Transfer Learning for Heterocycle Retrosynthesis

Ewa Wieczorek,[†,‡] Joshua W. Sin,[†,§] Matthew T. O. Holland,[‡,†] Liam Wilbraham,[¶] Victor Sebastián Pérez,[¶] Anthony Bradley,[¶] Dominik Miketa,[¶] Paul E. Brennan,[‡] and Fernanda Duarte[*,†]

†*Chemistry Research Laboratory, 12 Mansfield Road, Oxford, OX1 3TA*

‡*Alzheimer's Research UK Oxford Drug Discovery Institute, Centre for Artificial Intelligence in Precision Medicine, Centre for Medicines Discovery, Nuffield Department of Medicine, University of Oxford, Oxford OX3 7FZ, U.K.*

¶*Exscientia plc, The Schrödinger Building Oxford Science Park, Oxford OX4 4GE, U.K.*

§*Current address: Laboratory of Artificial Chemical Intelligence (LIAC), Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland*

E-mail: fernanda.duartegonzalez@chem.ox.ac.uk

Abstract

Heterocycles are important scaffolds in medicinal chemistry that can be used to modulate the binding mode as well as pharmacokinetic properties of drugs. The importance of heterocycles has been exemplified by the publication of numerous datasets containing heterocyclic rings and their properties. However, those datasets lack synthetic routes towards the published heterocycles. Consequently, novel and uncommon heterocycles are not easily synthetically accessible. While retrosynthetic prediction models could usually be used to assist synthetic chemists, their performance is poor for heterocycle formation reactions due to low data availability. In this work, we compare the use of four different transfer learning methods to overcome the low data availability problem and improve the performance of retrosynthesis prediction models for ring-breaking disconnections. The *mixed fine-tuned* model achieves top-1 accuracy of 36.5% and, moreover, 62.1% of its predictions are chemically valid and ring-breaking. Furthermore, we demonstrate the applicability of the *mixed fine-tuned* model in drug discovery by recreating synthetic routes towards two drug-like targets published last year. Finally, we introduce a method for further fine-tuning the model as new reaction data becomes available.

Introduction

Retrosynthesis, the iterative process of breaking down a molecule into simpler precursors, has traditionally been the domain of expert organic chemists.[1] However, even for experienced chemists, this approach presents challenges due to the vast chemical space of potential transformations and the incomplete understanding of reaction mechanisms and their dependence on reaction conditions. To overcome these challenges, efforts have persisted since the 1970s to integrate computation into synthetic planning by developing Computer-Aided Synthesis Planning (CASP) tools, with one of the earliest examples being the Logic and Heuristics Applied to Synthetic Analysis (LHASA) by Pensak *et al.*[2] Despite numerous attempts, CASP tools had limited success until recently.[3]

Significant progress in CASP tools has occurred in the last decade,[4] driven by advances in machine learning (ML) methodologies and the availability of chemical datasets, such as Lowe's US Patents Office (USPTO) reaction extracts.[5] Following the seminal work by Segler *et al.*[6] on the use of neural networks and search algorithms in the 3N-MCTS CASP tool, there has been a proliferation of new ML models for retrosynthesis prediction. These models can be broadly classified into two categories: template-based[6–9] and template-free methods.[10–14] Template-based methods rely on predefined reaction rules extracted from datasets, where algorithms match a target molecule with predefined templates. CASP tools utilising such models include ASKCOS,[7] AiZynthFinder,[8] and Retro*.[9] In contrast, template-free methods, such as graph-based[13,14] or sequence-to-sequence[10–12] (seq2seq) approaches, bypass the use of an external template database by directly training on raw reaction data. While early seq2seq models were based on long-short term memory networks (LSTMs),[12] the breakthrough in seq2seq reaction prediction came when Schwaller *et al.* applied the transformer model[15] commonly used in Natural Language Processing (NLP) for forward reaction prediction, creating the Molecular Transformer.[16] In this case, reaction prediction is treated as a translation problem using Simplified Molecular Input Line Entry System (SMILES)[17] strings to represent the chemical transformation. Since then, seq2seq retrosynthesis prediction models have shown high accuracies on public benchmarking test sets, with the Augumented Transformer[11] achieving 46.2% top-1 reactant accuracy on the USPTO-full dataset.[18] The recent developments have led to transformers emerging as a premier architecture for retrosynthesis planning utilised in platforms such as IBM RXN.[10]

Despite the high efficacy of CASP tools on general reaction datasets, predicting retrosynthetic disconnections for specific, less prevalent areas of chemistry remains a significant challenge due to dataset bias.[19,20] Heterocycle formation reactions are an example of under-represented reaction classes, accounting for only 5% of reported chemical reactions in the USPTO dataset.[19] However, heterocycles are key motifs in drug design, with 85% of the top 200 best-selling small molecule drugs of 2022 featuring heterocyclic rings,[21] where they act as bioisosteric replacements improving pharmacokinetic and toxicological properties of drug targets.[22–24] Although numerous virtual libraries document theoretically synthesisable heterocyclic scaffolds,[25] synthetic pathways towards novel heterocycles remain underexplored, with the focus in medicinal chemistry being on ring derivatisation rather than ring formation.[26,27] Enhancing the prediction capacity of CASP tools for reactions forming these crucial chemical motifs could stimulate the exploration of novel heterocyclic molecules, potentially fuelling new therapeutic breakthroughs.

This work aims to enhance the performance of CASP tools for heterocycle retrosynthesis by combining seq2seq models and transfer learning, where knowledge learned from one

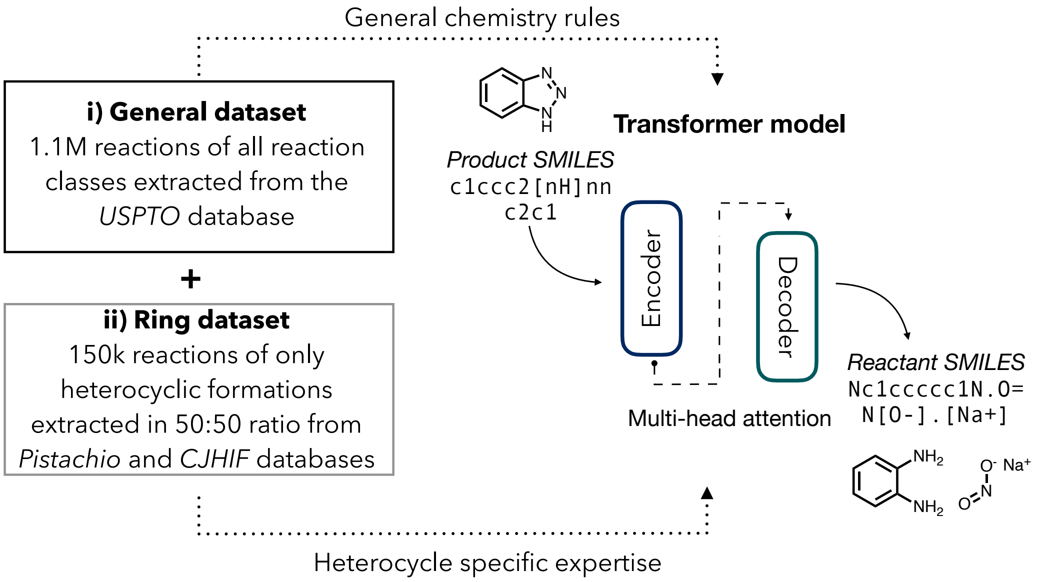
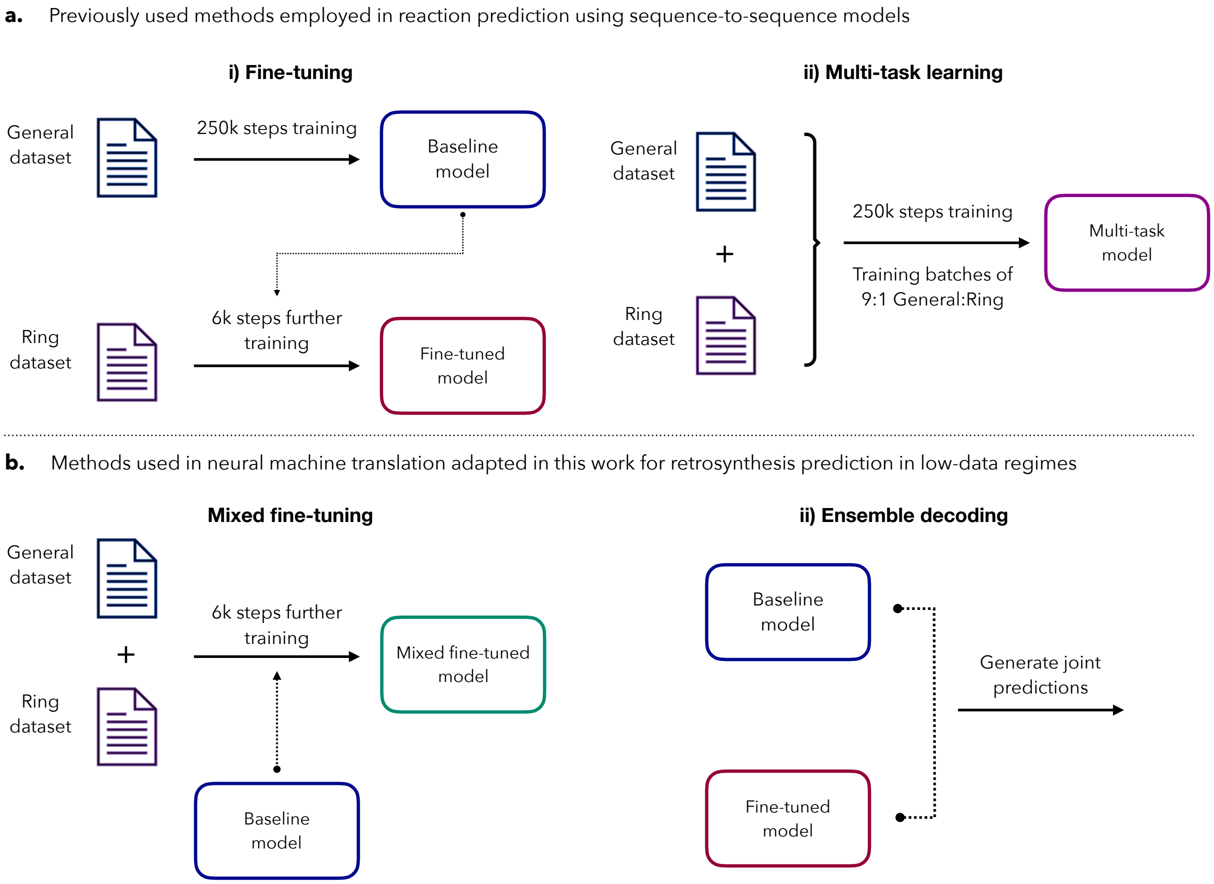

General chemistry rules
i) General dataset
1.1M reactions of all reaction classes extracted from the *USPTO* database
+
ii) Ring dataset
150k reactions of only heterocyclic formations extracted in 50:50 ratio from *Pistachio* and *CJHIF* databases
Heterocycle specific expertise
Transformer model
Product SMILES
c1ccc2[nH]nn c2c1
Encoder
Decoder
Multi-head attention
Reactant SMILES
Nc1ccccc1N.O= N[O-].[Na+]
N
N
N
H
NH2
NH2
O- Na+
N
O

Figure 1: Utilisation of general (i) and domain-specific (ii) data in transfer learning approaches for sequence-to-sequence retrosynthesis prediction.

task is used to boost the performance on a related task (Figure 1). Two transfer learning approaches, fine-tuning and multi-task learning, have been previously successfully applied for the forward reaction prediction of carbohydrate reactions[20] and Heck reactions,[28] as well as forward and retrosynthesis prediction of enzymatic reactions[29,30] (Figure 2A). However, both of those approaches come with limitations. For example, in the reported examples, fine-tuning has shown a quick training time and increased accuracy for reactions of interest but exhibited a large drop in performance for other, more common reaction types. Conversely, multi-task learning maintained good performance on all reaction types but required longer training time, making it less suitable for frequent retraining as new reaction data becomes available. To address these limitations, here we evaluate mixed fine-tuning[31] and ensemble decoding,[32] which have previously proven effective in language translation tasks but have not been used in retrosynthesis prediction (Figure 2B). We compare those methods to the template-based approach reported by Thakkar *et al.* specifically for ring-forming reaction prediction in the 'Ring Breaker'.[19] We use two datasets to train these models: a large dataset of all reaction types based on USPTO ("*General*") and a smaller dataset of just heterocycle formations ("*Ring*"). We show that the *mixed fine-tuned* model is the best for use in multi-step retrosynthesis, with 10% increase in accuracy over the baseline for heterocycle formations and only a marginal decrease in performance for other reactions. We then further demonstrate its applicability by predicting retrosynthetic routes towards two recently published heterocycle-containing drug-like targets. Finally, we test the *mixed-fine tuned* model on recently developed heterocycle formations and demonstrate how it can be further fine-tuned to improve its accuracy on this new data.



a. Previously used methods employed in reaction prediction using sequence-to-sequence models
i) Fine-tuning
General dataset
250k steps training
Baseline model
Ring dataset
6k steps further training
Fine-tuned model
ii) Multi-task learning
General dataset
+
Ring dataset
250k steps training
Training batches of 9:1 General:Ring
Multi-task model
b. Methods used in neural machine translation adapted in this work for retrosynthesis prediction in low-data regimes
Mixed fine-tuning
General dataset
+
Ring dataset
6k steps further training
Mixed fine-tuned model
Baseline model
ii) Ensemble decoding
Baseline model
Fine-tuned model
Generate joint predictions

Figure 2: Overview of transfer learning methods used in this work for heterocycle retrosynthesis prediction. (a) Methods previously used for forward reaction prediction and retrosynthesis. Fine-tuning consists of training a baseline model on a large dataset of all reaction classes, which is then fine-tuned on a smaller dataset of only reactions of interest. In multi-task learning, the model is trained on both datasets at the same time. (b) Methods only used in NLP tasks. In mixed fine-tuning, the *baseline* model is fine-tuned on both datasets. In ensemble decoding the prediction is made jointly with the *baseline* and *fine-tuned* model.

Methods

Datasets

In this study, we utilised the USPTO dataset pre-processed by Pesciullesi *et al.*,[20] which is henceforth referred to as the *General* dataset. Additionally, we curated a dataset of 165,216 ring formation reactions, referred here to as the *Ring* dataset, comprising about 80k reactions extracted from academic journals (CJHIF dataset[33]) and 80k reactions from additional patent data (Pistachio dataset).[34] The creation of the *Ring* dataset is described in more detail in SI§S4. A visualisation of the chemical space of the datasets is included in the SI (Figure S1), showing that ring-breaking reactions occupy distinct areas of the chemical space.

The *Ring* dataset was split into train, validation, and test sets with 90:5:5 ratio based on the Tanimoto similarity of reaction products[35] using DeepChem.[36] The *General* dataset splitting was retained from the work of Pesciullesi *et al.*[20] Additionally, we performed a random split of the *Ring* dataset and trained the mixed fine-tuned model on the randomly split dataset to assess the effect of dataset splitting (SI§S8).

Retrosynthesis prediction models

We trained the single-step retrosynthesis prediction models based on the seq2seq Transformer architecture using the OpenNMT-py package.[37] All hyperparameters used here are provided in the SI§S1 and are based on the work of Pesciullesi *et al.*[20] We trained the baseline model on only the *General* dataset. As fine-tuning and multi-task learning have been previously used for reaction prediction, we adopted the parameters previously reported for these models. For the *multi-task* model, we used a dataset weight ratio of 9 (*General*):1 (*Ring*). For the *fine-tuned* model, the number of fine-tuning steps was set to 6,000. For mixed fine-tuning, a 1:1 dataset weight ratio and 6,000 fine-tuning steps were chosen after a benchmark (SI§S7). Ensemble decoding was performed with in-built OpenNMT-py functionality using the *fine-tuned* model and the *baseline* model.

Furthermore, we trained a single-step template-based retrosynthesis prediction model on only ring-forming reactions based on the approach used by Thakkar *et al.* in 'Ring Breaker'.[19] Our dataset comprised reactions from the *Ring* dataset and ring formations extracted from the *General* (USPTO[5]) dataset. Atom-mapping of reaction data was conducted using RXNMapper,[38] and reaction templates were subsequently extracted using RDKit[39] and RDChiral.[40] We used TensorFlow[41] to construct the multilabel classification neural network for prediction. The selected hyperparameters are provided in the SI§S2.

To adapt the trained single-step retrosynthesis prediction models to multi-step route planning tools, we used a neural-based A* search algorithm based on Retro*.[9] Multi-step route planning tools were constructed for both the baseline and mixed fine-tuned single-step models. The stock molecule database chosen was eMolecules.

Model evaluation metrics

The single-step retrosynthesis prediction models were evaluated on both the *General* and *Ring* test sets using metrics based on top-N accuracy and round-trip accuracy.[10] For the *Ring* test set, we calculate reactant-only accuracy, where the prediction is considered accurate if all the ground truth reactants are present. For the *General* test set, due to the lack of separation between reactants and reagents, we calculate top-N accuracy by directly comparing the set of predicted precursors to the ground truth. We also consider the round-trip accuracy[10] of the suggested disconnections, which represent the "chemical validity" of predictions, i.e. what proportion of predicted reactant sets are expected to produce the desired product. Additionally, we introduce a new metric: the ring-breaking round-trip accuracy, calculated only for the "Ring" dataset. A disconnection is considered to be ring-breaking round-trip accurate when it is round-trip accurate and the number of rings in the product is higher than in predicted reactants. In this way, we consider not only whether the prediction is chemically valid but also whether it involves a ring disconnection, i.e. the reaction type we're aiming to improve the model's performance for.

All metrics reported in the main text are for top-1 predictions. However, metrics for top-3 and top-5 predictions are available in the SI§S9. A more detailed explanation of the metrics can be found in SI§S6.

Further fine-tuning

We extracted a set of 1,475 heterocycle formations from 47 scientific publications from 2022 reporting new methodologies for heterocycle synthesis(SI§S11). This dataset (referred to as the *Recent* dataset) was split randomly into a train, validation and test sets with a ratio of 80:10:10. Further fine-tuning was carried out using the mixed fine-tuning approach, starting from the *mixed fine-tuned* model and training it for 6,000 steps on the *General* , *Ring* and *Recent* datasets with a 4:4:1 dataset weight ratio.

Results

Optimisation of the single-step retrosynthesis model

Comparative analysis of transfer learning approaches

We commenced our study by comparing the performance of different transfer learning approaches, focusing on methods previously used for chemical reaction prediction (i.e., multitask learning and fine-tuning) and methods employed in the NLP domain (mixed fine-tuning and ensemble decoding) (Figure 2). The comparison is conducted on the *Ring* test set to assess their performance in predicting ring-breaking reactions compared to a *baseline* model trained only on the *General* dataset (Figure 3A). In addition to the commonly used reactant-accuracy, we also consider whether the prediction was chemically valid and corresponded to a ring-breaking reaction. This identifies predictions that differ from the ground truth disconnection present in the test set but still disconnect the ring.

Our results show that on the *Ring* test set, the *fine-tuned* model outperforms the other

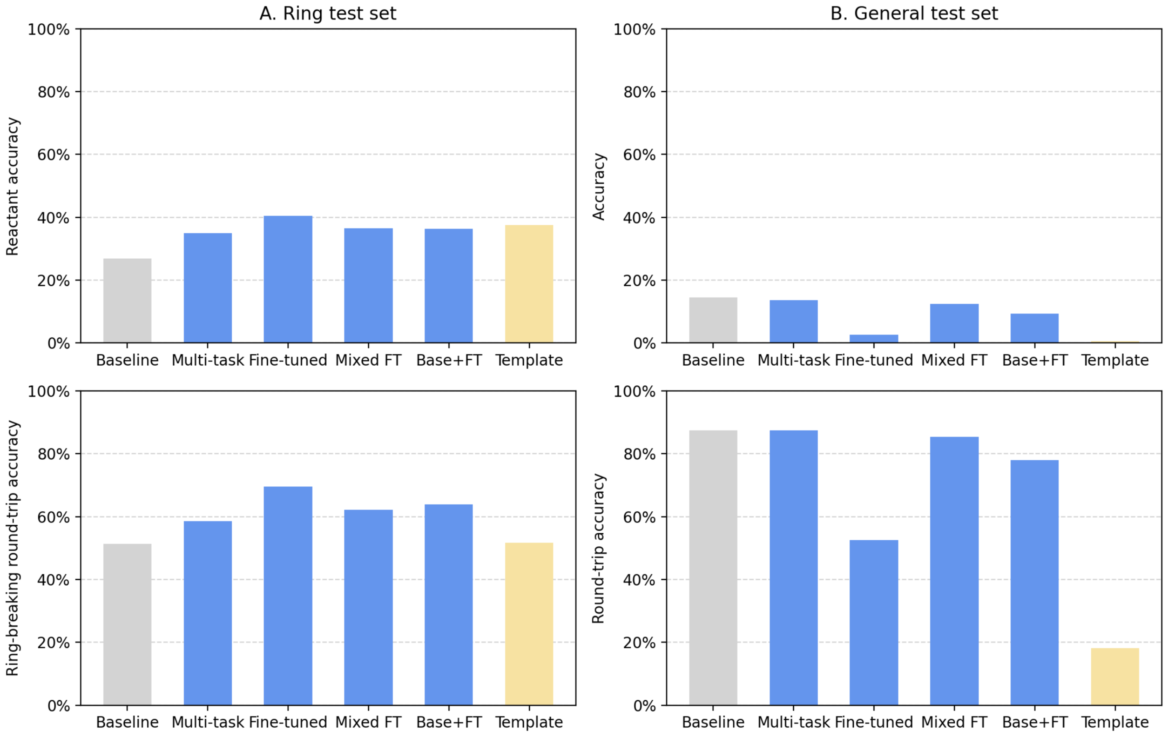


Figure 3: Comparison of model performance on the (A) *Ring* and (B) *General* test set. Top-1 reactant-accuracy and proportion of valid ring-breaking top-1 predictions are shown for the *Ring* test set. Top-1 accuracy and round-trip accuracy are shown for the *General* test set. *Multi-task, fine-tuned, mixed fine-tuned* (Mixed FT) models, ensemble decoding (Base+FT) and *template-based* (Template) model are compared to the *baseline*.

approaches, achieving a top-1 reactant accuracy of 40.5% (Figure 3A). Moreover, 69.5% of all its top-1 predictions are chemically valid and correspond to ring-breaking reactions. The three other approaches also show improvement over the *baseline* model with top-1 reactant-accuracies of around 36% and 62% valid ring-breaking top-1 predictions. While the observed improvement over *baseline* isn't as high as reported in previous studies[20,28] (13.6% increase in accuracy for *fine-tuned* model here vs 27.0% for carbohydrate reactions and 28.6% for Heck reactions), there are two key aspects to note. Firstly, the mentioned studies used transfer learning for forward reaction prediction, not retrosynthesis, which is considered to be a much easier task, only having one "correct" answer. Moreover, heterocycle formations are a much larger and more diverse class of reactions than Heck reactions or even carbohydrate reactions, making it more difficult for the model to learn all the different reactivity.

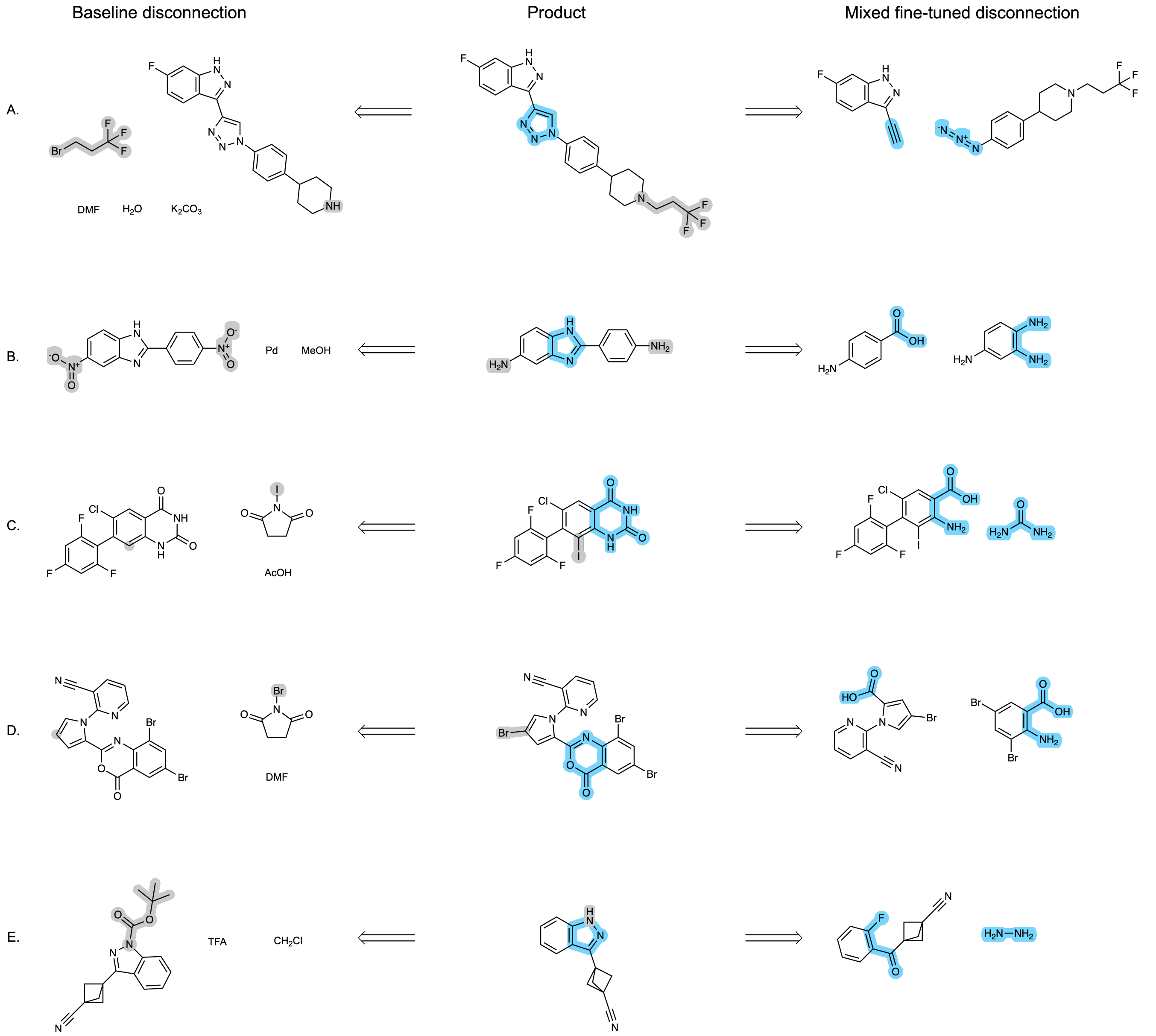


Figure 4: Example top-1 predictions of the *mixed fine-tuned* and *baseline* models for *Ring* test set molecules. For all the examples shown, the *mixed fine-tuned* prediction was accurate, whilst the *baseline* prediction was valid but not ring breaking. The disconnections suggested by the *mixed fine-tuned* model are highlighted in blue, while the disconnections suggested by the *baseline* model are highlighted in grey.

Interestingly, even though each of our approaches increases the proportion of top-1 valid ring-breaking predictions by at least 7% when compared to the *baseline* model, the same trend is not observed when considering just the top-1 round-trip accuracy of the predictions (SI§S9). For example, for the *mixed fine-tuned* model, the ring-breaking round-trip accuracy increases by over 10%, while the round-trip accuracy decreases by 1%. The same trend can be observed for all other approaches apart from the *multi-task* model, where the round-trip accurcay increases but not as much as the ring-breaking round-trip accuracy (SIS9). This indicates that the main improvement between the various models trained using transfer learning and the *baseline* model is in the type of disconnection suggested, i.e. ring-breaking versus more common reaction types, and not in turning chemically invalid disconnections into valid ones. It also suggests that while the molecules in the *Ring* test set were synthesized using ring formation reactions, there are other chemically viable disconnections available.

Indeed, comparing the predictions of the *baseline* and *mixed fine-tuned* model revealed that the former often suggested more common reaction types, such as functional group interconversions (FGIs) or protection/deprotections, instead of the ground-truth heterocycle formation predicted by the *mixed fine-tuned* (Figure 4). For instance, in example 4A. the *mixed fine-tuned* model correctly identifies a click reaction to generate the triazole from two fragments of similar complexity. In contrast, the *baseline* model only suggests a more trivial N-alkylation reaction. Similarly, for 4B. the *mixed fine-tuned* model suggests a condensation reaction to form the central benzimidazole ring, while the *baseline* model suggests a functional group interconversion, which would be more suitable earlier in the synthetic route. In 4C. and 4D. the *baseline* model predicts simple halogenation reactions rather than ring disconnections. Interestingly, although the *mixed fine-tuned* model's prediction is accurate for 4D., it was not counted as round-trip accurate due to the forward model predicting a condensation reaction with both the carboxylic acid and the nitro group instead of just a single condensation with the former. This highlights a limitation of metrics based on round-trip accuracy, where the model's prediction is only assessed by another model that is not 100% accurate instead of comparing the prediction to those reported in the literature or assessed by skilled organic chemists. Finally, in 4E. the *mixed fine-tuned* model correctly predicts the disconnection of indazole, while the *baseline* model suggests a Boc protection of the nitrogen without simplifying the molecule. While the ability of the model to suggest protection reactions is notable, as they are crucial parts of synthetic routes, this specific protection is unnecessary and might lead the model to predict a cycle of protection/deprotection reactions, preventing further disconnections of the molecule.

When tested on the *General* test set, the models exhibit almost the opposite trend (Figure 3B). Performance of the *fine-tuned* model drastically decreases compared to the *baseline* model, with the top-1 accuracy dropping from 14.5% to 2.7% and top-1 round-trip accuracy from 87.4% to 52.6%. Meanwhile, the metrics for the *mixed fine-tuned* and *multi-task* model only change marginally, dropping by at most 2%. Ensemble decoding falls in between, with a top-1 accuracy of 9.3% and round-trip accuracy of 77.9%. The drop in performance observed with the *fine-tuned* model can most likely be attributed to catastrophic forgetting,[42] the tendency of NNs to forget previously learned information when trained on new data. This drop can be disregarded if the model is only intended for one-step ring disconnection. However, it becomes problematic for multi-step retrosynthesis as the fine-tuned model will not be able to disconnect the linear intermediates obtained after disconnecting the ring. In that case, either the *mixed fine-tuned* or *multi-task* model would be more suitable.

Considering time and resources, mixed fine-tuning appears preferable due to its 40 times shorter training time and comparable performance to multi-task learning, especially if planning to frequently retrain the model as new data becomes available. Ensemble decoding employs two models to make the prediction, and therefore, it takes longer at inference than the other three methods.

Overall, both multi-task learning and mixed fine-tuning show improved performance for ring-breaking disconnections while retaining the ability to predict other reaction classes, with mixed fine-tuning being preferable due to shorter training time. While the *fine-tuned* model performs best for heterocycle disconnections, it is not suitable for multi-step retrosynthesis due to catastrophic forgetting. Ensemble decoding ranks in the middle, not being as good at ring disconnections as the *fine-tuned* model, but also performing worse for other reaction classes than the *mixed fine-tuned* model. Due to this, we perform all further experiments and comparisons with the *mixed fine-tuned* model, as the most versatile and best performing one.

Comparison to the template-based model

The *mixed fine-tuned* model was further benchmarked against 'Ring Breaker',[19] the template-based model trained specifically for heterocycle retrosynthesis. To allow for fair comparison, we re-trained 'Ring Breaker' with our additionally extracted ring formation data, using the whole *Ring* dataset and ring formation reactions from the *General* dataset. We compared the performance of the *mixed fine-tuned* model to this ring-breaking specific template-based model.

In terms of reactant-accuracy, both the *mixed fine-tuned* and the template-based model have similar top-1 reactant-accuracies (Figure 3A), with the template-based model's reactant-accuracy being slightly higher. However, the *mixed fine-tuned* model has significantly higher top-1 round-trip accuracy. These trends remain consistent across top-3 and top-5 predictions (SI§S9). Moreover, the round-trip accuracies for the template-based model decrease rapidly from top-1 to top-5, from 53.4% to 31.8%, while the *mixed fine-tuned* model maintains high round-trip accuracy from top-1 (74.6%) to top-5 (71.8%) (Table 1). The *mixed fine-tuned* model also suggests a higher overall proportion of chemically valid ring-breaking disconnections (defined in Methods), with 59.9% for the *mixed fine-tuned* model compared to 30.8% for the template-based model in the first 5 predictions (SI§S9). Additionally, the *mixed fine-tuned* model maintains considerable accuracy on the *General* test set, while the template-based model achieves a low top-1 accuracy of 0.5%.

Furthermore, we observe that the template-based model generates a larger proportion of non-admissible predictions of 'None', with 48.8% of top-5 predictions being inadmissible,

Table 1: Comparison of the template-based model to the mixed fine-tuned model. Top-N Round-trip accuracy refers to the proportion of predictions within the first N predictions for the test set considered chemically valid. The proportion of inadmissible predictions refers to the percentage of predictions in the first N predictions for the test set that did not output a viable SMILES string.

Metric	Mixed fine-tuned model			Template-based model		
	top-1	top-3	top-5	top-1	top-3	top-5
Round-trip accuracy	74.6%	73.3%	71.8%	53.4%	38.9%	31.8%
Inadmissible predictions	0.5%	0.7%	0.8%	27.6%	41.0%	48.8%

compared to only 0.8% of the *mixed fine-tuned* model's predictions corresponding to invalid SMILES strings (Table 1). For the template-based model, the increase in proportion of inadmissible predictions between top-1 and top-5 correlates with the decrease in round-trip accuracy, indicating that the low round-trip accuracy is due to the model's inability to apply multiple templates to one molecule. Hence, it is likely that the *mixed fine-tuned* model learns a wider range of chemistry than the template-based model, which is limited in diversity when it comes to disconnection strategies.

Overall, our results demonstrate that the *mixed fine-tuned* model significantly outperforms the template-based model in round-trip accuracy, suggesting more diverse disconnections for both general and ring-breaking disconnections, making it the preferred choice for multi-step retrosynthesis as discussed in the following section. However, it is important to note that the forward reaction prediction model used for calculating round-trip accuracy is of the same architecture as the *mixed fine-tuned* model and is trained on the same reaction data (but with reversed labels). This could be biasing the metric towards the *mixed fine-tuned* model and mean that the difference in round-trip accuracy between the *mixed fine-tuned* model and the template-based model is not as significant as it seems. A more objective way of calculating metrics such as round-trip accuracy could be to use a different model to predict reaction viability instead of the forward reaction prediction model, however we were not able to train such a model for this work due to lack of negative reaction data.

Mixed fine-tuned multi-step model

To assess the practical use of the *mixed fine-tuned* model in synthesis planning for drug-like targets, we constructed a multi-step retrosynthesis prediction tool using neural-guided A* Search, based on the algorithm used in Retro*. The two drug-like targets included **CZS-035** and **ADD** (Figure 5), for which syntheses were reported in 2023. The exact reactions employed in these synthesis are therefore absent in our training set, which contains reactions from patents and literature up to 2022. For comparison, we also built an analogous multi-step retrosynthesis tool employing the baseline single-step model, maintaining identical search settings.

The first case study, **CZS-035**, is a ligand for polo-like kinase 4 (PLK4) and a warhead component used to synthesise a therapeutic PROTAC for breast cancer treatment, discovered by Sun *et al.*[43] (Figure 5A). Both the *baseline* and *mixed fine-tuned* multi-step models successfully identify retrosynthetic routes for **CZS-035** from purchasable precursors in our stock molecule database. Both models accurately reproduce the protection of nitrogen with Boc (**A1**) as seen in the literature synthesis.[43] Both models also correctly identify the two S_NAr disconnections used in the literature to reproduce **B1** and **B2**. However, the *mixed fine-tuned* model uniquely identifies the final ring disconnection of pyrazole in **B1** to **C1** and **C2**, which aligns with the literature approach. In contrast, the *baseline* model suggests the more complex and more expensive pyrazolopyrimidine **C3** as the final purchasable precursor. This result showcases the enhanced performance of the *mixed fine-tuned* model for predicting key ring disconnections for multi-step routes, overcoming catastrophic forgetting and correctly identifying all non-ring breaking disconnections of **CZS-035**. We note that the ability of seq2seq models over template-based models to simultaneously suggest protections and S_NAr disconnections in different sites as in **A1** is a unique advantage.

The second case study was **ADD** (compound 15d in ref[44]), a merged human butyrylcholinesterase (hBChE) inhibitor/cannabinoid receptor 2 (hCB2R) ligand and a therapeutic target for preventing learning impairments in Alzheimer's disease (Figure 5B).[44] The *baseline*

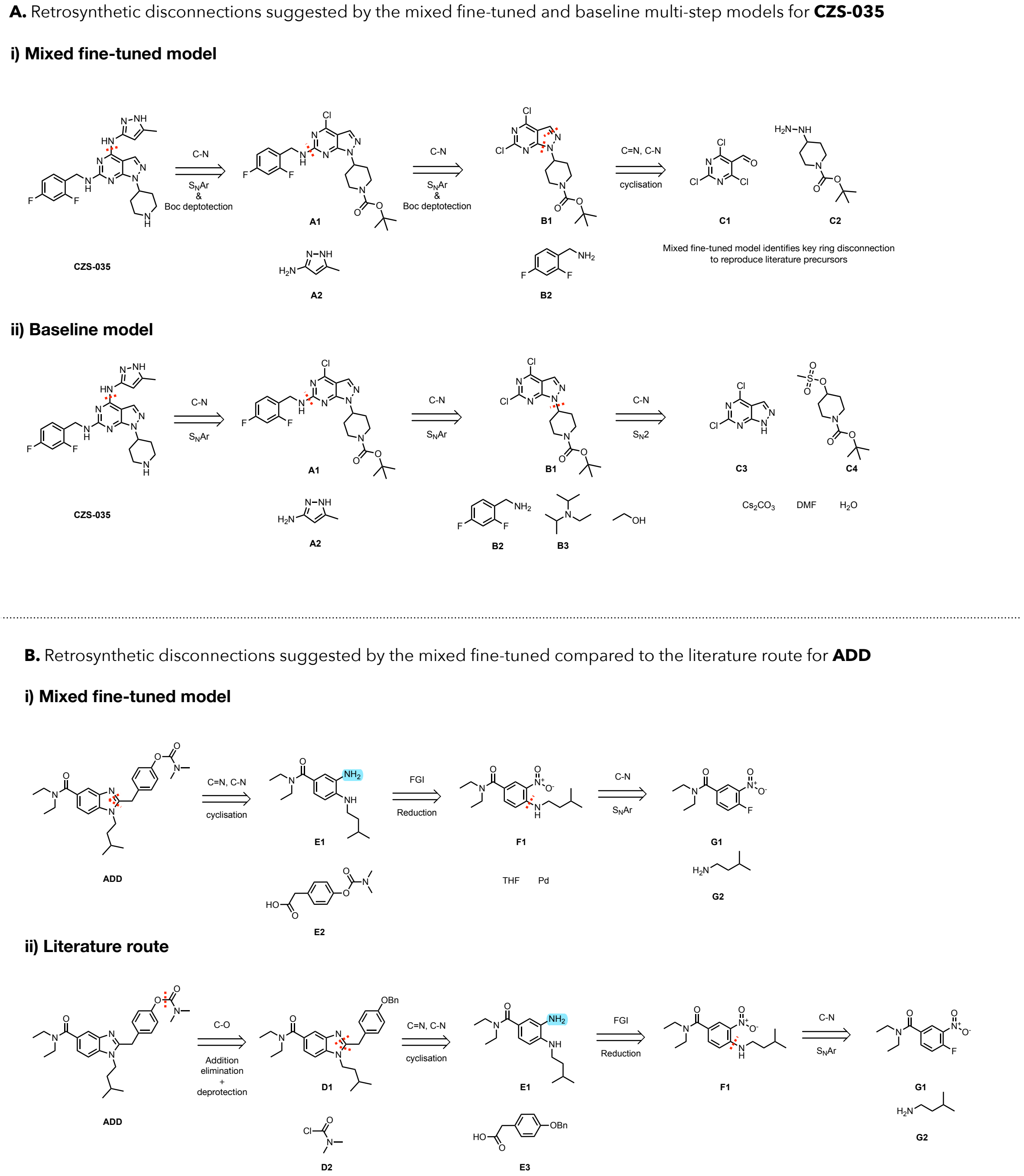

A. Retrosynthetic disconnections suggested by the mixed fine-tuned and baseline multi-step models for CZS-035
i) Mixed fine-tuned model
C-N
SNAr & Boc deptotection
C-N
SNAr & Boc deptotection
C=N, C-N
cyclisation
CZS-035
A1
A2
B1
B2
C1
C2
Mixed fine-tuned model identifies key ring disconnection to reproduce literature precursors
ii) Baseline model
C-N
SNAr
C-N
SNAr
C-N
SN2
CZS-035
A1
A2
B1
B2
B3
C3
C4
Cs2CO3
DMF
H2O
B. Retrosynthetic disconnections suggested by the mixed fine-tuned compared to the literature route for ADD
i) Mixed fine-tuned model
C=N, C-N
cyclisation
FGI
Reduction
C-N
SNAr
ADD
E1
E2
F1
G1
G2
THF
Pd
ii) Literature route
C-O
Addition elimination + deprotection
C=N, C-N
cyclisation
FGI
Reduction
C-N
SNAr
ADD
D1
D2
E1
E3
F1
G1
G2

Figure 5: Example synthetic routes found by the *mixed fine-tuned* model for molecules of clinical interest. (A) Comparison of the retrosynthetic routes for **CZS-035** predicted by (i) *mixed fine-tuned* and (ii) *baseline* models. (B) The retrosynthetic route for **ADD** (i) predicted by the *mixed fine-tuned* model compared to (ii) the literature route. The *baseline* model failed to predict a complete route for this compound.

multi-step model failed to identify a synthetic route, while the *mixed fine-tuned* model predicts retrosynthetic disconnections similar to the literature route (Figure 5). Reagents were omitted from the literature route to focus on the core synthons. While the *mixed fine-tuned* model deviated by not reproducing the carbamate disconnection of **ADD** to benzyl-protected phenol **D1**, instead using the pre-synthesised phenyl carbamate **E2**, it proposed subsequent disconnections featuring the same cyclisation, reduction, and S_NAr as the literature route to mutually predicted reactants **E1**, **F1**, **G1**, and **G2**. This further reaffirms the improved ring-breaking performance in multi-step retrosynthesis of the *mixed fine-tuned* model, where the *baseline* model failed for the benzoimidazole scaffold in **ADD**.

These results demonstrate the capability of the *mixed fine-tuned* multi-step model in suggesting tractable synthetic routes for newly-discovered, complex drug-like targets containing heterocycles. This highlights its potential as a tool for synthetic chemists, aiding them in designing synthetic routes towards novel heterocycle-containing therapeutics.

Recently developed heterocycle formation reactions

To evaluate whether the *mixed fine-tuned* model could extrapolate to unknown systems, we extracted 1.5k heterocycle ring-forming reactions from 47 papers published in 2022 (here referred to as *Recent* dataset). While the model was, unsurprisingly, unable to predict the exact reported reactions, it provided chemically valid ring-breaking predictions for 30.4% of the molecules. This indicates that while the reported reactions are new, potentially more efficient or greener routes than those reported already, many of the heterocycles formed were already synthetically accessible (Figure 6A). Interestingly, the routes suggested by our model often resembled the ground truth (6Ai.-iii.). For example, both the *mixed fine-tuned* model and literature suggested the same Friedländer synthesis for quinoline (Figure 6Ai.). In the literature synthesis, there is an additional oxime intermediate; however, the *mixed fine-tuned* model's prediction follows the direct approach previously taken for trifluoromethane-substituted quinolines by Jiang *et al.*[45]

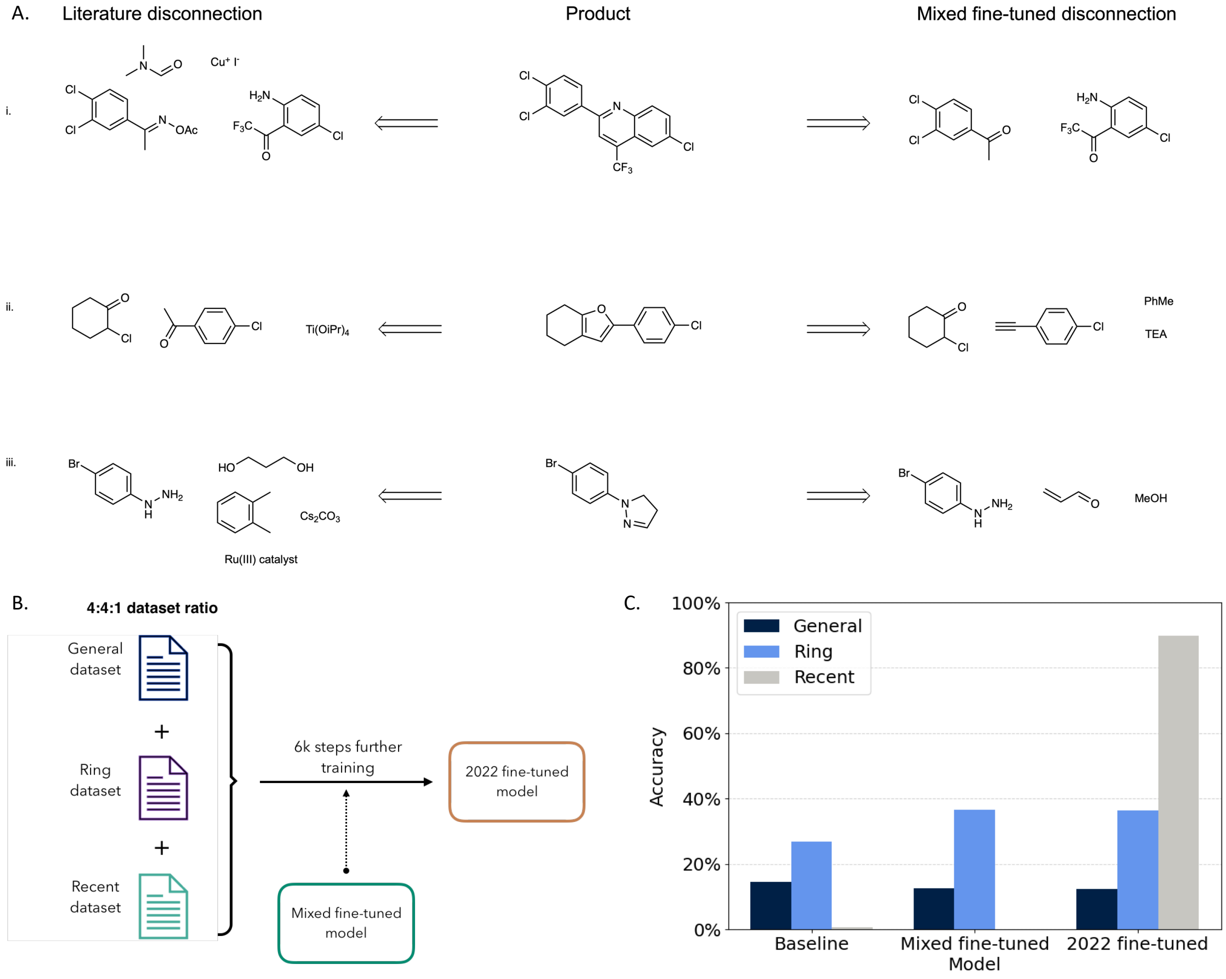


Figure 6: Recent reaction prediction. (A) Example valid predictions of the *mixed fine-tuned* model on the *Recent* test set. (B) The further fine-tuning approach: the *mixed fine-tuned* model is further trained on all three datasets. (C) Top-1 accuracy for the *baseline*, *mixed fine-tuned* and *further fine-tuned* model on *General*, *Ring* and *Recent* test sets. Reactant-only accuracy is reported for the *Ring* and *Recent* test sets.

Although the *mixed fine-tuned* model found valid ring-breaking disconnections for almost a third of the molecules in the *Recent* test set, when compared to the *Ring* test set, this proportion is lower by 30%. Therefore, this indicates that the *Recent* test set includes a higher number of heterocycles unknown to our model and therefore considered synthetically inaccessible. If the model was trained on those new heterocycle formations, it could potentially explore a new region of the chemical space. To address this, we further trained the *mixed fine-tuned* model using the *Recent* dataset. This updated *2022 fine-tuned* model was trained on the three datasets: *General*, *Ring* and *Recent* for another 6,000 steps starting from the *mixed fine-tuned* model (Figure 6B). The top-1 accuracy of this *2022 fine-tuned* model is shown in Figure 6C. This updated *2022 fine-tuned* model exhibited only a slight decrease in accuracy on the *General* and *Ring* test sets while showing an increased top-1 reactant accuracy on the *Recent* test set (89.9%). This illustrates that the model can be fine-tuned to incorporate new reaction data without significantly compromising performance on previously learned tasks. While we used a small dataset of heterocycle formations here, this approach could be applied to a larger dataset or reaction data for different reaction classes of interest.

Conclusion

In this work we compared four different transfer learning approaches: fine-tuning, multi-task learning, mixed fine-tuning and ensemble decoding. Our aim was to improve the performance of seq2seq retrosynthesis prediction models for ring-breaking disconnections. We have found that mixed fine-tuning performs best overall, with short training time, top-1 reactant-accuracy for ring formations increased by 10% compared to the *baseline* model and a barely decreased accuracy on other reaction classes. The accuracy for ring formations is comparable to the template-based model we trained based on Ring Breaker; however, the *mixed fine-tuned* model vastly outperforms the template-based model in other reaction classes. While the *fine-tuned* model performs best for ring formations, with top-1 reactant-accuracy of 40.5%, its performance significantly drops for other reaction classes due to catastrophic forgetting. This makes it unusable for multi-step retrosynthesis, which requires disconnection of both rings and linear intermediates. We have also introduced a new metric, the "ring-breaking round-trip accuracy", to assess the performance of the models for ring-breaking disconnections. By comparing the round-trip accuracy and ring-breaking round-trip accuracy of the *baseline* and *mixed fine-tuned* models, we have shown that both models suggested viable disconnections for a similar proportion of molecules. However, the key improvement in the *mixed fine-tuned* model was the type of disconnection suggested. While the *baseline* model suggests common reactions, such as protections/deprotections or functional group interconversions, which were either unnecessary or better suitable earlier in the synthetic route, the *mixed fine-tuned* model favoured ring formation reactions, with 62.1% of disconnections being ring-breaking round-trip accurate.

We then showcased the practical utility of the *mixed fine-tuned* model by using it for multi-step retrosynthesis of two newly-discovered, complex drug-like compounds containing heterocycles. This illustrates how the model can be used to assist synthetic and medicinal chemists, aiding them in designing synthetic routes towards novel heterocycle-containing therapeutics.

Finally, we have introduced a method for further fine-tuning the model on additional reaction data. By using this further mixed fine-tuning we have substantially improved the model's top-1 reactant-accuracy on ring formation reactions published in 2022 from 0% to 89.9% without significantly compromising performance for older ring formation reactions or other reaction classes. While this approach has been applied to a small dataset of less than 1.5k heterocycle formations, it has the potential to be scaled up for a larger dataset or a different reaction class.

Data availability

The General dataset (based on USPTO), the ring formation reactions derived from CJHIF, and the Recent dataset are available at: `https://github.com/duartegroup/Het-retro`

Code availability

The source code for single-step model training and inference is available at: `https://github.com/duartegroup/Het-retro`

Conflicts of interest

There are no conflicts to declare.

Author contributions

EW, PEB and FD conceptualised the study. EW and JWS carried out the calculations. All authors participated in data analyses and writing of the manuscript. EW, JWS, and FD wrote the first draft. LW, PEB and FD supervised the study.

Acknowledgements

This work was supported by funding from the Engineering and Physical Sciences Research Council (EPSRC) [grant number EP/S024093/1], Alzheimer's Research UK (grant reference: ARUK-2020DDI-OX) and the Centre of Artificial Intelligence in Precision Medicines (CAIPM), King Abdulaziz University, Jeddah, Saudi Arabia. The authors thank T. Watts and C. Wilson for insightful discussions. The authors acknowledge IBM Research for academic access to IBM Cloud.

References

(1) Corey, E. J. *The Chemistry of Natural Products*; Butterworth-Heinemann, 1967; pp 19–37.

(2) Pensak, D. A.; Corey, E. J. *Computer-Assisted Organic Synthesis*; ACS Symposium Series 61; American Chemical Society, 1977; Vol. 61; pp 1–32, Section: 1.

(3) Coley, C. W.; Green, W. H.; Jensen, K. F. Machine Learning in Computer-Aided Synthesis Planning. *Acc. Chem. Res.* **2018**, *51*, 1281–1289, Publisher: American Chemical Society.

(4) Schwaller, P.; Vaucher, A. C.; Laplaza, R.; Bunne, C.; Krause, A.; Corminboeuf, C.; Laino, T. Machine intelligence for chemical reaction space. *WIREs Computational Molecular Science* **2022**, *12*, e1604, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wcms.1604.

(5) Lowe, D. Chemical reactions from US patents (1976-Sep2016). `https://figshare.com/articles/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873`.

(6) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **2018**, *555*, 604–610.

(7) Coley, C. W. et al. A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science* **2019**, *365*, eaax1566, Publisher: American Association for the Advancement of Science.

(8) Genheden, S.; Thakkar, A.; Chadimová, V.; Reymond, J.-L.; Engkvist, O.; Bjerrum, E. AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning. *Journal of Cheminformatics* **2020**, *12*, 70.

(9) Chen, B.; Li, C.; Dai, H.; Song, L. Retro*: Learning Retrosynthetic Planning with Neural

Guided A* Search. 2020; http://arxiv.org/abs/2006.15820, arXiv:2006.15820 [cs, stat].

(10) Schwaller, P.; Petraglia, R.; Zullo, V.; Nair, V. H.; Haeuselmann, R. A.; Pisoni, R.; Bekas, C.; Iuliano, A.; Laino, T. Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chem. Sci.* **2020**, *11*, 3316–3325, Publisher: The Royal Society of Chemistry.

(11) Tetko, I. V.; Karpov, P.; Van Deursen, R.; Godin, G. State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis. *Nat Commun* **2020**, *11*, 5575, Number: 1 Publisher: Nature Publishing Group.

(12) Liu, B.; Ramsundar, B.; Kawthekar, P.; Shi, J.; Gomes, J.; Luu Nguyen, Q.; Ho, S.; Sloane, J.; Wender, P.; Pande, V. Retrosynthetic Reaction Prediction Using Neural Sequence-to-Sequence Models. *ACS Cent. Sci.* **2017**, *3*, 1103–1113, Publisher: American Chemical Society.

(13) Sacha, M.; Błaż, M.; Byrski, P.; Dąbrowski-Tumański, P.; Chromiński, M.; Loska, R.; Włodarczyk-Pruszyński, P.; Jastrzębski, S. Molecule Edit Graph Attention Network: Modeling Chemical Reactions as Sequences of Graph Edits. *J. Chem. Inf. Model.* **2021**, *61*, 3273–3284, Publisher: American Chemical Society.

(14) Somnath, V. R.; Bunne, C.; Coley, C. W.; Krause, A.; Barzilay, R. Learning Graph Models for Retrosynthesis Prediction. 2021; http://arxiv.org/abs/2006.07038, arXiv:2006.07038 [cs, stat].

(15) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. 2023; http://arxiv.org/abs/1706.03762, arXiv:1706.03762 [cs].

(16) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A.

Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Cent. Sci.* **2019**, *5*, 1572–1583, Publisher: American Chemical Society.

(17) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36, Publisher: American Chemical Society.

(18) Dai, H.; Li, C.; Coley, C.; Dai, B.; Song, L. Retrosynthesis Prediction with Conditional Graph Logic Network. Advances in Neural Information Processing Systems. 2019.

(19) Thakkar, A.; Selmi, N.; Reymond, J.-L.; Engkvist, O.; Bjerrum, E. J. "Ring Breaker": Neural Network Driven Synthesis Prediction of the Ring System Chemical Space. *J. Med. Chem.* **2020**, *63*, 8791–8808, Publisher: American Chemical Society.

(20) Pesciullesi, G.; Schwaller, P.; Laino, T.; Reymond, J.-L. Transfer learning enables the molecular transformer to predict regio- and stereoselective reactions on carbohydrates. *Nat Commun* **2020**, *11*, 4874, Number: 1 Publisher: Nature Publishing Group.

(21) McGrath, N. A.; Brichacek, M.; Njardarson, J. T. A Graphical Journey of Innovative Organic Architectures That Have Improved Our Lives. *J. Chem. Educ.* **2010**, *87*, 1348–1349, Publisher: American Chemical Society.

(22) Taylor, R. D.; MacCoss, M.; Lawson, A. D. G. Rings in Drugs. *J. Med. Chem.* **2014**, *57*, 5845–5859, Publisher: American Chemical Society.

(23) Dudkin, V. Y. Bioisosteric equivalence of five-membered heterocycles. *Chem Heterocycl Comp* **2012**, *48*, 27–32.

(24) Meanwell, N. A. Synopsis of Some Recent Tactical Application of Bioisosteres in Drug Design. *J. Med. Chem.* **2011**, *54*, 2529–2591, Publisher: American Chemical Society.

(25) Pitt, W. R.; Parry, D. M.; Perry, B. G.; Groom, C. R. Heteroaromatic Rings of the Future. *J. Med. Chem.* **2009**, *52*, 2952–2963, Publisher: American Chemical Society.

(26) Brown, D. G.; Boström, J. Analysis of Past and Present Synthetic Methodologies on Medicinal Chemistry: Where Have All the New Reactions Gone? *J. Med. Chem.* **2016**, *59*, 4443–4458, Publisher: American Chemical Society.

(27) Roughley, S. D.; Jordan, A. M. The Medicinal Chemist's Toolbox: An Analysis of Reactions Used in the Pursuit of Drug Candidates. *J. Med. Chem.* **2011**, *54*, 3451–3479, Publisher: American Chemical Society.

(28) Wang, L.; Zhang, C.; Bai, R.; Li, J.; Duan, H. Heck reaction prediction using a transformer model based on a transfer learning strategy. *Chem. Commun.* **2020**, *56*, 9368–9371, Publisher: The Royal Society of Chemistry.

(29) Kreutter, D.; Schwaller, P.; Reymond, J.-L. Predicting enzymatic reactions with a molecular transformer. *Chem. Sci.* **2021**, *12*, 8648–8659, Publisher: The Royal Society of Chemistry.

(30) Probst, D.; Manica, M.; Nana Teukam, Y. G.; Castrogiovanni, A.; Paratore, F.; Laino, T. Biocatalysed synthesis planning using data-driven learning. *Nat Commun* **2022**, *13*, 964, Number: 1 Publisher: Nature Publishing Group.

(31) Chu, C.; Dabre, R.; Kurohashi, S. An Empirical Comparison of Domain Adaptation Methods for Neural Machine Translation. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Vancouver, Canada, 2017; pp 385–391.

(32) Freitag, M.; Al-Onaizan, Y. Fast Domain Adaptation for Neural Machine Translation. 2016; `http://arxiv.org/abs/1612.06897`, arXiv:1612.06897 [cs].

(33) Jiang, S.; Zhang, Z.; Zhao, H.; Li, J.; Yang, Y.; Lu, B.-L.; Xia, N. When SMILES Smiles, Practicality Judgment and Yield Prediction of Chemical Reaction via Deep Chemical Language Processing. *IEEE Access* **2021**, *9*, 85071–85083, Conference Name: IEEE Access.

(34) NextMove Software, Pistachio. 2022; https://www.nextmovesoftware.com/pistachio.html.

(35) Kovács, D. P.; McCorkindale, W.; Lee, A. A. Quantitative interpretation explains machine learning models for chemical reaction prediction and uncovers bias. *Nat Commun* **2021**, *12*, 1695, Number: 1 Publisher: Nature Publishing Group.

(36) Ramsundar, B.; Eastman, P.; Walters, P.; Pande, V.; Leswing, K.; Wu, Z. *Deep Learning for the Life Sciences*; O'Reilly Media, 2019; https://www.amazon.com/Deep-Learning-Life-Sciences-Microscopy/dp/1492039837.

(37) OpenNMT-py. https://github.com/OpenNMT/OpenNMT-py.

(38) Schwaller, P.; Hoover, B.; Reymond, J.-L.; Strobelt, H.; Laino, T. Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. *Science Advances* **2021**, *7*, eabe4166, Publisher: American Association for the Advancement of Science.

(39) Landrum, G. RDKit: Open-source cheminformatics. http://www.rdkit.org.

(40) Coley, C. W.; Green, W. H.; Jensen, K. F. RDChiral: An RDKit Wrapper for Handling Stereochemistry in Retrosynthetic Template Extraction and Application. *J. Chem. Inf. Model.* **2019**, *59*, 2529–2537, Publisher: American Chemical Society.

(41) Abadi, M. et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. 2015; https://www.tensorflow.org/, Software available from tensorflow.org.

(42) McCloskey, M.; Cohen, N. J. In *Psychology of Learning and Motivation*; Bower, G. H., Ed.; Academic Press, 1989; Vol. 24; pp 109–165.

(43) Sun, Y. et al. Discovery of the First Potent, Selective, and In Vivo Efficacious Polo-like Kinase 4 Proteolysis Targeting Chimera Degrader for the Treatment of TRIM37-Amplified Breast Cancer. *J. Med. Chem.* **2023**, *66*, 8200–8221, Publisher: American Chemical Society.

(44) Spatz, P.; Steinmüller, S. A. M.; Tutov, A.; Poeta, E.; Morilleau, A.; Carles, A.; Deventer, M. H.; Hofmann, J.; Stove, C. P.; Monti, B.; Maurice, T.; Decker, M. Dual-Acting Small Molecules: Subtype-Selective Cannabinoid Receptor 2 Agonist/Butyrylcholinesterase Inhibitor Hybrids Show Neuroprotection in an Alzheimer's Disease Mouse Model. *J. Med. Chem.* **2023**, *66*, 6414–6435, Publisher: American Chemical Society.

(45) Jiang, B.; Dong, J.-j.; Jin, Y.; Du, X.-l.; Xu, M. The First Proline-Catalyzed Friedlander Annulation: Regioselective Synthesis of 2-Substituted Quinoline Derivatives. *European Journal of Organic Chemistry* **2008**, *2008*, 2693–2696, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ejoc.200800121.