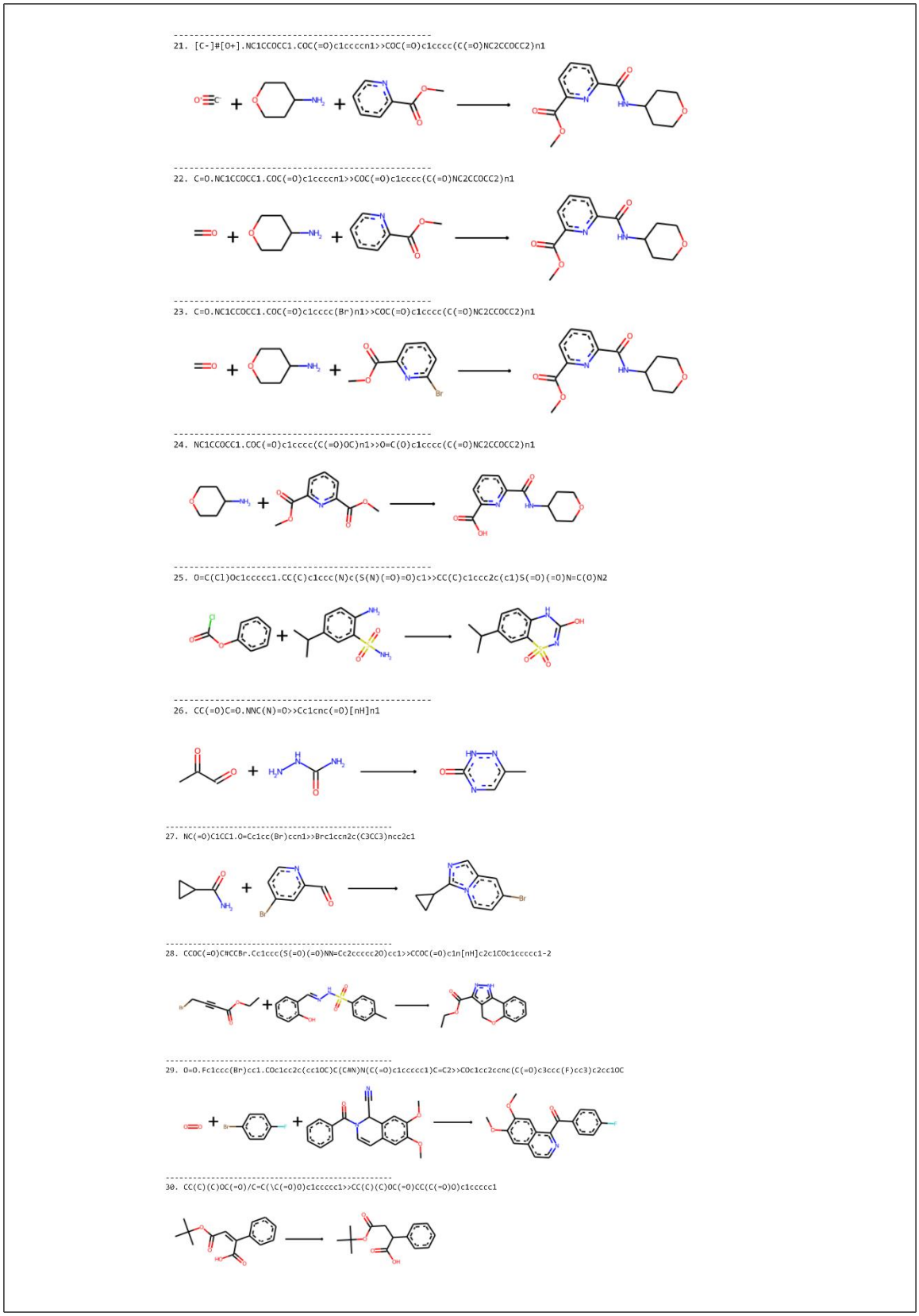

Figure S12 *Example reactions generated by Chemformer where the product molecule was found in the public Pubchem database.*

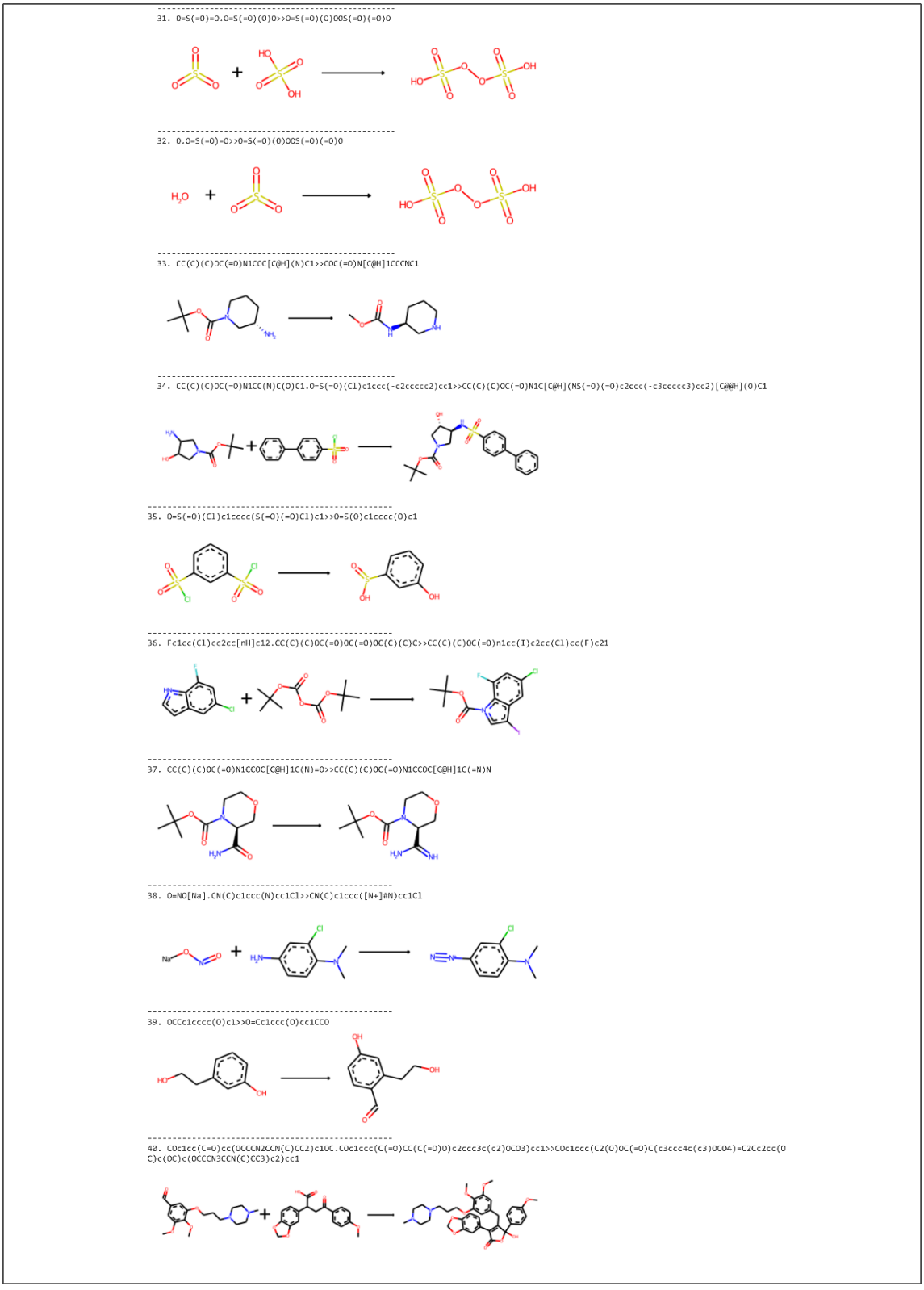

Figure S13 *Example reactions generated by Chemformer where the product molecule was found in the public Pubchem database.*

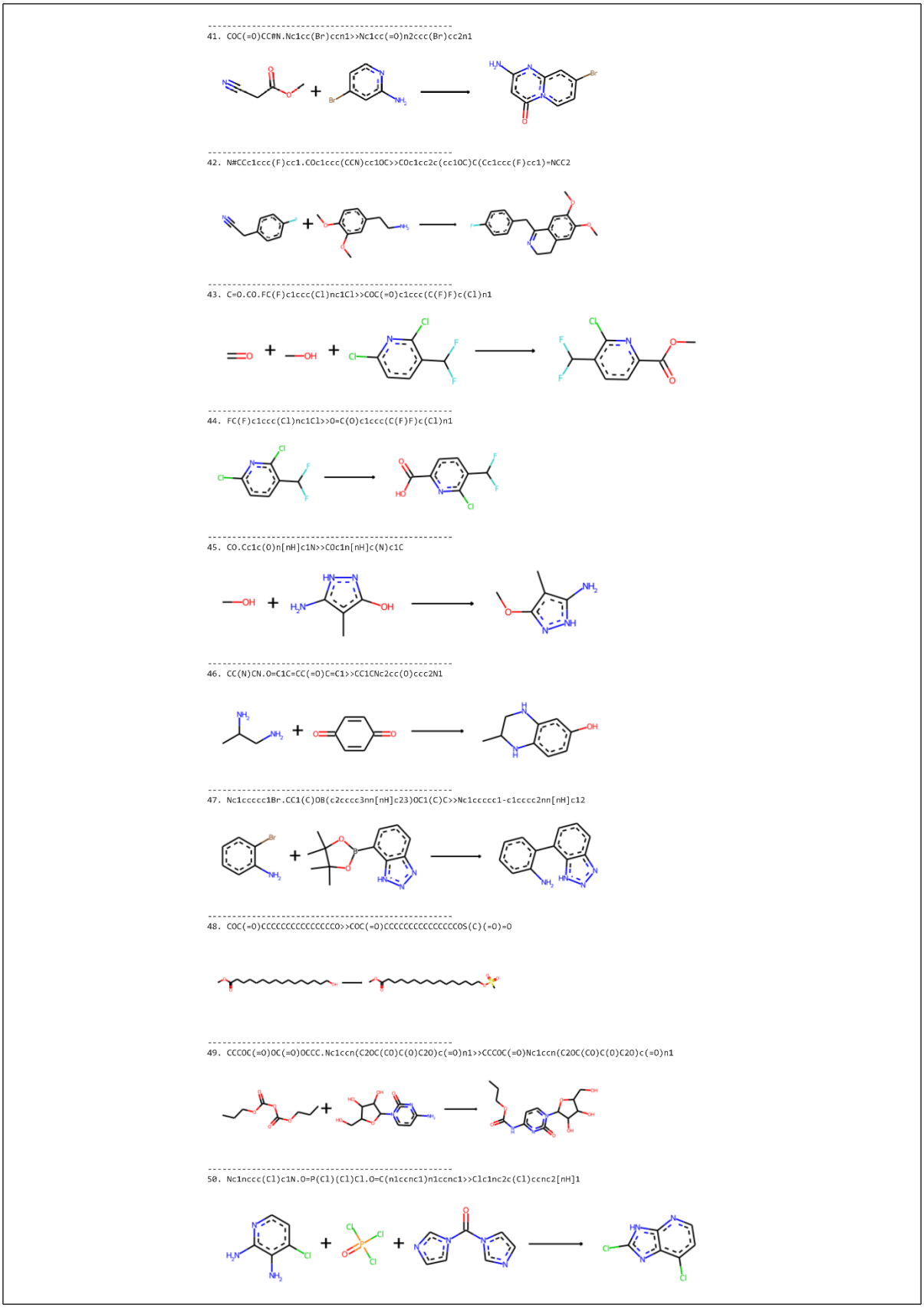

Figure S14 *Example reactions generated by Chemformer where the product molecule was found in the public Pubchem database.*

Supplementary tables

	Top-1	Top-3	Top-5	Top-10
Overall	0.3357	0.4917	0.5499	0.6065
N acylation to amide	0.5673	0.7932	0.8436	0.8805
Reductive amination	0.6637	0.8134	0.8548	0.8888
Suzuki coupling	0.3761	0.6210	0.7100	0.7926
Protections	0.4782	0.7017	0.7861	0.8513
Deprotections	0.3729	0.6028	0.6879	0.7702
N acylation to urea	0.6261	0.7904	0.8413	0.8878
N substitution with alkyl-X	0.3066	0.5080	0.6023	0.6777
N sulfonylation	0.7263	0.8736	0.9151	0.9457
N arylation with Ar-X	0.6051	0.8042	0.8589	0.9002
Ugi reactions	0.8043	0.8478	0.8913	0.8913
Sonogashira reaction	0.5510	0.8409	0.9092	0.9366
Buchwald-Hartwig and Ullman	0.4716	0.7012	0.7739	0.8447
Rearrangements	0.2481	0.3829	0.4429	0.4997
Ring-forming	0.3170	0.4597	0.5191	0.5771

Table S1 *Top-N exact matching accuracies of selected reaction classes, evaluated on the test set split of the proprietary data. The presented data are averages over inference batches.*

	Top-1	Top-3	Top-5	Top-10
Overall	0.7187	0.8542	0.8925	0.9248
N acylation to amide	0.9343	0.9775	0.9840	0.9891
Reductive amination	0.7307	0.8333	0.8593	0.8812
Suzuki coupling	0.9114	0.9749	0.9857	0.9910
Protections	0.8616	0.9614	0.9796	0.9850
Deprotections	0.7506	0.9082	0.9485	0.9744
N acylation to urea	0.9314	0.9721	0.9826	0.9874
N substitution with alkyl-X	0.8251	0.9253	0.9551	0.9722
N sulfonylation	0.9556	0.9880	0.9934	0.9971
N arylation with Ar-X	0.9204	0.9762	0.9834	0.9878
Ugi reactions	0.8913	0.9130	0.9130	0.9783
Sonogashira reaction	0.9166	0.9752	0.9836	0.9911
Buchwald-Hartwig and Ullman	0.9303	0.9802	0.9901	0.9924
Rearrangements	0.6210	0.8103	0.8626	0.9242
Ring-forming	0.6702	0.8133	0.8568	0.8949

Table S2 *Top-N round-trip matching accuracies of selected reaction classes, evaluated on the test set split of the proprietary data. The presented data are averages over inference batches.*