'Ring Breaker': Assessing Synthetic Accessibility of the Ring System Chemical Space

Amol Thakkar,¹ ¹ ¹ Nidhal Selmi,¹ Jean-Louis Reymond,² Ola Engkvist,¹ and Esben Jannik Bjerrum¹ ¹

¹ Hit Discovery, Discovery Sciences, R&D, AstraZeneca, Gothenburg, Sweden.
² Department of Chemistry and Biochemistry, University of Bern, Bern, Switzerland.

*Corresponding authors: amol.thakkar@dcb.unibe.ch, esben.bjerrum@astrazeneca.com

Abstract

Ring systems in pharmaceuticals, agrochemicals and dyes are ubiquitous chemical motifs. Whilst the synthesis of common ring systems is well described, and novel ring systems can be readily computationally enumerated, the synthetic accessibility of unprecedented ring systems remains a challenge. ‘Ring Breaker’ enables the prediction of ring-forming reactions, for which we have demonstrated its utility on frequently found and unprecedented ring systems, in agreement with literature syntheses. We demonstrate its performance on a range of ring fragments from the ZINC database and highlight its potential for incorporation into computer aided synthesis planning tools. Additionally, we generate a multi-label dataset using bipartite reaction graphs on which we train ‘Ring Breaker’ to model the relationship between one ring fragment and the multiple reactions recorded for its synthesis in the dataset; we thereby overcome the single-label approaches previously used. These approaches to ring formation and retrosynthetic disconnection offer opportunities for chemists to explore and select more efficient syntheses/synthetic routes.

Introduction

The recent wave of artificial intelligence (AI) within drug discovery has heavily impacted in the fields of de novo design, synthesis planning, and bioactivity prediction to name a few.¹ ² This holds the promise of accelerating Design, Make, Test, Analyze (DMTA) cycles, for which predictive models are desired to reduce failure rates in the drug discovery process.¹ ² Computer aided synthesis planning (CASP) has long been investigated as a means for predicting how to make a given compound.³⁻⁵ However, despite recent progress in the field,⁶⁻¹² synthetic planning tools based on neural network classifiers have failed to recognize reactions that are infrequently used or rare, due to to the heavily biased datasets available.⁷,¹³
As such, CASP tools have not yet focused on the synthesis of ring systems, the reactions for which often fall within the noise of datasets in question. The ability to deconstruct ring systems in novel ways, offers medicinal and process chemists alike, the opportunity to explore a wider range of chemical space and create more efficient synthetic routes, thereby leading to a competitive advantage.\textsuperscript{11}

Ring systems are key scaffold components in medicinal chemistry, and are fundamental motifs to a number of drugs on the market today.\textsuperscript{14} They vary greatly in nature – ring systems can be saturated, unsaturated, polycyclic and range in sizes from small heterocyclic rings to large macrocycles. In addition, they span over a range of chemical domains, from cyclic peptides to natural products, specialty chemicals, and dyes. As such, it is not surprising that many of the most frequently used reactions in organic synthesis pertain to the coupling of ring systems.\textsuperscript{15, 16} Although coupling reactions enable the synthesis of a wide range of structures, they are limited by the available building blocks. Ring-forming strategies, on the other hand, can enable the synthesis of novel building blocks containing ring systems, which can then be coupled to other fragments, thus allowing for the expansion of the synthetically feasible chemical space.

Ring systems play a role in the electronic distribution, three dimensionality, and scaffold rigidity of the small molecules they are part of.\textsuperscript{14, 17} They can directly interact with a protein target, such as in the well-defined example of the hinge binding motifs for kinase targets.\textsuperscript{18} In addition, they contribute to physiochemical properties such as lipophilicity or polarity and molecular reactivity, which in turn will determine a molecule’s absorption and distribution, metabolic stability, excretion and toxicity (ADMET) profile.\textsuperscript{14} Therefore, synthetic approaches to novel ring systems are desired in order to tune and exploit property profiles derived from their interaction with the target. As such, numerous publications have followed the exhaustive computational enumeration of heteroaromatic ring systems first described by Pitt \textit{et al}.\textsuperscript{19} These aim to enrich structure-activity relationship information, explore the chemical space of ring systems, and find motifs relevant for use in medicinal chemistry.\textsuperscript{17, 20-22} However, as of yet the synthetic accessibility of ring systems remains poorly explored.

Furthermore, the neural networks upon which CASP tools are built,\textsuperscript{7, 8, 23} are trained using the single label (templates) obtained from the dataset. As an analogy to retrosynthetic planning, this resembles a one compound to one reaction (template) situation, whereas in ‘truth’ a compound can be synthesized by multiple reactions at any given step in the pathway. In this study we propose a method for the extraction of multiple labels from the underlying dataset and demonstrate its use in the prediction of retrosynthetic ring disconnections. This was extended to the prediction of previously unseen fragments, such as the so called ‘Rings of the Future’ for which we examine the predictive performance.\textsuperscript{19, 24} We show how ‘Ring
Breaker’ can be viewed as a specialist for predicting ring formations, and used alongside current CASP tools, to guide route finding into avenues/tracks exploiting ring synthesis. The implications of predicting ring synthesis are far reaching and extend beyond the medicinal chemistry domain, to dyes, fragrances and agrochemicals to name a few.

Dataset Generation

Reaction datasets, in their current form, contain records of individual reactions whereby one compound can be the product of several different reaction classes or combination of reactants. In previous approaches these individual records have been used to train neural networks to either predict a retrosynthetic step or for reaction prediction. However, this strategy neglects the one to many nature of retrosynthetic analysis, where a given compound may be constructed in more than one way. To overcome the limitation imposed by direct use of the dataset entries, we first build a bipartite reaction graph to map the relationship between all compounds designated as products in the reaction dataset with their corresponding template or reaction rule. Using only templates that have been validated by applying them to the product and confirming that they regenerate the reactants recorded in the dataset, we ensure that the graph represents a ‘partial’ ground truth of the retrosynthetic space. For each compound designated as a product, the bipartite reaction graph is queried to obtain the neighboring connected nodes, from which we can extract a multi-label dataset for the subsequent training of neural networks. This approach allows us to train a multi-label multi-class classification neural network for the prediction of retrosynthetic steps, as opposed to the single-label multi-class classification network previously described (Figure 1). In doing so, the number of samples is limited to that of the number of products recorded in the dataset, rather than the number of individual reaction entries. This speeds up training of the network by reducing the number of samples, resulting in a more efficient way of scaling to the ever-growing chemical literature. Additionally, the label vectors better represent the nature of the problem and are closer to the ground truth as their sparsity is reduced. The ground truth defined as containing all possible retrosynthetic disconnections, and as such reaction templates that can be applied to any given product.
Figure 1: a) Schematic of multilabel generation. Previously machine learning approaches to retrosynthetic planning have been trained considering a one product to one template relationship. However, as multiple templates/reactions may be used on a given compound, it is desirable to train the model considering a one product to multiple template relationship. Here we build a bipartite reaction graph connecting compounds with their associated templates, which we subsequently query to extract a multi-label dataset. b) Examples of ring formations described in the USPTO dataset; these include ring closing metathesis and the Diels Alder reaction. The USPTO dataset was filtered using the crude measure of the difference in the number of rings between products and reactants to obtain a dataset describing ring formations.

In this work, we limited the reaction templates to those describing ring formations by using the crude measure of the difference in the number of rings between the products and reactants. We retain only reactions in which the difference is greater than one, thereby allowing multiple ring formations in one synthetic step. The bipartite reaction graph is then built, describing the retrosynthetic space corresponding to ring formations and queried to build a domain specific multi-label dataset (Figure 1, Table 1). Compared to the entirety of the datasets from which the ring formations were extracted, we found that ring-forming reactions constitute 4.5 % and 5.8 % of the USPTO and Reaxys® datasets, respectively. An even smaller percentage of all the templates extracted from these datasets correspond to ring formations (Table 1). Therefore, an all-encompassing classifier that considers all extracted

† Copyright © 2019 Elsevier Life Science IP Limited except certain content provided by third parties. Reaxys is a trademark of Elsevier Life Science IP Limited
templates to predict which can be applied in any given situation, has the difficult task of differentiating templates that can be applied. We propose a specialized ring formation classifier called ‘Ring Breaker’, that overcomes the current limitations of predicting ring syntheses. This can be injected as needed into a full retrosynthetic tool to enable access to well documented, as well as previously unreported ring systems.

Table 1: Breakdown of a ring formation specific dataset obtained from the USPTO and Reaxys datasets by considering all reactions in which there is a ring change greater than one between products and reactants. Compared to the entirety of each respective dataset, the percentage of reactions corresponding to ring formations is 4.5 % and 5.8 % respectively, and the percentage of corresponding templates even lower. This shows that ring-forming reactions could be poorly represented considering the whole dataset and could fall within the noise.

| Dataset          | Ring Formations | Percentage of Ring-forming Reactions in Dataset | Ring Formation Templates | Percentage of Ring Formation Templates |
|------------------|-----------------|-----------------------------------------------|--------------------------|----------------------------------------|
| USPTO 1976-2016  | 53,698          | 4.5 %                                         | 6,389                    | 2.1 %                                  |
| Reaxys®         | 265,716         | 5.8 %                                         | 15,662                   | 4.3 %                                  |

Results and Discussion

Prediction of well-known ring formations

To demonstrate the utility of ‘Ring Breaker’ on common ring formations, we retrieved examples from the organic chemistry literature which exemplify commonly used ring-forming reactions (Figure 2). In each case, the first applicable prediction (i.e. the first prediction for which application of the predicted template successfully generated a set of reactants) has been shown. The predictions were made by two models trained on the USPTO or Reaxys data respectively, to determine whether a difference in performance could be observed between the two datasets, considering their differing size and coverage as determined in a previous study.29
Figure 2: Substrates chosen from the literature for which named ring-forming reactions were known (substrates for which predictions failed are shown in Figures 3, 4, and 5). These were used to compare the performance of ‘Ring Breaker’ with our general retrosynthetic model. The substrates chosen were chosen such that there was limited functionality apart from the ring system, to emulate the simplified structures on which a ring-forming reaction may be necessary. Additionally, the simple substrates chosen allowed for evaluation of the two models ‘Ring Breaker’ and the general model for their performance on ring formations. The additional functionality present in more complex structures can detract from the ring-forming task as outlined in this paper.

We found that for the 20 substrates tested in this part of the study, ‘Ring Breaker’ performed better for the prediction of ring formations on average (Table 2). For each molecule, the top 50 predictions were restricted to those describing ring formations to compare the models. The general or standard model predicts several templates that do not encode ring formations, and these offer alternative routes to the functionalization of the ring system or entire fragment, as shown in the following examples (Figure 2-5). However, for comparative purposes, those shown in the table have been restricted solely to ring formations to determine whether: 1) the quantity of ring formations in the top 50 templates predicted
varied between the trained models and datasets, and 2) to determine the rank (i.e. the placement of the prediction) out of the top 50 predicted templates. Whilst one can argue that increasing the search space beyond the top 50 templates will give rise to more predicted templates that encode ring formations, this serves to increase the search breadth of the subsequent tree search. To this end, the computational expense associated with enumerating the tree to this extent must be balanced with the probability and accuracy of the prediction, where accuracy refers to the ability to predict a feasible set of reactants. In previous studies, the top 50 templates were used alongside a cumulative probability cutoff of 0.995 as a stopping criterion for further expansion of the search breadth. As such, it is unlikely that predictions beyond the top 50 templates will be enumerated even if considered due to their large cumulative probability, and there is no guarantee a predicted template can be successfully applied to yield a set of reactants. Therefore, this reflects to our measure of comparison outlined previously, the quantity of ring formations in the top 50 templates predicted and their rank.

Table 2: Prediction of named ring-forming reactions comparing the results from ‘Ring Breaker’ with that of our general retrosynthesis model for the top 50 single step predictions. The output was restricted to the top 50 reactions as this reflects the strategy used during the tree search to enumerate full retrosynthetic pathways, in which the network plays a guiding role. The results show that ‘Ring Breaker’ can predict ring-forming reactions 90 % of the time for the substrates tested using the Reaxys dataset; this is a significant improvement on the general model, which only predicts ring-forming reactions 45 % of the time. On average more ring-forming reactions are predicted and are ranked more highly for ‘Ring Breaker’ than the general model. Furthermore, we observed that the model trained on the Reaxys dataset was superior to that trained on the USPTO dataset for the substrates examined when using ‘Ring Breaker’.

Legend:
number of ring formations predicted (rank of first applicable ring formation)
e.g. the value 5(2) refers to five ring-formations predicted in the top 50 templates and first applicable one is the second prediction overall

**Bold** = first prediction by ‘Ring Breaker’ matches literature precedent

| Compound ID                  | Ring Breaker | General Model | Figure |
|------------------------------|--------------|---------------|--------|
|                              | Reaxys USPTO | Reaxys USPTO  |        |
| Paal-Knorr-furan synthesis   |              |               |        |
| 1                            | 1(1)         | 0(0)          | 0(0)   | Figure 4 |
| 2                            | 1(1)         | 0(0)          | 1(1)   | Figure 4 |
| Paal-Knorr-pyrrole synthesis |              |               |        |
| 3                            | 1(1)         | 1(3)          | 1(2)   | Figure 4 |
| 4                            | 1(2)         | 0(0)          | 0(0)   | Figure 4 |
| Paal-Knorr-thiophene synthesis |            |               |        |
| 5                            | 7(1)         | 3(1)          | 0(0)   | 3(8)   | Figure 5 |
| 6                            | 0(0)         | 0(0)          | 0(0)   | Figure 5 |
Across the substrates examined (Figure 2) we found that ‘Ring Breaker’ was able to suggest a ring-forming template in 98 % of cases using a model trained on Reaxys, compared to 45 % of cases when using the general model (Table 2). We also determined that on average the number of ring-forming templates predicted within the top 50 predictions by ‘Ring Breaker’ exceeded that predicted by the general model. These were also ranked higher than the first applicable ring-forming template predicted by the general model. Thereby, we established a clear benefit for the case of using a specialized ring-forming model, in conjunction with the general model, to increase the likelihood of predicting a ring-forming reaction during enumeration of the search tree. In addition, we found that the model trained on the Reaxys dataset outperformed that trained on the UPSTO dataset for the 20 cases examined using ‘Ring Breaker’ (Table 2). More comparable results were obtained for the general model for the prediction of ring formations, whereby the model can predict ring-forming templates in 45 % of cases compared to 55 % of cases for the Reaxys and USPTO datasets respectively; the reverse of the trend observed for ‘Ring Breaker’.
We further examined whether the first applicable template predicted by ‘Ring Breaker’ was able to regenerate the reactants shown in the literature or those expected. In 65 % of the cases examined (Table 2), the model trained on the Reaxys dataset was able to correctly identify the template leading to the expected reactants, in comparison to 50 % of cases when using the model trained on the USPTO dataset. The examples shown henceforth serve to illustrate: 1) the strengths and drawbacks of the template-based approach, 2) the differences between ‘Ring Breaker’ and the general model regarding the first applicable template.

The Diels-Alder reaction is one of the most well-known ring-forming reactions, and commonplace in an undergraduate chemist’s education. However, the general model fails to predict the template leading to the correct set of reactants as shown in Figure 3, for substrates 7 and 8 (Figure 2). The Diels-Alder approach cannot be predicted by either the USPTO or Reaxys models for the synthesis of quinolines (Figure 3, 8). However, when applied to the substituted cyclohexene (Figure 3, 7), the model is able to successfully identify the diene and dienophile used in the literature with a high probability for the USPTO (p=0.951) and Reaxys (p=0.998) ‘Ring Breaker’ models.\textsuperscript{30} In addition, the synthesis of dihydroisoquinolines (Figure 3, 9) via the Bischler-Napieralski reaction was successfully predicted by both the USPTO (p=0.997) and Reaxys (p=0.976) ‘Ring Breaker’ models, leading to β-arylethylamide, as reported in the literature (Figure 3).\textsuperscript{31} In the case of both, the Diels-Alder and Bischler-Napieralski reaction, the general model consistently predicts alternate, and feasible, strategies leading to the ring containing fragment; this demonstrates the utility of ‘Ring Breaker’ as a method of focusing the synthetic strategy toward ring formations at a given step. One exception we found to ‘Ring Breakers’ superiority, is the prediction of the dihydroisoquinoline (Figure 3, 9) using the USPTO standard model, where the standard model predicts the ring formation with a high probability (p=0.966).
Figure 3: Predictions for the Diels–Alder and Bischler–Napieralski reaction using ‘Ring Breaker’ and the standard model on compounds obtained from the literature employing commonly used ring formations. For each prediction, the patent id or Reaxys id corresponding to the template used has been given as a precedent. In all cases, the probability ‘p(x)’ of predicting the template for the given compound has been shown, where ‘x’ refers to the prediction’s rank (i.e. ‘p(1) = 0.983’ means the first prediction with an associated probability of 0.983). In cases where the precursors have been highlighted in a box, the predicted disconnection matches that reported in the literature. The literature reference from which the compound was obtained is given for each example. We have refrained from exhaustively showing all possible disconnections and have chosen the first prediction that can be applied to generate a set of reactants, regardless of whether they reflect the ‘ground truth’, to show the raw predictions. Models trained
on the USPTO and Reaxys datasets perform the same in all cases shown above with the exception of the Bischler-Napieralski reaction predicted for compound 9 using the USPTO dataset.

The Paal-Knorr series of ring synthesis can be used to provide access to substituted furans, pyrroles, and thiophenes (Figure 5). Its versatility and structural similarity between components, makes it an interesting case for testing retrosynthetic disconnections. The heteroaromatic ring varies by a single nitrogen, oxygen, or sulfur atom, and the ground truth disconnection in each case is almost the same. Figure 4 shows that the disconnections predicted by the model are dependent on the dataset. Both the Reaxys and USPTO datasets contain complementary templates, whereby the model trained on each dataset can predict retrosynthetic disconnections in some cases but not others. For the case of Paal-Knorr-furan synthesis, the USPTO data is not able to predict a disconnection, whereas the model trained on the Reaxys data predicts the literature disconnection with a high probability (0.983 and 0.998). The case of pyrrole synthesis further highlights an interesting problem, whereby the correct disconnection can be predicted by the USPTO model for a simplified ring system (Figure 4, 1). However, when the molecular complexity around the ring system was increased by replacement of the methyl groups with phenyl groups (Figure 4, 2), the model failed to respond to the change and was not able to predict an outcome. On the other hand, the model trained on Reaxys was able to correctly identify the ring system (Figure 4, 2) and predict the retrosynthetic disconnection reported in the literature. This highlights the underlying problem of template based approaches. The templates must be specific enough to yield a substructure match to the compound they are applied to and produce feasible reactants, whilst being general enough to be applicable across a broad range of suitable compounds without being promiscuous. Balancing these two requirements means that in cases such as the pyrrole synthesis, the template predicted for the simplified ring system cannot be applied to the more complex ring system shown. In contrast to ‘Ring Breaker’, the general model is only able to predict the correct set of precursors for compound 2, which uses the Paal-Knorr-furan synthesis.
Figure 4: Predictions for the Paal-Knorr-furan and pyrrole synthesis using ‘Ring Breaker’ and the standard model on compounds obtained from the literature employing commonly used ring formations. For each prediction, the patent id or Reaxys id corresponding to the template used has been given as a precedent. In all cases, the probability ‘p(x)’ of predicting the template for the given compound has been shown, where ‘x’ refers to the prediction’s rank (i.e. ‘p(1) = 0.983’ means the first prediction with an associated probability of 0.983). In cases where the precursors have been highlighted in a box, the predicted disconnection matches that reported in the literature. The literature reference from which the compound was obtained is given for each example where appropriate. We have refrained from exhaustively showing all possible disconnections and have chosen the first prediction that can be applied to generate a set of reactants, regardless of whether they reflect the ‘ground truth’, to show the raw predictions. Models trained on the USPTO and Reaxys datasets are complementary in this case, whereby predictions that can be made with one may not necessarily be made with the other.
In the case of thiophene synthesis, the model cannot identify a suitable template for the simple ring system (Figure 5, 6), regardless of the dataset used. However, in the more complex case (Figure 5, 5), it focuses its efforts on the morpholine ring, predicting the shown disconnection (Figure 5) with a high probability. Whilst this does not indicate that the system cannot predict thiophene formation, it alludes to the fact that these templates may be under represented underlying dataset.

**Paal–Knorr-thiophene synthesis**

| Compound | Reaxys | USPTO |
|----------|--------|-------|
| 5        |        |       |
|          |        |       |

**Figure 5**: Predictions for the Paal-Knorr-thiophene synthesis using ‘Ring Breaker’ and the standard model on compounds obtained from the literature employing commonly used ring formations. For each prediction, the patent id or Reaxys id corresponding to the template used has been given as a precedent. In all cases, the probability ‘p(x)’ of predicting the template for the given compound has been shown, where ‘x’ refers to the prediction’s rank (i.e. ‘p(1) = 0.983’ means the first prediction with an associated probability of 0.983). In cases where the precursors have been highlighted in a box, the predicted disconnection matches that reported in the literature. The literature reference from which the compound was obtained is given for each example where appropriate. We have refrained from exhaustively showing all possible disconnections and have chosen the first prediction that can be applied to generate a set of reactants, regardless of whether they reflect the ‘ground truth’ to show the raw predictions. The model fails to predict the synthesis of thiophenes in this case as it focuses on the morpholine ring for compound 5 and does not have template that matching the molecular environment in compound 6.

**Prediction of Fragments**

We performed a one-step retrosynthetic analysis to focus on the ring-forming step required to synthesize a range of ring containing subsets from the ZINC database (Figure 6). Examining a range of ring systems, from the most commonly occurring (in > 100 K substances) to the rarest (in < 1 K substances) we found that ‘Ring Breaker’ exhibited superior performance over the general models across all subsets examined,
regardless of the dataset used. The reason for this may be two-fold. First ‘Ring Breaker’ is exclusively limited to ring formations, so application of a promiscuous template may still lead to a result. However, this alone is not likely to lead to the large difference in performance observed. Second, the limited and domain specific training set better allows the model to learn in which context ring-forming templates can be predicted. This is in comparison to the general model in which ring-forming templates can be drowned out in the noise by more frequently-occurring templates, as there are several possible options for disconnections aside from the ring-forming templates.

Furthermore, we found that the Reaxys ‘Ring Breaker’ outperformed that trained on the USPTO dataset (Figure 7). This is in contrast to our previous observations, where we reported that the ability to generate synthetic routes for the general model did not depend on the training dataset.29 We have now determined that for the domain specific case of ring formations there is a clear effect arising from the training set used, attributed to the number and diversity of the samples available to the network for training.
Figure 6: Examining the predictive capability of ‘Ring Breaker’ to assess the synthetic accessibility of a range of ring systems obtained from the ZINC database. The heavy atom count for compounds in each subset is plotted against the number of compounds corresponding to the heavy atom count for the raw dataset and for each model to give the distribution of the subset before (blue) and after prediction with each model. The subsets correspond to a) common (occurring in greater than 100 K substances) b) present in drugs c) present in nature d) rare (occurring in under 1 K substances). The ‘Ring Breaker’ and general models were compared for each subset and between each training set, Reaxys and USPTO, for the prediction of one-step retrosynthesis for each subset. The ‘Ring Breaker’ trained on Reaxys (green) consistently outperformed all other models, and the ‘Ring Breaker’ far outperformed the standard model regardless of the training dataset. ‘Ring Breaker’ exhibits the best performance for ring systems between 5 and 20 heavy atoms in size, and the predictions follow the natural distribution with a slight divergence as not all ring systems can be predicted.

The performance of the model on ring systems classed as ‘rare’ in the ZINC database, is surprising (Figure 6). These rings systems can be assumed to be difficult to access synthetically, yet the model is able to
predict a one-step retrosynthetic disconnection in most cases. Examples are shown in Figure 7, with their corresponding patent precedent, which refers to the patent containing the reaction from which the predicted template was extracted. Whilst the retrosynthetic disconnection may not be used as described in the forward sense, we show that ‘Ring Breaker’ can act as an idea generator from which a trained synthetic chemist can build upon.

Figure 7: a) and b) Examples of ring systems from the ZINC database that could not be predicted by either model, trained on the USPTO or Reaxys data. The templates must be specific enough to yield a substructure match to the compound they are applied to and produce feasible reactants, whilst being general enough to be applicable across a broad range of suitable compounds without being promiscuous. Balancing these two requirements means that in cases such as the furan synthesis, the template predicted for more complex ring systems cannot be applied to the simple ring system shown. c) Predicted retrosynthetic disconnections are shown for a selection of compounds in the rare subset. For each prediction, the patent id corresponding to the template used has been given as a precedent.
In some cases (e.g. furan synthesis) that could not be predicted for the unsubstituted ring system (Figure 7), we have previously observed that a disconnection could be predicted from the substituted ring system (Figure 4). In such cases, it is a problem of template availability and the underlying dataset on which the model is trained. The template must be able to describe the changing atoms and bonds in the reaction, and therefore is specific to the reaction from which it was extracted in terms of the local molecular environment. Yet the template must also be able to be generally applied to a variety of compounds containing the same molecular environment from which the template was first extracted. Finally, the network is trained on the product of the reactions, and the uses corresponding templates as labels. Therefore, for the network to ‘learn’ in which context a given template can be applied, there must be a sufficient number of diverse examples containing the same local molecular environment to which the template has a substructure match. In this way, the network is better able to generalize to which compound a given template can be applied, and may explain why compounds, and by association templates that occur frequently within the dataset are better ‘understood’ by the network.

Accessing Virtual Fragments – ‘Rings of the Future’

Since the exhaustive computational enumeration of heteroaromatic ring systems first described by Pitt et al.,19 several articles have detailed the enumeration of ring systems,17, 20-22 yet none have proposed syntheses to access the motifs described. We examined ‘Ring Breaker’ in the context of novel ring systems, the so-called ‘Rings of the Future’.19 Having trained the model on patent data up to 2016, we selected two novel ring systems from the literature for which the synthesis were reported in 2016,24 and ensured that they were not present in the dataset. Rather than predicting the full synthetic route, we focused on the ring-forming step. We found the first applicable template in both cases corresponded to the disconnection reported in the literature (Figure 8). This further demonstrates the applicability of ‘Ring Breaker’ to previously unseen ring systems and shows how the approach can be used as an idea generator to explore novel ring-based scaffolds. Furthermore, the literature or patent precedent allows for researchers to lookup reaction conditions and experimental procedures.
Figure 8: Performance of ‘Ring Breaker’ on two ‘Ring of the Future’ compounds that were synthesized in 2016, for which the predicted disconnection matches that described in the literature synthesis (highlighted in the green boxes). The predictions are shown for the USPTO dataset, which contains reactions from patents up to 2016. The ‘Ring of the future’ compounds were not part of the USPTO dataset, and as such their syntheses were not part of the training of the model, thereby demonstrating that ‘Ring Breaker’ is capable of predicting and suggesting ideas for the synthesis of unprecedented ring systems. For each prediction, the patent id or Reaxys id corresponding to the template used has been given as a precedent. In all cases the probability ‘p(x)’ of predicting the template for the given compound has been shown, where ‘x’ refers to the predictions rank i.e. ‘p(1) = 0.983’ means the first prediction with an associated probability of 0.983. In cases where the precursors have been highlighted in a box, the predicted disconnection matches that reported in the literature.
Incorporation into Computer Aided Synthetic Planning Tools

In their current state, template-based synthetic planning tools, which rely on a classification network to predict which template can be applicable in a given context, struggle to differentiate ring-forming reactions from the multitude of other suitable reactions that can be applied to any given compound. This is due to the large number of templates available and the relatively low frequency of ring-forming reactions within the datasets (Table 1). As such, in cases where a ring disconnection may be suitable or may lead to a more efficient synthetic route, the network and subsequent tree search do not often prioritize, apply, and generate synthetic routes which proceed through ring formations. To overcome this problem, the ‘Ring Breaker’ model can be viewed as a specialist which can be consulted at various stages of the tree search to yield routes that proceed through ring formations, in addition to its use as a stand-alone tool. In Figure 9 we demonstrate one such use case, whereby the heterocycle is formed as an alternative to its functionalization to yield the queried compound. This methodology, using a domain specific model in conjunction with a general model, can be extended to other areas of synthetic chemistry in which the data is limited and domain specific knowledge (i.e. a specialist) is required.

Figure 9: Exemplary synthesis demonstrating the use of ‘Ring Breaker’ to augment synthesis planning tools. ‘Ring Breaker’ can be used to guide retrosynthetic tree search to explore areas exploiting ring-forming strategies, thereby allowing for the convergent synthesis of rings, rather than linear functionalization.
Conclusions

We have developed a methodology for proposing syntheses to and assessing the synthetic accessibility of ring systems using a specialized ring-forming neural network called ‘Ring Breaker’; remarkably, our ‘Ring Breaker’ exhibits superior performance over the general model. It can be used either as a stand-alone tool, or it can be incorporated into synthetic planning tools. In addition, we have described a scalable and representative method for the generation of labels for computer aided synthesis planning tasks. The model, when trained separately on either the USPTO or Reaxys datasets, shows that the model trained on Reaxys outperforms that trained on the USPTO dataset. Notably, in cases where one fails the other is often able to suggest a suitable disconnection, and in this sense the two datasets are complementary. We determined that for a series of common ring formations and ring fragment subsets obtained from the ZINC database, the models can suggest suitable disconnections for the ring-forming step in most cases. Furthermore, in all subsets examined ‘Ring Breaker’ far outperforms the predictive capability of the general model.

Although the models consistently underperform for ring systems smaller than 5 heavy atoms, for which fragments are not common because of torsional strain. The distribution of compounds for which ring-forming steps could be predicted closely follows the raw distribution of ring systems in the subsets. Our study was extended to previously unseen ring systems, for which the model trained solely on the USPTO dataset could predict the ring-forming step reported in the literature as the first applicable template. This highlights the utility of the method across the range of common, rare, and previously unseen ring systems, where the tool can be used as an idea generator. Given that the model varies in predictive capability depending on the substitution of the ring system, we established that this originates from the availability of a suitable template and by association the underlying dataset. Whilst suitable templates describing the reaction are suggested, they cannot be applied as they do not share an exact sub-structure match to the query compound.

We propose that the specialized model can be used alongside the current ‘all encompassing’ model currently used in synthetic planning tools, and as a stand-alone idea generator for proposing retrosynthetic disconnections to a wide range of ring systems, including those previously unseen. This has implications in the pharmaceutical, agrochemical, and dye industries, to name a few, where ring systems are an important and widely used motif at the center of many marketed compounds. Furthermore, we propose that this methodology can be extended to other specialized domains within synthesis planning tasks where the data may be limited and domain specific knowledge (i.e. a specialist) is required.
Methods

Reaction Datasets and Template Extraction

The United States Patent Office extracts (USPTO) ranging from the years 1976 to 2016 is publicly available. This is split into granted and applied patents and is openly available for use by the community. The Reaxys dataset is commercially available, provided by Elsevier under licensing agreements. The ring subsets described were obtained from the ZINC database and used as is.

All reactions were atom-mapped and classified using the commercially available Filbert and HazELNut packages (v. 3.1.8) provided by NextMove software. These were subsequently processed using RDKit and RDChiral for template extraction. The bipartite reaction graph was built using NetworkX and queried to yield the multi label dataset as described previously.

Classification Network

The template library was constructed by filtering the respective dataset for templates that occurred a minimum of 3 times. In all cases duplicate reactions were removed prior to filtering. Products were represented as extended connectivity fingerprints (ECFP) with a radius of 2, using the Morgan algorithm in RDKit. Whereas, templates were represented as binarized labels in a one-vs-all fashion using the scikit-learn library using the ‘LabelBinarizer’. Both the input ECFP4 and output vectors were precomputed. Training, validation, and test sets were constructed as a random 90/5/5 split of the datasets, using a random state of 42, where the datasets were shuffled prior to splitting. This was conducted using the scikit-learn library.

The network framed as a supervised multiclass classification problem was trained using Keras with Tensorflow as the backend, the Adam optimizer with an initial learning rate of 0.001, and categorical cross entropy as the loss function. The learning rate was decayed on plateau by a factor of 0.5, where the plateau was considered as no improvement of the validation loss after 5 epochs. The top 1, 5, 10, and 50 accuracies were monitored throughout the training process, and the loss on the validation set was used with early stopping (patience 10) to determine the number of epochs for which the model was trained. The standard model deployed within this study was trained as described in our previous work.
Figure 10: Architecture used to train the ‘rollout’ policy taking molecules represented as ECFP4 as input, through a fully connected layer of 512 nodes, ELU as the activation function. Followed by a dropout of 0.5 and softmax output layer.

Declarations

Availability of Data and Materials

Reaxys datasets were used with permissions. Filbert, NameRxn and HazelNut were used for atom-mapping and classification under license from NextMove software. All code used in the production of this work will be made available at: https://github.com/reymond-group/RingBreaker

Competing Interests

The authors declare that they have no competing interests.

Funding

Amol Thakkar is supported financially by the European Union’s Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement no. 676434, “Big Data in Chemistry” (“BIGCHEM,” http://bigchem.eu).

Author contributions

A. Thakkar designed and conducted the research and wrote the manuscript. A. Thakkar and E.J. Bjerrum designed the concept. A. Thakkar and N.Selmi designed and selected the compound sets. E.J. Bjerrum, O. Engkvist and J.L. Reymond supervised the project.
Acknowledgements

The authors would like to thank NextMove software for providing access to NameRxn for atom-mapping. Elsevier for providing the Reaxys dataset. The Molecular AI group at AstraZeneca and Reymond group at the University of Bern for their support.

References

1. Green, C. P.; Engkvist, O.; Pairaudeau, G., The convergence of artificial intelligence and chemistry for improved drug discovery. Future Medicinal Chemistry 2018, 10 (22), 2573-2576.
2. Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T., The rise of deep learning in drug discovery. Drug Discovery Today 2018, 23 (6), 1241-1250.
3. Corey, E.; Long, A.; Rubenstein, S., Computer-assisted analysis in organic synthesis. 1985, 228 (4698), 408-418.
4. Corey, E. J.; Wipke, W. T., Computer-Assisted Design of Complex Organic Syntheses. Science 1969, 166 (3902), 178.
5. Pensak, D. A.; Corey, E. J., LHASA—Logic and Heuristics Applied to Synthetic Analysis. In Computer-Assisted Organic Synthesis, AMERICAN CHEMICAL SOCIETY: 1977; Vol. 61, pp 1-32.
6. Segler, M. H. S.; Waller, M. P., Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction. Chemistry – A European Journal 2017, 23 (25), 5966-5971.
7. Segler, M. H. S.; Preuss, M.; Waller, M. P., Planning chemical syntheses with deep neural networks and symbolic AI. Nature 2018, 555, 604.
8. Coley, C. W.; Thomas, D. A.; Lummiss, J. A. M.; Jaworski, J. N.; Breen, C. P.; Schultz, V.; Hart, T.; Fishman, J. S.; Rogers, L.; Gao, H.; Hicklin, R. W.; Plehiers, P. P.; Byington, J.; Piotti, J. S.; Green, W. H.; Hart, A. J.; Jamison, T. F.; Jensen, K. F., A robotic platform for flow synthesis of organic compounds informed by AI planning. 2019, 365 (6453), eaax1566.
9. 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 Central Science 2017, 3 (10), 1103-1113.
10. Klucznik, T.; Mikulak-Klucznik, B.; McCormack, M. P.; Lima, H.; Szymkuć, S.; Bhowmick, M.; Molga, K.; Zhou, Y.; Rickershauser, L.; Gajewska, E. P.; Toutchkine, A.; Dittwald, P.; Startek, M. P.; Kirkovits, G. J.; Roszak, R.; Adamski, A.; Sieredzińska, B.; Mrksich, M.; Trice, S. L. J.; Grzybowski, B. A., Efficient Syntheses of Diverse, Medically Relevant Targets Planned by Computer and Executed in the Laboratory. Chem 2018, 4 (3), 522-532.
11. Molga, K.; Dittwald, P.; Grzybowski, B. A., Navigating around Patented Routes by Preserving Specific Motifs along Computer-Planned Retrosynthetic Pathways. Chem 2019, 5 (2), 460-473.
12. Engkvist, O.; Norrby, P.-O.; Selmi, N.; Lam, Y.-h.; Peng, Z.; Sherer, E. C.; Amberg, W.; Erhard, T.; Smyth, L. A., Computational prediction of chemical reactions: current status and outlook. Drug Discovery Today 2018, 23 (6), 1203-1218.
13. Jia, X.; Lynch, A.; Huang, Y.; Danielson, M.; Lang’at, I.; Milder, A.; Ruby, A. E.; Wang, H.; Friedler, S. A.; Norquist, A. J.; Schrier, J., Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis. Nature 2019, 573 (7773), 251-255.
14. Taylor, R. D.; MacCoss, M.; Lawson, A. D. G., Rings in Drugs. Journal of Medicinal Chemistry 2014, 57 (14), 5845-5859.
15. Roughley, S. D.; Jordan, A. M., The Medicinal Chemist’s Toolbox: An Analysis of Reactions Used in the Pursuit of Drug Candidates. Journal of Medicinal Chemistry 2011, 54 (10), 3451-3479.
16. Boström, J.; Brown, D. G.; Young, R. J.; Keserü, G. M., Expanding the medicinal chemistry synthetic toolbox. *Nature Reviews Drug Discovery* 2018, 17, 709.
17. Visini, R.; Arús-Pous, J.; Awale, M.; Reymond, J.-L., Virtual Exploration of the Ring Systems Chemical Universe. *Journal of Chemical Information and Modeling* 2017, 57 (11), 2707-2718.
18. Liu, Y.; Gray, N. S., Rational design of inhibitors that bind to inactive kinase conformations. *Nature Chemical Biology* 2006, 2 (7), 358-364.
19. Pitt, W. R.; Parry, D. M.; Perry, B. G.; Groom, C. R., Heteroaromatic Rings of the Future. *Journal of Chemical Information and Modeling* 2009, 52 (9), 2952-2963.
20. Mok, N. Y.; Brown, N., Applications of Systematic Molecular Scaffold Enumeration to Enrich Structure–Activity Relationship Information. *Journal of Chemical Information and Modeling* 2017, 57 (1), 27-35.
21. Tyagarajan, S.; Lowden, C. T.; Peng, Z.; Dykstra, K. D.; Sherer, E. C.; Krška, S. W., Heterocyclic Regioisomer Enumeration (HREMS): A Cheminformatics Design Tool. *Journal of Chemical Information and Modeling* 2015, 55 (6), 1130-1135.
22. Ward, R. A.; Kettle, J. G., Systematic Enumeration of Heteroaromatic Ring Systems as Reagents for Use in Medicinal Chemistry. *Journal of Medicinal Chemistry* 2011, 54 (13), 4670-4677.
23. Baylon, J. L.; Clifone, N. A.; Gulcher, J. R.; Chittenden, T. W., Enhancing Retrosynthetic Reaction Prediction with Deep Learning Using Multiscale Reaction Classification. *Journal of Chemical Information and Modeling* 2019, 59 (2), 673-688.
24. Silva Júnior, P. E.; Rezende, L. C. D.; Gimenes, J. P.; Maltarollo, V. G.; Dale, J.; Tossini, G. H. G.; Emery, F. S.; Ganesan, A., Synthesis of two ‘heteroaromatic Rings of the Future’ for applications in medicinal chemistry. *RSC Advances* 2016, 6 (27), 22777-22780.
25. Schreck, J. S.; Coley, C. W.; Bishop, K. J. M., Learning Retrosynthetic Planning through Simulated Experience. *ACS Central Science* 2019, 5 (6), 970-981.
26. Schwaller, P.; Gaudin, T.; Lányi, D.; Bekas, C.; Laino, T., “Found in Translation”: predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chemical Science* 2018, 9 (28), 6091-6098.
27. Szymkuć, S.; Gajewska, E. P.; Klucznik, T.; Molga, K.; Dittwald, P.; Startek, M.; Bajczyk, M.; Grzybowski, B. A., Computer-Assisted Synthetic Planning: The End of the Beginning. *Angewandte Chemie International Edition* 2016, 55 (20), 5904-5937.
28. Grzybowski, B. A.; Bishop, K. J. M.; Kowalczyk, B.; Wilmer, C. E., The wired universe of organic chemistry. *Nature Chemistry* 2009, 1, 31.
29. Thakkar, A.; Kogej, T.; Reymond, J.-L.; Engkvist, O.; Bjerrum, E. J., Datasets and Their Influence on the Development of Computer Assisted Synthesis Planning Tools in the Pharmaceutical Domain. *ChemRxiv* 2019.
30. Saunthwal, R. K.; Patel, M.; Verma, A. K., Regioselective Synthesis of C-3-Functionalized Quinolines via Hetero-Diels–Alder Cycloaddition of Azadienes with Terminal Alkynes. *The Journal of Organic Chemistry* 2016, 81 (15), 6563-6572.
31. Movassaghi, M.; Hill, M. D., A Versatile Cyclodehydration Reaction for the Synthesis of Isoquinoline and β-Carboline Derivatives. *Organic Letters* 2008, 10 (16), 3485-3488.
32. Rao, H. S. P.; Jothilingam, S.; Facile Microwave-Mediated Transformations of 2-Butene-1,4-diones and 2-Butyne-1,4-diones to Furan Derivatives. *The Journal of Organic Chemistry* 2003, 68 (13), 5392-5394.
33. Minetto, G.; Raveglia, L. F.; Sega, A.; Taddei, M., Microwave-Assisted Paal–Knorr Reaction – Three-Step Regiocontrolled Synthesis of Polysubstituted Furans, Pyrroles and Thiophenes. *European Journal of Organic Chemistry* 2005, 2005 (24), 5277-5288.
34. ZINC. [http://zinc.docking.org/rings/subsets/](http://zinc.docking.org/rings/subsets/) (accessed 27/08/2019).
35. Lowe, D. Chemical reactions from US patents (1976-Sep2016). https://figshare.com/articles/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873 (accessed 31-04-2018).
36. Reaxys®, Copyright © 2019 Elsevier Limited except certain content provided by third parties. Reaxys is a trademark of Elsevier Limited.
37. NextMove Software | HazELNut.
38. RDKit: Open-source cheminformatics; http://www.rdkit.org.
39. Coley, C. W.; Green, W. H.; Jensen, K. F., RDChiral: An RDKit Wrapper for Handling Stereochemistry in Retrosynthetic Template Extraction and Application. Journal of Chemical Information and Modeling 2019, 59 (6), 2529-2537.
40. Aric A. Hagberg, D. A. S. a. P. J. S., Exploring network structure, dynamics, and function using NetworkX. Proceedings of the 7th Python in Science Conference (SciPy2008) 2008.
41. Rogers, D.; Hahn, M., Extended-Connectivity Fingerprints. Journal of Chemical Information and Modeling 2010, 50 (5), 742-754.
42. Pedregosa, F. a. V., G. and Gramfort, A. and Michel, V. and Thirion, B. and Grisel, O. and Blondel, M. and Prettenhofer, P. and Weiss, R. and Dubourg, V. and Vanderplas, J. and Passos, A. and Cournapeau, D. and Brucher, M. and Perrot, M. and Duchesnay, E., Scikit-learn: Machine Learning in {P}ython. Journal of Machine Learning Research 2011, 12, 2825-2830.
43. Chollet, F. (2015) keras, GitHub. https://github.com/fchollet/keras.
44. Martín Abadi, A. A., Paul Barham, Eugene Brevdo,; Zhifeng Chen, C. C., Greg S. Corrado, Andy Davis,; Jeffrey Dean, M. D., Sanjay Ghemawat, Ian Goodfellow,; Andrew Harp, G. I., Michael Isard, Rafal Jozefowicz, Yangqing Jia,; Lukasz Kaiser, M. K., Josh Levenberg, Dan Mané, Mike Schuster,; Rajat Monga, S. M., Derek Murray, Chris Olah, Jonathon Shlens,; Benoit Steiner, I. S., Kunal Talwar, Paul Tucker,; Vincent Vanhoucke, V. V., Fernanda Viégas,; Oriol Vinyals, P. W., Martin Wattenberg, Martin Wicke,; Yuan Yu, a. X. Z., TensorFlow: Large-scale machine learning on heterogeneous systems. 2015.
45. Diederik P. Kingma, J. B., Adam: A Method for Stochastic Optimization. arXiv pre-prints 2014.