An exploration of dataset bias in single-step retrosynthesis prediction

Anonymous Author(s)
Affiliation
Address
email

Abstract

Single-step retrosynthesis models are integral to the development of computer-aided synthesis planning (CASP) tools, leveraging past reaction data to generate new synthetic pathways. However, it remains unclear how the diversity of reactions within a training set impacts model performance. Here, we assess how dataset size and diversity, as defined using automatically extracted reaction templates, affect accuracy and reaction feasibility of three state-of-the-art architectures – template-based LocalRetro and template-free MEGAN and RootAligned. We show that increasing the diversity of the training set (from 1k to 10k templates) significantly increases top-5 round-trip accuracy while reducing top-10 accuracy, impacting prediction feasibility and recall, respectively. In contrast, increasing dataset size without increasing template diversity yields minimal performance gains for LocalRetro and MEGAN, showing that these architectures are robust even with smaller datasets. Moreover, reaction templates that are less common in the training dataset have significantly lower top-*k* accuracy than more common ones, regardless of the model architecture. Finally, we use an external data source to validate the drastic difference between top-*k* accuracies on seen and unseen templates, showing that there is limited capability for generalisation to novel disconnections. Our findings suggest that reaction templates can be used to describe the underlying diversity of reaction datasets and the scope of trained models, and that the task of single-step retrosynthesis suffers from a class imbalance problem.

1 Introduction

Retrosynthesis is a key pillar of organic chemistry, requiring expert chemical knowledge to develop a sequence of reactions that lead to the synthesis of a target product. As research has progressed, so too has the space of possible transformations,[1,2] yet organic synthesis remains a bottleneck in drug discovery.[3] The pioneering work of Corey and Wipke[4] has since spawned a plethora of computer-aided synthesis planning programs,[5,6,7] in which a multi-step algorithm recursively calls on a single-step model to generate potential precursors. These single-step algorithms can be broadly categorised as template-based, where models learn to identify reaction centres and apply rules from an explicitly pre-defined library,[8,9,10,11] or template-free, where models learn reaction patterns implicitly from reaction SMILES[12,13,14] or molecular graphs.[15,16,17] The latter class of models is unconstrained by reaction templates and is thus expected to be able to propose novel transformations.[18,13,19,14,12]

These methods, as is the case with machine learning algorithms generally,[20,21] have previously been found to be sensitive to imbalanced data, often reinforcing biases rather than identifying important trends.[22,23,24] This is most clearly evidenced by template-based models, where retrosynthesis is formulated as a multi-class classification task[25] and thus model performance is heavily affected by the underlying distribution of the reaction templates in the training data. Within retrosynthesis, this bias

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

manifests as preferential prediction of specific reaction classes, regioselectivities, or stereoselectivities which are better represented in the training set.[22,23,24] The widely used open-source USPTO reaction dataset,[26] derived from US patent data, and its subsets have been extensively used for training and model comparison,[27,28,29] however its underlying biases have been often overlooked during model evaluation.[23] Torren-Peraire *et al.* train and test multiple models on a variety of datasets, but the lack of a common test set means that results and biases cannot be directly compared.[30] Thakkar *et al.* investigate the impact of template library size on the performance of template-based models, but do not use template-free models and do not discuss the impacts of bias.[31] Thus, it is unclear how training data impacts model predictions, and what future reaction databases should look like in terms of size and diversity.[24,32]

Despite many works evaluating and comparing retrosynthesis models, there is little consensus on the best way to realistically evaluate extrapolation to real world scenarios.[33,30] Often models are trained and evaluated on a particular random split of USPTO50k,[27] which is itself a cleaned random subset of the USPTO database[26], however this relatively small dataset cannot demonstrate how model performance would scale when trained and tested on much larger and more diverse in-house reaction libraries.[30] Recently, Bradshaw *et al.* have shown random splits of patent databases yield overly optimistic results, due to the similarity of reactions within the same patent or published by the same author.[34] Instead, they use patent- and author-based splits to simulate out-of-distribution (OOD) data and measure generalisation to reactions from unseen patents and authors, respectively. Other studies instead define generalisation as the ability to predict novel transformations defined by reaction templates.[35,36,37,38,39] However, these studies focus on how well different model architectures can generalise to new templates, but not how the underlying training data impacts generalisation.

Here, we investigate the effect that dataset size and diversity have on single-step model performance by training and testing on different subsets of a reaction database. We generate USPTO-retro, a retrosynthesis-specific dataset derived from USPTO,[26] analyse its diversity through local reaction templates,[11] and use it to train and test three state-of-the-art single-step architectures: LocalRetro[11] (template-based), MEGAN[17] (graph-based template-free), and RootAligned[14] (SMILES-based template-free). We show that top-*k* accuracy is correlated with the popularity of reaction templates in the training set for all models, regardless of architecture, suggesting that this metric can serve as a measure of reaction diversity. Finally, we evaluate performance on external test sets extracted from the Pistachio database[40] to demonstrate a protocol for measuring generalisation to seen and unseen reaction templates (Figure 1A).

2 Methods

Data Two databases are used in this work: the USPTO reaction database[26] for training and testing, and the commercial Pistachio reaction database[40] as an external test set. We apply a retrosynthesis preprocessing pipeline to both datasets based on recent efforts towards standardisation and open science,[41,33] and the Pistachio database is further filtered to ensure no overlap with the training data. This pipeline removes reagents and erroneous reactions to ensure data quality and is applicable to any reaction database. A detailed description of the data cleaning steps along with the codebase is provided in SI§S1.

This pipeline was applied to the USPTO reaction database[26] to generate USPTO-retro, which includes 1,103,781 atom-mapped reaction SMILES. Reaction templates were extracted using the LocalTemplate[11] algorithm, generating a total of 10,028 local reaction templates. Two external test sets were created from Pistachio: Pistachio ID, containing 10k reactions with in-distribution templates seen in USPTO-retro, and Pistachio OOD, containing 10k reactions with unseen out-of-distribution templates.

Splitting The USPTO-retro dataset was split into training, validation, and test sets using a random 90:5:5 split, consistent with established practice in retrosynthesis studies.[27,29,28,26] This is referred to as the **full** split. To prevent data leakage, all reactions sharing the same product were assigned to the same subset.

To investigate the effects of dataset size and diversity, the training set was further split into 10%, 25%, and 50% subsets using two splitting strategies (Figure 1B):

- **Narrow split**: This strategy selects a subset of reaction templates and includes all associated reactions in the training, validation, and test sets, sequentially increasing template diversity with training dataset size. The validation and test sets are similarly filtered to contain only templates seen during training. This split aims to measure how many reaction templates models can learn to predict, and the effect of increasing template diversity on model performance.
- **Broad split**: In contrast, this strategy randomly samples a fraction of reactions from all templates in the full training set. The validation and test sets are not altered. This split is designed to measure how much data per template is needed to learn these chemical transformations.

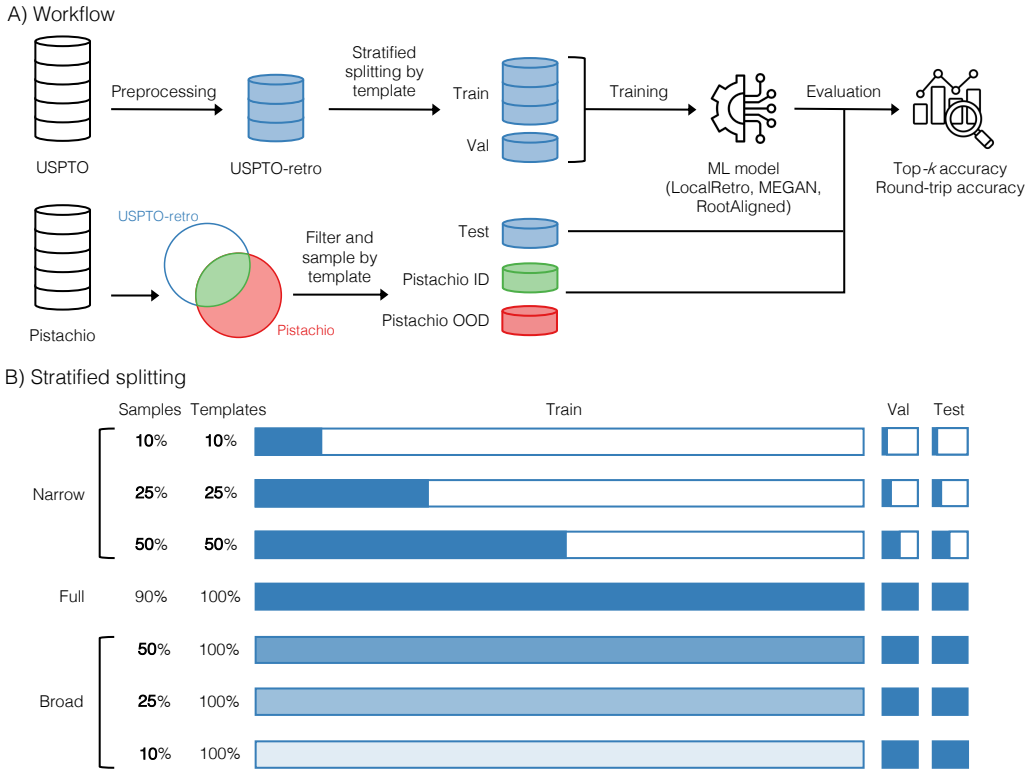

Figure 1: A) Workflow of data processing, training, and testing. The USPTO-retro dataset (blue) was randomly split into training, validation, and test sets, and then further split via stratified splitting by template. Two external test sets were created from Pistachio: Pistachio ID (green), containing 10k reactions with templates seen in USPTO-retro, and Pistachio OOD (red), containing 10k reactions with unseen templates. B) Visualisation of the splitting strategies used for training and testing. The sizes of the coloured bars indicate the number of templates sampled, while the opacity represents the proportion of reactions sampled.

Models Three model architectures were evaluated, each representing a distinct class of retrosynthesis algorithms. (i)) LocalRetro[11], a template-based algorithm; (ii) MEGAN[17], a semi-template algorithm, and (iii) RootAligned[14], a template-free algorithm. All models were trained using their respective repositories and evaluated using the Syntheseus platform,[33] which automatically removes duplicate and invalid predictions.

Evaluation While there are many evaluation metrics available to evaluate the performance of single-step models,[33,42,13] here we employed top-*k* accuracy and round-trip accuracy, which respectively measure recall and chemical feasibility.[33,42,13]

Top-*k* accuracy measures the proportion of test reactions for which the ground truth reactants appear among the model's top-k predictions. In this case, the ground truth is the reported reactants from the

test set. The top-10 accuracy metric is analysed in all experiments to mimic the desired breadth of a search tree in a multi-step algorithm.[33]

Top-*k* round-trip accuracy evaluates the proportion of top-*k* predicted reactants that satisfy back-translation.[13] This is done by checking whether they regenerate the original product via a forward reaction model (here RootAligned trained on the full USPTO-retro training set) to predict the top-1 product from each set of predicted reactants. If the predicted product matches the original target, the prediction is considered successful. We report top-1 and top-5 round-trip accuracy metrics to estimate the chemical feasibility of the top predictions.[13] It is important to note that the calculation of round-trip accuracy requires the use of a forward prediction model and is thus not 100% accurate, and should be interpreted as an approximation rather than an absolute measure of chemical validity.

3 Results and Discussion

3.1 Data analysis

We started our study by analysing the distribution of reaction templates within the newly generated USPTO-retro dataset, extracted using LocalTemplate.[11] Despite USPTO-retro containing over 1 million atom-mapped reaction SMILES, it shows a significant bias towards a small percentage of templates. Template frequency is used here to quantify the number of reactions a template describes in the training set, and, by extension, reaction classes (Figure 2A). The frequency of a template ranges from 2 to 78,922, with 50% of templates occurring fewer than 12 times. This bias underscores the inherent nature of open-source reaction databases, where certain reactions dominate. For example, the top 10 templates account for just 0.1% of all templates and together describe 30% of the training data.

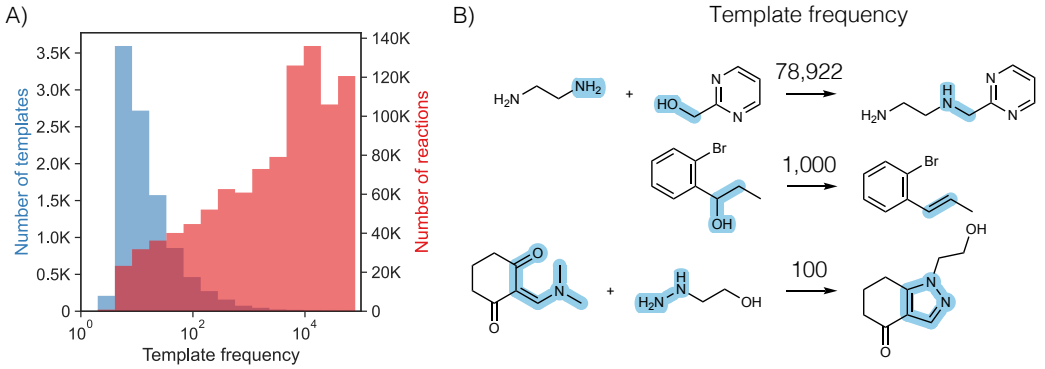

Figure 2: A) Histogram of templates (blue) and reactions (red) in the training set grouped by template frequency (on a log scale and with a box width of 0.3). Template frequency refers to the number of reactions in the training set described by a specific template. B) Example reactions from the training set with the template highlighted in blue and the template frequency labelled.

The most common reaction template, an example of which is shown in Figure 2B, corresponds to a C-N bond-forming S_N2 reaction, which accounts for >78k (8%) of all reactions in the training set. This template is similar to the next two most popular templates, which differ only in their leaving groups. Conversely, rarer templates include those with uncommon leaving groups or highly specific reaction centres. While these reactions are less common in the dataset, they are not necessarily less effective or harder to apply experimentally. Therefore, understanding the implications of this template imbalance on model performance is key for formulating better training and data curation strategies.

3.2 Impact of template distribution on model performance

To evaluate the impact of template distribution on model performance, we employed two splitting strategies to further partition the training set beyond the initial random split: the narrow and broad split. Both strategies sequentially increase the size of the training data, but differ in the diversity and distributions of their templates. The narrow split increases the number of unique reaction templates

in the training set as its size grows, allowing us to isolate the effect of increasing template diversity. In contrast, the broad split maintains template diversity while increasing the number of training examples, allowing us to assess the effect of increasing data volume per template. We analyse the resulting performances from these two strategies in the following subsections.

3.2.1 Narrow split

The narrow split is designed to evaluate how increasingly template-diverse datasets affect model performance. As expected,[31,25] models trained on less diverse datasets achieve higher top-k accuracy, as they have fewer competing reactions to choose from (Figure 3A). Increasing the number of templates from 1k to 10k results in a decrease in top-10 accuracy of 11.6% for LocalRetro, 14.5% for MEGAN, and 10.4% for RootAligned.

This decrease in top-k accuracy does not imply lower reaction feasibility; rather, it indicates the model's increased vocabulary of reactivity as a broader set of plausible reactions is suggested. Round-trip accuracy is used here to estimate the feasibility of the predicted reactions.[13] The top-1 round-trip accuracy remains roughly consistent across all splits and models, with over 89% of top predictions likely to be feasible reactions. In contrast, the top-5 round-trip accuracy increases by 14-21% across all models as template diversity increases, suggesting that lower-ranked predictions become more feasible when the model is exposed to more reaction types.

This behaviour differs from previous studies wherein top-k accuracy improves with additional randomly split training data.[43,25] In our case, increasing both the volume and diversity of training data leads to a decrease in top-k accuracy. This highlights the importance of explicitly reporting and accounting for reaction template diversity when comparing model performance across datasets with varying levels of diversity.

3.2.2 Broad split

The broad split aims to model the effect of increasing training set size while maintaining reaction diversity by using all available templates. Our results show that performance slightly improves for

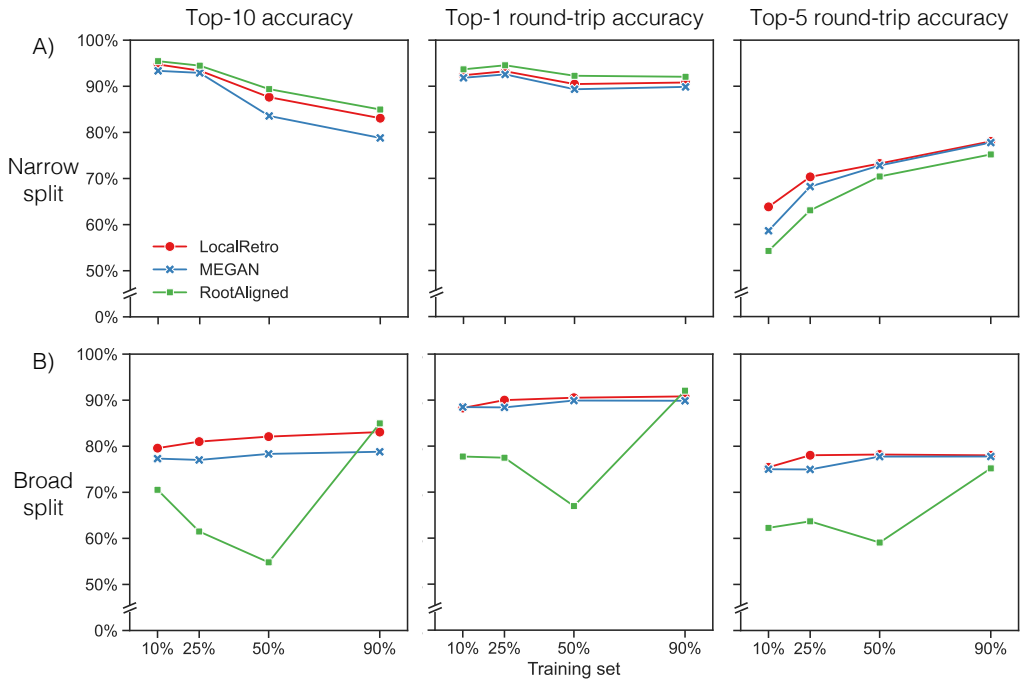

Figure 3: Top-10 accuracy (left), top-1 round-trip accuracy (middle) and top-5 round-trip accuracy (right) of models trained on the (A) narrow (increasing template diversity) and (B) broad splits (increasing data volume).

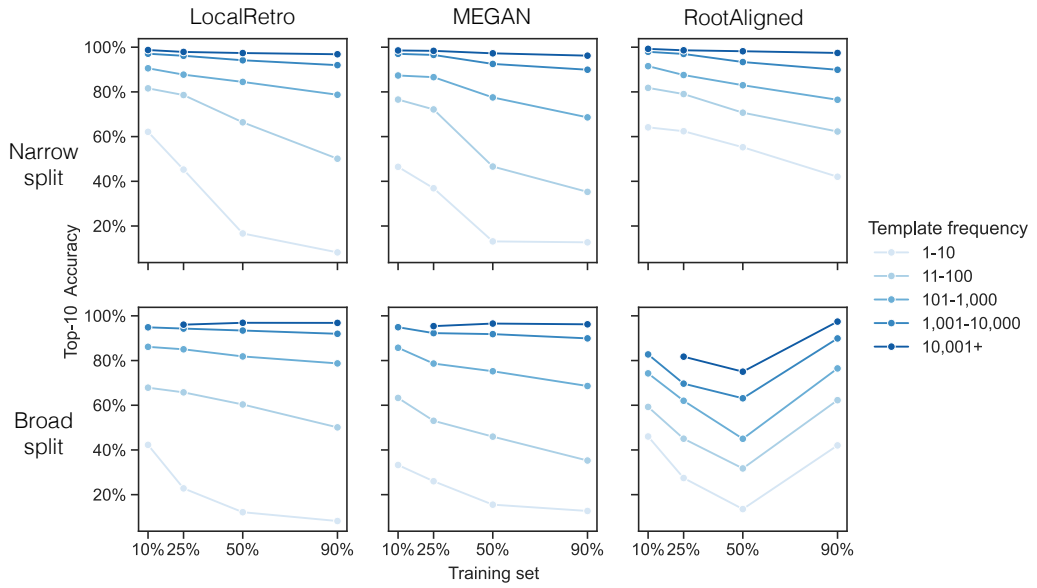

Figure 4: Top-10 accuracy of all trained models, as grouped by template frequency in the training set. The template frequency measures the number of times a particular template appears in the training set.

LocalRetro and MEGAN, with top-10 accuracy increasing by 3.5% for LocalRetro and 1.8% for MEGAN with a ninefold increase in training set size (Figure 3B). These results suggest that, with sufficient reaction diversity, these models are robust against variations in the size of the training set.

In contrast, the RootAligned model exhibits a substantial decrease in performance across the broad split. Its top-10 accuracy degrades by 15.7% between the 10% to 50% training sets, but recovers to 85.0% with the full training set. The consistent performances of LocalRetro and MEGAN indicate that the variations observed for RootAligned arise from the underlying transformer architecture rather than the size or nature of these training sets. This template-free approach attempts to implicitly learn chemistry directly from SMILES strings, whereas the template-based and semi-template methods provide a more structured way of learning reactions through predefined templates and graph edits. Consequently, the learning process of the RootAligned model may require more examples of the same reactions to fully utilise this chemistry. Models may also be more easily overfit on the smaller training datasets, leading to memorisation and pattern matching, which cannot generalise to the test set. Further investigation is needed to determine if this behaviour occurs with other template-free models.

3.3 Accuracy by template

Next, we investigated how template frequency bias in the training data affects model performance, focusing on top-10 accuracy across reaction templates (Figure 4). A clear trend emerges: templates that appear more frequently in the training set are predicted with significantly higher accuracy. The difference in top-10 accuracy between rare templates (frequency of 1-10) and popular templates (frequency of 10,001+) is at most 88.6% for LocalRetro, 83.5% for MEGAN, and 55.4% for RootAligned. A similar, though weaker, correlation is observed when considering Tanimoto similarities between the training and test sets (Figure S3). These trends persist even in models that do not explicitly use reaction templates, such as MEGAN and RootAligned, implying that template frequency reflects the underlying class distribution of reaction data.

In both the narrow and broad splits, increasing the training set size amplifies the spread of top-k accuracies across template frequencies. For the most frequent templates (with frequency > 10,001), LocalRetro and MEGAN consistently achieve top-10 accuracy above 95%, regardless of training set sizes. In contrast, rare templates (with frequency 1-10) show a marked drop in accuracy as training set size increases: top-10 accuracy decreases between the narrow 10% and full 90% training sets

by 53.9% for LocalRetro, 33.8% for MEGAN, and 22.1% for RootAligned. This behaviour is most pronounced for LocalRetro, which explicitly considers reaction templates and thus learns to prioritise more frequent classes during training. RootAligned, which implicitly encodes chemistry through SMILES strings, is less sensitive to these class imbalances. These results suggest that increasing both the number and imbalance of reaction templates contributes to performance disparities. To mitigate this, further work is needed to incorporate class balancing strategies during model training.

While the top-*k* accuracy measures how often a reaction template is correctly predicted, it does not describe how often that type of template is recalled. Thus, it is also important to understand if the models are oversampling from popular reaction classes as a way of mimicking the training set distribution. This behaviour is most easily studied in the LocalRetro model, as its algorithm readily outputs a ranked list of predicted templates. In all splits, the model oversamples the most popular template classes for its highest ranked prediction (Figure 5A). Rarer templates are undersampled compared to the true test distribution, which contributes to their low top-10 accuracy. These rarer templates are instead sampled more often at lower ranks as the model is less confident in their prediction (Figure 5B).

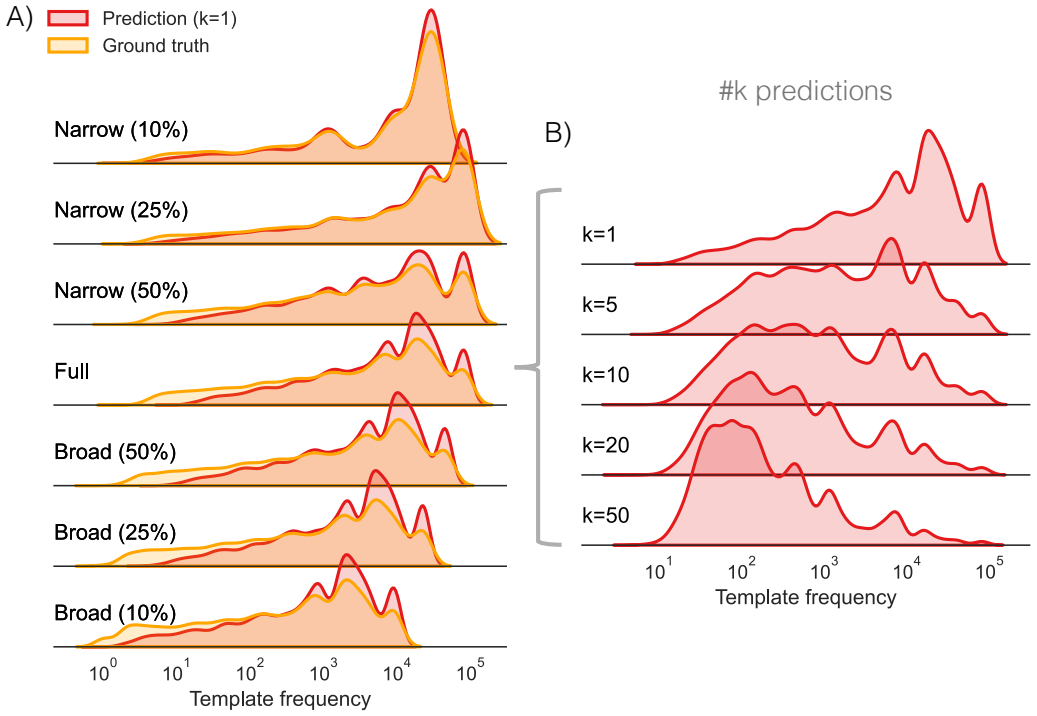

Figure 5: (A) Kernel density estimations (KDEs) of the training template frequency of the top prediction from LocalRetro (red) and ground truth (yellow). (B) The KDE distributions of the training template frequency of the #1, #5, #10, #20, and #50 predictions from the LocalRetro model trained on the full training set.

3.4 Generalisation to novel reactions

Generalisability in single-step retrosynthesis refers to a model's predictive capability for novel reactions. This can be assessed in multiple ways, for example, considering the prediction of previously unseen target products using known reaction templates or the prediction of novel disconnections not encountered during training. To systematically evaluate both aspects, we split our external test set from the Pistachio database into Pistachio ID (In-Distribution), which contains novel products with seen templates, and Pistachio OOD (Out-Of-Distribution), which contains novel products with unseen templates (Figure 6A). We use the broad split to evaluate generalisation to novel products (ID) and the narrow split to evaluate generalisation to novel disconnections (OOD).

On the Pistachio ID test set (Figure 6B), all models exhibit a moderate decline in top-10 accuracies when compared to their performance on the USPTO-retro test set (Section 3.3): 7-9% for LocalRetro, 6% for MEGAN, and 5-12% for RootAligned. This indicates that models successfully generalise to novel products using templates learnt during training, with similar performance trends to previous results. The slightly reduced performance on this test set is likely due to the lower structural similarity between Pistachio ID products and those in the USPTO-retro training sets (Figure S4).

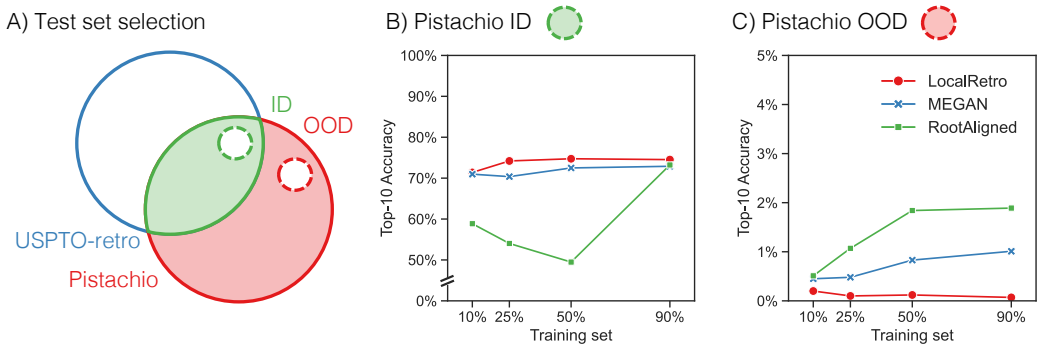

Figure 6: A) Diagrammatic representation of the overlap of templates between the USPTO-retro and Pistachio datasets. The Pistachio ID test set is selected from in-distribution templates (the intersection area shown in green), whereas the Pistachio OOD test set is selected from out-of-distribution templates (the exclusive area shown in red). B) Top-10 accuracy of all models trained on the broad split and tested on the Pistachio ID test set. C) Top-10 accuracy of all models trained on the narrow split and tested on the Pistachio OOD test set.

In contrast, performance on the Pistachio OOD test set (Figure 6C) reveals severe limitations in generalisability to novel disconnections, in agreement with previous findings.[35,36,37,38] LocalRetro exhibits near-zero top-10 accuracy, which is expected given its reliance on predefined templates. The non-zero accuracy suggests template ambiguity, where different templates from the training and OOD test sets occasionally yield the same sets of reactants. This occurs due to overlapping SMARTS patterns or errors in atom mapping. MEGAN and RootAligned models show modest generalisability, which increases with increased training diversity and peaks at top-10 accuracies of 1% and 2% respectively with the full training sets. Their low but non-zero accuracy implies that models prioritise recognising and applying patterns seen in the training data over utilising underlying chemical principles to generate novel, feasible disconnections.

These results highlight the differences in capabilities between ID and OOD generalisation, emphasising the need for distinct evaluations that distinguish between these two scenarios. Previous studies showing the traditional learning pattern of increasing top-*k* accuracy with increasing training data volume[43,25] may, in fact, be misattributing the effect of additional template coverage of the test set to additional data. This explanation may also apply to studies showing low generalisability to external datasets[33] or author-/patent-based splits[34], wherein their test sets possibly contain both seen and unseen templates. Furthermore, the extremely low generalisability of template-free models to novel templates suggests that these models are not yet sufficiently developed to warrant their use for predicting new chemistries.

4 Conclusion and future work

In this study, we presented a comprehensive assessment of the accuracy and feasibility of three state-of-the-art single-step retrosynthesis models – template-based LocalRetro and template-free MEGAN and RootAligned – exploring how dataset size and diversity, defined in terms of local reaction templates, affect performance.

Our results have highlighted the critical role of training set diversity in model performance. Increasing the diversity of the training set significantly increases top-5 round-trip accuracy, an indicator of prediction feasibility, while reducing top-10 accuracy, reflecting the ability of the model to recover the ground truth. This trade-off suggests that more diverse datasets enable the prediction of a broader range of plausible reactions, even if they differ from the ground truth. Interestingly, increasing dataset

size without increasing template diversity yields minimal performance gains for LocalRetro and MEGAN models, suggesting that template diversity has a greater impact on model performance than volume.

We also examined the impact of template frequency on model performance. All three models, regardless of whether they explicitly use templates, show a strong correlation between a template's frequency in the training set and the model's ability to predict it correctly. This indicates that all models implicitly rely on the distribution of reaction templates learnt during training, with rare templates consistently underperforming compared to more frequent ones.

Finally, to assess real-world applicability, we evaluated model performance on two external test sets derived from the Pistachio database: one containing novel products with known templates (Pistachio ID) and another with novel products and unseen templates (Pistachio OOD). While all models generalised reasonably well to new molecules involving known templates, their ability to predict novel disconnections was limited. These results highlight the differences in capabilities between ID and OOD generalisation. LocalRetro failed almost entirely on OOD reactions due to its reliance on predefined templates, while MEGAN and RootAligned achieved only 1–2% top-10 accuracy. These results highlight the need for evaluation protocols that clearly distinguish between in-distribution (ID) and out-of-distribution (OOD) generalisation.

These results also offer a new perspective on recent advances in transfer learning for retrosynthesis prediction, wherein fine-tuning effectively modifies the training template distribution. For instance, the mixed fine-tuning approach to bias predictions towards heterocyclic ring disconnections reported by Wieczorek *et al.* can be viewed as addressing the underlying class imbalance issues present in the initial training set.[44] Our results suggest that similar systematic approaches to class imbalance during training could improve representation across reaction classes. Similar challenges have been addressed in other domains, such as computer vision, through pre-training, data augmentation, and re-weighting strategies.[45]

The performance trends across the narrow and broad splits raise questions about what data should be used to train retrosynthesis models. Ideally, models would learn underlying physical principles to propose feasible reactions; however, evaluation shows that they are more likely to learn to mimic the template distribution of the training set. Furthermore, models do not necessarily exhibit worse accuracy when trained on less data; therefore, data curation efforts should prioritise quality and diversity over quantity. As such, it is clear that as chemists we cannot blindly train models with all available data and not consider the types of chemistry that data represents, and whether that chemistry suits our synthetic goals and targets.

Availability of data and materials

The code used to preprocess and split the datasets, as well as the model training configurations, are available at (attached zip).

Conflicts of interest

There are no conflicts to declare.

References

[1] S. Szymkuć, T. Badowski and B. A. Grzybowski, *Angew. Chem. Int. Ed.*, 2021, **60**, 26226–26232.

[2] D. G. Brown and J. Boström, *J. Med. Chem.*, 2016, **59**, 4443–4458.

[3] D. C. Blakemore, L. Castro, I. Churcher, D. C. Rees, A. W. Thomas, D. M. Wilson and A. Wood, *Nat. Chem.*, 2018, **10**, 383–394.

[4] E. J. Corey and T. W. Wipke, *Science*, 1969, **166**, 178–192.

[5] Z. Zhong, J. Song, Z. Feng, T. Liu, L. Jia, S. Yao, T. Hou and M. Song, *WIREs Comput. Mol. Sci.*, 2024, **14**, e1694.

[6] C. W. Coley, W. H. Green and K. F. Jensen, *Acc. Chem. Res.*, 2018, **51**, 1281–1289.

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

[8] F. Strieth-Kalthoff, S. Szymkuć, K. Molga, A. Aspuru-Guzik, F. Glorius and B. A. Grzybowski, *J. Am. Chem. Soc.*, 2024, **146**, 11005–11017.

[9] J. D. Shields, R. Howells, G. Lamont, Y. Leilei, A. Madin, C. E. Reimann, H. Rezaei, T. Reuillon, B. Smith, C. Thomson, Y. Zheng and R. E. Ziegler, *RSC Med, Chem,*, 2024, **15**, 1085–1095.

[10] P. Seidl, P. Renz, N. Dyubankova, P. Neves, J. Verhoeven, J. K. Wegner, M. Segler, S. Hochreiter and G. Klambauer, *J. Chem. Inf. Model.*, 2022, **62**, 2111–2120.

[11] S. Chen and Y. Jung, *JACS Au*, 2021, **1**, 1612–1620.

[12] R. Irwin, S. Dimitriadis, J. He and E. J. Bjerrum, *Mach. Learn.: Sci. Technol.*, 2022, **3**, 15022.

[13] P. Schwaller, R. Petraglia, V. Zullo, V. H. Nair, R. A. Haeuselmann, R. Pisoni, C. Bekas, A. Iuliano and T. Laino, *Chem. Sci.*, 2020, **11**, 3316–3325.

[14] Z. Zhong, J. Song, Z. Feng, T. Liu, L. Jia, S. Yao, M. Wu, T. Hou and M. Song, *Chem. Sci.*, 2022, **13**, 9023–9034.

[15] Z. Tu and C. W. Coley, *J. Chem. Inf. Model.*, 2022, **62**, 3503–3513.

[16] V. R. Somnath, C. Bunne, C. W. Coley, A. Krause and R. Barzilay, 35th Conference on Neural Information Processing Systems, 2021.

[17] M. Sacha, M. Błaż, P. Byrski, P. Dąbrowski-Tumański, M. Chromiński, R. Loska, P. Włodarczyk-Pruszyński and S. Jastrzębski, *J. Chem. Inf. Model.*, 2021, **61**, 3273–3284.

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

[19] P. Schwaller, T. Gaudin, D. Lányi, C. Bekas and T. Laino, *Chem. Sci.*, 2018, **9**, 6091–6098.

[20] B. van Giffen, D. Herhausen and T. Fahse, *J. Bus. Res.*, 2022, **144**, 93–106.

[21] R. Wang, P. Chaudhari and C. Davatzikos, *Proc. Natl. Acad. Sci.*, 2023, **120**, e2211613120.

[22] A. Thakkar, A. C. Vaucher, A. Byekwaso, P. Schwaller, A. Toniato and T. Laino, *ACS Cent. Sci.*, 2023, **9**, 1488–1498.

[23] D. P. Kovács, W. McCorkindale and A. A. Lee, *Nat. Commun.*, 2021, **12**, year.

[24] G. Durant, F. Boyles, K. Birchall and C. M. Deane, *Nat. Comput. Sci.*, 2024, 1–9.

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

[26] D. M. Lowe, *PhD thesis*, University of Cambridge, Cambridge, 2012.

[27] N. Schneider, N. Stiefl and G. A. Landrum, *J. Chem. Inf. Model.*, 2016, **56**, 2336–2346.

[28] W. Jin, C. W. Coley, R. Barzilay and T. Jaakkola, 31st Conference on Neural Information Processing Systems, 2017.

[29] H. Dai, C. Li, C. W. Coley, B. Dai and L. Song, 33rd Conference on Neural Information Processing Systems, 2019.

[30] P. Torren-Peraire, A. K. Hassen, S. Genheden, J. Verhoeven, D. A. Clevert, M. Preuss and I. V. Tetko, *Digit. Discov.*, 2024, **3**, 558–572.

[31] A. Thakkar, T. Kogej, J.-L. Reymond, O. Engkvist and E. J. Bjerrum, *Chem. Sci.*, 2019, **11**, 154–168.

[32] S. M. Kearnes, M. R. Maser, M. Wleklinski, A. Kast, A. G. Doyle, S. D. Dreher, J. M. Hawkins, K. F. Jensen and C. W. Coley, *J. Am. Chem. Soc.*, 2021, **143**, 18820–18826.

[33] K. Maziarz, A. Tripp, G. Liu, M. Stanley, S. Xie, P. Gainski, P. Seidl and M. Segler, *Faraday Discuss,*, 2024, –.

[34] J. Bradshaw, A. Zhang, B. Mahjour, D. E. Graff, M. H. S. Segler and C. W. Coley, *Challenging reaction prediction models to generalize to novel chemistry*, 2025, `http://arxiv.org/abs/2501.06669`, arXiv:2501.06669 [cs].

[35] M. H. Segler and M. P. Waller, *Chem. Eur. J.*, 2017, **23**, 6118–6128.

[36] Y. Yu, L. Yuan, Y. Wei, H. Gao, X. Ye, Z. Wang and F. Wu, *RetroOOD: understanding out-of-distribution generalization in retrosynthesis prediction*, 2023, `http://arxiv.org/abs/2312.10900`, arXiv:2312.10900 [cs] version: 1.

[37] H. Tu, S. Shorewala, P. T. Ma and V. Thost, *Retrosynthesis Prediction Revisited*.

[38] S. Chen and Y. Jung, *Assessing the Extrapolation Capability of Template-Free Retrosynthesis Models*, 2024, http://arxiv.org/abs/2403.03960, arXiv: 2403.03960.

[39] A. M. Westerlund, S. Manohar Koki, S. Kancharla, A. Tibo, L. Saigiridharan, M. Kabeshov, R. Mercado and S. Genheden, *J. Chem. Inf. Model.*, 2024, **64**, 3021–3033.

[40] J. Mayfield, I. Lagerstedt and R. Sayle, *Pistachio "Fantastic reactions and how to use them"*, 2021.

[41] V. S. Gil, A. M. Bran, M. Franke, R. Schlama, J. S. Luterbacher and P. Schwaller, NeurIPS 2023 AI for Science Workshop, 2023.

[42] F. Hastedt, R. M. Bailey, K. Hellgardt, S. N. Yaliraki, E. A. del Rio Chanona and D. Zhang, *Digit. Discov.*, 2024.

[43] J. Pang and I. Vulić, *Faraday Discuss.*, 2025, **256**, 413–433.

[44] E. Wieczorek, J. W. Sin, M. T. O. Holland, L. Wilbraham, V. S. Perez, A. Bradley, D. Miketa, P. E. Brennan and F. Duarte, *Transfer learning for heterocycle synthesis prediction*, 2024, https://chemrxiv.org/engage/chemrxiv/article-details/6617d56321291e5d1d9ef449.

[45] J. M. Johnson and T. M. Khoshgoftaar, *J. Big Data*, 2019, **6**, year.

Supporting Information

An exploration of dataset bias in single-step retrosynthesis prediction

Anonymous Author(s)

S1 Data preprocessing

The USPTO database was downloaded from https://figshare.com/s/5e57a3399c52701cbc15.[41] The 2023Q1 version of the commercial Pistachio database[40] was used, and reactions were deduplicated and mapped with RXNMapper.[?]

The preprocessing pipeline includes the following steps:

- Remove multi-product reactions by either filtering out small side products (<6 atoms) or removing reactions with large side products.
- Remove reactions with purely inorganic products.
- Remove reactions where the product is present as a reactant.
- Remove "stereoalchemy" by removing any stereochemistry tokens if present in the product but not in the reactants.
- Remove reagents (precursor species which do not contribute to the atom mapping).
- Remove reactions with >4 reactants.
- Canonicalise reaction SMILES.
- Remove reactions with over 512 tokens.
- Remove duplicate reactions.
- Extract templates using LocalTemplate[11] and filter out templates with under 6 occurrences.

The pipeline was applied to the USPTO and Pistachio databases, yielding 1,103,781 and 3,720,288 reactions, respectively.

The Pistachio database was further filtered by patent number to exclude all US patents and avoid overlap with the USPTO. Templates with fewer than 20 occurrences were removed, and the database was divided into two sets: one containing templates also present in USPTO-retro, and the other containing new templates. A subset of 10,000 reactions was randomly sampled from both sets to generate Pistachio ID and Pistachio OOD, respectively.

The code used for cleaning and splitting both datasets can be found in the Github repository: (attached zip).

S2 Splitting distributions

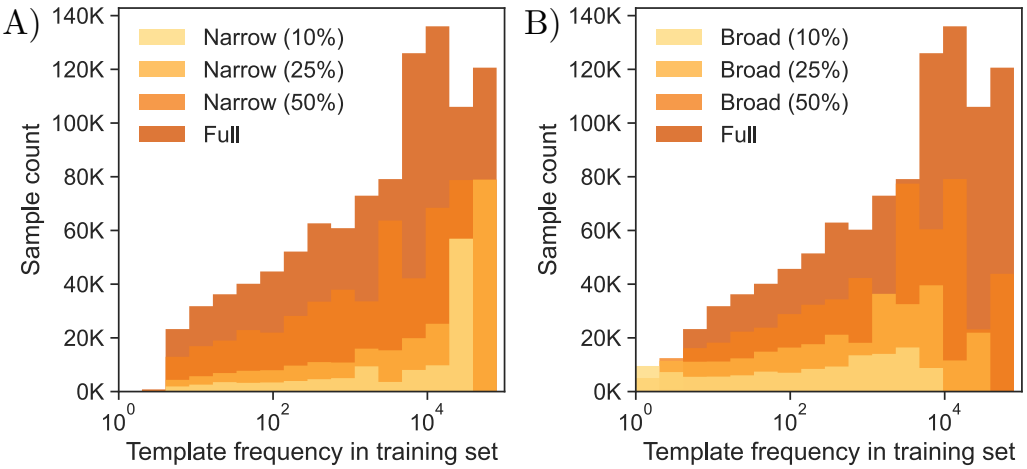

Figure S1: Histograms of sample counts in the training set by template frequency in the training set (on a log scale) across (A) the narrow split and (B) the broad split. Template frequency refers to the number of samples in the training set containing a specific template, while the sample count refers to the number of reactions with that template frequency.

S3 Training hyperparameters

The configuration files used to train the models can be found in the Github repository: (attached zip). Each model was trained using its respective repository.

The LocalRetro models were trained using default hyperparameters, and an early stopping patience of 5 epochs was implemented. The MEGAN models were trained with default hyperparameters. The RootAligned models trained on the narrow and full splits used the default hyperparameters; however, those trained on the broad split have a separate optimised set of hyperparameters. This optimisation was done to improve the suboptimal performance observed when training with the default hyperparameters.

S4 Evaluation metrics

The retrosynthesis platform *Syntheseus*[33] was used for evaluating all trained models. This involves automatically filtering out predictions which are invalid or duplicated, as shown in Figure S2. The evaluation metrics used include:

- **Top-*k* accuracy:** the proportion of test reactions with an exact ground truth match (i.e. the expected reactants from the test set) in the top-*k* predicted reactants.
- **Top-*k* round-trip accuracy:** the proportion of top-*k* predicted reactants which satisfy back-translation.[13] This is calculated by using a forward reaction model to predict the top-1 product of each set of predicted reactants, and this product is compared to the original target product.

A RootAligned forward prediction model is trained on the full USPTO-retro dataset for the purpose of back-translation and calculating top-*k* round-trip accuracy.

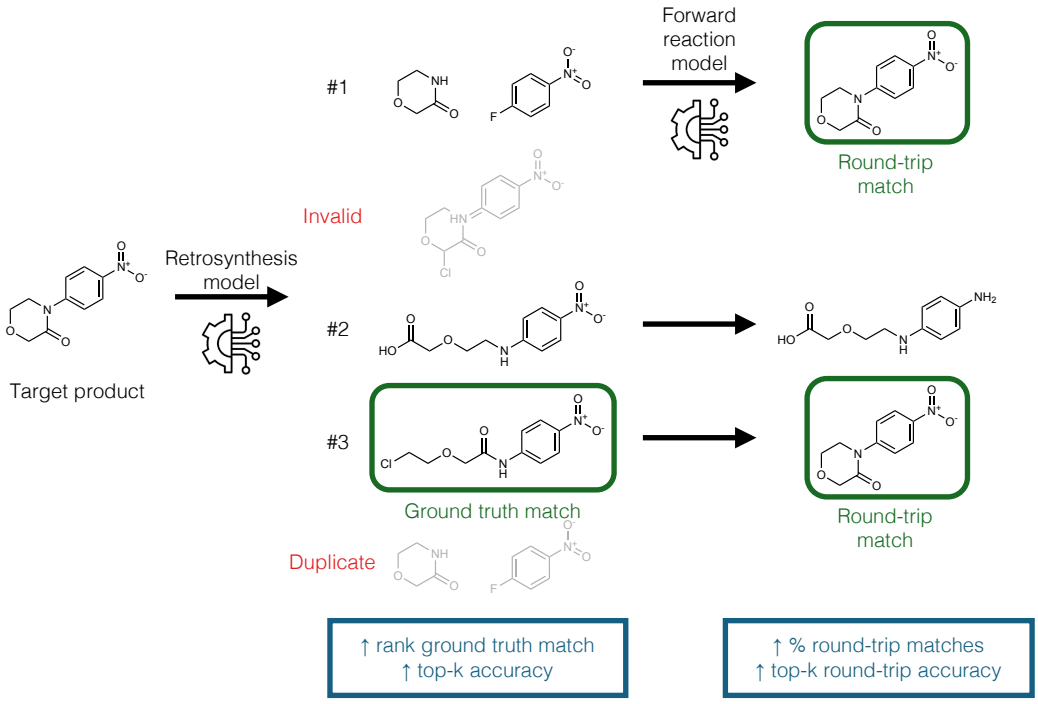

Figure S2: Visualisation of the retrosynthesis evaluation pipeline. Given a target product, the retrosynthesis model produces $n = 50$ sets of reactants, which are then filtered to contain only valid and unique predictions. The rank of the ground truth match determines the top-*k* accuracy. A forward reaction model is used to predict the product of each set of predicted reactants and this is compared to the target product to determine the top-*k* round-trip accuracy.

S5 Narrow split results

Table S1: Top-*k* accuracy and round-trip (RT) accuracy of models trained and tested on the narrow split.

		Top-*k* accuracy (%)					Top-*k* RT accuracy (%)	
Model	Training set (%)	*k*=1	5	10	20	50	1	5
LocalRetro	10	77.6	93.1	94.7	95.6	95.9	92.4	63.8
	25	74.2	91.6	93.4	94.6	95.1	93.2	70.3
	50	60.2	83.4	87.6	90.1	91.3	90.5	73.2
	90	50.5	77.4	83.1	86.7	88.5	90.8	78.0
MEGAN	10	76.9	91.8	93.3	94.6	95.1	91.8	58.6
	25	71.7	90.2	92.9	94.4	95.3	92.6	68.2
	50	53.2	78.0	83.6	87.1	89.6	83.6	61.2
	90	42.1	71.2	78.8	83.8	87.5	89.9	77.8
RootAligned	10	79.8	94.1	95.4	96.1	96.2	93.7	54.3
	25	76.2	92.8	94.5	95.4	95.5	94.5	63.1
	50	60.9	85.5	89.4	91.2	91.4	92.3	70.4
	90	49.1	78.8	85.0	87.6	87.8	92.1	75.2

S6 Broad split results

Table S2: Top-*k* accuracy and round-trip (RT) accuracy of models trained and tested on the broad split.

		Top-*k* accuracy (%)					Top-*k* RT accuracy (%)	
Model	Training set (%)	*k*=1	5	10	20	50	1	5
LocalRetro	10	42.1	71.8	79.6	84.8	87.4	88.3	75.4
	25	45.3	74.0	81.0	85.5	87.9	90.0	78.0
	50	48.3	75.9	82.1	86.3	88.3	90.5	78.2
	90	50.5	77.4	83.1	86.7	88.5	90.8	78.0
MEGAN	10	40.5	69.5	77.3	82.6	86.3	88.5	75.0
	25	40.2	69.2	77.0	82.5	86.3	88.4	75.0
	50	41.3	70.7	78.3	83.5	87.2	89.9	77.8
	90	42.1	71.2	78.8	83.8	87.5	89.9	77.8
RootAligned	10	29.8	58.9	70.5	77.7	79.1	77.7	62.3
	25	23.5	49.4	61.5	70.2	71.9	77.5	63.7
	50	18.3	43.1	54.8	62.6	63.9	67.0	59.1
	90	49.1	78.8	85.0	87.6	87.8	92.1	75.2

S7 Tanimoto similarity between training and test sets

Tanimoto similarity was used to calculate the similarity between product molecules present in the USPTO-retro training and test sets. Morgan fingerprints with default RDKit parameters were calculated for all product molecules using RDKit, and Tanimoto similarity was calculated between all pairs of fingerprints from the training and test sets. Molecules were deemed to be similar if the Tanimoto similarity score was over 0.4, and the total count of similar molecules in the training set for a given test set product molecule was collected. The counts were then divided by the total number of reactions in the training set to get the percentage similarity (Figure S3).

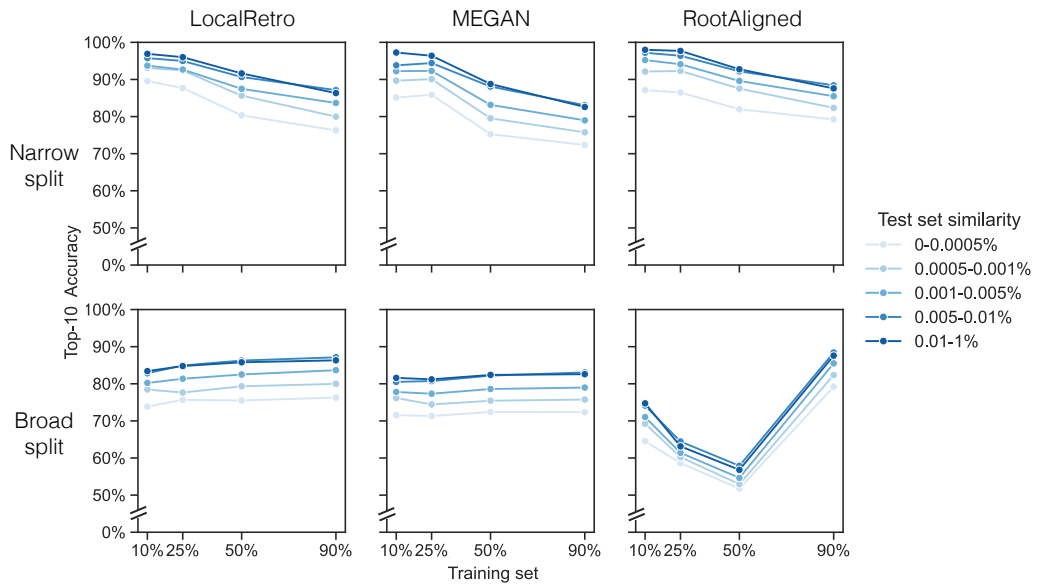

Figure S3: Top-10 accuracy of models trained and tested on splits of USPTO-retro, with test set reactions binned by their percentage similarity to the training set product molecules. Similarities are calculated as the proportion of pairwise Tanimoto similarities over a threshold score of 0.4.

S8 ID results

Tanimoto similarity was used to calculate the similarity between product molecules present in the full USPTO-retro training and the USPTO-retro test set and Pistachio ID test set, using the same method as discussed in Section S7. Figure S4 shows that Pistachio ID has a high proportion of products with 0% similarity to the training set, and is overall less similar to the training set than the USPTO-retro test set. Figure S5 shows the resulting top-10 accuracy of models tested on Pistachio ID, which suggests that the lowered accuracy of these models is due to the increased proportion of dissimilar test products.

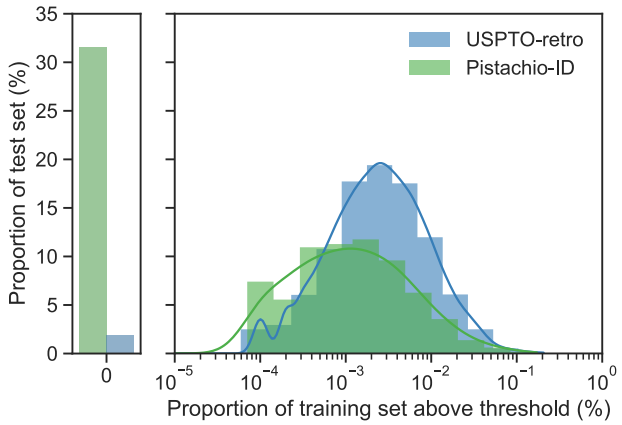

Figure S4: Similarity of the USPTO-retro (blue) and Pistachio ID (green) test sets to the USPTO-retro full training set. Similarities are calculated as the proportion of pairwise Tanimoto similarities over a threshold score of 0.4.

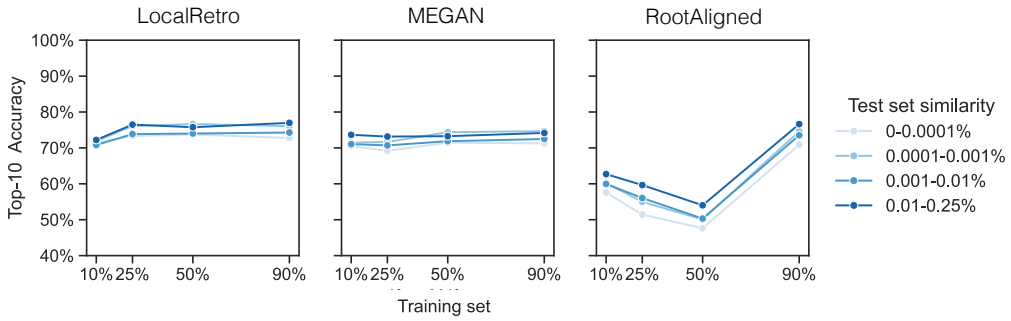

Figure S5: Top-10 accuracy of models trained on the broad splits of USPTO-retro and tested on Pistachio ID, with test set reactions binned by their percentage similarity to the training set product molecules. Similarities are calculated by the proportion of pairwise Tanimoto similarities over a threshold score of 0.4.

Table S3: Top-*k* accuracy of models trained on the broad split and tested on the Pistachio ID test set.

		Top-*k* accuracy (%)				
Model	Training set (%)	*k*=1	5	10	20	50
LocalRetro	10	35.6	62.9	71.4	77.0	80.4
	25	38.1	66.1	74.2	79.1	82.2
	50	38.7	67.0	74.7	79.9	82.9
	90	39.5	67.5	74.5	79.7	82.8
MEGAN	10	35.4	62.9	71.0	76.2	79.8
	25	35.5	62.7	70.4	75.9	79.9
	50	37.4	64.7	72.5	77.9	81.5
	90	37.6	65.1	72.9	78.1	82.2
RootAligned	10	21.6	47.6	58.9	66.2	67.7
	25	20.1	43.0	54.0	62.4	64.0
	50	17.2	39.1	49.5	56.9	58.1
	90	38.8	66.5	73.2	76.5	76.8

S9 OOD results

Table S4: Top-*k* accuracy of models trained on the narrow split and tested on the Pistachio OOD test set.

		Top-*k* accuracy (%)				
Model	Training set (%)	*k*=1	5	10	20	50
LocalRetro	10	0.08	0.20	0.20	0.22	0.23
	25	0.02	0.08	0.10	0.13	0.20
	50	0.05	0.10	0.12	0.18	0.20
	90	0.03	0.07	0.07	0.15	0.20
MEGAN	10	0.03	0.27	0.45	0.71	1.03
	25	0.03	0.22	0.48	0.93	1.40
	50	0.04	0.40	0.83	1.32	2.22
	90	0.02	0.59	1.01	1.53	2.28
RootAligned	10	0.05	0.22	0.51	0.79	0.99
	25	0.21	0.69	1.07	1.43	1.62
	50	0.10	1.14	1.84	2.51	2.71
	90	0.28	1.08	1.89	2.97	3.23