Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal



Cite this: DOI: 10.1039/d6dd00072j

# Assessing the extrapolation capability of template-free retrosynthesis models

Jonghwi Choe,†[a] Shuan Chen†[ab] and Yousung Jung *[abc]

Template-free retrosynthesis models offer the potential to extrapolate beyond established chemical reaction spaces, addressing inherent limitations of template-based approaches. However, it remains unclear whether these models can reliably predict accurate, novel, and chemically feasible pathways outside their training distribution. In this study, we rigorously assess the extrapolation ability of state-of-the-art template-free models using carefully constructed out-of-distribution (OOD) benchmarks derived from USPTO datasets. While these models can generate novel synthetic routes, their exact-match accuracy on OOD reactions is remarkably low (typically <1%). Moreover, round-trip performance (≈5–30%) is influenced by the performance of the forward model and may not fully capture some chemically reasonable predictions. Complementary manual inspection mitigates this limitation by revealing that the surrogate forward model produces false negatives, where chemically feasible reactions are incorrectly predicted as infeasible, and *vice versa* for false positives. These results underscore a critical challenge: current models may exhibit little creative extrapolation yet lack mechanisms to ensure chemical feasibility. Addressing this gap is essential for developing retrosynthesis models that are not only innovative, but also reliable for real-world synthesis planning.

Received 12th February 2026
Accepted 11th May 2026

DOI: 10.1039/d6dd00072j

rsc.li/digitaldiscovery



## Introduction

Retrosynthesis analysis aims to identify suitable precursors for synthesizing a target molecule.[1] Given the combinatorial complexity of possible synthetic routes, manually planning a single synthesis route for a target molecule often takes days to weeks for a typical chemist. To accelerate this process, computer-aided synthesis planning (CASP) systems were first introduced by E. J. Corey.[2] In recent decades, machine learning (ML)-based models have been increasingly applied to enhance retrosynthesis prediction.[3] Early ML models were trained to select proper reaction templates[4,5] from a predefined set of synthesis rules extracted from training data. While this restriction ensures experimental reliability by staying within well-established chemical domains, effectively acting as a safeguard, recent studies have claimed it as a limitation that simultaneously creates a bottleneck for exploring novel chemical spaces that deviate from existing knowledge.[6–8]

Unlike template-based methods that rely on predefined reaction rules, template-free approaches treat the retrosynthesis as either a graph editing[9] task or a SMILES generation[10–12] task. These models are, in principle, capable of proposing chemical transformations that go beyond the predefined synthesis rules. However, it remains unclear whether they can reliably predict chemically feasible synthetic routes that lie outside their training reaction data—referred to as out-of-distribution (OOD) reactions—as opposed to in-distribution (ID) reactions[13,14] that are similar to those seen during training.

To address this, RetroOOD[15] formalizes label- and covariate-shift scenarios, and reassesses state-of-the-art models primarily *via* top-*k* accuracy. Tanović *et al.*[16] analysed template-frequency skew and proposed "narrow" *versus* "broad" training-set partitions to disentangle template diversity from examples-per-template effects, evaluating performance mainly *via* top-*k* and round-trip accuracy. Beyond these task-specific studies, broader benchmarking efforts have emphasized that reported retrosynthesis performance can be highly sensitive to evaluation design itself. For example, Syntheseus provides a standardized benchmarking framework for both single-step and multi-step synthesis planning, and shows that the ranking of state-of-the-art methods can change under more carefully controlled evaluation settings.[17] Likewise, Hastedt *et al.* introduced an automated benchmarking and interpretability pipeline and showed that chemical validity, feasibility, and interpretability can differ substantially across retrosynthesis frameworks, with purely data-driven approaches often producing unfeasible or invalid predictions.[18] However, these studies still do not provide an in-depth analysis of the chemical quality of out-of-distribution (OOD) reactions generated by template-free

[a]*Department of Chemical and Biological Engineering, and Institute of Chemical Processes, Seoul National University, 1 Gwanak-ro, Seoul, South Korea*

[b]*Institute of Engineering Research, Seoul National University, 1 Gwanak-ro, Seoul, South Korea*

[c]*Interdisciplinary Program in Artificial Intelligence, Seoul National University, 1 Gwanak-ro, Seoul, South Korea. E-mail: yousung.jung@snu.ac.kr*

† These authors contributed equally to this work.

models, which is central to assessing whether such models genuinely extrapolate beyond the reaction patterns represented in the training data.

In this work, we investigate whether template-free retrosynthesis models can genuinely extrapolate to novel reaction spaces, or whether their apparent performance primarily reflects memorization of transformation patterns seen during training. Beyond exact-match accuracy, we evaluate template-free retrosynthesis models by examining the novelty, chemical validity, and synthetic feasibility of OOD reactions. In addition to standard round-trip accuracy, we conduct manual inspection to assess chemical plausibility and reveal surrogate biases intrinsic to round-trip metrics. Further comparison between language-based and graph-based template-free models further reveals that chemistry-aware inductive biases can substantially improve feasibility without sacrificing their novelty. (We note that an earlier version of this work was presented at the NeurIPS 2023 ELLIS Workshop on Molecule Discovery and archived on arXiv (arXiv:2403.03960); the present manuscript expands upon that preliminary study.)

## Materials and methods

### Dataset curation and dataset splitting

In this paper, we define in-distribution (ID) and out-of-distribution (OOD) reactions according to their associated reaction templates,[19] which specify the atom-level bond changes and transformation patterns (Fig. 1A). To rigorously evaluate extrapolation beyond the training distribution, we extracted local reaction templates (LRTs) using LocalMapper.[19] LRTs focus on the reaction core and capture only the essential bond rearrangements involved in each transformation (Fig. 1B). Compared with general reaction templates that incorporate extended molecular contexts, LRTs reduce over-fragmentation of chemically similar reactions caused by peripheral substituent variations. This property enables a cleaner separation between ID and OOD reactions with minimal template overlap. Therefore, we performed template-based dataset splitting using LRTs.

We sorted all extracted LRTs by their frequency of occurrence and selected the most frequent templates that collectively account for 80% of all reactions. Reactions associated with these templates were assigned to the training set and considered as ID reactions, representing well-established synthetic knowledge. The remaining 20% of reactions, corresponding to less frequent and often more diverse templates, were reserved for validation and testing and treated as OOD reactions. Templates appearing exclusively in these subsets were designated as test templates.

Following this protocol, the USPTO-50k[20] split contains 39 982/5016/5018 reactions for training, validation, and test sets, associated with 45/40/835 unique templates, respectively. Similarly, the USPTO-480k[20] split consists of 383 784/46 519/48 730 reactions using 232/667/19 321 templates.[21] The pronounced imbalance between reaction counts and template diversity reflects the intrinsically long-tailed distribution of chemical reaction space and establishes a stringent benchmark for OOD extrapolation. The statistics and distributions of the curated data can be found in Table S1 and Fig. S1.

### Evaluation metrics

We assessed the performance of template-free retrosynthesis models using four metrics designed to capture different facets

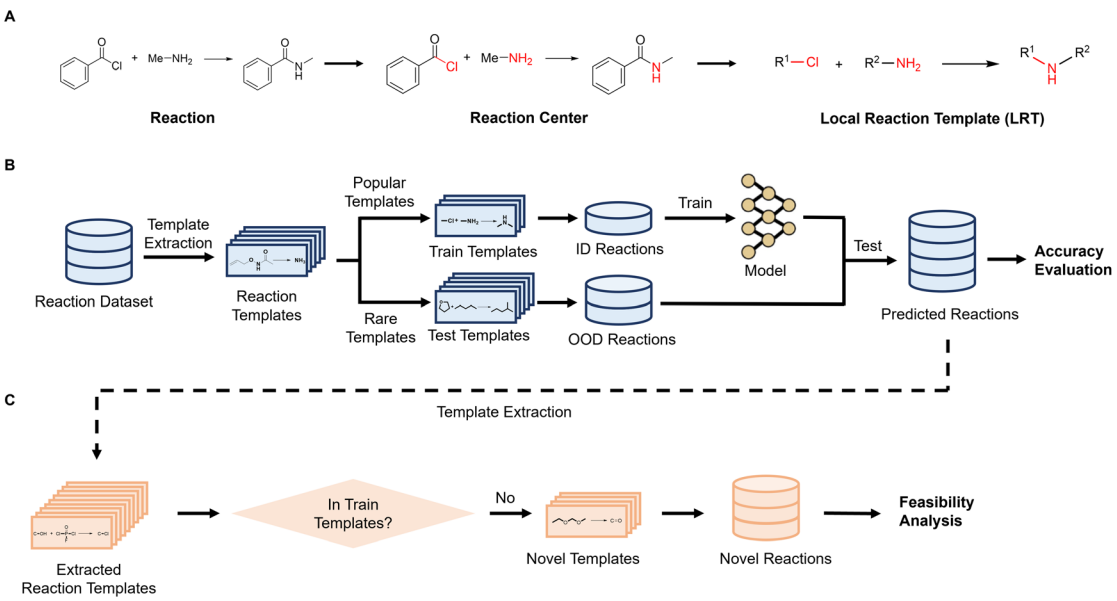


Fig. 1 Workflow for curating in-distribution (ID) and out-of-distribution (OOD) reaction datasets to evaluate template-free retrosynthesis models. (A) Illustration of the process of extracting a reaction template from a specific chemical transformation. (B) Reaction templates are extracted from the reaction dataset to form unique reaction templates. The templates with the highest popularity are selected as the ID training set, while the remaining templates are designated as OOD for validation and testing. (C) The decision tree for defining novel reactions. The process involves extracting reaction templates from the predicted reactions and checking whether the extracted templates appear in the training set.

of model capability: exact-match accuracy, reaction validity, reaction novelty, and round-trip accuracy.

Exact-match accuracy measures how accurately a model predicts reactants by comparing the predicted reactant set against the ground-truth reactants after SMILES canonicalization using RDKit.[22] For each target product $P_i$, the retrosynthesis model produces a ranked list of $K$ candidate reactant sets $\{R_i^{(k)}\}_{k=1}^{K}$. We compute top-$K$ exact-match accuracy by checking whether any of the top-$K$ candidates matches the ground-truth reactants.

Reaction validity evaluates whether the predicted reactions can be successfully converted into molecules using RDKit[22] and if the resulting reaction satisfies atom balance with respect to the target product $P_i$. We report the fraction of valid predictions among the top-$K$ candidates.

Reaction novelty evaluates whether a predicted reaction corresponds to a transformation unseen in the training set. We analysed the novelty of each prediction generated by each model by extracting their reaction templates. Predicted reactions whose extracted templates were distinct from the training templates were defined as novel reactions and subjected to further feasibility analysis (Fig. 1C).

Round-trip accuracy evaluates the feasibility of predicted retrosynthetic outputs using a reaction outcome prediction model (surrogate forward model) that predicts product candidates from the predicted reactants.[23] For each target $P_i$, each candidate reactant set $R^{(k)}{}_i$ is passed to the surrogate model $f_{\text{surr}}$, which returns the top-$n$ product predictions $\{\hat{P}_{i,k}^{(j)}\}_{j=1}^{n}$ A prediction is considered cycle-consistent if $P_i$ appears among these outputs, and round-trip accuracy is computed as the fraction of targets for which at least one of the top-$K$ candidates is cycle-consistent. After comparing the performance of different forward synthesis prediction models, including Transformer,[24] Chemformer,[10] MEGAN,[9] and LocalTransform[25] on the USPTO-480k dataset, we selected LocalTransform[25] as the surrogate model in this work (Table S2 and S3) as it demonstrated the best performance across all splitting regimes considered over other alternatives. In addition, to avoid architectural bias, we deliberately chose surrogate models with different architectures from the inspected template-free models, ensuring that the evaluation does not favour models with similar inductive biases.

## Results and discussion

In this study, we implemented and evaluated four template-free models, divided into two categories based on their input representation: language-based models and graph-based models. For language models, we selected the Transformer[24] and its pretrained variant, Chemformer.[10] These models use a transformer architecture to capture complex sequence dependencies and are well-suited for processing SMILES-based representations of molecules. For graph-based models, we used MEGAN[9] and GraphRetro,[6] both of which are designed to operate on molecular graph representations. By learning chemistry from molecular graphs, these models learn molecular structures at the atomic and bond level, enabling them to predict retrosynthesis pathways through direct manipulation of molecular graphs. These models were selected as archetypal baselines to cover the major distinct mechanisms of retrosynthetic prediction currently in use.

Table 1 Top-$k$ exact-match accuracy (%) on USPTO-50k and USPTO-480k

| Model | USPTO-50k | | | | USPTO-480k | | | |
|---|---|---|---|---|---|---|---|---|
| Top-$k$ | $k=1$ | 3 | 5 | 10 | $k=1$ | 3 | 5 | 10 |
| Transformer[24] | 0.64 | 1.51 | 1.91 | 2.01 | 0.17 | **0.47** | **0.64** | 0.74 |
| Chemformer[10] | **1.05** | **2.13** | **2.47** | 2.84 | **0.22** | **0.47** | 0.60 | **0.83** |
| MEGAN[9] | 0.08 | 0.71 | 1.04 | 2.35 | 0.12 | 0.31 | 0.46 | 0.80 |
| GraphRetro[6] | 0.58 | 1.22 | 1.71 | 2.36 | 0.08 | 0.15 | 0.20 | 0.31 |

### Exact-match accuracy

The exact-match accuracy of the evaluated models on the OOD reactions shown in Table 1 is remarkably low across all models compared to the random split shown in literature (Table S4). On the USPTO-50k test reactions, top-1 exact-match accuracies struggled to exceed 1%, with Chemformer achieving 1.05% and Transformer 0.64%. Even considering the top-10 predictions, accuracy remained below 3% for all the template-free models. Performance deteriorated further on the larger and more diverse USPTO-480k test reactions, where top-1 accuracies dropped to around 0.1–0.2% for all the template-free models. Interestingly, we observed that language-based models generally performed better than graph-based models in this evaluation.

### Validity and novelty of predicted reactions

Despite the low exact-match accuracy of the evaluated models on OOD test reactions, which frequently falls below 1% (Table 1), we found they were still capable of generating novel reactions. To explore this further, we analyzed the novelty and validity of predictions for each model, as shown in Fig. 2. The $x$-axis corresponds to the top-$k$ prediction rank ($k = 1$–10) and the $y$-axis indicates the percentage of predictions categorized as train, novel, or invalid templates. Template-free models, particularly the Transformer, demonstrated the ability to propose novel reaction templates, with novelty rates reaching up to ∼30% among the top-10 predictions on the USPTO-50k dataset (Fig. 2A).

Our results reveal a consistent novelty–validity trade-off whose origin depends strongly on model inductive biases. Overall, higher novelty tended to coincide with a larger fraction of invalid outputs, reflecting a trade-off between exploration and chemically valid generation. However, this relationship is not uniform across architectures: GraphRetro produces a level of novelty comparable to the Transformer while maintaining substantially higher validity, suggesting that chemistry-aware inductive biases (*e.g.*, functional-group-aware edits) can mitigate validity loss even when novelty remains high. For end-to-end models such as Transformer and MEGAN, limited validity on smaller datasets suggests that additional data is required to

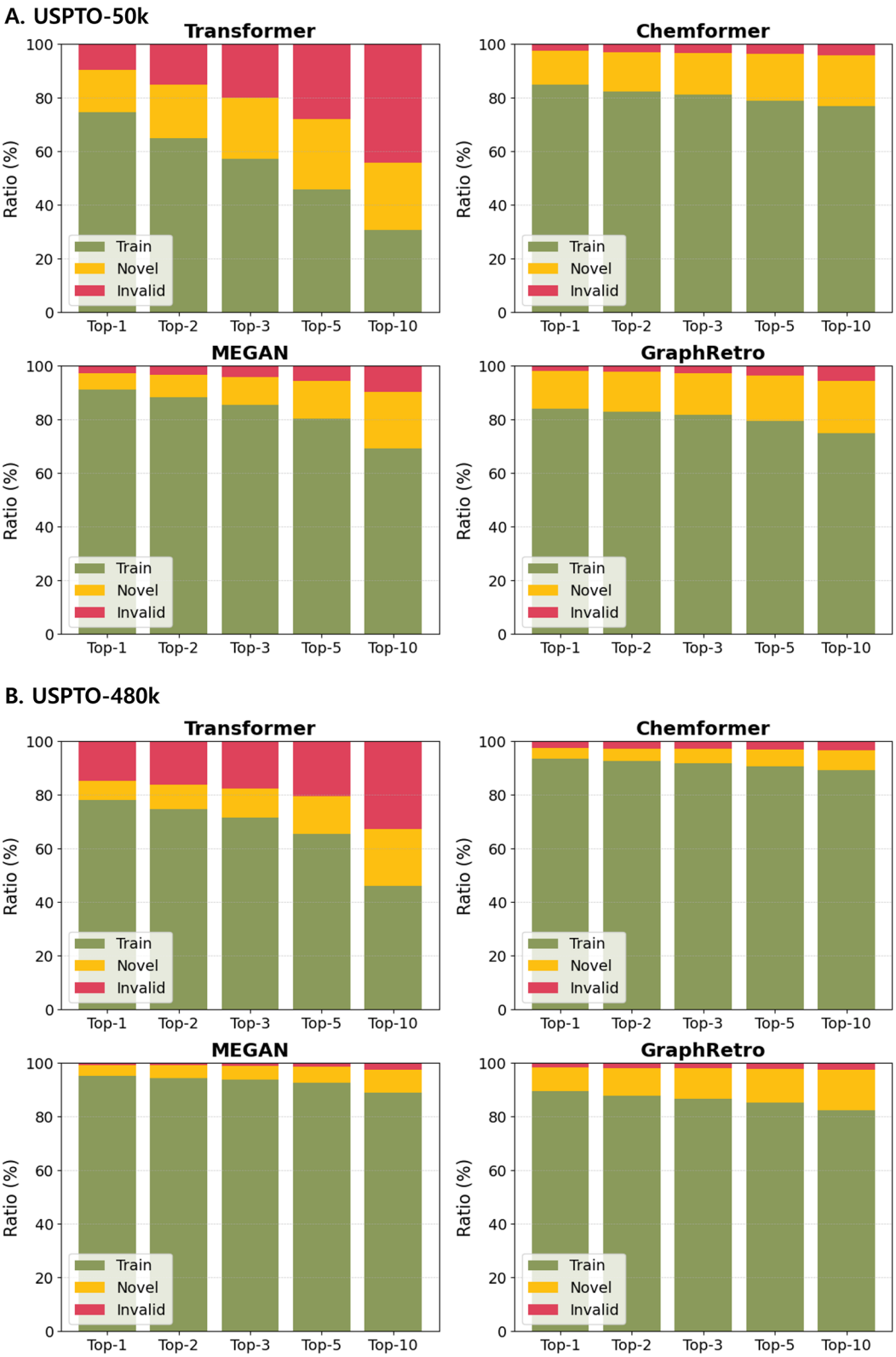


Fig. 2 Distribution of extracted reaction template types for the evaluated template-free models on (A) USPTO-50k and (B) USPTO-480k dataset across top-1 to top-10 predictions.

Table 2 Top-$k$ round-trip accuracy (%) on USPTO-480k

| Model | | | | |
|---|---|---|---|---|
| Top-$k$ | $k = 1$ | 2 | 3 | 5 |
| Transformer[22] | 5.95 | 5.95 | 5.94 | 5.91 |
| Chemformer[10] | 8.25 | 8.87 | 9.25 | 9.74 |
| MEGAN[9] | 17.61 | 17.75 | 17.90 | 17.95 |
| GraphRetro[6] | **29.15** | **28.80** | **28.51** | **28.05** |

reliably learn chemically plausible reaction representations. In contrast, Chemformer and GraphRetro maintain high validity even at smaller scales, indicating that their behaviour is not governed by data insufficiency.

Notably, for these chemistry-aware models, increasing dataset size primarily reduces prediction novelty rather than improving validity. We interpret this effect not as overfitting or correction of undertrained behaviour, but as an implicit, data-driven regularization toward canonical reaction patterns that are repeatedly reinforced during training. As larger datasets expose models to a broader yet more unevenly distributed set of reaction templates, predictions become increasingly concentrated around statistically dominant transformation modes.

At the same time, larger reaction corpus such as USPTO-480k also exhibit a more fragmented template landscape, in which unseen or weakly represented transformations are structurally farther from the dominant training distribution. As a result, extrapolation to unseen templates becomes more challenging despite increased data volume. Together, these observations suggest that the observed trade-off between novelty and validity arises from the interaction between dataset structure and model inductive biases, rather than from insufficient training on smaller datasets.

## The chemical feasibility of novel reactions

Finally, we evaluated the chemical feasibility of retrosynthesis predictions, particularly the novel ones selected from the predictions, using round-trip accuracy as a feasibility metric and complemented it with manual inspection. The round-trip accuracy for these novel reactions on USPTO-480k shown in Table 2 revealed that GraphRetro achieved significantly higher round-trip accuracy compared to its exact-match performance and the round-trip accuracies of other models. Analysis of the correlation between round-trip accuracy and novel templates suggested this performance stemmed from the focused and highly feasible reaction templates recognizable by the surrogate model. The round-trip accuracy for these novel reactions on USPTO-50k is given in Table S5.

As shown in Fig. 3, cumulative average round-trip (RT) accuracy shows a strong correlation with the template popularity. For all models, this accuracy is highest for reactions associated with frequently used templates, but it decreases markedly as the evaluation expands toward less common templates. While language-based models (Transformer and Chemformer) show a continuous decline in accuracy when considering the predictions corresponding to rarer templates, graph-based models (MEGAN and GraphRetro) show a more stable profile.

This trend is consistent with Table S6, where language-based models show a steep drop in average RT accuracy as the template rank range moves from popular templates (around 30% for top 100 templates) toward the rarest ones (less than 5%

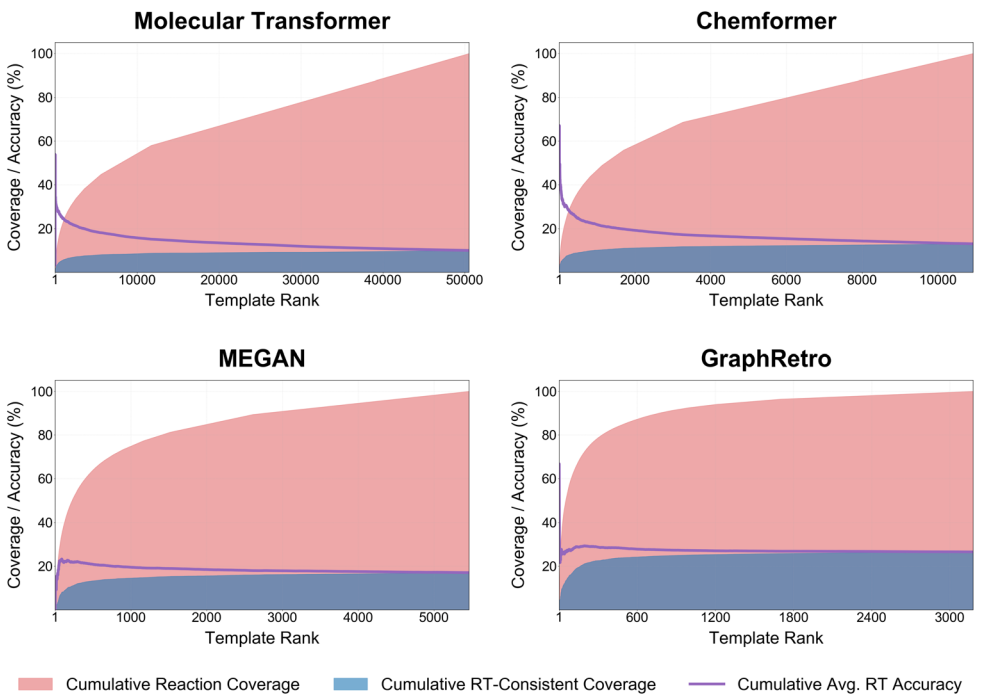


Fig. 3 Cumulative reaction coverage, cumulative round-trip consistent coverage and cumulative average round-trip accuracy as a function of template rank of the novel reactions generated by the template-free models trained on USPTO-480k dataset.

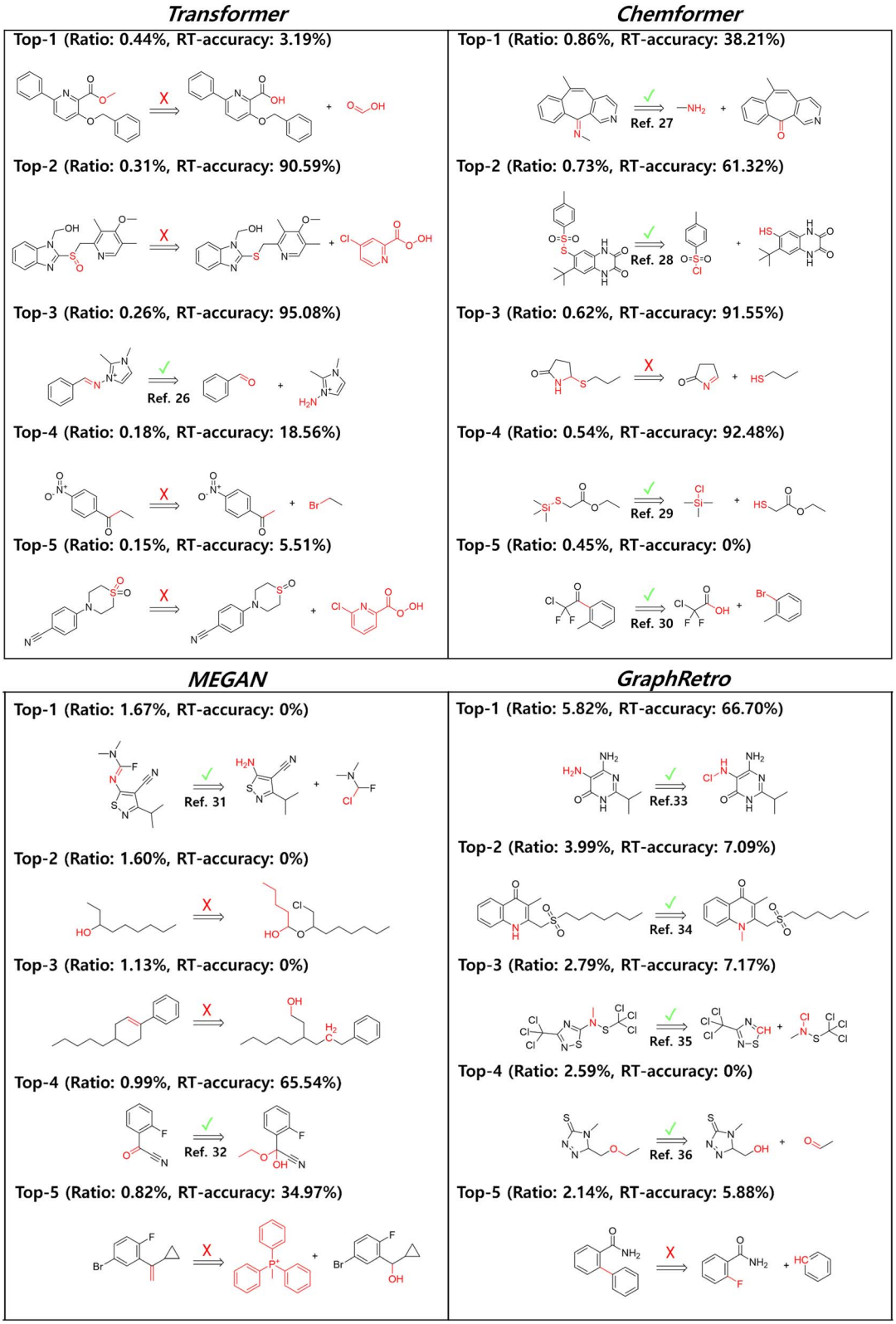


Fig. 4 Representative novel reactions generated by template-free models based on the most frequently found reaction templates. The templates are listed in descending order of frequency within each model, from top to bottom. The specific atoms and bonds corresponding to the local reaction templates (LRTs) are highlighted in red. The ratio of each novel reaction template within the entire dataset and the round-trip accuracy (RT-accuracy) for each novel reaction template are provided. Additionally, the chemical feasibility of each reaction is manually inspected and indicated above the arrow (check for feasible and cross for infeasible).

for templates ranked over 10 000), while nearly half (45.7%) of Transformer's total predictions fall into the rarest template range. In contrast, graph-based models exhibit highly concentrated template usage and a comparatively stable average RT profile, with >20% RT accuracy for popular (top 100) templates to ∼10% RT accuracy for the rarest template range. Similar analysis for the USPTO-50k dataset can be found in Fig. S2 and Table S6.

Despite the computationally evaluated metrics provided in Table 2 and Fig. 3, interpreting round-trip accuracy requires caution, as it serves more as a useful heuristic for evaluating the self-consistency of a model rather than a definitive metric for true chemical feasibility. For instance, GraphRetro shows relatively high RT accuracy but extremely low exact-match accuracy, indicating that while the surrogate model (used for round-trip evaluation) recognizes the generated reactions, these do not necessarily correspond to the literature-recorded reactions. To better assess chemical novelty and feasibility, we manually examined the reactions associated with the five most frequently proposed novel templates (Fig. 4). Using SciFinder (https://scifinder-n.cas.org/), we systematically evaluated each reaction;[26–36] if a similar transformation is found in the database, we consider it chemically feasible; otherwise infeasible.

Although the evaluated models can generate novel reaction templates with high round-trip (RT) accuracy, our manual inspection found that most of these novel templates reflect only limited chemical novelty. In many cases, the templates are not exact matches to predefined ones, but they still follow similar patterns in the training set. For instance, the Transformer's top-3 template[24] (95.08% RT accuracy) corresponds to the formation of an imine through the reaction of an aldehyde with a primary amine. This is a fundamental transformation in organic chemistry but absent in the training set. Similarly, Chemformer's top-4 template[27] (92.48% RT accuracy) represents a TMS protection reaction where a thiol nucleophilically attacks the silicon atom of trimethylsilyl chloride (TMS-Cl), displacing a chloride ion. This reactivity directly parallels reactions in the training dataset that use *tert*-butyldimethylsilyl chloride (TBS-Cl) as their reagents, where an oxygen atom attacks a silicon center. Both cases share the common motif of heteroatom (sulfur or oxygen) attack on silicon, demonstrating that the model has transferred the concept of silyl protection from oxygen to sulfur. Overall, these examples show that the models' novel templates with high RT accuracy are often conservative, mechanism-preserving extrapolations rather than genuinely new chemical reactivity.

For reaction feasibility, manual inspection reveals that high RT accuracy does not guarantee chemical validity. A representative example is Transformer's top-2 template, which achieved a high RT accuracy of 90.59%. This template proposes the synthesis of a sulfoxide product *via* the oxidation of a sulfide precursor. While sulfide oxidation as a general reaction type is feasible, the model predicted a specific, unconventional chlorinated pyridine–peroxyacid as the oxidant. This reagent appears to be a hallucinated analogue of mCPBA (*meta*-chloroperoxybenzoic acid) lacking chemical precedent or stability, yet the surrogate model accepted it based on its structural similarity to known oxidants. Similarly, Chemformer's top-3 template (91.55% RT accuracy) suggests synthesizing a sulfide-substituted lactam from an $\alpha$, $\beta$-unsaturated lactam and a thiol *via* a Michael addition. However, the required starting material, which includes a strained 5-membered unsaturated lactam, is thermodynamically unstable and likely inaccessible as a stable reagent. These cases demonstrate that surrogate models could misclassify infeasible reactions as correct because they resemble common reaction patterns (*e.g.*, oxidation, conjugate addition) in the training set.

In contrast, manual inspection also indicates that round-trip evaluation can penalize chemically reasonable OOD predictions when they are not covered by the surrogate model's learned chemical space. For example, Chemformer's top-5 template[28] (0% RT accuracy) represents acylation reactions for aryl ketone synthesis. These reactions typically proceed *via* metal–halogen exchange to generate an organometallic nucleophile, which attacks the electrophilic carbonyl carbon of the acid derivative to yield a ketone. Similarly, MEGAN's top-1 template[29] (0% RT accuracy) corresponds to the reaction of an amine with a chloro-activated electrophile, converting the amino group into an amidine derivative. In this transformation, the amine acts as a nucleophile and attacks the electrophilic carbon bearing the chloride leaving group, followed by chloride displacement and proton transfer to furnish the C=N bond, with HCl (or its salt) as a byproduct.

Overall, manual inspection suggests that round-trip evaluation does not clearly separate different types of errors. In particular, surrogate models can penalize correct OOD predictions (*i.e.* false negatives) while endorsing incorrect ones (*i.e.* false positives). At the same time, the novel yet valid predictions were mostly simple extensions of known reactions, such as replacing the atoms of known functional groups, indicating that truly new kinds of chemistry extrapolated by template-free models are rare in practice. These findings highlight the limitations of relying solely on round-trip accuracy to assess prediction quality, particularly for OOD reactions. Also, these results emphasize that while round-trip accuracy is a practical tool for high-throughput screening of model logic, it possesses inherent limitations as a true feasibility metric, particularly in out-of-distribution (OOD) spaces where surrogate bias is most pronounced.

Despite the surrogate model's shortcomings, our inspection reveals that Chemformer and GraphRetro tend to produce more chemically feasible predictions than other template-free models, respectively. This observation suggests the role of inductive biases, such as SMILES pretraining or functional-group-based molecule generation, may ensure the models' extrapolative behaviours within the bounds of chemical realism while performing relatively reasonable extrapolation.

## Conclusions

In summary, we present a feasibility-aware evaluation of the extrapolation capability of template-free retrosynthesis models under explicitly defined out-of-distribution (OOD) settings

using reaction templates. By analysing the exact-match accuracy, novelty, validity, and chemical plausibility of generated reactions across multiple datasets and model classes, we show that while models designed with inductive bias through pre-training or chemical-aware generation constraints show marginal meaningful extrapolation capabilities in the chemical space, apparent extrapolation does not reliably correspond to meaningful chemical discovery.

Our findings establish important boundaries for interpreting the prediction outputs of template-free retrosynthesis models. While these approaches can produce novel reactions, such novelty alone is insufficient evidence of genuine exploration of creative and feasible reaction space. We therefore emphasize that claims of discovery or extrapolation of future template-free models should be supported by chemically grounded evaluation beyond accuracy-based metrics and evidence.

More broadly, this work highlights the need for alignment between computational evaluation protocols and chemical reasoning and provides a foundation for more responsible interpretation of AI-generated synthesis proposals.

Recent single-step retrosynthesis studies have introduced newer generative architectures, including Markov-bridge, diffusion-based, GFlowNet-based, and flow-matching approaches, as well as ensemble frameworks that combine complementary inductive biases.[36–40] Although these models were not the present study, extending feasibility-aware OOD evaluation to such architectures will be an important direction for future work. In particular, it will be valuable to test whether improved top-*k* or round-trip performance in these newer frameworks translates into genuinely chemically plausible, feasible, and extrapolative predictions under explicit OOD settings.

## Author contributions

Jonghwi Choe – data curation, methodology, formal analysis, software, validation, investigation, visualization, and writing original draft. Shuan Chen – conceptualization, data curation, methodology, project administration, supervision, investigation, visualization, and writing – original draft. Yousung Jung – conceptualization, supervision, funding acquisition, resources, and writing – review & editing.

## Conflicts of interest

There are no conflicts to declare.

## Data availability

The supplementary information (SI) is available free of charge on the RSC Publications website and includes supplementary methods, experimental details, and additional figures and tables. The datasets and prediction results used in this study are publicly available on Figshare at https://doi.org/10.6084/m9.figshare.30843134. The code and instructions for reproducing the analyses are archived and permanently available on Zenodo at https://doi.org/10.5281/zenodo.20053801, and the actively maintained repository can be found at https://github.com/snu-micc/Retrosynthesis-Extrapolation.

Supplementary information: supplementary methods, experimental details, extra figures/tables. Methods and materials, ablation study, and the additional details of the experimental results. See DOI: https://doi.org/10.1039/d6dd00072j.

## Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-00514706 and RS-2023-00283902) and the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2021-II211343).

## Notes and references

1 E. J. Corey, *Angew Chem. Int. Ed. Engl.*, 1991, **30**, 455–465.
2 E. J. Corey and W. T. Wipke, *Science*, 1969, **166**, 178–192.
3 F. Strieth-Kalthoff, F. Sandfort, M. H. S. Segler and F. Glorius, *Chem. Soc. Rev.*, 2020, **49**, 6154–6168.
4 S. Chen, J. Noh, J. Jang, S. Kim, G. H. Gu and Y. Jung, *Acc. Chem. Res.*, 2024, **57**, 1964–1972.
5 M. H. S. Segler and M. P. Waller, *Chem.–Eur. J.*, 2017, **23**, 5966–5971.
6 V. R. Somnath, C. Bunne, C. Coley, A. Krause and R. Barzilay, in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2021, vol. 34, pp. 9405–9415.
7 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.
8 Y. Deng, X. Zhao, H. Sun, Y. Chen, X. Wang, X. Xue, L. Li, J. Song, C.-Y. Hsieh, T. Hou, X. Pan, T. S. Alomar, X. Ji and X. Wang, *Nat. Commun.*, 2025, **16**, 7012.
9 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.
10 R. Irwin, S. Dimitriadis, J. He and E. J. Bjerrum, *Mach. Learn. Sci. Technol.*, 2022, **3**, 015022.
11 I. V. Tetko, P. Karpov, R. Van Deursen and G. Godin, *Nat. Commun.*, 2020, **11**, 5575.
12 B. Liu, B. Ramsundar, P. Kawthekar, J. Shi, J. Gomes, Q. Luu Nguyen, S. Ho, J. Sloane, P. Wender and V. Pande, *ACS Cent. Sci.*, 2017, **3**, 1103–1113.
13 A. S. Tadanki, H. S. P. Rao and U. D. Priyakumar, *Digit. Discovery*, 2025, **4**, 831–845.
14 J. Bradshaw, A. Zhang, B. Mahjour, D. E. Graff, M. H. S. Segler and C. W. Coley, *ACS Cent. Sci.*, 2025, **11**, 539–549.
15 Y. Yu, L. Yuan, Y. Wei, H. Gao, F. Wu, Z. Wang and X. Ye, in *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth*

*Symposium on Educational Advances in Artificial Intelligence*, AAAI Press, 2024, vol. 38, pp. 374–382.
16 S. Tanovic, E. Wieczorek and F. Duarte, *Digital Discovery*, 2026, **5**, 793–802.
17 K. Maziarz, A. Tripp, G. Liu, M. Stanley, S. Xie, P. Gaiński, P. Seidl and M. H. S. Segler, *Faraday Discuss.*, 2025, **256**, 568–586.
18 F. Hastedt, R. M. Bailey, K. Hellgardt, S. N. Yaliraki, E. A. del Rio Chanona and D. Zhang, *Digit. Discovery*, 2024, **3**, 1194–1212.
19 S. Chen, S. An, R. Babazade and Y. Jung, *Nat. Commun.*, 2024, **15**, 2250.
20 D. M. Lowe, *PhD thesis*, University of Cambridge, 2012, DOI: 10.17863/CAM.16293.
21 N. Schneider, N. Stiefl and G. A. Landrum, *J. Chem. Inf. Model.*, 2016, **56**, 2336–2346.
22 *RDKit: Open-Source Cheminformatics*, https://www.rdkit.org.
23 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.
24 P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas and A. A. Lee, *ACS Cent. Sci.*, 2019, **5**, 1572–1583.
25 S. Chen and Y. Jung, *Nat. Mach. Intell.*, 2022, **4**, 772–780.
26 Y. Tamura, H. Hayashi, Y. Nishimura and M. Ikeda, *J. Heterocycl. Chem.*, 1975, **12**, 225–230.
27 D. G. Brenner, W. Halczenko and K. L. Shepard, *J. Heterocycl. Chem.*, 1985, **22**, 555–559.
28 J.-P. Mahieu, M. Gosselet, B. Sebille and Y. Beuzard, *Synth. Commun.*, 1986, **16**, 1709–1722.
29 X. Sun, Z. Song, H. Li and C. Sun, *Chem.–Eur. J.*, 2013, **19**, 17589–17594.
30 M. Kajino, A. Hasuoka and H. Nishida, *World Intellectual Property Organization Pat.*, WO2007026916A1, 2007.
31 S. Kobayashi, R. Akiyama and H. Kitagawa, *J. Comb. Chem.*, 2000, **2**, 438–440.
32 G. Rueedi, P. Panchaud, A. Friedli, J.-L. Specklin, C. Hubschwerlen, A.-C. Blumstein, P. Caspers, M. Enderlin-Paput, L. Jacob, C. Kohl, H. H. Locher, P. Pfaff, C. Schmitt, P. Seiler and D. Ritz, *J. Med. Chem.*, 2024, **67**, 9465–9484.
33 G. S. Borovikova, E. S. Levchenko, E. M. Dorokhova, Zh. O. Khim., B. M. Baughman, P. Jake Slavish, R. M. DuBois, V. A. Boyd, S. W. White and T. R. Webb, *ACS Chem. Biol.*, 2012, **7**, 526–534.
34 K. G. Liu, K. L. Olszewski, J.-I. Kim, M. V. Poyurovsky, K. Morris, X. Yu and C. Lamarque, *World Intellectual Property Organization Pat.*, WO2020005935, 2020.
35 Y. Fujii, H. Furugaki, S. Yano and K. Kita, *Chem. Lett.*, 2000, **29**, 926–927.
36 I. Igashov, A. Schneuing, M. Segler, M. M. Bronstein and B. Correia, in *Proc. Int. Conf. Learn. Represent.* (ICLR), 2024.
37 S. Current, Z. Chen, D. Adu-Ampratwum, X. Ning and S. Parthasarathy, *arXiv*, 2025, preprint arXiv:2505.23721, DOI: 10.48550/arXiv.2505.23721.
38 P. Gaiński, M. Koziarski, K. Maziarz, M. Segler, J. Tabor and M. Śmieja, *arXiv*, 2025, preprint arXiv:2406.18739, DOI: 10.48550/arXiv.2406.18739.
39 R. Yadav, Q. Yan, G. Wolf, A. J. Bose and R. Liao, *arXiv*, 2026, preprint arXiv:2506.04439, DOI: 10.48550/arXiv.2506.04439.
40 K. Maziarz, G. Liu, H. Misztela, A. Tripp, J. Li, A. Kornev, P. Gaiński, H. Hoefling, M. Fortunato, R. Gupta and M. Segler, *arXiv*, 2025, preprint arXiv:2412.05269, DOI: 10.48550/arXiv.2412.05269.