

Computer Aided Synthesis Prediction to Enable Augmented Chemical Discovery and Chemical Space Exploration

Inaugural dissertation
of the Faculty of Science,
University of Bern

Presented by
Amol Thakkar
From London, United Kingdom

Supervisor of the doctoral thesis:
Prof. Dr. Jean-Louis Reymond

Department of Chemistry, Biochemistry, and Pharmaceutical Science

Original document saved on the web server of the University Library of Bern.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) licence. To see the licence go to <https://creativecommons.org/licenses/by-nc-nd/4.0/> or write to Creative Commons, PO Box 1866, Mountain View, CA 94042, USA.

Computer Aided Synthesis Prediction to Enable Augmented Chemical Discovery and Chemical Space Exploration

Inaugural dissertation
of the Faculty of Science,
University of Bern

Presented by
Amol Thakkar
From London, United Kingdom

Supervisor of the doctoral thesis:
Prof. Dr. Jean-Louis Reymond

Department of Chemistry, Biochemistry, and Pharmaceutical Science

Accepted by the Faculty of Science.

Bern, 11th February 2022

The Dean
Prof. Dr. Zoltan Balogh

"You have to solve a problem, right? If you have to become a mathematician or an engineer to solve the problem, so what? The last thing the piece of paper that is a Ph.D. should tell you is what field you're in; it should tell you only that you can learn. You have to pull up your sleeves and do something real."

-Enrico Clementi, IBM Research, 1985

Acknowledgements

First and foremost, I would like to thank my family, friends, and colleagues for their support and encouragement. I have had the privilege of living in both Switzerland and Sweden during my PhD, working in both industry and academia. I would like to thank the countless people I have interacted with along the way, be it professional or personal. You have all shaped my view of the world, taught me about different ways of life, and inspired me in your own unique ways.

In earnest my doctoral journey started before my arrival in Bern. I would like to thank my friends, colleagues, and supervisors at the University of St Andrews, particularly Prof Michael Buehl and Dr John Mitchell for their advice, early education in the field of computational chemistry, and helping me realise the power of taking learning into your own hands. To my colleagues at Pfizer who challenged my interests in the digitization of organic synthesis and predictive modelling, and their support during my time there. Prof Johannes Kaestner with whom I started my doctoral journey, showed me kindness and support during my short stint in Stuttgart and the mutual realisation of my desire for an industrially based doctorate led to the decision to leave shortly after joining.

Prof Jean-Louis Reymond for giving me the chance to carry out a PhD in Cheminformatics without any prior knowledge of the field or programming. I am grateful for his support and willingness to take a chance. In addition, the interdisciplinary and diverse research environment that he has fostered within the group is a continuous source of ideas and challenging opinions, as well as a great learning ground for developing communication between researchers in different fields. A special thanks to Sacha, Giorgio, Alice, Kris, Daniel, and Mahendra, who were extremely welcoming and helped me settle into the group. Aline, Kris, Giorgio, and Kleni for the constant reminder of how organic synthesis really works in the lab and helping me better understand where synthesis planning can be applied to their projects. Thank you to Philippe, David, Hippolyte, Sven, Josep, Dina, Marion, Marc, Xingguang, and the rest of the group, past and present, for the support you have shown me during the PhD, and especially for being there when we all needed a release from our daily lives. From hiking, climbing, and cycling, to the several beers, rakija, and dinners we have shared together.

Sandra Zbinden, thank you for all the administration you have handled on our behalf. I have been inspired by your efficiency, and the way that you shield us from all bureaucratic matters, allowing us to focus on the research.

Dr Ola Engkvist and Dr Esben Jannik Bjerrum for hosting and supervising me on secondment at AstraZeneca, Gothenburg, Sweden. This marked a period of tremendous personal growth for me. The work presented in this thesis was a result of the contribution made by the community in AstraZeneca, from the chemists who helped give feedback on the tools we developed, and constantly provided us with a list of new features to be added, to the computational teams who provided a wealth of technical expertise. I would especially like to thank my colleagues in the

Molecular AI team, the post-docs, and graduate students with whom I had the privilege of working and supervising. The open discussions at lunch, in the pub, or at various dinners allowed me to develop my ideas in an open constructive environment.

All of those who laid the groundwork for the research carried out in this thesis. From the researchers that came before, the groups currently working in the field, and most importantly providers of open-source community driven software such as RDKit and Tensorflow. Roger, Noel, and John at NextMove software for providing their data and being on hand to provide a wealth of support.

Dr Igor Tetko and Prof Philippe Renaud for agreeing to examine the thesis, the time they have taken and valuable feedback they have given me. An additional thanks to Dr Igor Tetko for organising the BigChem project, and the research fellows for creating such a great community.

Ann-Britt Albinson and Josef Böhm for their hospitality during my time in Sweden, and the members of Zähringerstrasse 19, for making me feel at home in Switzerland.

I am especially grateful for the unwavering support provided by my family, especially my parents, grandparents, and sister. Ida you and your family have provided an exceptional support in this time. Thank you for being there and tolerating me when I was lost in my thoughts, constantly changing my views and perhaps being a nuisance to you all. Your encouragement, love, and understanding has helped me to keep going.

General Abstract

The drug-like chemical space is estimated to be 10^{60} molecules, and the largest generated database (GDB) obtained by the Reymond group is 165 billion molecules with up to 17 heavy atoms. Furthermore, deep learning techniques to explore regions of chemical space are becoming more popular. However, the key to realizing the generated structures experimentally lies in chemical synthesis. The application of which was previously limited to manual planning or slow computer assisted synthesis planning (CASP) models. Despite the 60-year history of CASP few synthesis planning tools have been open-sourced to the community. In this thesis I co-led the development of and investigated one of the only fully open-source synthesis planning tools called AiZynthFinder, trained on both public and proprietary datasets consisting of up to 17.5 million reactions. This enables synthesis guided exploration of the chemical space in a high throughput manner, to bridge the gap between compound generation and experimental realisation.

I firstly investigate both public and proprietary reaction data, and their influence on route finding capability. Furthermore, I develop metrics for assessment of retrosynthetic prediction, single-step retrosynthesis models, and automated template extraction workflows. This is supplemented by a comparison of the underlying datasets and their corresponding models.

Given the prevalence of ring systems in the GDB and wider medicinal chemistry domain, I developed ‘Ring Breaker’ - a data-driven approach to enable the prediction of ring-forming reactions. I demonstrate its utility on frequently found and unprecedented ring systems, in agreement with literature syntheses. Additionally, I highlight its potential for incorporation into CASP tools, and outline methodological improvements that result in the improvement of route-finding capability.

To tackle the challenge of model throughput, I report a machine learning (ML) based classifier called the retrosynthetic accessibility score (RAscore), to assess the likelihood of finding a synthetic route using AiZynthFinder. The RAscore computes at least 4,500 times faster than AiZynthFinder. Thus, opens the possibility of pre-screening millions of virtual molecules from enumerated databases or generative models for synthesis informed compound prioritization.

Finally, I combine chemical library visualization with synthetic route prediction to facilitate experimental engagement with synthetic chemists. I enable the navigation of chemical property space by using interactive visualization to deliver associated synthetic data as endpoints. This aids in the prioritization of compounds. The ability to view synthetic route information alongside structural descriptors facilitates a feedback mechanism for the improvement of CASP tools and enables rapid hypothesis testing. I demonstrate the workflow as applied to the GDB databases to augment compound prioritization and synthetic route design.

Table of Contents

1	Thesis Outline	1
1.1	Thesis Outline and Contributions	1
1.2	Publications	3
1.3	Conference Presentations	5
1.4	Awards	5
1.5	Funding Acknowledgement	5
2	Introduction	6
2.1	Artificial Intelligence in Chemistry	7
2.2	Chemical Data Representation and Formats	9
2.3	Chemical Space	13
2.4	Chemical Reaction Data	15
2.5	Chemical Synthesis and Computer Aided Synthesis Planning	16
2.5.1	A Brief History of Computer Aided Synthesis Planning	17
2.5.2	Deep Learning in Computer Aided Synthesis Planning	20
2.5.3	Building Systems for Computer Aided Synthesis Planning	23
3	Datasets and their influence on the development of computer assisted synthesis planning tools in the pharmaceutical domain	26
3.1	Introduction	27
3.2	Results and Discussion	28
3.2.1	Template Specification	28
3.2.2	Reaction Datasets and Template Coverage	28
3.2.3	Neural-Network Guided Template-Based Retrosynthetic Planning	30
3.2.4	Template Size and Policy Network Accuracy	31
3.2.5	The Effect of Template Library Size on Performance	32
3.2.6	Datasets and Performance	35
3.2.7	Comparison of Test and Reaction Datasets	38
3.2.8	Exemplary Synthetic Routes	38
3.3	Conclusions	40
3.4	Methods	40
3.4.1	Reaction Datasets and Template Extraction	40
3.4.2	Policy Networks	44

3.4.3	Assessing the Number of Successfully Applied Templates of The Top N Predictions	45
3.4.4	Tree Search with 1N-MCTS	45
3.4.5	Implementation	47
3.4.6	Stocks	47
3.4.7	Template Library Size and Performance	48
3.4.8	Datasets and Performance	48
3.5	Availability of data and materials	48
3.6	Author Contributions.....	48
4	“Ring Breaker”: Neural Network Driven Synthesis Prediction of the Ring System Chemical Space	49
4.1	Introduction	50
4.1.1	Overview of Template Based Retrosynthesis	51
4.2	Results and Discussion.....	52
4.2.1	Brute Force Application, Filtering, and Prioritization	52
4.2.2	Diels–Alder and Bischler–Napieralski	57
4.2.3	Paal–Knorr	59
4.2.4	Prediction of ZINC Fragments and DrugBank	62
4.2.5	Accessing Virtual Fragments: “Rings of the Future”	66
4.2.6	Incorporation into Computer Aided Synthetic Planning Tools	68
4.3	Conclusions	69
4.4	Methods.....	71
4.4.1	Reaction Data Sets and Template Extraction.....	71
4.4.2	Dataset Generation.....	71
4.4.3	Classification Network.....	73
4.4.4	Brute Force Application, Filtering, and Prioritization	75
4.4.5	Prediction of ZINC Fragments and DrugBank	75
4.5	Availability of Data and Materials.	75
4.6	Author Contributions.....	75
5	Retrosynthetic accessibility score (RAscore) – rapid machine learned synthesizability classification from AI driven retrosynthetic planning	76
5.1	Introduction	77
5.2	Methods.....	78

5.2.1	AiZynthFinder – a tool for computer aided synthesis planning	78
5.2.2	Retrosynthesis prediction for training set generation	79
5.2.3	Machine learning classifiers for estimation of retrosynthetic accessibility	79
5.2.4	Average linkage as a method for evaluating machine learning based classifiers ...	80
5.3	Results and discussion.....	81
5.3.1	Route statistics from the generation of labels for machine learning classifiers	81
5.3.2	Attempts at using SAscore, SCscore, and SYBA	83
5.3.3	Machine learning classifiers for estimation of retrosynthetic accessibility	83
5.3.4	Prediction time	85
5.3.5	Applicability domain	85
5.3.6	Examples – Limitations of RAscore arising from CASP	87
5.4	Conclusions	90
5.5	Availability of data and materials	90
5.6	Author Contributions.....	90
6	Linking Navigation of Chemical Libraries to Synthetic Route Prediction using Browser Based Visualisation Tools	91
6.1	Introduction	92
6.2	Methods and Implementation	92
6.2.1	AiZynthFinder – A Tool for Computer Aided Synthesis Planning	92
6.2.2	Precomputing Synthetic Routes	94
6.2.3	GDBRouteBrowser – Linking Chemical Library Visualisation to CASP	94
6.2.4	Compound Selection	97
6.3	Results and Discussion.....	98
6.4	Conclusion.....	103
6.5	Availability of data and materials	103
6.6	Author Contributions.....	104
7	Conclusion and Outlook	105
7.1	Summary	105
7.2	Outlook.....	108
7.2.1	Data	108
7.2.2	Modelling Chemical Reactions and Pathways.....	110
7.2.3	Education and Collaboration.....	111
7.2.4	A Personal Vision for the Future	112

8	Appendix	117
9	Bibliography	156
10	Declaration of Consent	185

1 Thesis Outline

1.1 Thesis Outline and Contributions

This thesis is focused on developing artificial intelligence (AI) driven computer aided synthesis planning (CASP) tools, to facilitate experimental realisation of theoretical molecules obtained from computational methods or otherwise. The following outline describes the contributions I have made to the field towards solving the overarching problem.

- Chapter 2 introduces the concepts of AI in chemistry and CASP. This is followed by a brief history of CASP from manual heuristics-based approaches, to the current machine learning based developments that have shaped the field. I then discuss chemical data representations of relevance to the thesis and introduce the concept of chemical space.
- Chapter 3: **Datasets and their influence on the development of computer assisted synthesis planning tools in the pharmaceutical domain**, is the first description of the development of the AiZynthFinder CASP tool which I co-led the development of. To this end, the underlying reaction data is first examined, and methods established for data pre-processing workflows as applied to reaction datasets. Improvements to existing template extraction algorithms are outlined, and new metrics have been developed for their assessment. An analysis of the largest compiled reaction dataset is conducted, and new metrics developed for assessing single- and multi- step retrosynthetic models. The neural network-based models are combined with Monte-Carlo tree search to assess the influence of training data for each step in the workflow. Overall, this has resulted in one of the first open-source retrosynthetic planning tools and can be used to predict synthetic routes to any compound of interest.
- Chapter 4: **“Ring Breaker”: Neural Network Driven Synthesis Prediction of the Ring System Chemical Space**, tackles the issue of predicting the formation of ring systems. These currently pose a bottleneck for CASP approaches. The development of domain specific models to target chemical motifs, and a switch to multi-label approaches for retrosynthetic prediction is outlined as applied to ring systems. This improvement allows for faster model training and improved scalability. Strategies to improve multi-step retrosynthetic prediction are demonstrated, and the model has been deployed for interactive use.
- Chapter 5: **Retrosynthetic accessibility score (RAscore) – rapid machine learned synthesizability classification from AI driven retrosynthetic planning**, confronts the issue of the scalability of the prediction of full retrosynthetic pathways. The computation of retrosynthetic pathways can be rate limiting owing to the computational expense that is incurred. This is not always required, as in the case of screening large datasets of molecules. Thus, this chapter describes a proxy model, framing the estimation of synthetic

accessibility as a binary classification task. To this end an automatic machine learning framework is introduced enabling more efficient model building. Where existing synthetic accessibility scores fail to separate compounds that are predicted as synthesisable by CASP, this chapter outlines the assessment of several machine learning classifiers using a metric called average linkage. The established score computes at least 4,500 times faster than full retrosynthetic prediction, enabling the screening of large virtual libraries.

- Chapter 6: **Linking Navigation of Chemical Libraries to Synthetic Route Prediction using Browser Based Visualisation Tools**, aims to provide one tool for synthetic chemists combining compound data, visualisation, and synthetic route predictions to aid in the prioritisation of molecules. A molecule cannot be thought of on its own with disregard for its properties, or its synthesis. However, at present different tools are used to provide insight into a given molecule to aid decision making. This complicates matters by enforcing a high barrier of entry for the required information. This chapter establishes methods for consolidating molecular descriptors with synthetic route information in an interactive visualisation. I demonstrate use cases for the prioritisation of molecules as applied to GDB subsets. Additionally, I show how such tools can be used to test hypothesis relating to CASP tools to identify current bottlenecks and obtain feedback from chemists.
- Chapter 7: **Conclusion and Outlook**, concludes the thesis by summarising the contributions made, and outlines challenges and opportunities that remain in the field.

1.2 Publications

The thesis consists of first author publications as listed below, presented throughout the thesis as separate chapters, or where otherwise indicated.

1. A. Thakkar, E.J. Bjerrum, O. Engkvist, J.-L. Reymond, Neural Network Guided Tree-Search Policies for Synthesis Planning. *Artificial Neural Networks and Machine Learning – ICANN 2019: Workshop and Special Sessions. ICANN 2019. Lecture Notes in Computer Science*, **2019**, vol 11731. Springer, Cham. https://doi.org/10.1007/978-3-030-30493-5_64
2. A. Thakkar, T. Kogej, J.-L. Reymond, O. Engkvist, E. J. Bjerrum, Datasets and their influence on the development of computer assisted synthesis planning tools in the pharmaceutical domain. *Chem. Sci.* **2019**, DOI: 10.1039/C9SC04944D
3. A. Thakkar, N. Selmi, J.-L. Reymond, O. Engkvist, E. J. Bjerrum, “Ring Breaker”: Neural Network Driven Synthesis Prediction of the Ring System Chemical Space. *J. Med. Chem.* **2020**, 63, 8791–8808. DOI: 10.1021/acs.jmedchem.9b01919
4. A. Thakkar, S. Johansson, K. Jorner, D. Buttar, J.-L. Reymond, O. Engkvist, Artificial intelligence and automation in computer aided synthesis planning. *React. Chem. Eng.* **2021**, 6, 27–51. DOI: 10.1039/D0RE00340A
5. A. Thakkar, V. Chadimová, E. J. Bjerrum, O. Engkvist, J.-L. Reymond, Retrosynthetic accessibility score (RAscore) – rapid machine learned synthesizability classification from AI driven retrosynthetic planning. *Chem. Sci.* **2021**, 12, 3339–3349. DOI: 10.1039/D0SC05401A
6. A. Thakkar, P. Schwaller, How AI for Synthesis Can Help Tackle Challenges in Molecular Discovery. *CHIMIA International Journal for Chemistry*, **2021**, 75 (7-8), 677-678

The following is a list of publications that were co-authored during the thesis. Some elements have been incorporated where indicated based on author contributions.

1. E.J. Bjerrum, A. Thakkar, O. Engkvist, Artificial Applicability Labels for Improving Policies in Retrosynthesis Prediction. *Mach. Learn. Sci. Technol.* **2020**, 2
2. S. Genheden, A. Thakkar, V. Chadimová, J.-L. Reymond, O. Engkvist, E.J. Bjerrum, AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning. *Cheminform.* **2020**, 12, 70. DOI: 10.1186/s13321-020-00472-1
3. L. David, A. Thakkar, R. Mercado, O. Engkvist, Molecular representations in AI-driven drug discovery: a review and practical guide. *J. Cheminform.* **2020**, 12, 56. DOI: 10.1186/s13321-020-00460-5
4. S. Johansson, A. Thakkar, T. Kogej, E. Bjerrum, S. Genheden, T. Bastys, C. Kannas, A. Schliep, H. Chen, O. Engkvist, AI-assisted synthesis prediction. *Drug Discov. Today Technol.* **2019**, 32–33, 65–72. DOI: 10.1016/j.ddtec.2020.06.002

5. D. Sumner, J. He, A. Thakkar, O. Engkvist, E. J. Bjerrum, Levenshtein Augmentation Improves Performance of SMILES Based Deep-Learning Synthesis Prediction. *ChemRxiv*. **2020**, DOI 10.26434/chemrxiv.12562121.v2.

1.3 Conference Presentations

The following is a list of conferences attended and presentations given during the period of the thesis.

1. Cheminformatics Strasbourg Summer School, Strasbourg, France, June/July 2018 (oral presentation)
2. Machine Learning in Drug Discovery Summer School, KU Leuven, Belgium, August 2018
3. Swiss Chemical Society Fall Meeting, Lausanne, Switzerland, 2018
4. German Cheminformatics Conference, Mainz, Germany, 2018
5. BigChem Spring School, Gothenburg, Sweden, May 2019 (oral presentation)
6. International Conference on Artificial Neural Networks, Munich, Germany, September 2019 (oral presentation)
7. Biopharmaceuticals R&D Science Symposium, Gothenburg, Sweden, October 2019 (poster/oral presentation)
8. German Cheminformatics Conference, Mainz, Germany, November 2019 (oral presentation)
9. AI Powered Drug Discovery and Manufacturing, Boston, USA, February 2020 (poster presentation)
10. Swiss Chemical Society Fall Meeting, Bern, Switzerland, August 2020 (oral presentation)
11. Swiss Cheminformatics Workshop, Lausanne, Switzerland, September 2021 (oral presentation)
12. Swiss Chemical Society Fall Meeting, Bern, Switzerland, September 2021 (oral presentation)
13. RSC AI in Chemistry, London, United Kingdom, September 2021 (poster presentation and lightning talk)
14. RDKit User Group Meeting 2021, Virtual, October 2021 (oral presentation)

1.4 Awards

1. SCS Fall Meeting 2021 – Best Oral Presentation (computational chemistry)

1.5 Funding Acknowledgement

This thesis was 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>).

2 Introduction

Herein, I introduce a range of topics that are of relevance to the chapters contained within this thesis. I start by introducing the topic of artificial intelligence in chemistry, and the factors that are important to its integration within the chemical community. This is followed by an introduction to representations of chemical data that are used throughout the subsequent chapters. Having covered some of the common data representations used, I go onto introduce chemical datasets, starting with an introduction to chemical space. While chemical space refers to individual molecular entities, these must be first be synthesized. In this thesis I investigate a computational approach to the synthesis of chemical compounds, thus I introduce chemical reaction data as a separate topic. I provide a brief history of reaction data as well as outlining the available datasets and limitations with each, which we discuss in more detail in chapter 3. With an understanding of the role of artificial intelligence in chemistry, how data is represented, and molecular and reaction data, I introduce the topic of computer aided synthesis planning, the subject of the chapters in this thesis. I introduce the history of synthesis planning technologies and how they have developed to the present day and highlight the introduction of deep learning to the field. Finally, I bring the chapter together by illustrating an overview of how computer aided synthesis planning systems are built to enable the reader to understand the context of each of the subsequent developments.

2.1 Artificial Intelligence in Chemistry

Parts of this section have been reproduced from the following article with permission from the Royal Society of Chemistry.

A. Thakkar, S. Johansson, K. Jorner, D. Buttar, J.-L. Reymond, O. Engkvist,
Artificial intelligence and automation in computer aided synthesis planning.
React. Chem. Eng. **2021**, 6, 27–51. DOI: 10.1039/D0RE00340A

The use of artificial intelligence (AI) and automation to augment drug discovery and development has been the subject of several reviews in recent years and promises to accelerate both discovery and development in an effort to deliver medicines to patients faster.^{1–3} The subject has once again gained popularity, with key drivers being the accessibility of improved methods, increased computational power, and larger datasets. Artificially intelligent systems have the potential to transform chemistry by conducting or assisting with tasks previously reserved for humans. In the brief history of the field, the definitions of what is deemed an ‘intelligent’ system have continued to change as technologies are outdated and new ones take their place.^{4–6} What once constituted artificial intelligence and automation no longer rouses interest among the chemical community as they have become routine tasks. For example, consider the collection of NMR spectra – a chemist is now able to submit samples for NMR analysis and await the result, with the machine carrying out automated sampling, recording of the spectra, and subsequent processing of the raw free induction decay data. This has recently been extended to the assignment of NMR spectra.⁷ As can be seen, these technologies have now become deeply embedded into chemical workflows and augment the ability of the chemists using them, allowing them to focus their time on analysis and the design of future experiments. Computer aided synthesis planning (CASP), the topic of this thesis, has not yet reached the stage where it is an integral part of a chemist's workflow, but there has been much discussion about how best to integrate it, at which stage, what to expect, and what it will deliver.¹ This ongoing debate signifies the beginning of a period of development by which members of several distinct research communities, ranging from biology, chemistry, mathematics, physics, robotics, and computer science must come together to build ‘intelligent’ and automated solutions that work for the chemist.

Over the last 60 years, artificial intelligence has been used as a tool to find solutions to a plethora of chemical problems, from *de novo* design of compounds,⁸ and the reactions required to make them,⁹ to bioactivity prediction,¹⁰ and safety assessment.¹¹ However, despite attempts to create platforms for CASP, none have experienced widespread adoption, with the exception of chemical search engines such as Reaxys and SciFinder. There are of course reasons other than their potential limitations and performance that contributed to lack of adoption during the early years, for instance the accessibility of computers, the internet, and barriers to entry in the form of steep learning curves. However, in the last few years, the tools have become more accessible, which in the context

of the time taken for development of the underlying mathematical frameworks, is a relatively short period of time. Furthermore, there is a behavioral element that has limited adoption which is well-summarized by the late Carl Djerassi:¹²

“Symbolic manipulations by computers are in principle important in two areas of chemistry – synthesis and structure elucidation. It is the former where the use of computers has not been widely accepted because of the fear that thinking man will simply be reduced to an appendage to a machine. The synthetic chemist wishes to be both architect and building contractor – the former function being the intellectually and aesthetically more pleasing one – and it is precisely this architectural role that the computer is perceived partially to usurp”.

These behavioral aspects toward adoption have been discussed by Griffen *et al.* and provide a view of the problems the community and companies face, and have faced regarding the adoption of computational tools.¹³ At present, however, AI and automation cannot carry out the actions or higher-level reasoning required to run discovery and development cycles autonomously. Whilst technical improvements have been made toward this end, the behavioral aspect should not be overlooked.

As such, I believe that – in their current and future state – the algorithms presented henceforth, should be viewed as augmenting the ability of a human chemist to arrive at the desired solution. Thereby, they will act as tools to inspire and inform the decision maker rather than to replace or fully automate the design, make, analyze, and test (DMTA) cycle. In this regard the goal for the computational tools outlined herein is to improve the productivity of chemists, especially with regards to well-established practices, thus allowing more time to focus on novel or more difficult chemistries.

Whilst fully automated chemistry is one goal towards which AI and automation is being developed, this should be with the end goal of facilitating the work of a wet-lab chemist, rather than with the aim of replacing lab-based chemists. I emphasize that synthetic chemistry is not necessarily the bottle neck in drug discovery and is only one contributing factor in the process. Bender has discussed this in more detail with a view on efficacy and safety in drug development,¹⁴ and there are several ongoing works in the clinical phase to improve the whole process.^{15,16} Nevertheless, to facilitate development in any of the highlighted areas, an interdisciplinary approach bringing together experts from different fields is required. In addition, emphasis should be placed on ease of use and accessibility of the tools that are developed. Successful approaches may be characterized as those with a shallow learning curve for the experimentalist, a rich data source for the theoretician or data scientist, and tight-knit integration throughout the community from discovery to development. The approach should also be scalable, adaptable, reliable, and most importantly, meet the needs of the end user.

2.2 Chemical Data Representation and Formats

Parts of this section have been adapted from the following article where appropriate based on the contributed material.

*L. David, **A. Thakkar**, R. Mercado, O. Engkvist, Molecular representations in AI-driven drug discovery: a review and practical guide. J. Cheminform. 2020, 12, 56. DOI: 10.1186/s13321-020-00460-5*

A reaction is often represented graphically with the *reactants* written to the left of a *reaction arrow*, and a set of resulting *products* written to the right of the arrow. The *conditions* under which the transformation occurs are written above or below the arrow, including information such as reagents, catalysts, solvents, temperature, and so forth. The graphical illustrations of reaction schemes often found in publications are, however, not easily machine-readable. Therefore, there exist a series of reaction data exchange formats that enable reactions to be represented in a machine-readable format. There is no inherent requirement for one format or another, as this is dependent on the application, toolkit, or software package used. Commonly used formats include the RXN and RD files, as well as several XML based file formats.^{17,18}

2.2.1.1 SMILES, SMARTS, Reaction SMILES and SMIRKS

The Simplified Molecular Input Line Entry System, more commonly referred to as SMILES was developed by Weininger *et al.* in 1988 to represent molecular structures.¹⁹ Since its inception it has grown to become one of the most widely used linear notations, and has been incorporated into several commonly used cheminformatics toolkits. SMILES are a non-unique and unambiguous representation obtained by traversing the molecular graph in the direction of node ordering. In the case of RDKit the graph traversal algorithm used is depth-first search,²⁰ thus depending on the toolkit used, the graph traversal algorithm and node ordering may differ, therefore leading to several representations for the same molecule. In an effort to create a unique representation, different canonicalization methods have been proposed, however they may not be identical across toolkits.^{21–23} A further modification to the specification of SMILES, so called isomeric SMILES was later introduced enabling the encoding of configurations around double bonds (Z or E isomers) using the '\' and '/' symbols, and configuration around tetrahedral centers through the use of the '@' and '@@' symbols. However, the specification of organometallic compounds and ionic salts often used in the reagents of reactions are often poorly supported as the parent species cannot be easily identified. To overcome this issue ChemAxon Extended SMILES (CXSMILES) were proposed that store special features, one of which is fragment grouping enabling grouping of ions and salts to their parent species.²⁴

The SMILES format used for describing molecules has been extended to so-called Reaction SMILES by Daylight Chemical Information Systems.²⁵ Each molecule in the reactants, agents,

and products is represented by a SMILES string, and disconnected structures are separated by a period ('.'); this includes the individual molecules, ions and ligands, which are listed in an arbitrary manner. Reactants, agents, and products are separated by either the '>' or '>>' symbol (the latter used when agents are not given). Atom-mappings (i.e. mappings of atoms in the reactants to their equivalent atoms in the products) can be stored in Reaction SMILES as a non-negative integer following the character ':' within an atom expression. Atom mappings do not apply to agents. Furthermore, the storage of additional textual information such as the reaction center (i.e. the atom and bonds that change during a transformation) or reaction conditions is not supported. Nonetheless, formats such as the RXN and RD file formats, especially the latter, can store this additional metadata, as can other file formats or databases.

The SMILES Arbitrary Target Specification (SMARTS) is an extension of SMILES enabling substructure searching and specification.²⁶ In addition to the SMILES specification which encodes the underlying molecular graph, the symbols available in SMARTS allow the encoding of patterns, thus a more general specification of the molecular graph. In analogy to computer science the SMARTS pattern can be likened to a regular expression and allows some of the same operations. These operations include the specification of logical operators such as "OR", "AND" and "NOT", as well as wild cards. In contrast to SMILES, SMARTS additionally enable generalization of bond types as well as isotopes. These additional features enable specification of detailed atom environments within a molecule, thus enable substructure search in databases. While it is true that all SMILES can be valid SMARTS, as SMARTS inherit language features from SMILES, the reverse does not hold true.

SMIRKS belong to the same family as SMILES and SMARTS. Where SMARTS describe molecular patterns or substructures generically, SMIRKS patterns can be used to define generic reaction transformations. They can be used to describe the reaction center, to enumerate virtual libraries, and to form the knowledge base for reaction and retrosynthetic prediction systems. If one considers that a reaction is a set of atoms and bonds that change during a reaction and the reactant or substrate upon which that change occurs, then SMIRKS must encode the same set of atoms and bonds that change during the reaction, and the site at which that change occurs in the substrate as specified by a SMARTS pattern. The SMARTS pattern is used to specify both the site at which the atom and bond changes occur, and to capture any indirect effects that may influence the reaction. The atomic expressions must be defined such that (a) for any part of a molecule that is to be considered in a generic transformation for which the bonding does not change, SMARTS are to be used, and (b) in cases where bonds change, SMILES are to be used. In this sense, SMIRKS is a hybrid approach between SMILES and SMARTS. There are some rules that must be followed to ensure that SMIRKS patterns can be applied. The two sides of the transformation, the reactant(s) and product(s), must contain the same number of mapped atoms, and they must correspond on either side of the reaction. Additionally, any explicit hydrogens must appear explicitly on either side of the reaction and have corresponding atom mapping numbers. SMIRKS are converted into

a reaction graph for their subsequent use. The reaction SMILES and corresponding SMIRKS are shown in Figure 1.

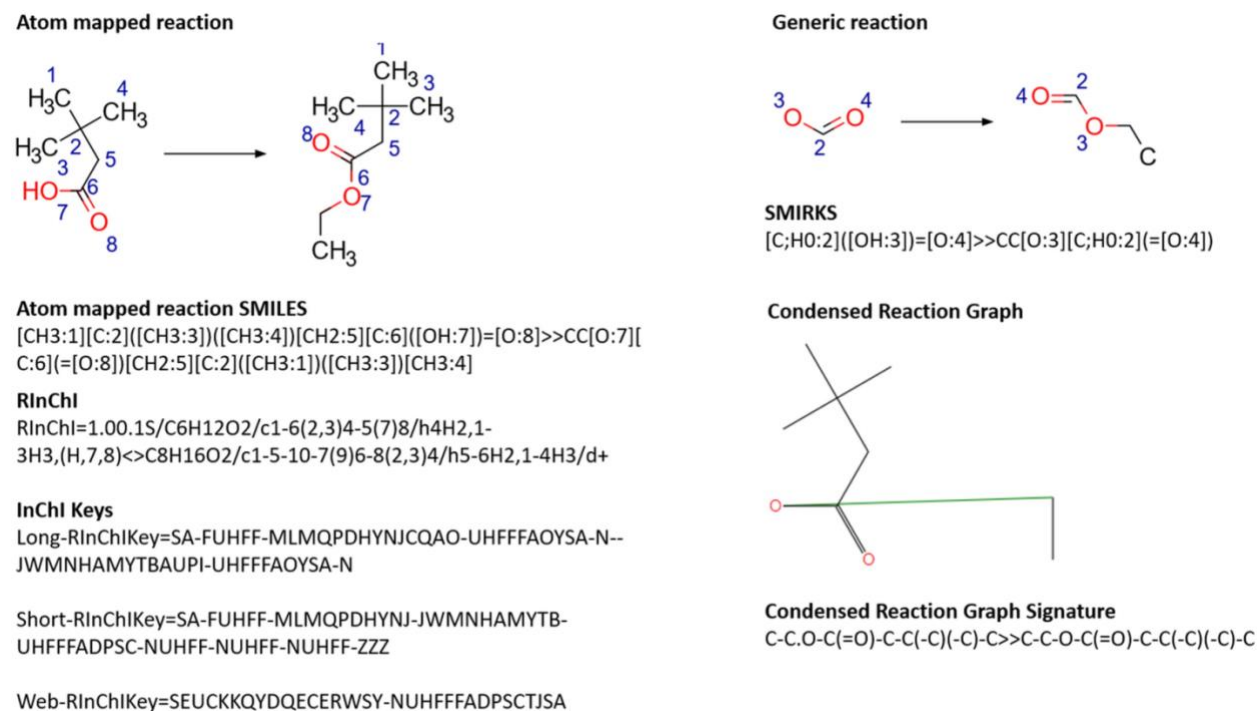


Figure 1: A selection of representations for a simple esterification reaction. The atom mapped reaction is shown in the top left as a structural diagram. The atom maps are consistent between reactant and product as shown. The atom maps in the SMIRKS do not correspond to the atom maps in the full reaction. Rather, they are used to keep track of the atoms within the SMIRKS. The condensed reaction graph and corresponding signature was generated using CGRtools.²⁷

2.2.1.2 InChI, InChIKey, and RInChI

InChI were introduced in 2006 by NIST, and are composed of multiple layers.²⁸ The layers encode different parts of the molecular graph, for instance the *main* layer encodes the chemical formula, atom connections, and hydrogen atoms and their corresponding connections. Charge, stereochemistry, and isotopic information are also encoded in a series of sublayers. In contrast to SMILES all InChI are a unique representation, and can be hashed to form the InChIKey, although hash collisions may occur when more than one InChI produces the same hash.²⁹

An extension of the InChI, RInChI,^{30,31} was developed between 2008 and 2018 and introduced a unique, order invariant identifier for reactions. It was developed in response to the growing size of reaction data to aid reproducibility, to consider more information than just the participating molecules, and to provide enough information such that practically identical reactions would be represented the same way. RInChI grammar, however, is relatively more complicated than that of Reaction SMILES.

RInChIs use InChIs to describe each molecule. Where InChIs cannot be generated for a molecule, the RInChI tracks the number of “structureless” entities that are present in each of the reactants, agents, and products. In addition to specifying each molecule and reaction role, the RInChI must

include information about equilibrium, unbalanced, or multi-step reactions. The RInChI employs a layering system, whereby each layer can describe a different aspect of the chemical reaction. Solvents and catalysts may be accounted for in a similar manner as in Reaction SMILES; however, RInChIs additionally allow for the direction of the reaction to be described. This is particularly useful, as different labs may conduct the same reaction under slightly different conditions, potentially reaching different conclusions about the direction of the reaction. The RInChI generated in this case would be the same, except for the direction flag. This aids in the identification of reactions that are in practical terms identical.

A proposed further extension to RInChI, ProcAuxInfo, enables the storage of metadata relating to yields, temperature, concentration, and other reaction conditions.³² RInChI offers an alternative to Reaction SMILES that enables the identification of duplicate reactions, as the order in which molecules are listed in Reaction SMILES is arbitrary. Hashing the RInChI to yield the RInChI key provides a powerful tool for efficiently indexing and searching reaction data.^{30,32} However, there is no SMARTS or SMIRKS equivalent for RInChI, limiting its use in substructure searching and in encoding generic chemical transformations. The RInChI and corresponding keys are shown in Figure 1.

2.2.1.3 Condensed graph of reaction (CGR)

Varnek and co-workers have developed the CGR approach,³³ whereby molecular structures are encoded in a matrix containing the occurrence of fragments of a given type. The CGR is a superposition of the reactant and product molecules, and additionally defines what atoms and bonds have changed as well as their properties. This builds on the description of organic reactions using imaginary transition states as described by Fujita.³⁴ In analogy to SMIRKS, the CGR can be used to describe a reaction transformation. An example CGR is shown in Figure 1.

With the renewed interest in chemical reactions within cheminformatics in recent years, Varnek and co-workers have developed an open source toolkit enabling the wider use of CGR.²⁷

2.2.1.4 Extended Connectivity Fingerprints (ECFP)

While there are a multitude of fingerprinting methods available, they broadly fall into four distinct categories: circular (also known as radial), atom-pair, descriptor/property based, and more recently learned fingerprints from deep neural networks. In this thesis the ECFP variant is frequently used, so will be covered briefly. Fingerprints are a numerical representation of a molecule and take the form of an n -dimensional vector. The calculation of the vector is dependent on the type of fingerprint used, and in the case of the ECFP variant is computed by the Morgan algorithm.^{23,35}

The ECFP is computed by iterating around the atoms in the molecule as defined by the Morgan algorithm. At each atom the environment within a predefined radius is considered, commonly a radius of two or three is used. A tuple of the resulting substructure is extracted considering the element, number of heavy atom neighbors, number of hydrogens, charge, isotope, whether the atom is in a ring, and a variety of other chemical features. The resulting circular substructures are

then hashed and converted into an integer value which acts as an identifier. The identifiers define the ECFP and can be used in two different ways. First an ECFP can be considered as an ordered list of hash keys or identifiers, where each identifier corresponds to a substructure, or circular environment as computed by the Morgan algorithm, where the list is sorted in ascending order. The identifiers constitute a virtual bit string, where each bit corresponds to the presence or absence of the substructure in a molecule. Instead of encoding the presence or absence of the substructure, one may also consider the frequency by which the substructure occurs, known as the frequency counted ECFP (EFCF). The second way of computing the ECFP is by converting the virtual bit string into a fixed length vector in a process known as "folding", commonly to 1024 bits. This simplifies subsequent computations and is easier to store on disk owing to its lower memory requirements, however the "folding" operation can induce bit collisions as two or more circular environments may be represented by the same bit. Therefore, information regarding the topology of the molecule may be lost in the process.³⁵

2.3 Chemical Space

The chemical space is a mathematical construct in cheminformatics which refers to the property space spanned by all possible chemical compounds. The compounds are defined by the number, type, topological connectivity, and spatial orientation of the constituent atoms. By one estimate the "drug-like" chemical space obeying Lipinski's rule-of-five for oral bioactivity is 10^{60} molecules.^{36,37} However, with the introduction of new modalities outside the Lipinski domain, such as oligonucleotides, hybrids, molecular conjugates, as well as new use of small molecules for instance proteolysis targeting chimeras (PROTACS), the "drug-like" chemical space is likely underestimated.³⁸ A recent study has found that of the 94 million entries in the PubChem database, 7 million break at least one of four Lipinski constraints for oral bioavailability.³⁹ Whilst the beyond Lipinski space is one interest of the Reymond group, of relevance to this thesis is the "chemical space project", which asks the question: how many molecules are possible in total?

To answer the question, the group has focused on the enumeration of all possible molecules up to a given size, made up of the elements C, N, O, S, and the halogens. The molecules are enumerated starting from mathematical graphs produced by the GENG program.⁴⁰ Graphs suited towards building saturated hydrocarbons were selected accounting for ring strain and topology. Unsaturation was then introduced following rules for valency, aromaticity, and ring strain. The resulting skeletal structures were then modified by replacement of N, O, S and the halogens for C atoms in the structures, taking into account functional group stability. This enumeration process has resulted in the databases GDB11,^{41,42} GDB13,⁴³ and the largest generated database to date has been GDB17, consisting of 166.4 billion molecules up to 17 heavy atoms (Table 1).^{44,45} The realisation of GDB20 is currently underway.

Table 1: Outlines of some of the available databases to demonstrate the relative size of the GDB databases in comparison to databases of known compounds and the virtual or enumerated space obtained by other means. It can be seen from the table that the largest database GDB17 is at least 8 times the size of the Enamine REAL space considered to be synthetically achievable, and even greater than the known space as shown by Reaxys, CAS, and PubChem. For a more extensive list readers are referred to a list by Coley et. Al.⁴⁶

Database	Size	Description
GDB-11 ^{41,42}	26 M	Enumerated drug-like organic molecules up to 11 heavy atoms (C, N, O, F)
GDB-13 ⁴³	970 M	Enumerated drug-like organic molecules up to 11 heavy atoms (C, N, O, F)
GDB-17 ^{44,45}	165 B	Enumerated drug-like organic molecules up to 11 heavy atoms (C, N, O, F)
FDB-17 ⁴⁷	10 M	Uniformly sampled selection of GDB-17 containing fragment-like molecules.
GDBChEMBL ⁴⁸	10 M	Uniformly sampled selection of GDB-17 containing ChEMBL-like molecules as computed by the ChEMBL likeness score.
GDBMedChem ⁴⁹	10 M	Uniformly sampled selection of GDB-17 filtering for medicinal chemistry like compounds.
PubChem ⁵⁰	111 M	Known compounds and experimental properties with bioassays
ChEMBL ⁵¹	2.1 M	Known compounds and measured bioactivity
CAS ⁵²	188 M	Compounds registered with the chemical abstracts service, and measured bioactivity
Reaxys ⁵³	148 M	Compounds obtained from literature and bioactivity data
Enamine REAL ⁵⁴	20 B	Enumerated synthetically accessible structures from known building blocks
SAVI ⁵⁵	1.5 B	Enumerated synthetically accessible structures from known building blocks

Other approaches for the enumeration of chemical space have also been examined. Genetic algorithms such as SPROUT were previously investigated,⁵⁶ and have once again gained traction as deep learning based methods have started to arise.⁵⁷ Following the combinatorial chemistry era, several approaches emerged assembling molecules from a set of known building blocks using coupling reactions.^{58,59} This led to the advent of large virtual collections such as the Pfizer Global Virtual Library,⁶⁰ Enamine REAL,⁵⁴ and SAVI.⁵⁵

More recently, deep learning approaches using techniques derived from natural language processing (NLP) have been able to generate compound libraries.^{8,61,62} In one approach the 68.9 % of the GDB13 database was able to be reproduced by training on 1 million structures corresponding to 0.1 % of the database.⁶³ In this sense, the model can be viewed as a compression of the virtually accessible space negating the need to store on disk. Furthermore, techniques such as particle swarm optimization have been employed to traverse the latent space (mathematical representation learned by the model) to generate only molecules with the desired properties.^{64,65}

Other types of models combining generation and synthetic accessibility have also been investigated, similarly to the combinatorial approaches that were previously used. However instead of coupling the fragments together by known rules, deep learning approaches estimate the probability that a synthetic step will work, or generate a set of reactants on the fly to build up the target molecule.^{66,67}

Given that the size of the chemical space is so large, one needs to ask the question as to how to filter the space and understand its contents. To this end several techniques have been developed that are categorically known as "virtual screening" (VS).^{68–73} In addition to VS, visualization techniques, primarily using dimensionality reduction have been employed as a means of exploratory data analysis.^{74–80} These include techniques such as the widely used principal component analysis (PCA),^{81,82} t-distributed stochastic neighbor embedding (t-SNE),⁸³ uniform manifold approximation and projection (UMAP),⁸⁴ self-organising maps (SOM),⁸⁵ generative topographic maps (GTM),⁸⁶ and more recently developed in our group the tree map (TMAP).⁸⁷ I use the TMAP in chapter 6 as an interface to link chemical library visualization to synthesis prediction.

2.4 Chemical Reaction Data

Parts of this section have been reproduced from the following article according to the rights and permissions outlined by Elsevier.

*S. Johansson, **A. Thakkar**, T. Kogej, E. Bjerrum, S. Genheden, T. Bastys, C. Kannas, A. Schliep, H. Chen, O. Engkvist, AI-assisted synthesis prediction. Drug Discov. Today Technol. **2019**, 32–33, 65–72. DOI: 10.1016/j.ddtec.2020.06.002*

Whether through physics based or statistical modelling, a dataset is required that can be parsed by a computer. To this end, journals and publishing houses have made their datasets available under licensing agreements in computer readable format, by means of text mining and manual data entry. These include texts such as ‘Beilstein’s Handbook of Organic Chemistry’ later known as the Beilstein database,⁸⁸ and now distributed by Elsevier under the name of Reaxys which contains over 57 million reactions as of 2021.⁵³ The size of Reaxys is continuously growing, and as of 2019, Elsevier and the University of Melbourne, Australia initiated a project called ChEMU to develop natural language processing-based (NLP) models in order to text mine patent literature.⁸⁹ The Chemical Abstracts Service (CAS), on the other hand, cover approximately 140 million reactions from 1840 to the present day, and thus is the largest provider of reaction data.⁵² Smaller datasets include SPRESI by InfoChem containing 4.6 million reactions between 1974–2014,⁹⁰ and Pistachio by NextMove Software containing patent data up to the present day with 9.2 million

reactions and growing.⁹¹ A smaller subset of the patent data containing 3.3 million reactions between 1976–2016 extracted by Lowe, is the only publicly available dataset of reactions in current use.^{92,93} The Lowe patent dataset is most commonly known as USPTO, from which various subsets have been created for training and benchmarking predictive models, such as the USPTO 50K.

Whilst the above datasets contain reactant/product structures, reaction conditions (solvent/catalyst/reagent) and yield information, the data is not always consistent and does not contain negative data. The often lack of annotation, both in public and commercial datasets, is a hinderance to the wider applicability of synthetic planning algorithms and their generalizability. This is compounded by the prevalence of bias towards specific reaction types, and literature reporting of only positive data.^{94,95} The lack of negative examples is arguably more important in forward synthetic planning tasks, especially concerning regio- and chemo-selectivity. To overcome these inconsistencies, there has been a drive towards the formation of an open access reaction database, bringing together stakeholders from industry and academia to support efforts related to synthesis prediction and experimental design.⁹⁶ Both Merck and Pfizer have already published HTE data to be included, and efforts to generate more consistent data are underway in the wider community.^{97,97,98} Cronin and co-workers have published the chemical descriptive language (XDL), which sets out experimental procedures for use on robotic systems.⁹⁹ Similarly, IBM published a method using NLP to extract experimental procedures from patents and the scientific literature to create a structured automation friendly format.¹⁰⁰ These schemas are hoped to aid reproducibility, and in the acquisition of more consistent data for use subsequent modelling tasks. Additionally, the Pistoia Alliance has partnered with Elsevier in order to define a unified data model (UDM) for the exchange of reaction information.¹⁰¹

2.5 Chemical Synthesis and Computer Aided Synthesis Planning

Parts of this section have been reproduced from the following article with permission from the Royal Society of Chemistry.

A. Thakkar, S. Johansson, K. Jorner, D. Buttar, J.-L. Reymond, O. Engkvist,
Artificial intelligence and automation in computer aided synthesis planning.
React. Chem. Eng. **2021**, 6, 27–51. DOI: 10.1039/D0RE00340A

To begin our foray into CASP I first define it as encompassing, but not limited to: (1) retrosynthetic analysis, the task of breaking a given compound down into simpler precursors; (2) reaction prediction, the task of predicting the product of a reaction given a set of precursors; (3) reaction condition prediction, the task of predicting a set of conditions (*e.g.* catalyst, temperature, solvent) under which a given reaction takes place; (4) reaction optimization, improving a pre-defined

objective such as yield or purge of impurities by adjusting the conditions under which the reaction is carried out (Figure 2). I do not refer to reaction discovery as being part of CASP, as the definition of a new reaction is not well defined. For instance, a novel reaction could be thought of as a new set of conditions for a known transformation, consider coupling reactions for example, for which there are a plethora of catalysts demanding specific substrate choices or reactions which are mechanistically different. Another crucial aspect of chemical synthesis is the role artificial intelligence can play in optimizing isolation and purification techniques; this has also been omitted in the following discussions. All the highlighted tasks come together to form a system capable of predicting and optimizing synthetic pathways to target molecules. As such, CASP tools have many possible areas of application within drug discovery and development, as well as in parallel functions in the agrochemicals and specialty chemicals industries.

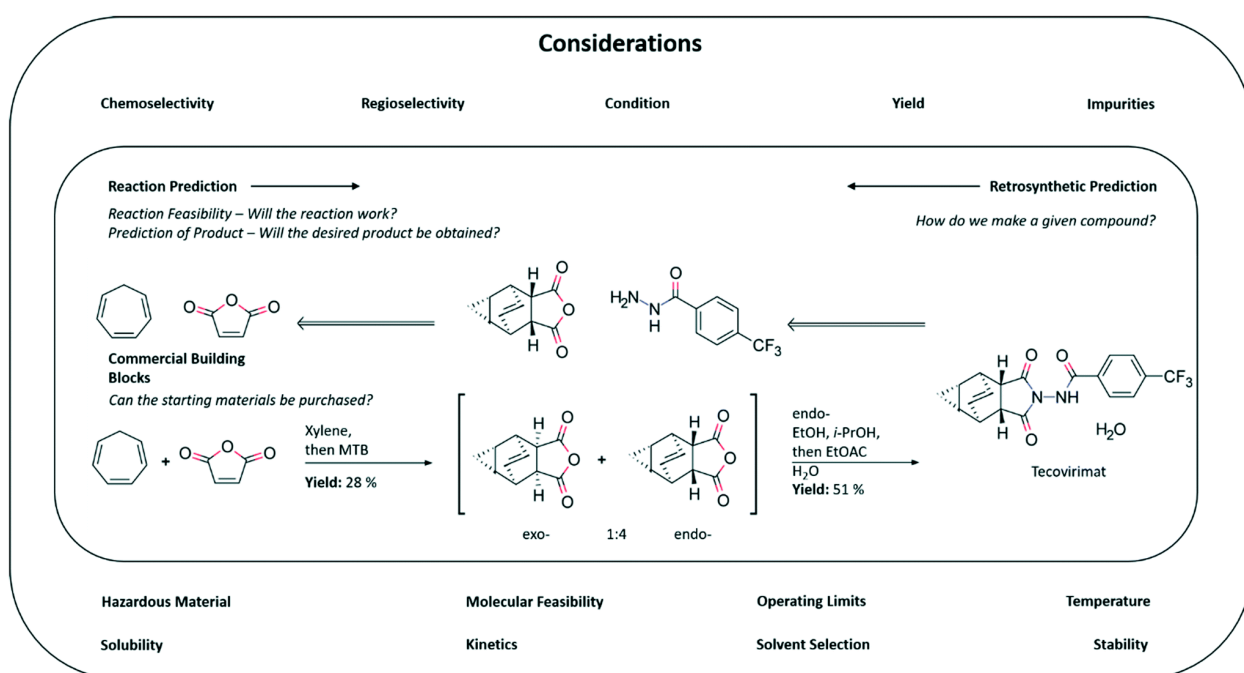


Figure 2: An efficient synthesis to tecovirimat annotated to exemplify potential application of CASP tools. There are several questions that CASP may be able to help answer using a mixture of statistical and physics-based modelling from available datasets and from first principles. Models built with the aim of answering the questions outline in the figure, have the potential to augment the ability of the bench chemist.

2.5.1 A Brief History of Computer Aided Synthesis Planning

Synthesis planning is the task of how to make molecules. When planning the synthesis of a molecule, we know the structure, and given this we want to know the steps required to obtain the structure through experiment. Whilst the molecule is made up of atoms, in the vast majority of cases the final molecule is pieced together from constituent smaller molecules. The problem then becomes which of the smaller molecules does one choose, and how should they be chosen?

Robert Robinson was one of the first to think about the subject, as shown in the synthesis of tropinone in 1917, where the concept of an "imaginary hydrolysis" was proposed.¹⁰² Decades later E. J. Corey picked up on the idea, in one of the first attempts at digitising organic chemistry and CASP in an approach known as Logic and Heuristics Applied to Synthetic Analysis (LHASA).^{103–106} While LHASA no longer survives it built the foundation for future attempts at codifying organic synthesis, and formed the framework for thinking about planning organic synthesis known as retrosynthesis or the disconnection approach as used today.¹⁰⁷ The steps outlined in LHASA and the disconnection approaches can be summarised as: 1) Recognition of functional groups in the target molecule, 2) Identification of known reactions acting on the functional groups as disconnection sites, 3) Repeating the process on the constituents as necessary to find available starting materials, 4) Determination of reagents and conditions to realise each disconnection, and 5) Realise the plan in the laboratory and modify according to experimental results.

In the 1960s Corey and co-workers attempted to codify and organize the rules of organic chemistry *via* a language called PATRAN (Pattern TRANslator), and while the language did not extend beyond LHASA, it inspired the codification that is still in use today.^{103,104,106} The language was used in the original LHASA approach to planning synthetic routes which proposed that the codified rules could be used as a chemical knowledge base encoding the disconnection and possible functional group incompatibilities. Building on the initial LHASA approach in 1973, Wipke determined that a topological description of molecular structure was not sufficient for a complete representation of chemistry; rather he proposed that one should take into account the spatial arrangement of atoms, thus stereochemistry. The proposal was incorporated into a program called Simulation and Evaluation of Chemical Synthesis (SECS), and used a connection-table based language called ALCHEM to encode stereochemical requirements and selection rules of reactions.¹⁰⁸ During the same period, SYNCHEM was being investigated by Gelernter and co-workers.¹⁰⁹ SYNCHEM was later abandoned in favor of SYNCHEM2 owing to stereochemistry considerations in line with the SECS approach. The program used Wiswesser line notation (WLN) to represent molecules,¹¹⁰ although matrix like representations were also accepted. A heuristics driven search for multi-step synthetic routes was then conducted by identifying functional groups in the molecule, applying heuristics categorised based on their applicability to the functional group, and repetition of the procedure for each precursor molecule generated until a set of pre-defined building blocks was reached.¹⁰⁹ Variations of this procedure are the foundations of synthetic route planning tools today, and follow the steps highlighted by the disconnection approach mentioned earlier.

EROS was another such example of a pioneering synthesis planning system by the Gasteiger group. Reactions were encoded to represent electron shifts, in likeness to the curly arrows drawn by chemists to represent reaction mechanisms.^{111–113} The mechanistic view of reactions lent itself well to reaction prediction and was later complemented by the workbench of organic synthesis (WODCA) which dealt with the task of finding appropriate precursor molecules and starting materials for the target structure.¹¹⁴ In another approach named CAMEO, Jorgensen formalised

the task of reaction prediction and retrosynthesis by examining estimated reactivity parameters covering a wide scope of reactions without relying on predefined heuristics for bond rearrangement.¹¹⁵ A key benefit of predicting from calculated reactivity parameters was that a wider albeit more fuzzy coverage of organic chemistry was now possible as it overcame the pre-specified precedents used by the EROS system.¹¹⁶ Other formalisations based on bond-electron matrices using the Dugndji-Ugi model were also investigated, and applied to reaction classification and reaction network generation as implemented in to programs called IGOR and RAIN, which assisted in the identification of new reactions, reagents, and reaction mechanisms.^{117–120} In addition the 1990's marked a transition to knowledge base-guided systems derived from reaction databases, of which a key example is the system for organic reaction prediction by heuristic approach (SOPHIA).¹²¹

Stemming from the aforementioned CASP approaches, three main paradigms for CASP have emerged. The first paradigm comprises synthesis planning systems based on encoding heuristics according to the disconnection approach or similar principles as outlined previously, and a second paradigm of data driven CASP, which creates a model of chemical reactivity from a reaction dataset consisting of reactions that have been performed in the laboratory. The third paradigm may be categorised as symbolic artificial intelligence, which is a combination of the first two paradigms. The combined outcome is a system that extracts heuristics from reaction datasets and learns how they should be applied in a data-driven manner, which will be the focus of this thesis.

As a leeway into data driven CASP, one of the most popularized approaches in recent years has come from the Grzybowski group, in the form of a program called Synthia (formerly Chematica). The team has curated a list of over 100,000 organic reaction rules which took over 10 years of hand coding by expert chemists for incorporation into Synthia.¹²² The encoding of reaction rules is still ongoing as new chemistry is being discovered and older rules are refined. The approach has been validated in the laboratory on medicinally relevant targets.¹²³ In addition to the transforms, functional group compatibilities and conditions under which the transformations were applicable were also encoded. While the approach is inspired from the earlier LHASA efforts, Synthia distinguishes itself in the number of rules that were encoded, and the modern computational efforts that could be applied.¹²⁴ These include accounting for stereoselectivity through encoding stereochemistry information into heuristics, evaluation of the correctness of a prediction using machine-learning based molecular mechanics and quantum mechanics routines, examination of side reactions, scoring the manually encoded heuristics, improved search algorithms combining a broad and deep search, strategies to reach less reactive precursors, navigation around reactivity conflicts, and the ability to perform two different reactions simultaneously.^{125–128} What started out as a heuristics based approach, now combines modern machine learning, network analysis, theoretical chemistry, and computer science algorithms to create a powerful tool for synthesis planning. However, one of the bottlenecks of the Synthia approach is the need to encode rules manually. To overcome the manual encoding process we will now examine the different

approaches to data-driven CASP and the reader is referred to reviews for further information regarding the historical developments of the field.^{116,122,129–131}

2.5.2 Deep Learning in Computer Aided Synthesis Planning

There are two predominant approaches to data-driven CASP: 1) rule or template-based approaches (heuristics), whether machine extracted or human-curated and 2) rule or template-free approaches (non-heuristic). These extremes lie on a continuous spectrum, with some studies combining the two. The main approaches used will be defined and outlined briefly below.

2.5.2.1 Rule/template-based methods – Symbolic AI

Given the extent of the task, and the growing size of the chemical literature, another approach to encoding reaction rules was to automatically extract them from reaction SMILES in the form of SMIRKS patterns.^{132–135} These approaches may be faster but have been the subject of much debate concerning the accuracy with which they represent reactions and is discussed comprehensively by Molga *et al.*¹²⁴ Our recent study comparing a variety of proprietary and public databases found that approximately 2% of templates were common between the datasets (chapter 3).¹³⁶ Whilst these are not necessarily different reactions, different structural variants are captured that artificially inflate the size of the rule set. To account for this Baylon *et al.* take a two-step approach. They first predict the reaction class or group, and subsequently a rule within the group which is used to enumerate the reactants from the given product.¹³⁷

Reaction rules or transformations are primarily used by expert system approaches to CASP,^{106,115,121,122,132,133} or more recently neural network classifiers for both the retrosynthesis and reaction prediction tasks.^{9,138,139} Neural network based systems are significantly faster than their predecessors such as the retrosynthesis tool ICSYNTH for finding full retrosynthetic pathways.¹³³ However, because of the number of variables that must be accounted for when benchmarking one tool against another, including but not limited to: the reaction data underlying the tool, the scoring functions used, the availability of building blocks, and the implementation of the search algorithm, it is not immediately clear where one method is better than another. Rather each tool has the potential to excel in specific areas depending on the developers and end users' priorities.

Segler and Waller use a neural network trained to predict which rule to apply in the retrosynthetic direction for a given compound from hundreds of thousands of possible rules.¹³⁵ The network is employed as a 'policy' to enumerate potential synthetic routes represented as a tree to which Monte-Carlo tree search is applied (MCTS).⁹ The methodology inspired from game AI has been used to predict moves in games such as Go and Chess, as well as stock market prices.¹⁴⁰ The approach combines historical ideas in CASP with developments in deep learning, resulting in the prediction of synthetic routes in seconds. ASKCOS, developed by Coley and co-workers, takes inspiration from this approach for retrosynthetic route prediction, however they employ graph

neural networks for predicting chemical reactivity and a NN classifier for selecting reaction conditions including catalyst, solvent, reagent and temperature.^{134,138,141,142} Furthermore, the neural networks used for prioritization fail to account for infrequently used reactions. They therefore do not prioritize templates (reaction rules) that could be used *in silico* but have not been used in the underlying reaction dataset. For example, consider a Suzuki coupling that can be used to join two fragments together. If there are no examples of Suzuki couplings that have been used to join two fragments to form a ring, the model will be unable to predict such a reaction although it is possible. This has been partly addressed by domain specific modelling (chapter 4),¹⁴³ and training NNs that account for template applicability.^{144,145}

Notably, after the availability of datasets, the encoding or representation of chemical transformations is a bottle neck in predictive modelling. However, rules offer the advantage that predictions may be traced back to the underlying data, which is a feature that the end user wants.

2.5.2.2 Template free approaches – inspired by natural language processing (NLP)

The treatment of chemistry as a language has been explored both as a means of understanding chemical space, and codifying reaction transformations.^{146–149} The various encoding strategies are covered comprehensively by Öztürk *et al.*¹⁵⁰ In contrast to rule-based approaches which predict a set of products or reactants by applying a transformation, NLP inspired approaches learn the syntax of the reactants or products depending on the task to be solved, most commonly from reaction SMILES. The problem is framed as a translation task, translating the reactants to products or *vice versa*. In one approach reaction SMILES are tokenized to give a vocabulary, much like a sentence may be split into its constituent words. The tokens are one-hot encoded into an n -dimensional binary vector, where the presence of a token is signified by a 1, where n is the size of the vocabulary. The vectors are fed to a neural network which learns to predict the next character/token in the sequence given a set of products or reactants, thereby reconstructing the original reaction or predicting a new one. Whilst these methods have shown promising results and improvements in line with developments in NLP, from sequence–sequence to transformer architectures within computer science,^{151,152} they lack the link back to the original data. However, they are potentially more interpretable than rule-based methods owing to the advent of attention, which can highlight areas of the reaction on which the algorithm focuses. This was recently demonstrated by Schwaller *et al.*, whereby they were able to show that the algorithm implicitly learns atom-atom mapping.¹⁵³ Thus the model is able to learn which atoms are changing during a reaction. Bort *et al.* employed similar approaches using an autoencoder and generative topographic mapping to sample novel reactions from reaction space learnt by the model.¹⁵⁴

Baldi and coworkers have taken an alternate approach to reaction prediction based on mechanistic information. They use an existing expert system to label their dataset with the required mechanistic information, thereby overcoming problems with poor data availability and annotation. Having defined a molecular orbital (MO) based reaction unit to model reactions as flows of electrons from sources to sinks, they use a two-stage machine learning approach to rank reactions that correspond to the most productive for a set of reactants and conditions.¹⁵⁵ Recently this has been expanded to

use NLP, specifically an architecture using long-short term memory (LSTM), which while less accurate includes more contextual information, and is able to predict reactive sites based solely on SMILES strings.¹⁵⁶

2.5.2.3 Graph neural networks

Matrix representations of reactions were pioneered by Dugundji and Ugi, in the early 1970's where the reaction was described as an 'R' matrix, corresponding to the bond changes or changes of non-bonded valence electrons.¹²⁰ In this respect, the 'R' matrix can be considered to be like a rule or template representing the transformation taking place. Similar ideas have now been extended and applied for both the retrosynthesis and reaction/condition prediction tasks using graph convolutional neural networks.^{134,157} More recently Shi *et al.* have used a graph to graphs (G2G) framework for the retrosynthesis task.¹⁵⁸ The first step is reaction center identification which is common among rule-based methodologies, however rather than enumerate sets of precursors given a rule, the product is first broken into synthons (hypothetical units resembling reactants, in analogy to the formulation by Corey *et al.*).¹⁰⁵ The reactants are then generated *via* a series of graph transformations from the synthons, thus taking into account that one synthon may correspond to multiple reactants. The graph transformations only affect small localized parts of the reactant/product as recognized by Somnath *et al.* who postulate that the graph topology is largely unaltered during the course of a reaction.¹⁵⁹

2.5.2.4 Reaction networks

Chemical reactions naturally lend themselves to representation as a graph or network, that is a set of vertices or nodes, molecules in this case, connected by directed edges, reactions. Typically, many studies concerned with route predictions deal with tree like structures, which can be considered sub-graphs of the overall reaction graph. However, several works have studied the statistics of reaction graphs at scale.^{122,160–162} Grzybowski *et al.* mapped the 'Universe of Organic Chemistry' and charted its evolution over time. In the process they identified a core set of organic compounds contributing to over 35% of known reactions.¹⁶² Furthermore, they frame the prediction of synthetic routes as a network optimization problem, whereby for a given set of products, they aim to find the set of substrates minimizing the cost. Similarly, Lapkin *et al.* use graph networks for the identification of strategic molecules in supply chains.¹⁶¹ They too have analyzed the statistics of the network of organic chemistry,¹⁶⁰ and reach a consensus with the work of Grzybowski *et al.*¹⁶² Both found that on average six synthetic steps were required to synthesize any given compound from another in the network on average.

Jacob and Lapkin additionally use a stochastic block model based on the network of organic chemistry to predict and discover new reaction pathways.¹⁶³ Likewise, Segler and Waller identify complementary molecules in their graph. By doing so, they identify potential reaction partners for which the same reaction rules apply, thereby proposing reactions that appear to be novel.¹⁶⁴

2.5.3 Building Systems for Computer Aided Synthesis Planning

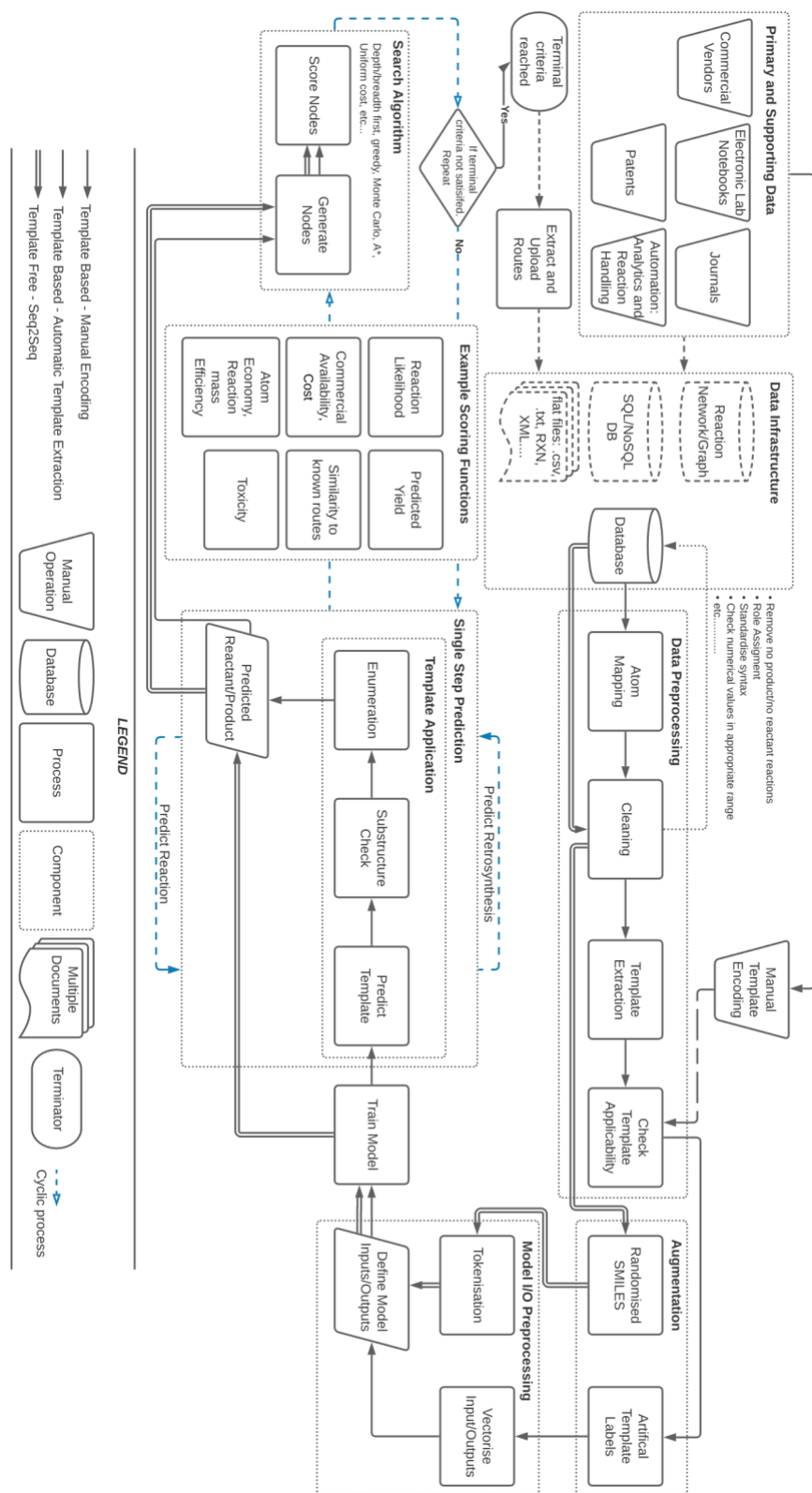


Figure 3: Schematic showing the different components required to build a complete computer aided synthesis planning system.

Systems for computer aided synthesis planning can be considered as modular (Figure 3). The key components required to build a complete synthesis planning system have not changed much during the history of the field (chapter 2.5.1). Where heuristics-based systems relied on manual encoding of reactions from publications and databases, modern data-driven systems can automatically extract patterns from reaction datasets. A variety of public and private reaction datasets are available as outlined in chapter 2.4. The reaction data must first be curated, and a data pre-processing workflow created for this purpose. I outline methods for doing so in chapters 3 and 4. Once a so-called knowledge base has been curated its representation must be chosen depending on the downstream prediction task. In this thesis reactions are represented as reaction SMILES and templates as SMIRKS patterns as outlined in chapter 2.2.1.1. However, template-free models consider reaction SMILES, and network-based models consider reaction networks as highlighted in chapter 2.5.2. The representation chosen influences the choice of model for the prediction task. Thus, consideration must be given to the type of neural network to be trained, the task to be modelled, and which input and output representations are appropriate.

For retrosynthesis, the neural networks are used to prioritize which reaction to use for a given product or intermediate compound. The type of network that can be used, can be chosen between those outlined in chapter 2.5.2, and the method by which reactants are generated differs. Template-free models can generate a SMILES representation for the reactants through sampling of the neural network. Whereas, template-based models require an intermediary enumeration step, whereby a substructure match is required between product and template prior to enumeration of the reactant following the SMIRKS pattern. In both cases, the neural networks are sampled such the top N most likely predictions are used.

Having chosen a model for prioritizing reactions at the single step level, a mechanism is need for conducting multi-step synthesis planning. This can be achieved by considering retrosynthesis as a search problem. Search problems are well known in computer science and game AI, having led to successes in games such as chess and Go.¹⁴⁰ In this thesis I opt to use Monte-Carlo tree search, however other strategies are possible such as the A* search algorithm or simpler depth first or breadth first search algorithms. To aid the search algorithm a scoring scheme is required. Here I use the availability of commercially available precursors, however more complex scoring schemes based on reaction prediction, convergent and divergent synthesis schemes, or estimated cost can be used instead. Combining the neural network, a search strategy, and a scoring scheme enables the prediction of multi-step synthetic pathways.

Finally, a method for validating synthetic pathways can be incorporated in the form of reaction prediction models. These can be used to score the likelihood of individual reaction steps, ascertain the conditions that could be used, predict the yield, or examine selectivity issues to name a few examples. These models can also be fed into a route scoring scheme to aid the search strategy according to a user's preferences. However, consideration must be taken when applying reaction prediction models to the outcome of retrosynthesis prediction as the models may not share the same applicability domain owing to their origin. The same applies for the use of reaction prediction models in scoring functions, thus it is vital to understand whether the information being lost due to application of these models is important for the individual use case.

3 Datasets and their influence on the development of computer assisted synthesis planning tools in the pharmaceutical domain

Computer Assisted Synthesis Planning (CASP) has gained considerable interest as of late. Herein we investigate a template-based retrosynthetic planning tool, trained on a variety of datasets consisting of up to 17.5 million reactions. We demonstrate that models trained on datasets such as internal Electronic Laboratory Notebooks (ELN), and the publicly available United States Patent Office (USPTO) extracts, are sufficient for the prediction of full synthetic routes to compounds of interest in medicinal chemistry. As such we have assessed the models on 1731 compounds from 41 virtual libraries for which experimental results were known. Furthermore, we show that accuracy is a misleading metric for assessment of the policy network, and propose that the number of successfully applied templates, in conjunction with the overall ability to generate full synthetic routes be examined instead. To this end we found that the specificity of the templates comes at the cost of generalizability, and overall model performance. This is supplemented by a comparison of the underlying datasets and their corresponding models.

This chapter has previously appeared as a scientific article in Chemical Science as part of the themed collections "Most popular 2019-2020 physical and theoretical chemistry articles" and "Accelerating Chemistry Symposium Collection".

This section has been reproduced from the following article with permission from the Royal Society of Chemistry.

A. Thakkar, T. Kogej, J.-L. Reymond, O. Engkvist, E. J. Bjerrum, *Datasets and their influence on the development of computer assisted synthesis planning tools in the pharmaceutical domain. Chem. Sci.* **2019**, DOI: 10.1039/C9SC04944D.

3.1 Introduction

Developments in computer assisted synthesis planning (CASP), specifically retrosynthetic analysis have gained considerable interest in recent years.¹⁶⁵ The resurgence of artificial intelligence (AI) in computer aided drug design (CADD) has driven the shift from more traditional expert systems, built around a manually encoded set of reactions as templates,¹²² to data-driven approaches,^{9,138} Recent successes have been reported coupling neural networks to Monte-Carlo tree search (MCTS),⁹ and within reinforcement learning frameworks,¹⁶⁶ deviating from more traditional expert systems.^{103,104,106,116,122,132,133,167} Their ability to rationalize a set of promising synthetic routes from reaction data, has been realized in the framework of Design, Make, Test, Analyze (DMTA) cycles, in which they have played an integral role for coupling to automation platforms.¹³⁸ However, despite recent achievements in the field to advance predictive capability, little attention has been paid to the underlying datasets, the size of the dataset required, an assessment criteria specific to the template prioritization method and overall model performance.¹⁶⁸

Retrosynthetic planning or analysis refers to the technique used by chemists to recursively deconstruct a compound into its simpler precursors, until a set of known or commercially available building blocks is reached.¹⁰⁵ After an initial pattern recognition step, a chemist works in the reverse direction, using a knowledge-base of synthetic transformations ('synthetic tool-box') obtained through years of experience and exposure to a variety of both successful and failed chemistry,^{169,170} to intuitively identify and prioritize a promising set of forward transformations required to synthesize a given compound. To complement this process, computer assisted synthesis planning (CASP) tools are desired that can rapidly consider a vast body of chemical knowledge, effectively prioritize a set of reactions, and develop synthesis plans that can be tailored for the domain in which they will be applied. These have been reviewed extensively elsewhere.^{103,104,106,116,129,130,165,171–173} With the rise of automation,^{97,138,174} *de novo* design,¹⁷⁵ and more extensive virtual libraries,⁶¹ such a tool has the added requirement that it must be able to pre-filter compounds prior to synthesis, thus reducing experimental failure and accelerating Design, Make, Test, Analyze (DMTA) cycles prevalent in molecular design.^{138,165,176,177}

Herein, we investigate the role of the template prioritization method and the tree search algorithm derived from the work of Segler and Waller.⁹ Template prioritization is framed as a multi-class classification problem, for which we employ a neural network which outputs the probability of applying any given template, henceforth referred to as the policy network. This constitutes the machine learning (ML) part of the process, which we couple to a search strategy and decision-making process in the form of a tree search. Together these constitute an AI driven model for retrosynthetic planning. We examine this model in the context of the underlying datasets, pooling from internal AstraZeneca ELN, publicly available USPTO,⁹² proprietary Reaxys¹⁷⁸ and Pistachio data.⁹¹ The overlap and relations between the datasets are examined. The final model's performance is tested on a set of 1731 compounds from a set of 41 virtual libraries designed at

AstraZeneca between October 2017 and January 2019, in relation to policy network accuracy, percentage of routes found, and the number of compounds synthesized experimentally. Thereby, demonstrating the potential use for such tools in DMTA cycles, and how datasets with known experimental results can be used to assess model performance and improvement of CASP tools. As such, we relate our findings of model performance to the underlying datasets, thereby demonstrating that models built on datasets such as internal or publicly available data can predict synthetic routes in line with the literature.

3.2 Results and Discussion

3.2.1 Template Specification

Templates were extracted using an adaptation of Coley *et al.*'s implementation for rule extraction,¹³⁴ which only contain the immediate neighborhood of the reaction centers, thus do not capture the extended environment required to account for leaving and protecting groups. In addition, the algorithm failed to account for reactive species, without specification of which, the reactants would not be regenerated. This has since been corrected by Coley *et al.* in RDChiral and has been extended in this study to encompass *ca.* 75 functional and protecting groups commonly used in organic synthesis.⁹¹ These were determined by analysis of frequently used reactions in the underlying datasets. We found that half of the top 10 templates across all datasets, and 12% of the Pistachio dataset accounted for protections and deprotections. This value is similar across all datasets examined in this study and demonstrates the utility of protecting group strategies in organic synthesis. Furthermore, we determined that these improvements translate into the model being able to account for the extended molecular environment for the groups specified. However, whilst the model can employ protections and deprotections, their use is not necessarily strategic. Further work is required to allow the model to learn their most appropriate use and incorporate them for maximal effect into synthetic route planning. The model is also limited in that it cannot learn the form of new protecting and functional groups from additional data and is restricted to those specified.

3.2.2 Reaction Datasets and Template Coverage

Given the variety of data sources, patents (USPTO and Pistachio), literature and patents (Reaxys), and industrial data (AstraZeneca ELN), it is interesting to note that a comparable number of templates were extracted from the Reaxys and patent datasets (Table 3). However, whilst both template sets are similar in size they differ in their coverage of the reaction space as highlighted in Figure 4. The inclusion of the Reaxys data offers a greater breadth of unique reaction templates, accounting for 41.1% of the overall combined dataset. The comparably high number of unique templates extracted from the combined patents data (32.5%), suggests that a considerable portion of patents data covered are not present in Reaxys (7.4% overlap), or that the structural components that make up the templates are unique to Reaxys. The exact differences between the patent coverage of the patent datasets (USPTO and Pistachio) and Reaxys is not clear with regards to the

templates that can be obtained. Furthermore, the increased number of structural components and templates unique to the Reaxys dataset may be a residual artefact of multi-step reaction pathways. In this regard, we have filtered for all multi-step reactions, such that they have been removed from the dataset to the best of our knowledge.

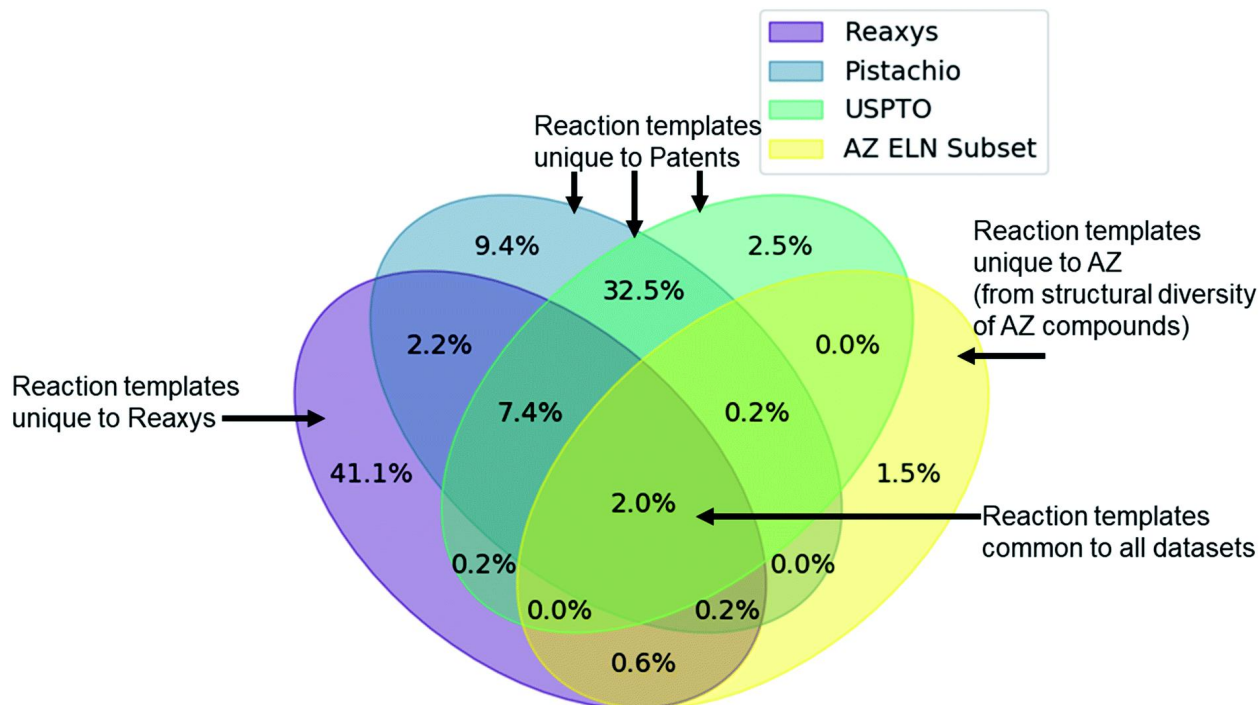


Figure 4: Venn diagram showing the overlap of the patent datasets (USPTO, Pistachio), Reaxys and a subset of AstraZeneca ELN data. Percentages are expressed as being part of the combined dataset. Only 2% of the extracted templates are common between all datasets, and 11.6% between Reaxys and patent data. All datasets add a unique component to the overall dataset, where the subset of AstraZeneca ELN data is the smallest contributor (4.5%) owing to the comparably lower dataset size. The two patent sets differ in content and coverage of the reaction space owing to the different time periods covered and the algorithms used for mining the data. These observations and the calculated overlap are dependent upon the template extraction strategy used, the specificity of the template (radius 1 in this case), and the subsequent procedure for the identification of duplicates/redundancies. Therefore, the percentages expressed hold true for the strategy used in this study and a template radius of 1.

The discrepancy between the two patent sets can be rationalized by the time-period over which the data was collected. The USPTO dataset accounts for reactions published up to September 2016 whereas Pistachio includes reactions until 17th Nov 2017. Further differences in the Pistachio and the public USPTO set arise from the inclusion of ChemDraw sketch data, and text-mined European patent office (EPO) patents which are included in Pistachio. The sketch data may be missing agent and condition details, as they are ‘as drawn’, and do currently not incorporate information from the accompanying text. Therefore, species that contribute a changing atom or bond may be absent and would not be incorporated in the template extraction. As this information cannot be included in the templates, the reaction is discarded, and no template is extracted.

The subset from the AstraZeneca ELN data accounts for 1.5% of unique templates. Additionally, we observe that there is a greater overlap with Reaxys than the patent data. These do not necessarily

correspond to novel reactions, but rather are an artefact of the structural diversity present in the AstraZeneca collection. For instance, the synthesis of a novel lead compound could have different atomic environments around the reaction center compared to the literature or patent precedent on which it was based, thus leading to a new reaction template. Similarly, 2% of all templates are common between the datasets, thus there is a small degree of structural overlap as might be expected. These observations and the calculated overlap are dependent upon the template extraction strategy used, the specificity of the template (radius 1 in this case), and the subsequent procedure for the identification of duplicates/redundancies. Therefore, the percentages expressed hold true for the strategy used in this study and a template radius of 1. Additionally, they are an upper bound estimate for the template overlap given the template extraction strategy used in this study, and the error associated with the redundancy identification method, as not all duplicates may have been removed.

3.2.3 Neural-Network Guided Template-Based Retrosynthetic Planning

Neural-network guided template based retrosynthetic planning methodologies were first pioneered by Segler and Waller.^{9,135} They trained three separate networks: an expansion policy which predicted a set of templates to be applied for a given compound, a rollout policy which predicted a stricter and more specific set of templates to be applied for a given compound, and an in-scope filter trained on positive reactions and a virtually enumerated set of negative reactions. In contrast, this study eliminates the expansion and in-scope filter policies, and focuses on a “naive” baseline retrosynthetic model using only a network inspired by that termed rollout policy by Segler and Waller.⁹

The network predicts which template to use given a compound, and a set of precursors is generated from the application of the template. This is then recursively applied to generate a retrosynthetic tree. The three primary conditions that must be fulfilled for a retrosynthetic route to be valid in this study are as follows. Firstly, there must be a template that has been extracted from the dataset which can be predicted for a given context.

Secondly, the predicted template can be successfully applied. Where successfully applied is defined as: the application of a template *in silico* that generates a set of precursors/reactants. The “success” is in reference to there being subgraph match between product and template, which enables the generation of a set of precursors, and does not reflect whether a reaction will be successful (that the reactants generated by application of the template will form the product) in the wet lab. Additionally, the set of precursors are required to be valid SMILES. It is native to the template-based approach that application of a template to the product or queried compound preserves the global structure of the compound and only alters that of the reactive site, therefore in this context it is implied that a valid SMILES also constitutes a valid set of reactants sharing the same structural features as the product. However, these are not necessarily viable precursors in the sense that they are devoid of selectivity issues and will work in the wet lab. This is a limitation we

have found that is inherent to the template-based methodology and in some cases originates from the underlying dataset from which the templates were extracted, as this “error” is carried forward.

While the ultimate task is to predict synthesis that will work in the wet lab, we draw a distinction in this study by attempting to first determine what can be predicted *in silico*. To this end, we view the goal of the neural network policy as being the maximization of the number of templates that can be applied. Thereby, enumerating all possible disconnections that fall within the top 50 predicted templates for a given compound. Finally, the terminal state of a route is determined by checking if the enumerated precursors are commercially available. However, this is not to say that they are devoid of reactivity conflicts, the identification of which is left to reaction prediction models that are not implemented in this study.

3.2.4 Template Size and Policy Network Accuracy

In previous studies, accuracy has been used as a metric to gauge the network's performance for the task of retrosynthetic planning.^{9,135,137} The accuracy of the policy network reflects its ability to correctly predict a reaction template. However, for the task of retrosynthetic planning the aim is to predict several applicable templates, not just the one recorded in the dataset. Given the underlying data describes a one-to-one mapping of product to template and the task is to predict a one product to many templates' relationship. High accuracy values are associated with the model's ability to predict the template or reaction center from which it was originally extracted, thus overfitting the data by creating a like for like mapping to the underlying dataset. Additionally, the accuracy does not account for the applicability of the predicted template, for which we and others have found high failure rates owing to an inability to match the template substructure to the target for which it was predicted.¹⁶⁶ This is illustrated in Figure 5, whereby the increased specification of the molecular environment surrounding the reaction center (radius) leads to a higher rate of failure for its application, and translates to decreased model performance. In contrast, the test accuracy does not highlight the extent of the performance decrease, but rather increases as more of the environment surrounding the reaction center is considered, thus is misleading.

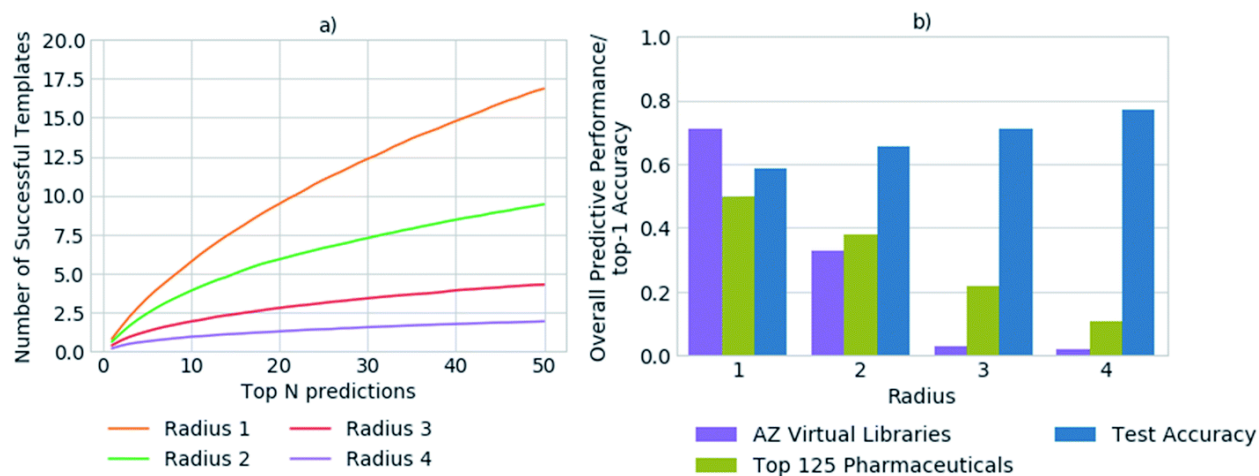


Figure 5:(a) The number of predicted templates that can be successfully applied to generate suitable precursors, as determined for a set of 20 000 randomly selected compounds from ChEMBL. The number of predicted templates that can be successfully applied decreases with increasing template specificity. Only ca. 34% of the top 50 templates are applicable on average in the best case, for the most general templates with a radius of 1. (b) Comparison of the top-1 accuracy on the test set, to overall performance with respect to the ability to generate full synthetic routes, for a set of 1731 compounds from 41 virtual libraries (AZ Virtual Libraries), and the top 125 small-molecule therapies of 2018 by sales (top 125 pharmaceuticals). The top-1 accuracy on the test set is not reflective of overall model performance and increases with template specificity. In contrast, the overall performance of the model decreases with increased template specificity as demonstrated for the virtual library and top 125 pharmaceuticals datasets.

We propose that in conjunction with the accuracy, the more task-specific measure of the number of applicable templates be used for policy assessment, and a more holistic view be taken of overall model performance. In all datasets examined, on average less than 1% of all templates were applicable for any given compound. Whereby, only ca. 0.00035% of all templates were applicable and in the top 50 templates prioritized by the network for any given compound. Increasing template specificity further reduces the number of templates that can be applied in a given context. Therefore, to balance specificity with generalizability we propose that templates considering the reaction center and the first degree nearest neighbors be used, in conjunction with the specification of a variety of functional and protecting groups, to maintain chemical integrity.

3.2.5 The Effect of Template Library Size on Performance

Figure 6 shows the top-1 accuracy computed for the hold out test set for a range of library sizes using templates obtained from the USPTO dataset, as compared to the ability to predict full synthetic routes to 1731 compounds in a series of 41 virtual libraries designed at AstraZeneca. We observed that the accuracy decreases with increasing template library size, where the size of the template library reflects the top N templates in the USPTO dataset. In comparison the average predictive ability of the model increases, reflecting a more task specific measure of model performance. Where predictive ability refers to the ability of the baseline retrosynthetic model (policy network combined with tree search) to generate a retrosynthetic route. In this context the predicted route is not assessed for ‘quality’ by use of more powerful reaction prediction models,¹⁵¹ or comparison to existing literature in an automatic fashion, but rather is a reflection of whether a retrosynthetic route can be proposed *in silico* from reaction datasets.

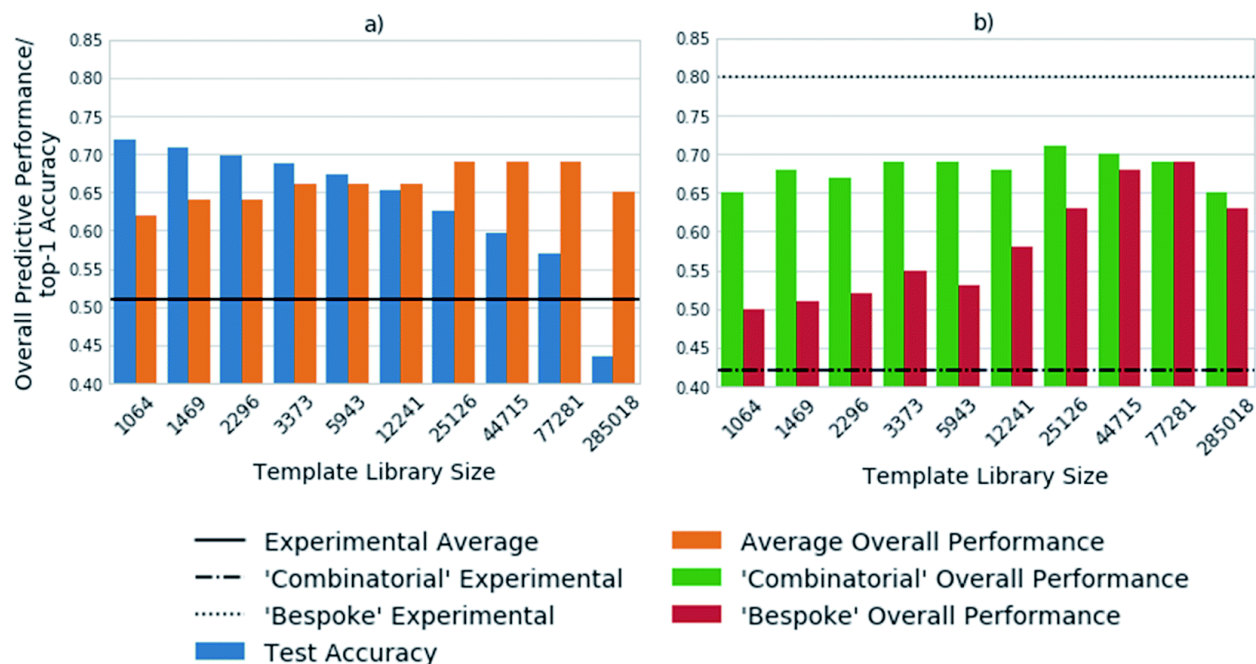


Figure 6: The template libraries were obtained by filtering the USPTO dataset for templates occurring a minimum of 1, 2, 3, 5, 10, 20, 35, 50, 75, and 100 times. A model was trained on each library and the results are shown for: (a) the top-1 accuracy on the test set, as compared to the overall performance. The overall performance is with respect to the ability to predict full synthetic routes to a set of 1731 compounds from 41 virtual libraries designed at AstraZeneca. The experimental average refers to the percentage of compounds synthesized out of those sent for synthesis after refinement of the virtual library. The accuracy decreases with increasing template library size, whereas the overall predictive performance increases up to a library size of the 77 281 most frequently occurring reactions. (b) The virtual library set can be further broken down into libraries designed using a 'combinatorial' approach, and a broader set of reactions using more 'bespoke' chemistry. The overall model performance increases marginally for the 'combinatorial' libraries with increasing template library size. Whereas, the libraries requiring more 'bespoke' chemistry for their synthesis benefit from the inclusion of additional reactions.

Of note is the increasing difference between the accuracy and overall predictive performance as the library size increases. Whilst the test accuracies have been measured for a baseline template-based CASP tool, template-free models are also prone to misleading accuracy values. In both cases the task is to predict a series of viable outcomes, however the accuracy reflects the ability to predict the 'ground truth' from the underlying dataset, which inherently accounts for only one 'true' value, thus is partially known. In a similar work, Segler and Waller used the top 1, 10 and 50 accuracies to gauge the performance of their network, and showed that a model trained on 17 134 rules extracted from Reaxys, covering 52% of the dataset, was able to predict the reaction center with accuracies of 50.1%, 89.1%, and 96% respectively.⁹ In an extension of the work considering only single step reactions Baylon *et. al.* reported an accuracy of 81% on 129 rules compared to 83% on 137 rules by Segler and Waller.^{135,137} However, we have found that accuracy can be misleading when used for the assessment of overall model performance as shown in Figure 5 and Figure 6, and specifically for the assessment of whether the network is able to correctly predict applicable reaction templates for single step reactions.

The virtual library set can be further broken down into libraries designed using a 'combinatorial' approach, and a broader set of reactions using more 'bespoke' chemistry, which covers the reaction

space more extensively. This enabled consideration of domain dependency with respect to template library size. We found that virtual libraries designed using a combinatorial approach benefited marginally from increasing the template library size. With the 1064 most frequently occurring templates in the USPTO dataset, routes could be found for 65% of the compounds in the virtual libraries designed using a combinatorial approach. This increased to a maximum of 72% when the 25 126 most frequently occurring templates were used. This is in line with what would be expected, as combinatorial libraries employ frequently used and robust reactions in their design.

In contrast, route predictions for libraries designed with a broader range of chemistry in mind, denoted ‘bespoke’, benefit from a larger template library size which covers the reaction space more extensively. Using the 1064 most frequently occurring templates in the USPTO dataset, the model predicted synthetic routes to 50% of the compounds in the ‘bespoke’ library, increasing by 19% to a maximal value of 69% when using 77 281 reaction templates. This alludes to the point that increasing the number of templates increases the chemical diversity of the templates, thus more synthetic routes can be found than with smaller template library sets. The increase in diversity of the templates originates from the fact that no two templates are the same, as they account for different sub-structural patterns. Increasing the template library size, also increases the probability of finding a sub-structural match to the product to which the template is applied. On the other hand, the ‘combinatorial’ libraries are less diverse, arising from the fact that a limited number of reactions were used to make them. Therefore, templates matching sub-structural patterns occurring within ‘combinatorial libraries’ are also limited. There is a balance between the number of reaction templates and the reaction space they represent, which is specific to the domain in which the tool is applied. However, increasing the number of reaction templates also introduces noise. This can be seen in Figure 6, where the overall predictive performance falls by 4% and 6% for the ‘combinatorial’ and ‘bespoke’ libraries respectively, when increasing the template library size from 77 281 to 285 018 reaction templates. Furthermore, increasing the number of reaction templates to those that occur less frequently (less than 3 times), increases the difficulty of identifying suitable templates. The increased difficulty more than offsets the increased coverage of the reaction space (Figure 6).

Compared to the experimental results for each virtual library, we found that the model consistently over-predicted the number of compounds that could be synthesized for the ‘combinatorial’ library. Whereas, the number of compounds that could be synthesized for the ‘bespoke’ library was consistently under-predicted. This highlights that only considering the number of compounds for which routes can be predicted does not afford enough granularity for the assessment of synthetic routes, and CASP tools. For instance, it is likely the baseline retrosynthetic model examined in this study may over predict the number of compounds that can be synthesized from the ‘combinatorial’ library, because some of the predicted steps may not translate to the wet lab. Further still, the conditions required to carry out the reaction in the forward direction are not predicted by the model, nor is there any certainty that they would yield an outcome in the wet lab

if predicted. This task is left to separate models that have not been implemented in this study, that attempt to predict conditions for a queried set of substrates and a given transformation.¹⁴¹

The under-prediction of retrosynthetic routes to compounds that were experimentally obtained in the ‘bespoke’ libraries, raises questions as to the coverage of the reaction space covered by the templates, and the ability of the policy network to prioritize suitable templates. Figure 6 examines the performance for a model trained on the USPTO dataset, thus it can be envisaged, based on Figure 4 that inclusion of the Reaxys dataset may improve the result obtained by enabling the prediction of templates missing from the USPTO data. However, as alluded to by Figure 6, this may increase the difficulty in identifying suitable templates, therefore improvements in the policy networks may be required for a higher number of routes to be found. The number of routes suggested by this methodology will be an upper bound estimate, which will decrease as measures are taken to increase the ‘quality’ of the suggested routes through incorporation of reaction and condition prediction models.

Furthermore, the reasons for a ‘failed’ synthesis are not always known and can be dependent on the nature of the project, the skill of the chemist, and the conditions used, to name a few factors influencing the outcome of a synthesis. These factors cannot always be quantified or considered qualitatively, thus both the predictions and ‘true’ experimental results have an associated degree of uncertainty which proves difficult to measure.

3.2.6 Datasets and Performance

We compared the predictive performance of models trained on each reaction dataset, and combinations thereof, on 1731 compounds from 41 virtual libraries at AstraZeneca and the top 125 small molecule therapies of 2018 (Figure 7). The models, regardless of reaction dataset, consistently over-estimate the number of compounds that can be synthesized in the case of the virtual libraries, and under-estimate with regards to the top 125 small molecule therapies. For both cases, the average number of steps taken to synthesize a molecule is 4, however the average time taken to solve each molecule varies considerably with the dataset size (Figure 7). The smaller datasets are faster at finding routes to a given compound (<4 seconds) owing to a smaller search space in comparison to the larger search spaces associated with the larger datasets (Pistachio and Reaxys). The simple architecture used is not able to handle the large search space and is biased towards frequently occurring reactions, which are augmented by the additional data in the larger sets. In the case of the top pharmaceutical compounds, the lower predictive performance may arise from more sophisticated ring systems, and natural product like structures upon which the final compound is based. Reactions of this nature are not prioritized by the network as they are infrequent, thus become difficult to separate from the noise. Whereas predictive performance on the virtual library dataset is higher than that for the top 125 small molecule therapies of 2018 across all datasets, as they make use of the most frequently employed reactions.

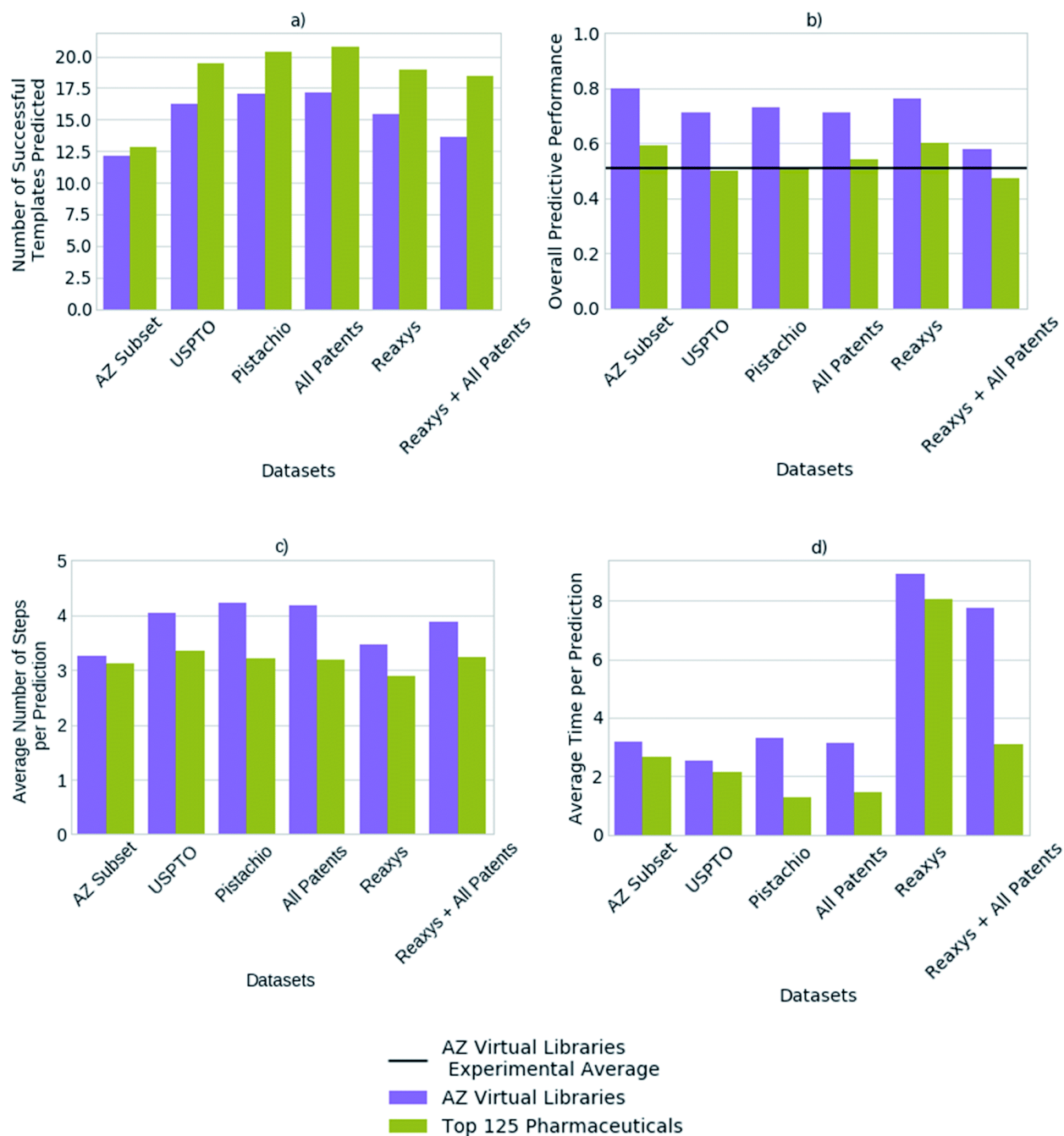


Figure 7:(a) The average number of successfully applied templates of the top-50 predicted templates for one-step synthesis per compound (b) the overall predictive performance with respect to the ability to generate full synthetic routes (c) the average number of steps taken per prediction per compound (d) the average time taken to predict a full synthetic route per compound, as found for each reaction dataset, for a set of 1731 compounds from 41 virtual libraries designed at AstraZeneca and the top 125 small molecule therapies by sales in 2018. The number of predicted templates that can be successfully applied for one-step synthesis does not correlate to the model's overall ability to generate full synthetic routes, when comparing between different template library sources (datasets). Whilst the model built on the subset of the AstraZeneca ELN suggests the lowest number of possible options at each step, the overall performance is comparable to, or exceeds models built on the larger reaction datasets. Thus, a model built on 4.5% of all templates considering all the datasets combined, can predict synthetic routes to compounds equally as well as the larger datasets examined.

The average number of successfully applied templates of the top 50 predicted templates for one-step synthesis per compound varies considerably across the reaction datasets examined (Figure 7). The model built on a subset of the AstraZeneca ELN appears to be worse than the models built on other reaction datasets by this measure. However, we have found that the number of options the network suggests for one-step synthesis does not impact overall model performance in this case. Thus, as Segler and Waller suggested in a previous study examining training set size,¹³⁵ models competitive with those built on larger reaction sets can be obtained with datasets as small as an internal ELN. The subset of the AstraZeneca ELN accounts for 4.5% of the template library obtained from a combination of all datasets examined yet is capable of providing sufficient training data to train policy networks and resulting models which are competitive with those of larger proprietary datasets. However, we expect that this is domain specific and reflects that the subset of the AstraZeneca ELN is tailored to the medicinal chemistry domain in comparison to the patent and Reaxys datasets, which are more extensive in their coverage (Figure 4). This further demonstrates that there is a balance between the type of chemistry covered by the template library set, and the size of the template library. An optimal set would be domain specific, and cover enough examples of sufficient diversity, that the output space would be manageable by the policy network. In the current approach we have found that as the dataset size increases, so does the output space of the policy network (Table 3). This increases the time taken to train the network, and makes it increasingly difficult for the network to prioritize appropriate reactions as seen when increasing template library size in Figure 6.

Previous studies have demonstrated that models built on the USPTO dataset, can predict one-step synthesis. We show that despite the seemingly lower amount of data in the USPTO dataset compared to Reaxys (Table 3), the USPTO dataset accounts for 44.8% of the template library obtained from a combination of all datasets examined. In comparison to 53.7% which comes from Reaxys, the largest of all the datasets examined. Whilst there is an 8.9% difference and the coverage of the reaction space that the templates encode varies (Figure 4), this does not appear to be a limiting factor for route prediction in the medicinal chemistry domain. Figure 7 shows that the model trained on Reaxys marginally outperforms that trained on the USPTO dataset, at the expense of longer prediction times. Furthermore, we show that as the size of the dataset increases to a combination of both Reaxys and the combined patents data (USPTO and Pistachio), the overall performance of the model decreases with regards to both time and number of routes identified. This may reflect the decrease in performance observed in Figure 6b, whereby increasing the number of templates increased the difficulty for the network to prioritize suitable templates.

We noted that the fingerprint size used to encode the product had a marginal effect on the ability of the model to predict full synthetic routes for the internal virtual library dataset (Input ECFP4 fingerprint size and performance). In addition, we found that increasing the size of the stock library to include the ACD catalogue, increased the ability of the model to predict full synthetic routes to compounds in the virtual library. For both the ‘Combinatorial’ and ‘Bespoke’ libraries, the model

was able to reduce the average time taken to predict full synthetic routes with the ACD catalogue, as well as reduce the average number of steps by one. The reduction in the average number of steps is more pronounced for the ‘Bespoke’ libraries, whereby it is consistent over both the USPTO and Reaxys datasets. This is in comparison to the ‘Combinatorial’ libraries whereby the reduction in the number of steps is not observed for the combined Reaxys and patent data (Performance and stock set of compounds).

3.2.7 Comparison of Test and Reaction Datasets

Figure 7 compared the performance of models built on a range of reaction datasets with two compound sets. A set of 1731 compounds obtained from internal AstraZeneca virtual libraries, and a set of the top 125 pharmaceutical compounds by sales in 2018. The former AstraZeneca virtual libraries can be viewed as general medicinal chemistry targets, given that there is no or little overlap with the reaction datasets (Table 2), to which the algorithm is able to generalize as shown in Figure 7. Whereas, the top 125 pharmaceuticals are well-known targets in the training domain given the much greater overlap with the underlying datasets (Table 2).

We found that the baseline retrosynthetic model examined in this study can generate retrosynthetic routes for compounds outside its training domain. While these routes may not necessarily be feasible in the wet lab, they can be viewed as ideas upon which a trained chemist can build. Alternatively, the algorithm may help to identify building blocks and precursors to a target compound that were previously not considered. In this regard, the quality of the retrosynthetic routes generated has not been assessed and is left to manual inspection.

Table 2: Percentage overlap of compounds in each of two compound datasets, AZ virtual libraries and top 125 pharmaceutical compounds by sales in 2018, with those reported as products in each of the reaction datasets. As expected, the top 125 pharmaceuticals have a much greater overlap with the products in each of the reaction datasets in comparison to the AZ virtual library compounds. This is because they are patented compounds with a literature precedence where both the patent and literature examples predate the most recent timepoints in the underlying dataset. Furthermore, the AZ virtual library compounds do not overlap with the literature and patent datasets and lie outside the training data.

Dataset	AZ Virtual Libraries (%)	Top 125 Pharmaceuticals (%)
USPTO 1976-2016	0	47
Pistachio Nov 2017	0	58
Combined Patents	0	58
Reaxys	0	70
Reaxys + Patents	0	78
AZ ELN Subset	2	4

3.2.8 Exemplary Synthetic Routes

Comparison to existing literature in the domain showed that the model trained solely on the USPTO dataset was competitive with that reported in the literature (Figure 8), and was able to find a route to the target compound in 4.26 seconds.⁹ This was also observed for models trained on the subset of the AZ ELN, Pistachio and Reaxys datasets. We found that the model was able to suggest an alternative route in addition to that reported, involving a ring formation (Figure 8). Furthermore, we show that the model can predict routes to the top 125 pharmaceutical products, where the

performance is dependent on the stock set of compounds. Examples of which have been given in the Appendix (Exemplary Synthetic Routes). The route predicted using the model trained on the USPTO dataset to Amenamevir is compared to the literature route.⁹¹ Both routes vary in the order of the steps they take, with the predicted route preferring a standard amide coupling over the amide Schotten–Baumann. However, the predicted route displays reactivity conflicts as deprotonation of the amine in the second step competes with the amide coupling. A further selectivity issue is present in the first disconnection step predicted for Amenamevir, as there will be competition between the nitrogen in the secondary amine and the amide. This is not the case for the literature route due to the ordering of the steps. Selectivity issues are also observed in Figure 8a for the last retrosynthetic step (first step in the forward synthesis) where there is competition between the –OH and alkyne C–H in the aromatic nucleophilic substitution. While we know the model to be capable of using protecting groups, these are not necessarily used in a strategic way, nor is their appropriate use always identified.

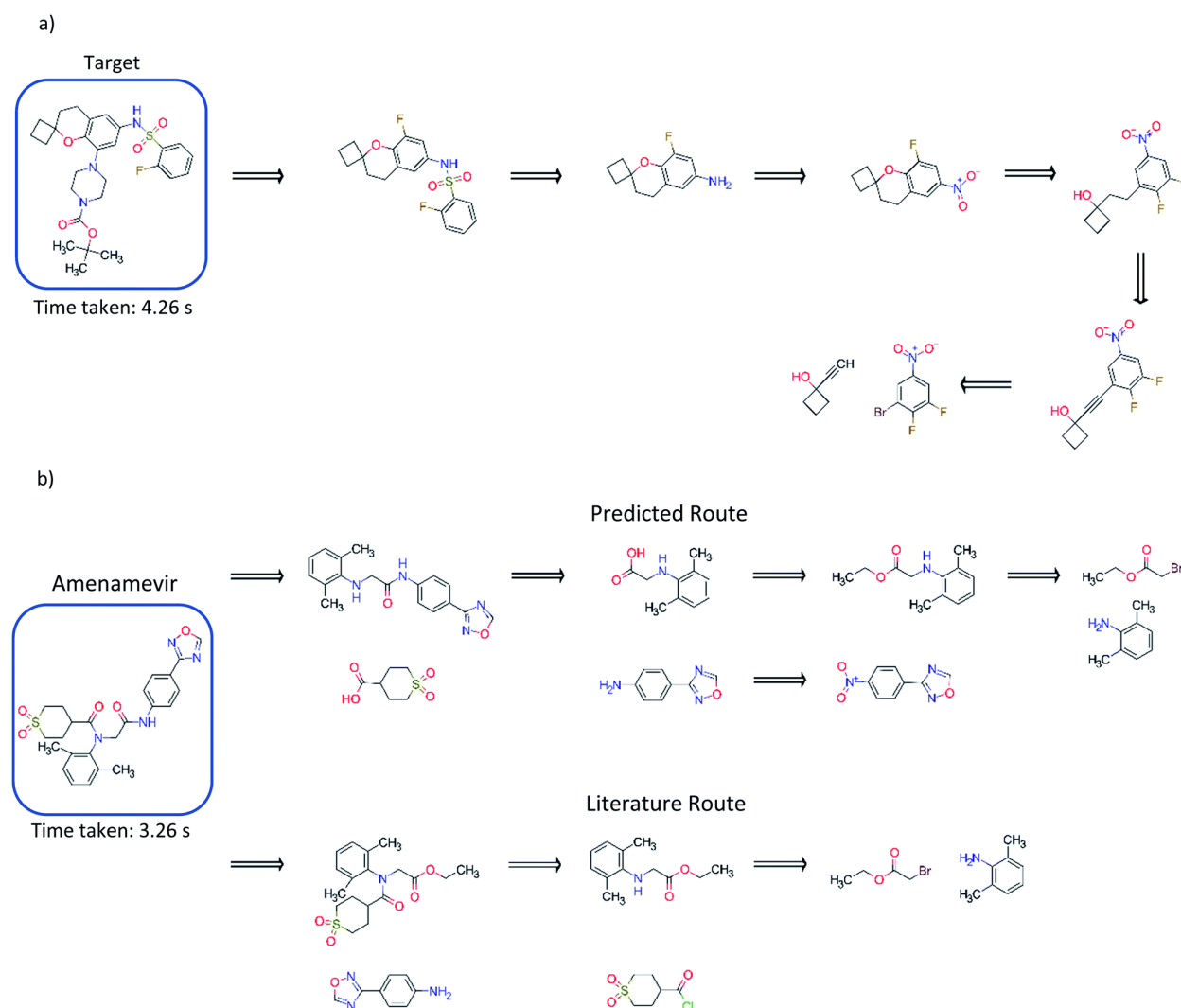


Figure 8: (a) Comparison to the exemplary synthesis shown by Segler and Waller.⁹ The model trained on the USPTO dataset, finds an alternative route to that in the previous study, and finds synthetic routes to the target compound in 4.26 seconds. The model can

prioritize and apply ring formations as demonstrated in step 4. (b) Comparison of the route found by the model trained on the USPTO dataset with the literature route for Amenamevir.¹⁷⁹ The model can suggest a route comparable to the literature, differing in the sequence of steps and using similar reactions to those in the literature. The predicted route is found in 3.26 seconds.

3.3 Conclusions

We have developed and implemented a baseline retrosynthetic tool with only a single neural network, to investigate the role of the ML template prioritization method in the tree search algorithm derived from the work of Segler and Waller.^{9,135} We have found that models trained on datasets as small as the internal ELN (4.8% of all templates) and USPTO datasets (44.8% of all templates), are sufficient for the prediction of synthetic routes to compounds found in medicinal chemistry pipelines. Furthermore, we demonstrated the potential use for such tools in compound selection and prioritization in DMTA cycles and suggest that datasets with known experimental results can be used to assess model performance.

In addition, we demonstrate that accuracy can be a misleading measure for the performance of the policy network and final tree-search model. Thus, we propose an alternative approach to assessing the ability of the policy network to identify and maximize the number of templates that can be applied, based on the number of templates that can be successfully applied in the top N predictions, for a given context. We demonstrate that the specificity and generalizability of the extracted templates must be balanced such that, the first degree nearest neighbors to the reaction center, are used in conjunction with the specification of functional and protecting groups that are common in organic chemistry.

We have found there is a dependence between the size and content of the template library used, and the domain in which it is applied. We found that syntheses of compounds originating from combinatorial libraries could be predicted using the most frequently occurring reactions. In contrast, compounds originating from libraries requiring more complex syntheses, required an expanded template set for their successful prediction. Further work is required to make use of the broad selection of reactions available to improve the variety and complexity of routes suggested. Further investigations into the template extraction process are also required to determine their descriptive limits and how this translates into route prediction.

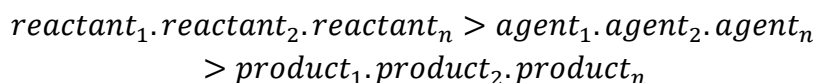
3.4 Methods

3.4.1 Reaction Datasets and Template Extraction

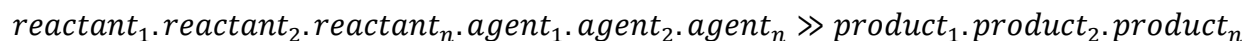
Of the datasets used, only the United States Patent Office extracts (USPTO) ranging from the years 1976 to 2016 is publicly available.⁹² This is split into granted and applied patents and is openly available for use by the community. A subset of the AstraZeneca Electronic Notebooks (ELN)

were mined (May 2019) to yield the internal proprietary dataset, considering only positive reactions, classified as those with a yield greater than 1% and having a conclusion statement. The Pistachio (2017-11-17)⁹¹ and Reaxys¹⁷⁸ datasets are commercially available, provided by NextMove software and Elsevier respectively under licensing agreements. The Reaxys dataset was filtered for multi-step reactions to yield only the intermediate single step records for which templates were extracted. Full details of the number of reactions and unique extracted templates can be found in Table 3.

All reactions were atom-mapped and classified using the commercially available Filbert and HazELNut packages (v. 3.1.8) provided by NextMove software.¹⁸⁰ These were subsequently processed using RDKit and RDChiral for template extraction,^{20,181} in conjunction with a custom reaction class developed by the authors to facilitate reaction processing. The reactions are parsed as reaction SMILES,¹⁹ along with the ID linking back to the data source, and classification code or textual classification obtained from the NameRxn software.¹⁸² The reaction SMILES are of the form:



where the reactants, agents, and products are separated by ‘>’ and the individual non-covalently bound species represented by a ‘.’ according to the Daylight SMILES specification.¹⁸³ The definition of reactant and agent is ambiguous, as agents may participate in the reaction and contribute mass to the products. Additionally, as the templates are extracted based on atom-mapping, only the species contributing to the product or changing during the reaction were considered in the process. Thus, we have moved all agents into the reactants to give a reaction SMILES of the form:



Through string manipulations, the reaction SMILES were split into their component parts on the ‘>>’ ensuring that the number of parts did not exceed three, one for each, reactants, agents, and products. Reactions leading to more than one product, incomplete reactions (*i.e.* missing reactants or products), or reactions in which the reactants and product were equivalent were removed. Equivalence was determined by converting the reactants and products to InChI and comparing.²⁸ Permutations in the ordering of reactants and products were accounted for. However, this was not significant in this case as we only account for reactions with one product.

Reaction templates were extracted as SMIRKS patterns using RDChiral,¹⁸¹ which we modified to consider an additional *ca.* 70 commonly occurring functional and protecting groups as determined by an analysis of the underlying datasets and extended to commonly used protecting groups in the wider literature.^{184,185} These are automatically identified through a substructure search of the encoded protecting groups and included in the templates alongside the reaction center and first degree nearest neighbor atoms. The reaction center is defined as atoms and bonds that change

during the reaction. Owing to the number of variations sharing the same core structure for some protecting groups *i.e.* silyl ethers, esters, but varying in alkyl chain length, we have refrained from an exhaustive encoding of all possible protective groups. Rather, we have focused on those we found to be commonly occurring in the dataset and cover the main form of the protecting group, leaving the decision of the exact form to the chemist.

The extracted templates were parsed and checked for validity in RDKit,²⁰ following which the template was applied to the product of the reaction from which it was extracted to determine if an outcome could be generated. The outcomes were assessed using the definitions shown in Figure 9, and the quality of the template extraction process quantified.

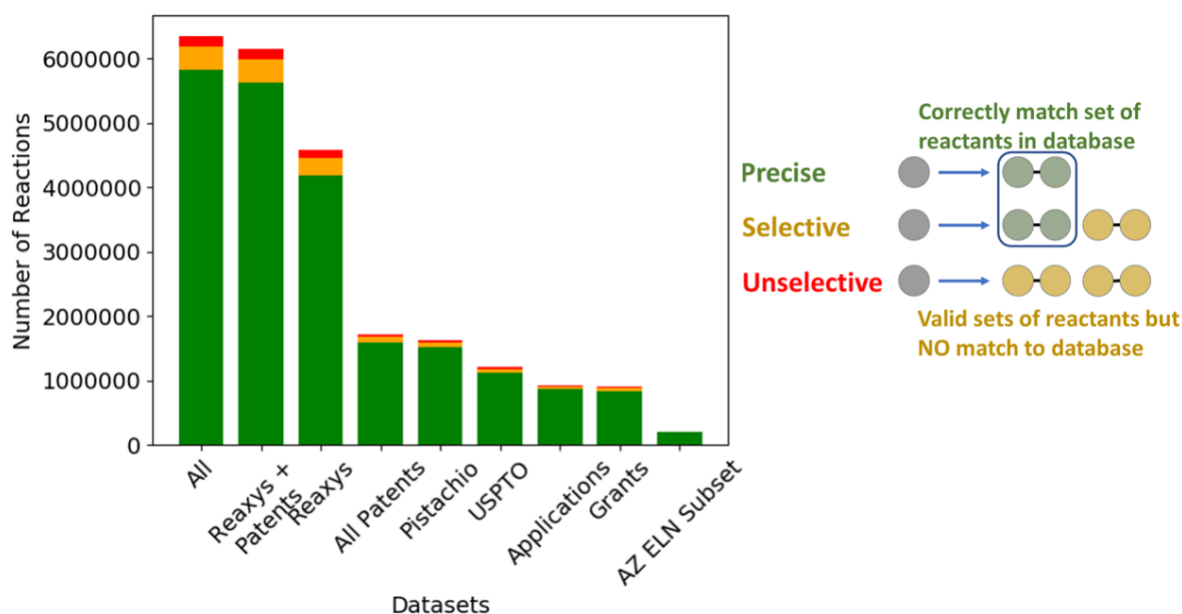


Figure 9: Left) Comparison of the quality of the extracted templates across the available datasets with respect to their ability to regenerate the reactants of the reaction from which the template was extracted. Right) Schematic of the categorization criteria used for determining the reaction templates selectivity, which we use as an initial measure of quality. The categories are defined as: Precise) The template can generate only the reactants from the reaction from which the template was extracted. Selective) The template generates the reactants from the reaction from which the template was extracted in addition to other possible precursors that are not part of the original reaction. Unselective) The template generates reactants that do not correspond to any of the reactants in the reaction from which the template was extracted. These may or may not be viable reactants.

The reactions and resulting templates were hashed individually following a hashing scheme developed by the authors inspired by the reaction InChI (Figure 10).³⁰ This was also used to identify duplicate reactions and templates and can be used as an identifier for database lookups.

The datasets used in this study and their respective sizes, given as the raw dataset size without filtering are shown in Table 3. To our knowledge, the combined dataset is the largest reported to date. To enable clarity in the task specific curation process, the reduction in size through extraction and validation, followed by duplicate removal has been shown. Extraction refers to the extraction of reaction templates from the reaction SMILES,¹⁹ and validation refers to the application of the

extracted template to the product of the reaction from which it was extracted, to determine if the corresponding reactants can be generated. Duplicates were identified as reaction SMILES consisting of identical reactants, agents, and products, using an order invariant hashing scheme accounting for variance in atom-mapping as developed by the authors. Unique reaction templates were also identified in the same manner.

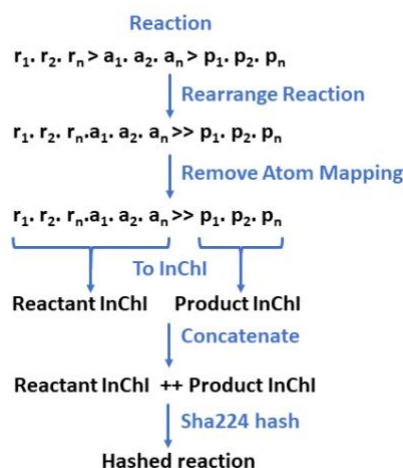


Figure 10: The reaction is initially rearranged to overcome the need for classification between reactants and agents, as the line is often blurred, and their definitions are often the source of debate. Atom-mapping is subsequently removed to overcome the discrepancies between toolkits, and variances in the positioning of the reactants and agents at the point of atom-mapping. The reactants and products are converted to a RDKit mol objects in without separation of the individual species. Conversion to InChI for the reactants and products respectively is carried out in RDKit.^{43,46} This is order invariant and overcomes the issue of having multiple SMILES representing the same molecular structure. The resulting InChIs are concatenated and hashed.

Table 3: Datasets used in this study and their respective sizes, given as the raw dataset size without filtering.

Dataset	Dataset size	Extracted and Validated	Without Duplicates	Templates Extracted
USPTO 1976-2016 ^{92 a)}	3,748,191	3,079,351	1,201,602	302,282
Grants ^{a)}	1,808,938	1,471,088	895,436	239,895
Applications ^{a)}	1,939,254	1,608,263	923,765	223,871
Pistachio Nov 2017 ^{b)}	6,836,027	4,897,300	1,627,792	367,488
Combined Patents	10,587,618	7,976,651	1,711,330	358,307
Reaxys ^{b)}	6,540,786 ^{d)}	5,071,074	4,571,364	361,603
Reaxys + Patents	17,128,404	13,047,725	6,141,875	665,288
AZ ELN Subset ^{b), c)}	398,779 ^{d)}	254,468	207,868	30,805
All Combined	17,523,783	13,302,193	6,342,331	675,530

a) Publicly available b) Proprietary c) Only successful reactions have been considered d) Values reported are those after an initial internal data curation step. Dataset size: refers to the number of reactions available as reaction SMILES before curation or subsequent filtering, unless otherwise specified. Extracted and Validated: refers to the number of reactions that remain after curation, automatic template extraction, and validation of the extracted template by application to the product of the reaction from which it was extracted, to determine if the corresponding reactants can be regenerated. Duplicates were identified as identical reaction SMILES considering variations in the ordering of different entities and atom-mapping. Duplicate templates were identified in the same manner. The number of products refers to products of single step reactions, where duplicates have been removed.

The overlap of reaction templates extracted from the respective datasets was ascertained by using the in-built set methods in Python. We have observed that some of the noise associated with

automatic template extraction originates from incorrect mapping, text-mining errors, and human-error from manual curation. There are several variations of these cases including, incorrect recording of functional groups, incorrect mapping of reactive components (*i.e.* substructures present in the reactive center may also be present in the solvent or reagents, for instance the incorrect mapping of an amine in both the reactant and base), accidental extension of alkyl chains, representation of catalysts and incomplete reactions, examples of which can be found in the Appendix (Data Inconsistencies). Whilst our approach to curation can identify such inconsistencies and disregard their associated reactions, further efforts are required to improve catalyst representation, text-mining, template SMIRKS generation and atom-mapping.

3.4.2 Policy Networks

Template libraries were constructed by filtering the respective dataset for templates that occurred a minimum of N 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.³⁵ Whereas, templates were represented as binarized labels in a one-vs-all fashion using the scikit-learn library using the ‘LabelBinarizer’.¹⁸⁶ 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.¹⁸⁶

The policy networks framed as supervised multiclass classification problems were trained using Keras¹⁸⁷ with Tensorflow¹⁸⁸ as the backend, the Adam optimizer with an initial learning rate of 0.001,¹⁸⁹ and categorical cross entropy as the loss function (Figure 11). 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.

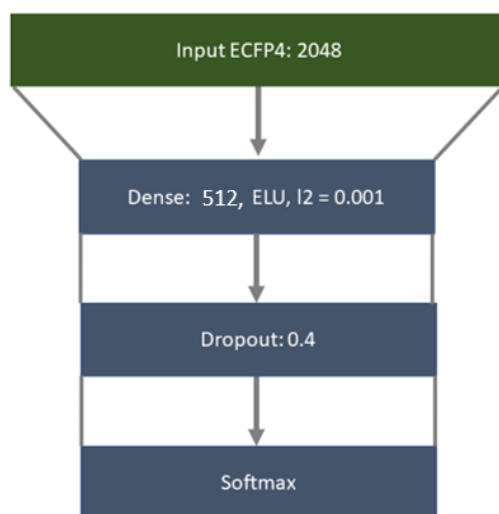


Figure 11: Architecture used to train the 'rollout' policy taking molecules represented as ECFP4 as input, through a fully connected layer of 512 nodes, ELU as the activation function, and L2 regularization set at 0.001. Followed by a dropout of 0.04 and softmax output layer.

3.4.3 Assessing the Number of Successfully Applied Templates of The Top N Predictions

A random subset of 200 and 20 000 compounds from ChEMBL (v. 24.1)¹⁹⁰ were used to assess the baseline number of applicable templates and the applicability of the top N templates respectively, unless otherwise stated. Salts were removed from the ChEMBL dataset using RDKit.²⁰ Random subsets were drawn from the resulting dataset using a random state of 1.

The model to be assessed was loaded into Keras and the compounds to be queried converted into ECFP4 fingerprints prior to passing to the model for prediction. The top N predictions sorted in order of decreasing probability were used for each compound. The templates were applied to the compound in turn using RDChiral to determine if an outcome was generated. Templates leading to an outcome were classed as successful.

3.4.4 Tree Search with 1N-MCTS

The tree search was implemented as a simplification of the algorithm described by Segler *et al.*⁹ The MCTS algorithm was simplified with regards to the policy network. The same network was used for both the expansion and the roll-out. The prior probabilities were not used by default during the selection of leaf nodes for expansion, but the Q value was initialized at 0.5 and N at 1, as expansion counts as a first visit.

3.4.4.1 Algorithm

The search tree is built up from nodes that contain states with current molecules of the route. The root node contains one molecule, which is the target molecule of the algorithm. Other nodes can

contain states with one or more molecules. Each node is bound to others in a directed way as parent-child nodes, with actions as edges. The action is the retrosynthetic reaction performed on one of the molecules of the parent state, to yield the molecules of the child node state. The search algorithm starts with the expansion of the root node (see below).

3.4.4.2 Selection of leaf node

In each iteration the search tree is traversed using the upper confidence bound (UCB) scores of the nodes (Equation 1).¹⁹¹ Starting from the root node, the UCB scores of the children are calculated.

$$UCB = \frac{Q}{N} + C * \sqrt{2 * \frac{\ln N_{-1}}{N}} \quad (\text{Eq 1})$$

Here Q is the current sum of previous rewards. N is the number of times the child state has been visited, N_{-1} is the number of times the parent state has been visited. C is a tunable parameter balancing exploitation and exploration which was set to 1.4 by default. If the selected child is already expanded (*i.e.* has child nodes), the UCB scores of these are then calculated and the next child selected in an iterative way until an unexpanded leaf node is selected. Actions are stored at the parent level, and the child nodes are first instantiated as node objects by applying the associated action when visited (see below).

3.4.4.3 Expansion of node

Expansion is performed by employing the expansion policy neural network for each of the molecules present in the state of the selected node. The top scored reaction templates are filtered to retain the top 50 or until a cumulative policy network score of 0.995 is reached. The possible actions (molecule + reaction) for all molecules are stored at the parent level, and vectors of associated Q and N values initialized (0.5 and 1 respectively).

The action with the highest UCB score is selected for the roll-out. In case of multiple actions sharing the largest score, random selection is performed. The child state is instantiated and added to the search tree by employing the associated reaction template to the molecule specified in the action using RDKit.²⁰ In case the reaction did not give any output, the action Q is given a value of -10^6 , effectively preventing reselection. If no actions are available, the state is marked terminal and the state evaluated with the reward function (see below).

3.4.4.4 Roll out

No in-scope policy was employed after the expansion phase. The roll out policy was identical to the expansion policy and thus allowed for reuse of the previous roll-outs during tree building and searching. Expansion of new child nodes during roll out is similar to the above, except the selection is done by random among the available actions. After each roll-out step the state was evaluated

and the roll out stopped if either the state was solved (all compounds found in stock) or the maximum tree depth reached, or no valid actions are available.

3.4.4.5 Reward calculation and back propagation

The reward function for the final state is then calculated (Equation 2) and the score back propagated through the tree, updating the Q and N values of all parent states between the final state and the root state (target compound).

$$reward = 0.95 * \frac{N_{in_stock}}{N} + 0.05 * \max(transforms) \quad (Eq \quad 2)$$

N is the total number of compounds in the state, N_{in_stock} is the number of compounds that are in stock. Transforms is the number of transforms each compound has undergone with respect to the root compound.

3.4.4.6 Iteration and stop of search

Selection of the next leaf node to expand is then instantiated from the root node, until the maximum number of iterations or the time limit has been reached. If early stopping is wanted, the algorithm can stop if any state contains a solved state with all compounds in stock.

3.4.5 Implementation

The algorithm was implemented in an object-oriented architecture, with a range of global objects for handling the search tree, the stock, the neural network predictions, settings of parameters and a logging object. The global objects were implemented using a Borg pattern that ensures singleton status and easy access though re-instantiation anywhere in the code. NetworkX was used to keep track of the parent-child relations during building of the search tree.¹⁹² The stock object keeps the stock as a set of InChIKeys for fast, hashed tests if compounds are contained in the stock. InChIKeys were calculated through the RDKit API for the InChI software.²⁸ Nodes and states are regular python classes that can have several different object instances. The state object contains information about the current molecules in that state as well as the number of conversions each molecule has undergone from the root states compound. Nodes contain vectors of possible actions and child Q and N values as well as methods expansion, traversing the tree and node expansion.

3.4.6 Stocks

A subset of the AstraZeneca internal catalogue and enamine building block sets were used as the stock set of compounds in all calculations unless specified. InChIKeys were computed for all compounds and duplicates removed. The subset of the AZ internal catalogue was obtained from a database dump of available compounds (January 2019) and contains 60 530 compounds. The enamine building blocks list was provided by enamine, January 2019, and consists of 162 194 compounds after preprocessing and filtering. The ACD catalogue was additionally used to provide a more extensive set of stock compounds.¹⁹³ The compounds which had a CHIME defined where

an InChIKey could be generated was extracted from ACD giving a final stock set of nearly 12.5 million compounds.

3.4.7 Template Library Size and Performance

To study the effect of library size on model performance, a filtering criterion of templates occurring a minimum of 1, 2, 3, 5, 10, 20, 35, 50, 75, and 100 times was applied to generate the appropriately sized libraries, and a policy network trained on each set.

1731 compounds spanning 41 virtual libraries designed at AstraZeneca between October 2017 and January 2019, and the top 125 small molecule therapies by sales in 2018 were used to test the algorithm.¹⁹⁴ The virtual library set can be further broken down into libraries designed using a ‘combinatorial’ approach, and a broader set of reactions using more ‘bespoke’ chemistry. Knowledge of the number of compounds sent for synthesis and the number of compounds successfully synthesized was contained within the dataset. The aim was to couple the policy network to the tree search to determine for how many of the compounds a synthetic route could be predicted, and whether it was reflective of experimental results.

3.4.8 Datasets and Performance

Each dataset was filtered for templates occurring a minimum of three times, and a policy network trained on each set. The policy network was assessed for the number of successfully applied templates of the top N predictions, where N was 50. Subsequently the policy network was coupled to the tree search to form the overall model, which was assessed using the virtual library dataset and the top 125 small molecule therapies by sales in 2018.¹⁹⁴

3.5 Availability of data and materials

AstraZenca, Pistachio and Reaxys datasets were used with permissions. Filbert, NameRxn and HazelNut were used for atom-mapping and classification under license from NextMove software. The implementations source code were made available at <https://github.com/reymond-group/CASP-and-dataset-performance>.

3.6 Author Contributions

Amol Thakkar and Esben Jannik Bjerrum designed and conducted the research. Thierry Kogej, Jean-Louis Reymond, and Ola Engkvist contributed ideas and provided scientific advice. Esben Jannik Bjerrum, Ola Engkvist and Jean-Louis Reymond supervised the project.

4 “Ring Breaker”: Neural Network Driven Synthesis Prediction of the Ring System Chemical Space

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’ uses a data-driven approach to enable 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 the performance of the neural network on a range of ring fragments from the ZINC and DrugBank databases and highlight its potential for incorporation into computer aided synthesis planning tools. These approaches to ring formation and retrosynthetic disconnection offer opportunities for chemists to explore and select more efficient syntheses/synthetic routes.

This chapter has previously appeared as a scientific article in the Journal of Medicinal Chemistry as part of the "Artificial Intelligence in Drug Discovery" special issue. Reprinted with permission Copyright 2020 American Chemical Society.

A. Thakkar, N. Selmi, J.-L. Reymond, O. Engkvist, E. J. Bjerrum, “Ring Breaker”: Neural Network Driven Synthesis Prediction of the Ring System Chemical Space. *J. Med. Chem.* **2020**, 63, 8791–8808. DOI: 10.1021/acs.jmedchem.9b01919.

4.1 Introduction

The recent wave of artificial intelligence (AI) within drug discovery has heavily impacted the fields of de novo design, synthesis planning, and bioactivity prediction, to name a few.^{8,177} 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.^{8,177} Computer aided synthesis planning (CASP) has long been investigated as a means for predicting how to make a given compound.^{103,104,106} However, despite recent progress in the field,^{9,123,127,135,138,146,165} synthetic planning tools based on neural network classifiers have failed to recognize reactions that are infrequently used or rare, due to the heavily biased data sets available.^{9,95} As such, CASP tools have not yet focused on the synthesis of ring systems, the reactions for which often fall within the noise of the datasets in question when the entire range of transformations are considered. Furthermore, current tools lack the ability to target specific sets of transformations. These are useful when the general CASP tool fails, or a chemist wants to target a specific set of transformations. This is particularly useful when conducting an interactive or stepwise search for synthetic routes. One such instance is the ability to deconstruct ring systems in novel ways, which 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.¹²⁷

Ring systems are key scaffold components in medicinal chemistry and are fundamental motifs to a number of drugs on the market today.¹⁹⁵ They vary greatly in nature; ring systems can be saturated, unsaturated, polycyclic and range in size 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.^{196,197} Although coupling reactions enable the synthesis of a wide range of structures, they are limited by the commercial availability of 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.^{195,198} They can directly interact with a protein target, such as in the well-defined example of the hinge binding motifs for kinase targets.¹⁹⁹ 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.¹⁹⁵ 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 et al.²⁰⁰ These aim to enrich structure–activity relationship information, explore the chemical space of ring systems, and find motifs relevant for

use in medicinal chemistry.^{198,201–203} However, as of yet the synthetic accessibility of ring systems remains poorly explored.

Furthermore, the neural networks upon which CASP tools are built^{9,137,138} 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.^{200,204} 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 pathways 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.

4.1.1 Overview of Template Based Retrosynthesis

Template based retrosynthesis has its roots in the first approaches to computer assisted synthesis planning (CASP) by Corey.¹⁰³ More recently, the works of Segler and Waller, and Coley et al. have pioneered this approach, coupling it with neural networks.^{9,138} This approach has been utilized in “Ring Breaker” for the prediction of ring formations in the retrosynthetic direction, starting from reaction data, namely, Reaxys¹⁷⁸ and the publicly available U.S. Patent Office extracts (USPTO)⁹² which describe the relationship between reactants, reagents, and products. The reaction center is extracted, the so-called reaction template or rule. The reaction template describes the atoms and bonds changing in the reaction and captures all changes occurring one bond away from the atoms involved in the reaction center. This can be viewed as a generalized form of the overall reaction from which the template was extracted.

The extracted templates can be applied to a given compound to enumerate a set of outcomes/reactants that return the retrosynthetic options available. For a template to be applicable and generate a set of outcomes/reactants, there must be a substructure match between the template and queried compound. As not all templates have a substructure match with the queried compound, not all templates can be applied to any given compound. Therefore, it is necessary to predict which templates can be applied to prevent an exhaustive search across all templates. The problem of predicting retrosynthetic pathways is formally described as a tree search, whereby the templates are applied recursively to generate a tree of possible options. In practice, the number of templates applied is the top 50 predicted by the model. As mentioned previously, not all templates will yield a set of reactants; therefore, the number of options at each step will be less than the 50 predicted. An additional constraint is applied such that when the cumulative sum of the probabilities associated with each template reaches a cutoff threshold, no further templates will be applied; therefore the number of options will be less than 50 in this instance.⁹ This limits the computational

expense of enumerating all possible options given by the extracted templates. The tree is then searched to find the most efficient synthetic pathways.^{9,122,123,162}

This study is focused on how prioritization of templates, specifically those that are associated with the formation of rings, can be used to predict synthesis of ring systems. Having determined which templates have been used in the context of forming rings, we examine three approaches for prioritization and compare them: (1) prediction using a model trained on all templates extracted from the reaction data, including those that do not form rings, to determine whether such a model is able to prioritize ring formations, termed the standard model; (2) applying a filter at inference to the standard model trained on all templates to select only those that correspond to ring formations, termed the filtered model; (3) a model trained on only the templates that have been used in the context of forming rings, termed “Ring Breaker”. In each instance, the model architectures are held constant, and only the inference method or training data are varied. For each of the models we train on either the templates obtained from the Reaxys or USPTO datasets. We then examine and compare the predictive capability of selected models on a set of commonly used ring formations, fragments from the ZINC database,²⁰⁵ approved drugs from DrugBank,²⁰⁶ and two compounds from the “rings of the future” set.

4.2 Results and Discussion

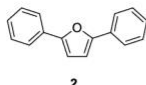
4.2.1 Brute Force Application, Filtering, and Prioritization

To exemplify the predictions of “Ring Breaker”, we retrieved examples from the organic chemistry literature for which commonly used ring-forming reactions were used (Figure 12). We assessed each substrate against each of the three methods of prioritization mentioned previously, for both the USPTO and Reaxys datasets (Figure 13). As mentioned in the overview of template-based retrosynthesis planning, not all templates are applicable for a given compound. Therefore, using brute force application, we exhaustively applied all the templates underlying each model to determine the maximum number of templates that were applicable. Figure 13a shows that the results obtained are as expected; the standard and filtered models have the same underlying set of templates and have not been constrained to those corresponding to ring formations. Therefore, both models have a higher number of applicable templates than the “Ring Breaker” model, for which only templates that have been used in the context of ring formations were considered. The same pattern is observed for the Reaxys dataset as shown in Figure 13c. More importantly, the maximum number of applicable templates that correspond to ring formations is a much smaller fraction of the total amount as shown in Figure 13b and Figure 13d.

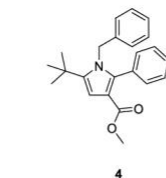
Paal-Knorr-furan synthesis



J. Org. Chem. 2003, 68, 5392-5394

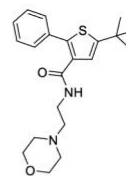


Paal-Knorr-pyrrole synthesis



Eur. J. Org. Chem. 2005, 5277-5288

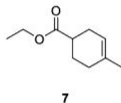
Paal-Knorr-thiophene synthesis



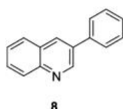
Eur. J. Org. Chem. 2005, 5277-5288



Diels-Alder reaction

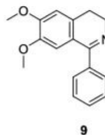


Org. Lett. 2006, 8, 12, 2487-2489



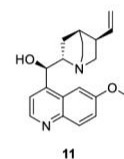
J. Org. Chem. 2016, 81, 656376572

Bischler-Napieralski reaction

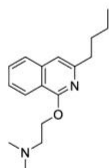


Org. Lett. 2008, 10, 16, 3485-3488

Skraup reaction



Pomeranz-Fritsch reaction



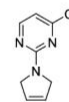
Robinson annulation



Ring-closing metathesis

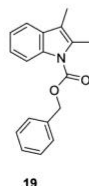
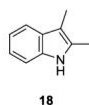


Chem. Rev. 2004, 104, 5, 2199-2238



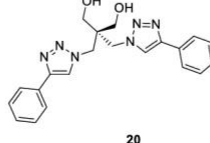
J. Org. Chem. 2008, 73, 7417-7419

Fischer indole synthesis



Org. Lett. 2009, 11, 23, 5454-5456

Azide-alkyne Huisgen cycloaddition



Angew. Chem. Int. Ed. 2002, 41, 2596-2599

Figure 12: Substrates from the literature for which named ring-forming reactions were known (substrates for which predictions failed are shown in Figure 15, Figure 16, Figure 17). These were used to compare the performance of “Ring Breaker” with our standard retrosynthetic model. The substrates 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 standard 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.

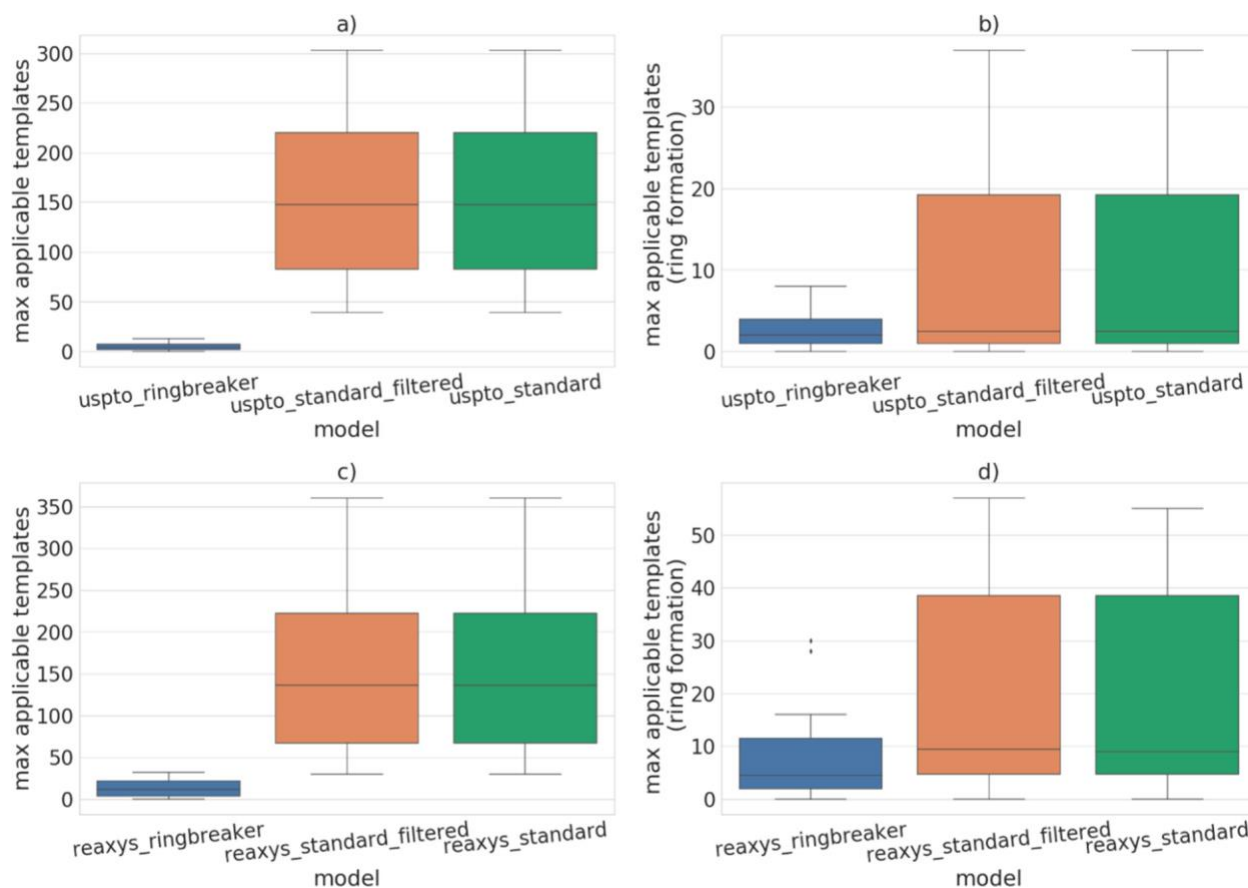


Figure 13: Application of all templates underlying each of the models, “Ring Breaker”, standard, and filtered models, to the substrates in Figure 12. (a) For the USPTO dataset, the maximum number of templates that can be applied from the standard and filtered models far exceeds that from the “Ring Breaker” model. This is because of the larger number of templates available to the standard and filtered models. The number of templates that can be applied in the top 50 predictions (b), which also correspond to ring formations, is a fraction of those that can be applied by brute force in silico. The same is observed for models built on the Reaxys dataset as shown in (c) and (d). The discrepancy between the top 50 predictions of the “Ring Breaker” model and standard and filtered models is described in Discrepancy between predictive and exhaustive search.

Comparing the three prioritization approaches across the substrates in Figure 12, we found that the “Ring Breaker” model consistently predicted the first applicable ring formation with a lower rank than both the standard and filtered models across the Reaxys and USPTO datasets (Figure 14a and Figure 14c). Notably, there appears to be missing values for the filtered model on the Reaxys dataset (Figure 14c). This is because the values lie off the range covered by the axis. The lowest ranked prediction for the filtered model on the Reaxys dataset was 133, and the median prediction rank was 736. This highlights an interesting discrepancy between the standard and filtered models trained on the Reaxys dataset. The difference can be explained by considering that the standard model was trained on all templates, and the filtered model is obtained by applying a filter to the standard model, to leave only templates which in the reaction dataset had been recorded as contributing to a ring formation. In the case of the standard model, the templates that correspond to ring forming reactions in the reaction dataset cannot be prioritized by the model. Therefore, once the predictions from the standard model are filtered, none of the remaining templates that

correspond to ring formations in the underlying reaction dataset are applicable within the top 50 predictions (Figure 14d). The standard model is therefore prioritizing and applying templates that can be used in the context of ring formations in silico but have never been recorded as a contributing to a ring formation in the underlying reaction dataset. Thus, this explains why the standard model can predict ring forming reactions in the top 50 whereas the filtered model cannot (Figure 14d). Comparatively, the “Ring Breaker” model is able to prioritize and apply more templates in the top 50 predictions than either the standard or filtered models as applied to the substrates in Figure 12 (Figure 14b and Figure 14d).

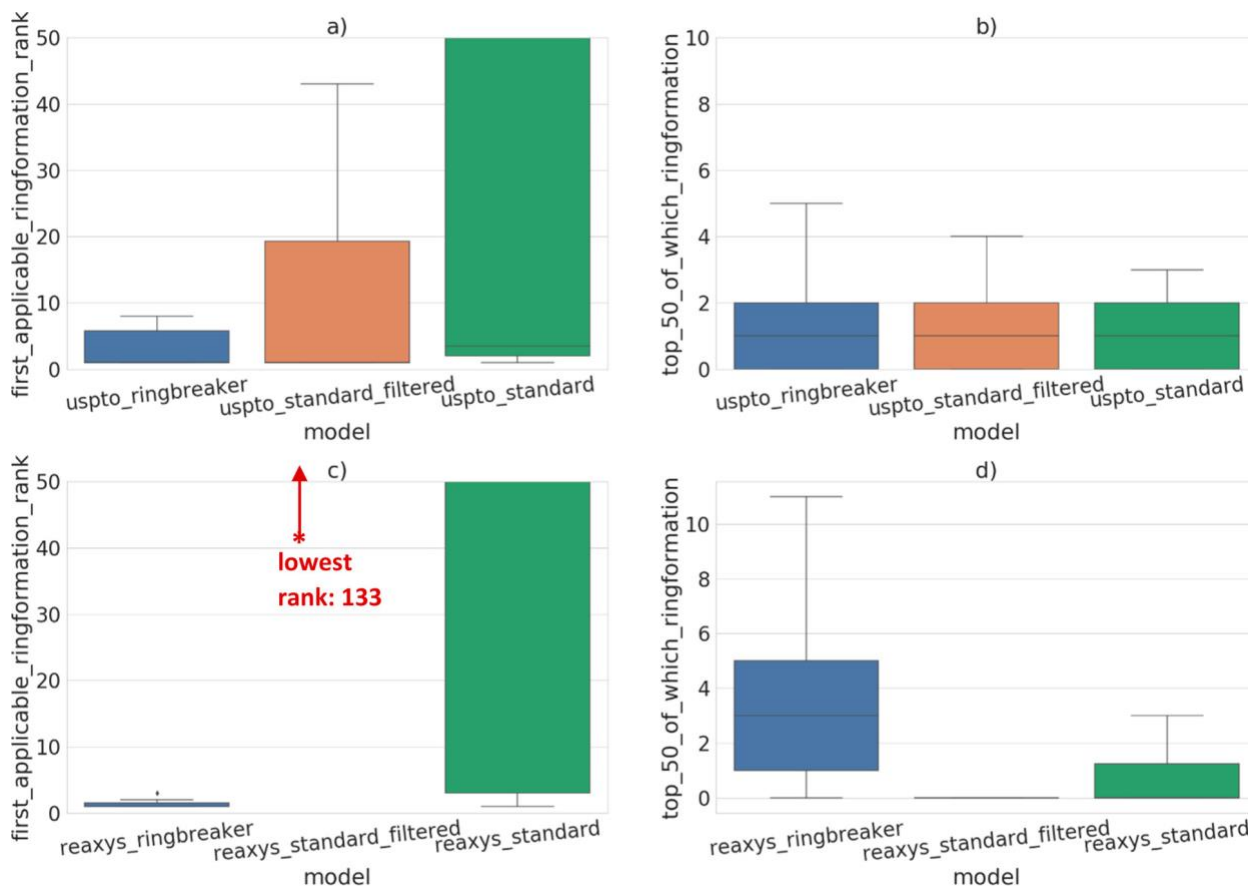


Figure 14: The “Ring Breaker” model consistently predicts ring formations with a lower rank than the standard and filtered models across both the USPTO and Reaxys datasets, as shown in (a) and (c). Additionally, the “Ring Breaker” model is able to predict more templates in the top 50 predictions that result in ring formations, than both the standard and filtered models, as shown in (b) and (d). The standard model is not able to prioritize templates that have been used in ring forming reactions as effectively as “Ring Breaker”, as demonstrated by filtering the standard model. Thus, for the Reaxys dataset, the first applicable template for the filtering approach is ranked as 133, and the median rank is 736, even after filtering all templates that have not been recorded as forming a ring in the reaction dataset.

The discrepancy observed in Figure 14c and Figure 14d can be further explained by considering a simplified case, as shown in Table 4. This simplified case shows that there were three ring forming templates as found from the reaction dataset. However, the total number of templates that result in a ring formation is five. This is context dependent, as a template that is not formally a ring formation or has not been observed as forming a ring in the underlying reaction dataset can be

applied and form a ring in silico. Thus, the number of templates corresponding to a ring formation in silico varies with the substrate, while the number of ring forming templates found from the reaction dataset is fixed. By predicting the ranks of the templates, we can see that for the case of the standard model, there are five templates that are applicable in silico. Yet only three of them correspond to a ring forming template; the others are context dependent. Of the three that are ring forming templates, only two are applicable for the given case. This leads to a discrepancy between the number of templates that are applicable in silico and predicted by the standard model (varies depending on substrate), and the number of templates that correspond to ring formations as found from the reaction data set (fixed).

Table 4: Simplified Example to Explain the Discrepancy between the Standard Model and Its Subsequent Filtering.^{a)}

Template	Ring-forming? (as found from the reaction dataset)	Ring-forming? (as found from in silico application)	“standard” rank	“filtered” rank	“applicability filtered” rank
T1	No	No	1	-	
T2	No	Yes	2	-	1
T3	No	Yes	3	-	2
T4	No	No	4	-	
T5	No	No	5	-	
T6	Yes	Yes	6	1	3
...					
T51	No	Yes	7	-	4
T52	Yes	No	8	-	
T53	Yes	Yes	9	2	5
Total Applicable	3	5	5	2	5

^{a)} The number of templates corresponding to ring formations is fixed, as they originate from the underlying reaction dataset. Many more templates are applicable in silico, and these are not considered by the filtering process, as the corresponding templates receive a value of zero. Thus, the filtered model is not aware of their applicability.

Therefore, for any given substrate, the number of templates that can be applied and happen to correspond to ring formations because of context dependency is greater than the number of templates that have been used for ring formations in the underlying reaction dataset, of which not all are applicable. Thus, the filtered model will not always have at least as many ring-forming templates in the top-50 predictions as the standard model. For the filtered model to have at least as many ring-forming templates in the top-50 predictions as the standard model, the model would have to consider the applicability of templates.

It can be argued that increasing the number of predictions beyond the top 50 templates will increase the number of predicted templates that encode ring formations which can be applied. However, it is clear from Figure 14 and the previous explanation that in the case of the standard model applied to the substrates in Figure 12, this would have to be extended to at least the top 150 predictions. This extension would enable at least one template that has been recorded as a ring formation in the reaction data to be applied. When the standard model is considered as part of the tree search algorithm that searches for and selects synthetic routes, rather than as a standalone model, this

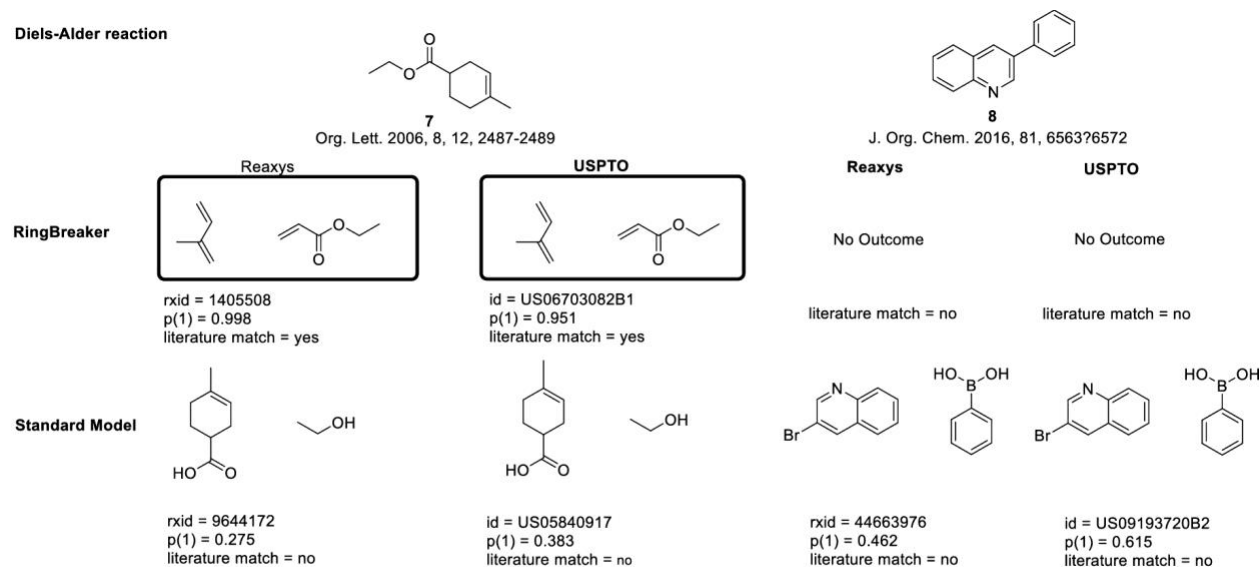
serves to increase the search breadth of the subsequent tree search. To this end, the computational expense associated with enumerating the tree to this extent must be balanced between the speed of prediction and the number of retrosynthetic options the model suggests. In the case of the prediction of ring formations, “Ring Breaker” balances both these criteria by predicting ring forming reactions with both a lower rank and higher number of applicable predictions in the top 50, as compared to the standard model or its subsequent filtering.

In addition, we found that “Ring Breaker” trained on the Reaxys dataset outperformed that trained on the USPTO dataset across the substrates in Figure 12 (Figure 14b and Figure 14d). The difference in performance may arise from the differing template space covered by the models as we have found in our previous studies.¹³⁶ This is evidenced by the greater number of applicable templates that correspond to ring formations in the Reaxys dataset as found by exhaustive application (Figure 13b and Figure 13d). Therefore, both the type and quantity of templates available to the Reaxys model differ from the USPTO model, as the two models did not differ in architecture but in the template and training set used. The diversity of the products in the training set for each of the models could additionally influence how well the model is able to learn; however, this has not been examined in this study.

4.2.2 Diels–Alder and Bischler–Napieralski

The Diels–Alder reaction is one of the most well-known ring-forming reactions and commonplace in an undergraduate chemist’s education. However, the standard model fails to predict the template leading to the correct set of reactants, as shown in Figure 15, for substrates 7 and 8. The hetero Diels–Alder approach cannot be predicted by either the USPTO or Reaxys models for the synthesis of quinolines (Figure 15, substrate 8). However, when applied to the substituted cyclohexene (Figure 15, substrate 7), the “Ring Breaker” models are able to successfully identify the diene and dienophile used in the literature with a high probability for the USPTO (template rank 1, $p = 0.951$) and Reaxys (template rank 1, $p = 0.998$) “Ring Breaker” models.²⁰⁷ In addition, the synthesis of dihydroisoquinolines (Figure 15, substrate 9) via the Bischler–Napieralski reaction was successfully predicted by both the USPTO (template rank 1, $p = 0.997$) and Reaxys (template rank 1, $p = 0.976$) “Ring Breaker” models, leading to β -arylethylamide, as reported in the literature (Figure 15, substrate 9).²⁰⁸ In the case of both the Diels–Alder and Bischler–Napieralski reactions, the standard model consistently predicts alternative 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 the increased performance of “Ring Breaker” over the standard model is the prediction of the dihydroisoquinoline (Figure 15, substrate 9) using the USPTO standard model (template rank 1, $p = 0.966$), where the standard model predicts the ring formation with a high probability.

Diels-Alder reaction



Bischler-Napieralski reaction

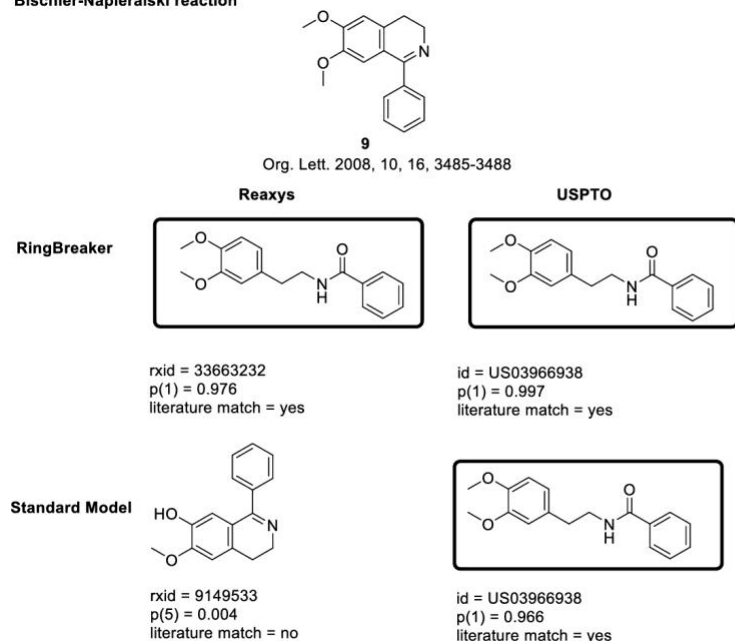
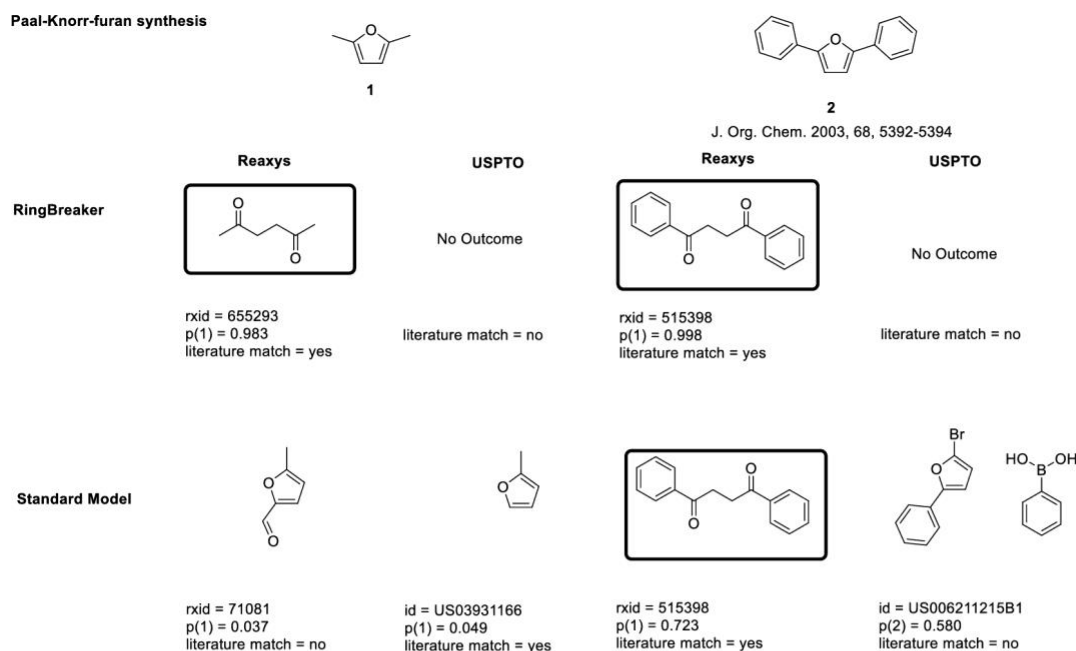


Figure 15: Predictions for the Diels–Alder and Bischler–Napieralski reaction using “Ring Breaker” and the standard model on compounds obtained from the literature employing commonly used ring formations. For each prediction, the patent identifier or Reaxys identifier corresponding to the template used 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 data set.

4.2.3 Paal–Knorr

The Paal–Knorr series of ring synthesis can be used to provide access to substituted furans,²⁰⁹ pyrroles, and thiophenes (Figure 16 and Figure 17).²¹⁰ Its versatility and structural similarity between components make 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 16 shows that the disconnections predicted by the model are dependent on the dataset. Both the Reaxys and USPTO datasets contain complementary templates, whereby the “Ring Breaker” 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 “Ring Breaker” is not able to predict a disconnection for substrates 1 and 2 in Figure 16, whereas the “Ring Breaker” model trained on the Reaxys data predicts the literature disconnection for both substrate 1 (template rank 1, $p = 0.983$) and substrate 2 (template rank 1, $p = 0.998$) in Figure 16 with a high probability.²¹⁰ The case of pyrrole synthesis further highlights an interesting problem, whereby the correct disconnection can be predicted by the USPTO “Ring Breaker” model for a simplified ring system (Figure 16, substrate 3). However, when the molecular complexity around the ring system was increased by replacement of the methyl groups with phenyl groups (Figure 16, substrate 4),²¹⁰ the model failed to respond to the change and was not able to predict an outcome. On the other hand, the “Ring Breaker” model trained on Reaxys was able to correctly identify the ring system (Figure 16, substrate 3) and predict the retrosynthetic disconnection reported in the literature.²¹⁰ 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 while 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 and is further exemplified in Discrepancy between predictive and exhaustive search. In contrast to “Ring Breaker”, the standard model is only able to predict the correct set of precursors for compound 2, which uses the Paal–Knorr furan synthesis.

Paal-Knorr-furan synthesis



Paal-Knorr-pyrrole synthesis

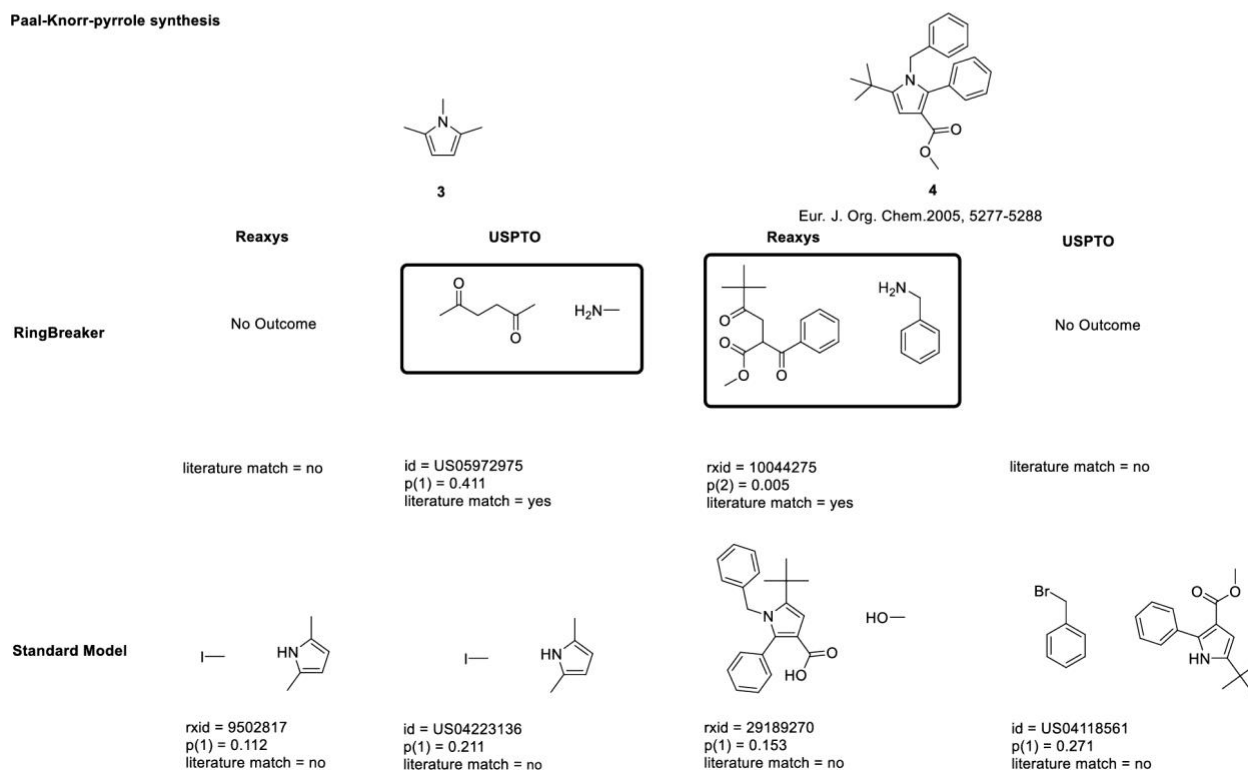


Figure 16: Predictions for the Paal-Knorr furan and pyrrole synthesis using “Ring Breaker” and the standard model on compounds obtained from the literature employing commonly used ring formations. For each prediction, the patent identifier or Reaxys identifier corresponding to the template used has been 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 or is the correct one as identified by expert chemists; the notation is equivalent

for consistency. The literature reference from which the compound was obtained is given for each example where appropriate. In cases where the literature reference has not been given, the compound was purposefully simplified to determine if the correct disconnection could be predicted. 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.

Paal-Knorr-thiophene synthesis

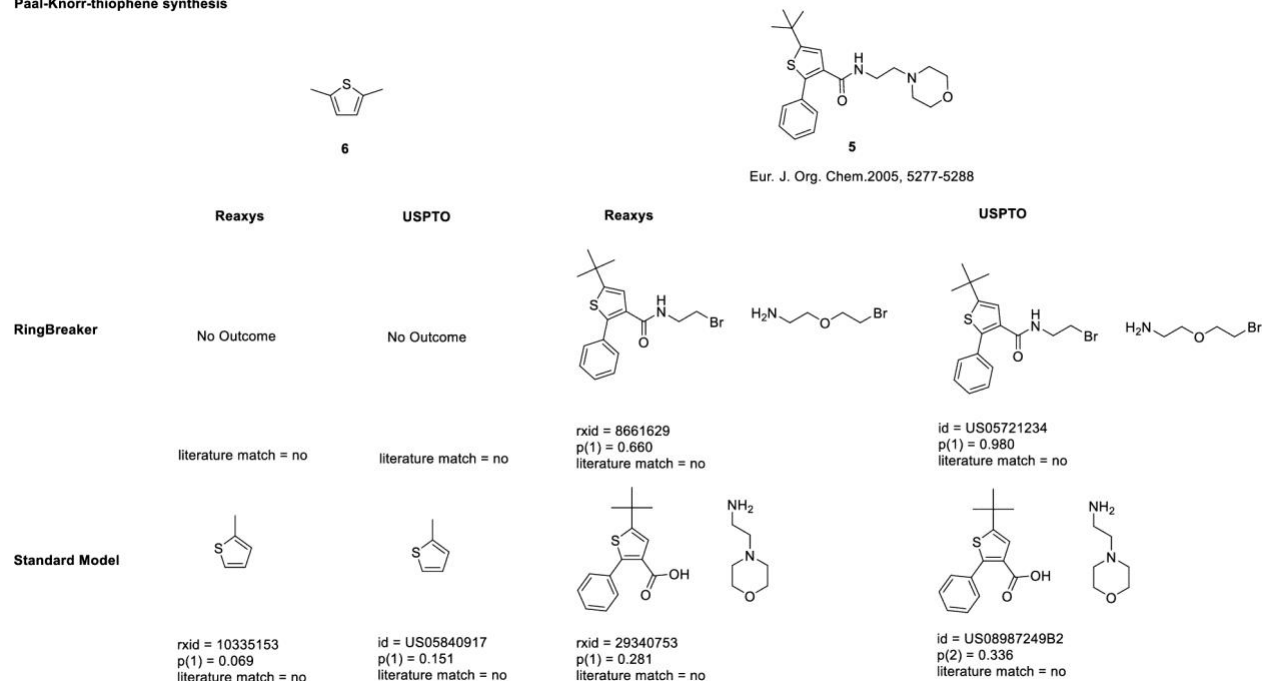


Figure 17: Predictions for the Paal-Knorr thiophene synthesis using “Ring Breaker” and the standard model on compounds obtained from the literature employing commonly used ring formations. For each prediction, the patent identifier or Reaxys identifier corresponding to the template used has been given 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 or is the correct one as identified by expert chemists; the notation is equivalent for consistency. The literature reference from which the compound was obtained is given for each example where appropriate. In cases where the literature reference has not been given, the compound was purposefully simplified to determine if the correct disconnection could be predicted. 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 any template matching the molecular environment in compound 6 for the desired outcome.

In the case of thiophene synthesis, the model cannot identify a suitable template for the simple ring system (Figure 17, substrate 6), regardless of the dataset used. However, in the more complex case (Figure 17, substrate 5), it focuses its efforts on the morpholine ring, predicting the shown disconnection (Figure 17, substrate 5) with a high probability. While this does not indicate that the models cannot predict thiophene formation, it alludes to the fact that these templates may be under-represented in the underlying dataset.

4.2.4 Prediction of ZINC Fragments and DrugBank

We performed a one-step retrosynthetic analysis considering only the top 50 predictions to focus on the ring-forming step required to synthesize a range of ring containing subsets from the ZINC database (Figure 18).²⁰⁵ Examining a range of ring systems, from the most commonly occurring (in >100K substances) to the rarest (in <1K substances), we found that “Ring Breaker” was able to predict ring formations for more substrates than that of the standard models across all subsets examined, regardless of the dataset used. The reason for this may be 2-fold. First “Ring Breaker” is exclusively limited to ring formations or reactions for which ring formations may result depending on context, 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 without distractions from non-ring forming templates. Therefore, it can prioritize the ring forming templates on which it is trained better than the standard model. This was found by filtering the predictions from the standard model for ring formations, as demonstrated in Figure 14. This is in comparison to the standard 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 18). This is in contrast to our previous observations, where we reported that the ability to generate synthetic routes for the standard model did not depend on the training dataset.¹³⁶ We have now determined that for the domain specific case of ring formations there is a clear effect arising from the training set used in the template space, attributed to the number and diversity of the samples available to the network for training. The difference in performance between the “Ring Breaker” models trained on the USPTO and Reaxys dataset can in part be attributed to the prevalence of some fragments in the Reaxys training set (Overlap of ZINC fragment sets with the training sets).

The performance of the model on ring systems classed as “rare” in the ZINC database is surprising (Figure 18). 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 19, with their corresponding patent precedent, which refers to the patent containing the reaction from which the predicted template was extracted. While 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.

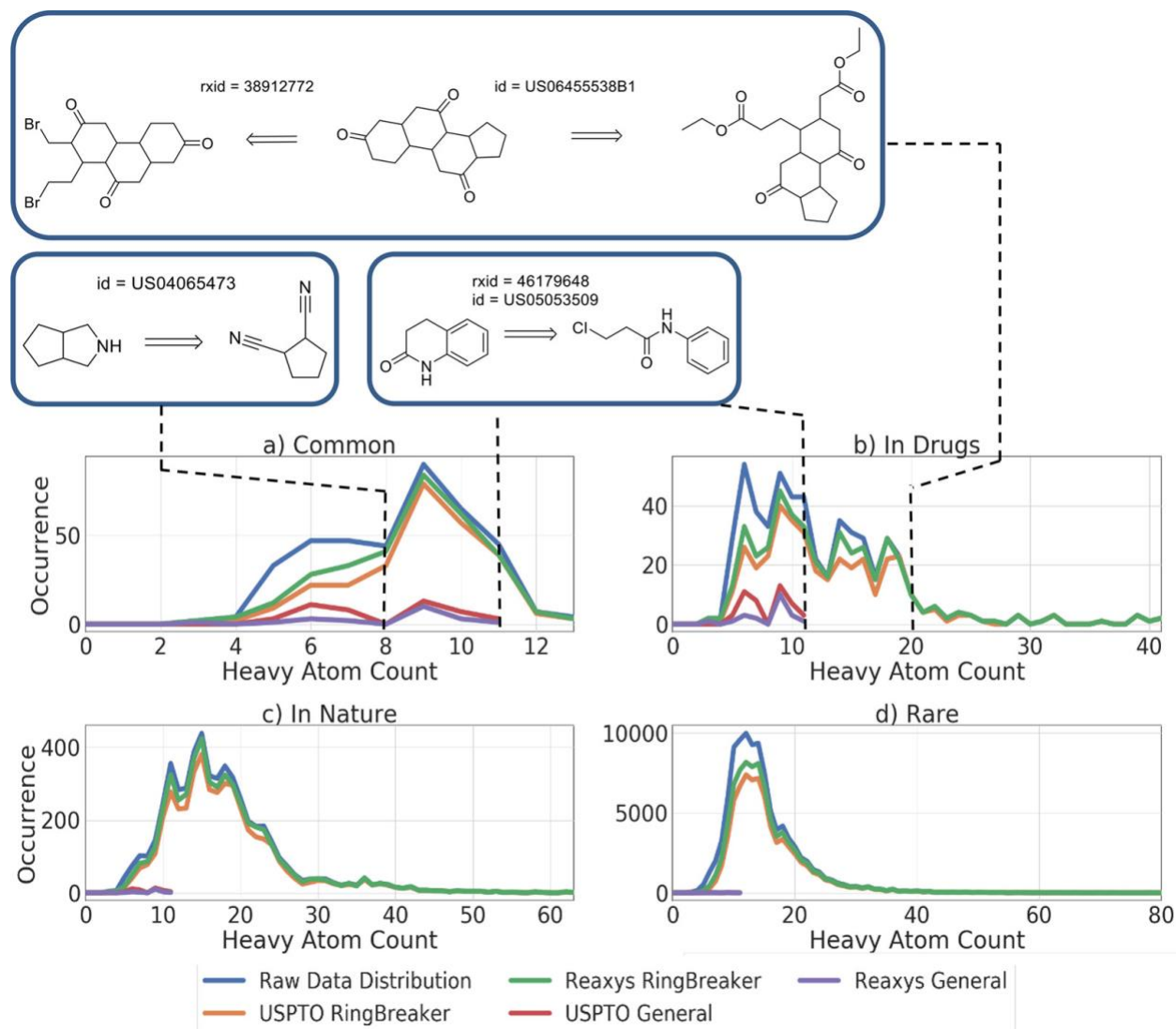
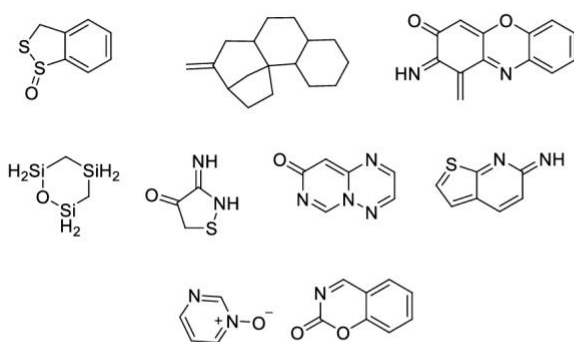
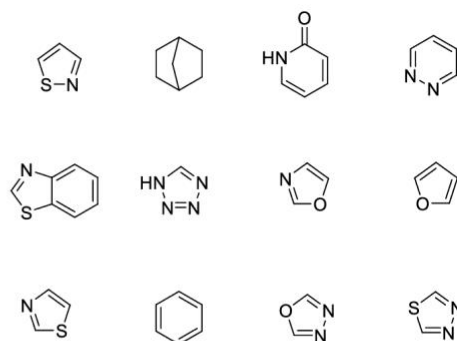


Figure 18: Examining the top 50 predictions of "Ring Breaker" to assess the synthetic accessibility of a range of ring systems obtained from the ZINC database. The top 50 predictions were considered, and templates were applied. If the predicted templates generated a set of outcomes, the prediction was successful and the ring system synthetically accessible *in silico*. The heavy atom count for compounds in each ZINC subset is plotted against the number of compounds corresponding to the heavy atom count. The raw dataset distribution is the distribution of compounds with a given heavy atom count as found for each ZINC subset. The distributions from each model show the distribution of compounds remaining after prediction with a given heavy atom count. The difference between the raw distribution and that of the models corresponds to compounds that could not be predicted. The subsets correspond to (a) common (occurring in greater than 100K substances), (b) present in drugs, (c) present in nature, and (d) rare (occurring in under 1K substances). The "Ring Breaker" and standard models were compared for each subset and each training set, Reaxys and USPTO. Only a one-step retrosynthesis was predicted for each ZINC 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 data set. "Ring Breaker" exhibits the best performance for ring systems between 5 and 20 heavy atoms in size, and the predictions follow the raw distribution with a slight divergence as not all ring systems can be predicted.

a) Zinc Rare subset - No Prediction



b) Zinc Frequent subset - No Prediction



c) Zinc Rare subset - Predictions

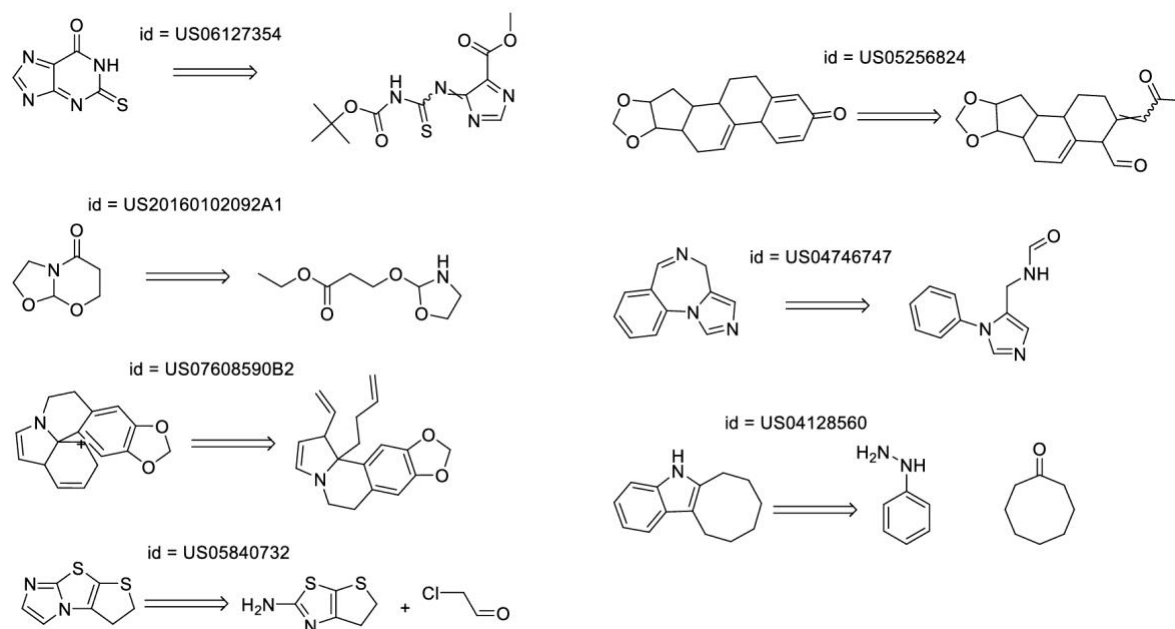


Figure 19: (a, b) Examples of ring fragments from the ZINC database that could not be predicted by “Ring Breaker” model trained on the USPTO or Reaxys data. The templates must be specific enough to yield a substructure match to the compound they are applied to and produce feasible reactants while being 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 using the USPTO “Ring Breaker” are shown for a selection of compounds in the rare subset. For each prediction, the patent identifier 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 19b), we have previously observed that a disconnection could be predicted from the substituted ring system (Figure 16, substrates 1 and 2). 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 be 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 corresponding templates are 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.

Given that the “Ring Breaker” models cannot always predict the synthesis of unsubstituted fragments owing to template specificity for the unsubstituted furan and benzene (Figure 19b), we additionally applied the “Ring Breaker” models to a set of 2039 approved drugs obtained from the DrugBank database.²⁰⁶ By doing so, we exemplify that in the case of complete structures, and not only unsubstituted fragments obtained from ZINC, “Ring Breaker” is able to predict ring forming steps for their synthesis. Figure 20 shows that the “Ring Breaker” model trained on the Reaxys dataset consistently predicts more applicable templates in the top 50 predictions and with a lower rank than the model trained on the USPTO dataset. This is in line with that observed previously as shown in Figure 14. Additionally, as expected, the number of predictions in the top 50 increases with the number of rings in the substrate (Figure 20c). Again, we observed that the “Ring Breaker” model trained on the Reaxys dataset suggests more ring formations in the top 50 predictions than that trained on the USPTO dataset. The exception is for compounds containing 10 rings, in which case the median number of ring formations suggested is the same for model trained on both datasets; however the number of suggestions exhibits a wider range.

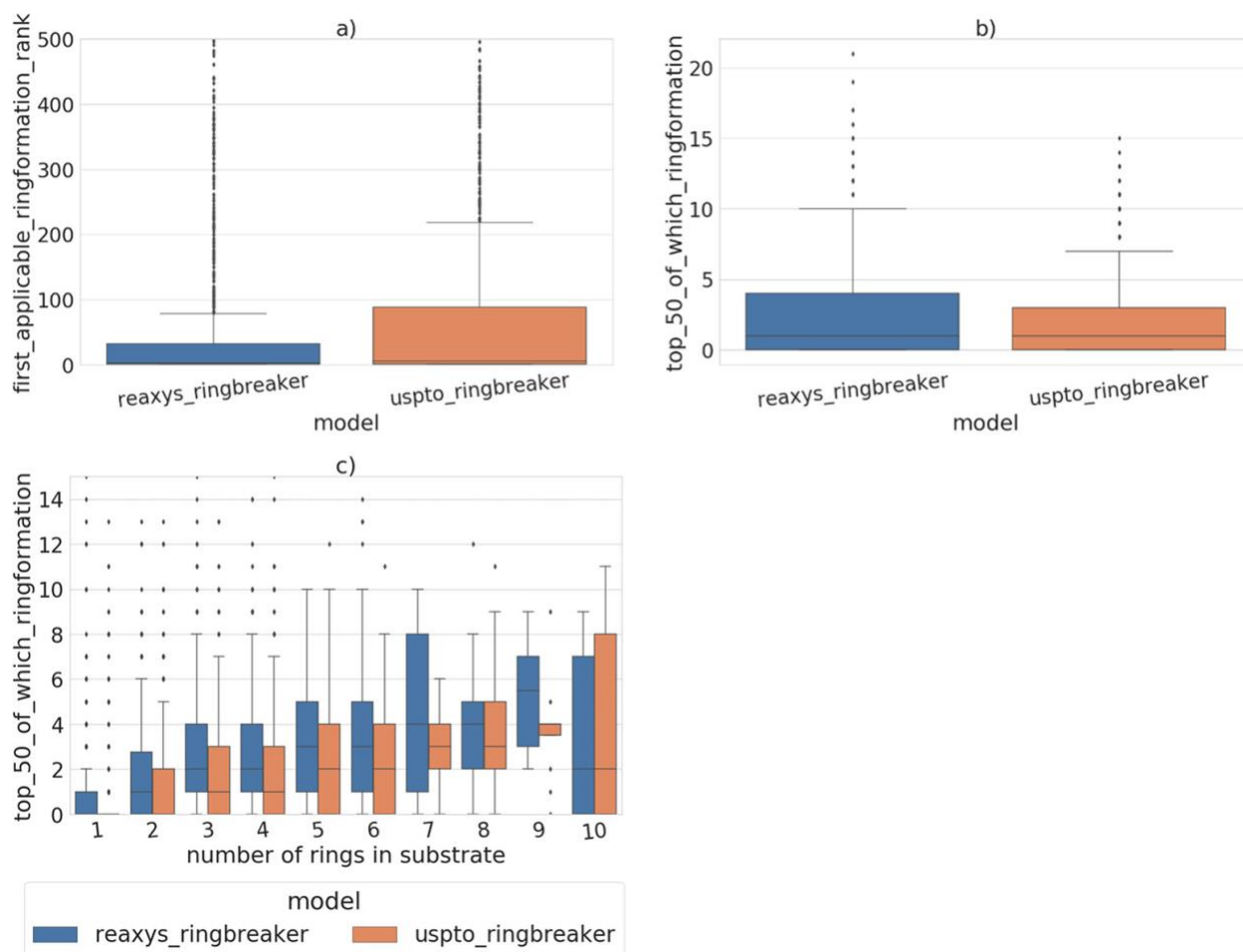


Figure 20: Statistics for the “Ring Breaker” models applied to 2039 approved drugs from DrugBank. (a) The “Ring Breaker” model trained on the Reaxys dataset predicts ring forming reactions with a lower rank than the model trained on the USPTO dataset. (b) The “Ring Breaker” model trained on the Reaxys dataset exhibits a wider range of templates predicted in the top 50 predictions than that trained on the USPTO dataset. However, the median number of ring formations is the same. (c) The “Ring Breaker” models predict more ring formations in the top 50 predictions as the number of rings in the substrate increases, as one would expect. The Reaxys model consistently predicts more ring formations in the top 50 predictions than the USPTO model. An exception is observed for substrates containing 10 rings, whereby the median values are the same, but the number of predictions by USPTO model exhibits a wider range.

4.2.5 Accessing Virtual Fragments: “Rings of the Future”

Since the exhaustive computational enumeration of heteroaromatic ring systems first described by Pitt et al.,²⁰⁰ several articles have detailed the enumeration of ring systems,^{198,201–203} yet few follow-up articles 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”.²⁰⁰ Having trained the model on patent data up to 2016, we selected two novel ring systems from the literature for which the syntheses were reported in 2016²⁰⁴ and ensured that they were not present in the training 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 21). 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 look up reaction conditions and experimental procedures.

RSC Adv., 2016, 6, 22777-22780

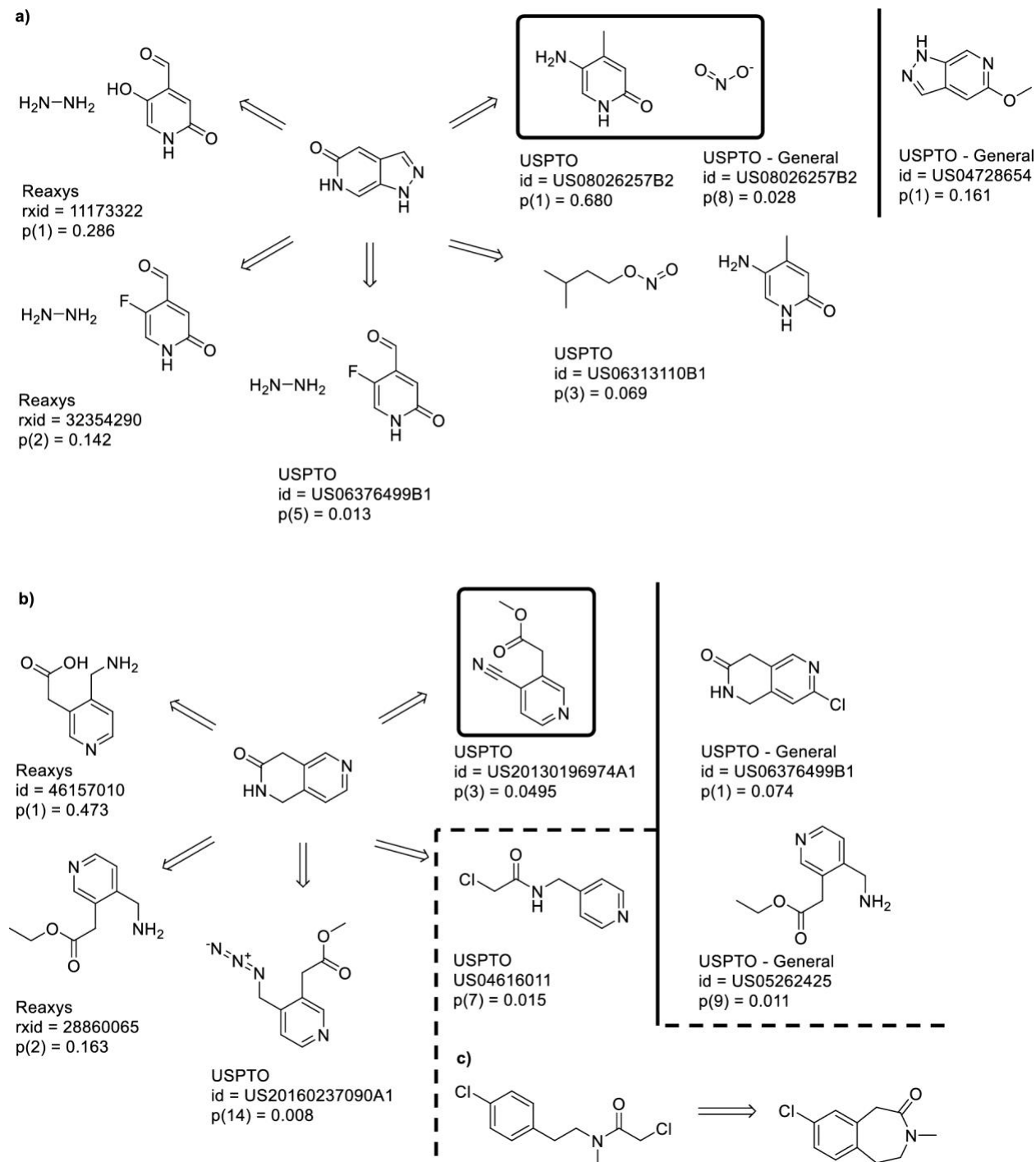


Figure 21: Performance of “Ring Breaker” on two “Ring of the Future” compounds that were synthesized in 2016, for which the predicted disconnection matches that described in the literature synthesis (highlighted in the green boxes). The predictions are

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 identifier or Reaxys identifier 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. (c) shows one of the reactions that are precedented for a benzene ring system but is applied in a different context, on the pyridine. The N atom of the pyridine is outside the template’s scope, and so the template can be applied in this context *in silico*.

The standard model predicts some of the ring forming reactions that appear in the “Ring Breaker” predictions with a lower rank, as would be expected due to the greater number of transformations on which the standard model is trained. The additional reactions on which the standard model is trained offer alternative approaches to access ring systems that do not necessarily involve ring formations, such as functionalization of the core scaffold (Figure 21). These approaches can be used for known ring systems; however, for novel systems that have not been observed before, the ring system must be built from basic building blocks. The prediction ranked number seven for the ring system in Figure 21b is based on a template extracted from the reaction shown in Figure 21c. The template predicted by the model is precedented for a benzene ring and applied on a pyridine for which the nitrogen atom falls outside the templates scope; thus the template can be applied in a different context *in silico*.

It is inherent to the template-based approach that novel chemistries are not able to be predicted, as the precedent arises from extraction of a specific subgraph from the underlying dataset. Therefore, the model learns to predict existing chemistry exhibiting a similar subgraph to the queried compound.

4.2.6 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 5). 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 that proceed through ring formations. To overcome this problem, the “Ring Breaker” model can be viewed as a specialist that can be consulted at various stages of the tree search to yield routes that proceed through ring formations. In addition, the model may be used as a stand-alone tool to target a specific set of transformations. This is particularly useful when building synthetic trees interactively in a stepwise manner. In Figure 22 we demonstrate one such use case, where the standard model fails to predict a disconnection from the bicyclic ring system. To counteract these problems, we apply the “Ring Breaker” model. This yields a prediction from which the search for a synthetic route can be continued. In addition to using “Ring Breaker” to kick-start the model from a point at which it becomes stuck on a ring system, the predictions can be used to supplement

the standard model throughout the search for a synthetic route. Thereby, the possibilities for routes going through ring disconnections could increase, leading to more convergent synthetic routes, as opposed to a linear series of functionalization. Incorporation of “Ring Breaker” into general route planning tools in each of the two ways described could help to maximize the chances of finding a more efficient route. This in turn could be used by synthetic chemists as an idea generator to discover existing disconnections but applied in a different context.

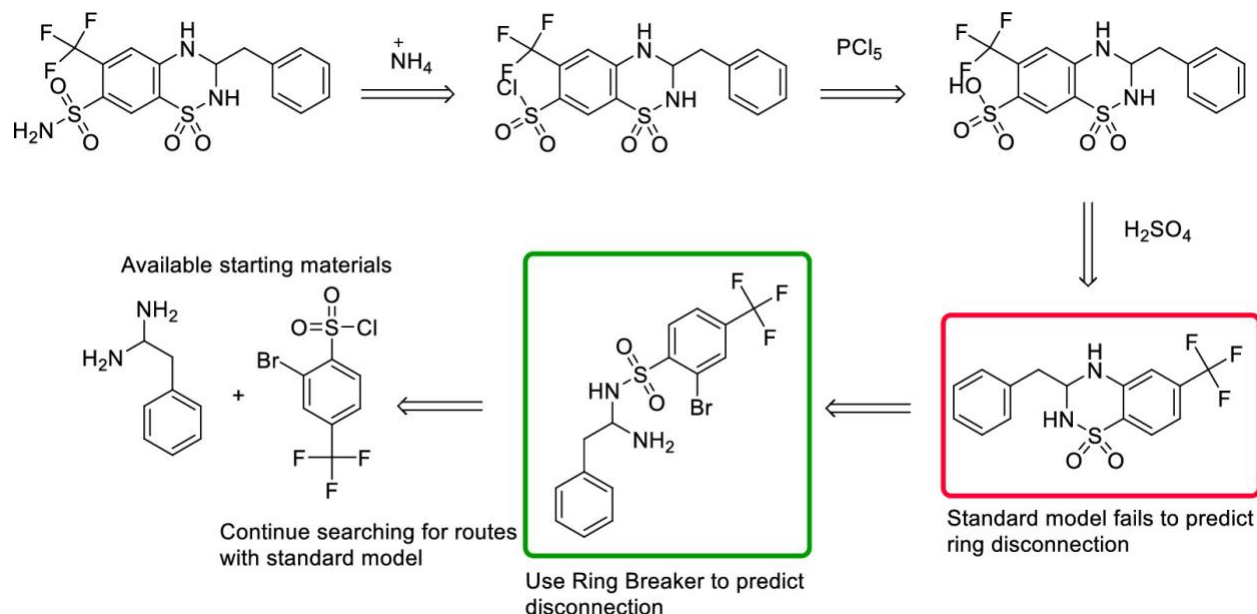


Figure 22: Exemplary synthesis demonstrating the use of “Ring Breaker” trained on the USPTO dataset to augment synthesis planning tools. “Ring Breaker” can be used to guide retrosynthetic tree search to explore areas exploiting ring-forming strategies. In some cases, as that shown above, the standard model fails to predict the disconnection necessary to break apart the ring system; thus it becomes stuck. “Ring Breaker” overcomes this bottleneck by targeting the ring system to suggest an appropriate disconnection, thus enabling the standard model to continue the retrosynthetic search to a set of building blocks. The building blocks shown are considered “in stock” when using the ACD catalog.

This methodology, using a domain specific model in conjunction with a standard model, can be extended to other areas of synthetic chemistry in which the data are limited and domain specific knowledge (i.e., a specialist) is required. Furthermore, the data on which the methodology relies could be augmented by integrating different data sources in order to increase the coverage of ring forming reactions.

4.3 Conclusions

We have developed a methodology for proposing syntheses and assessing the synthetic accessibility of ring systems using a specialized ring-forming neural network called “Ring Breaker”; remarkably, our “Ring Breaker” can predict more templates in the top 50 predictions than the standard and filtered models. It can be used 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 data sets 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” outperforms the predictive capability of the standard and filtered models. This is because the standard model fails to prioritize ring forming reactions, as found by applying a filter to the standard model to consider only templates that were used in a ring formation in the underlying reaction dataset.

Although the models consistently underperform for ring systems smaller than five heavy atoms, 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.²⁰⁴ 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. While suitable templates describing the reaction are suggested, they cannot be applied as they do not share an exact substructure match to the query compound. To verify that the “Ring Breaker” model is predictive on full structures and not only fragments, we applied the model to 2039 approved drugs from DrugBank. The “Ring Breaker” model trained on the Reaxys dataset was able to predict more ring formations in the top 50 predictions and with a lower rank than the corresponding USPTO model.

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.¹⁹⁵ 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. The methodology could also be extended to combine various data sources to increase domain specific coverage, in addition to data augmentation techniques published at the time of writing this manuscript.¹⁴⁵

4.4 Methods

4.4.1 Reaction Data Sets and Template Extraction

The United States Patent Office (USPTO) extracts ranging from the years 1976 to 2016 are publicly available.⁹² They are split into granted and applied patents and are openly available for use by the community. The Reaxys¹⁷⁸ dataset is commercially available, provided by Elsevier under licensing agreements.

All reactions were atom-mapped and classified using the commercially available Filbert and HazELNut packages (version 3.1.8) provided by NextMove software.¹⁸⁰ These were subsequently processed using RDKit and RDChiral for template extraction.^{20,181} The bipartite reaction graph was built using NetworkX and queried to yield a multilabel dataset.¹⁹²

4.4.2 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.^{9,135,138,149,166} However, this strategy neglects the one to many types 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 data set 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.^{122,162} 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 multilabel dataset for the subsequent training of neural networks. This approach allows us to train a multilabel multiclass classification neural network for the prediction of retrosynthetic steps, as opposed to the single-label multiclass classification network previously described (Figure 23). In doing so, the number of samples is limited to 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 is defined as containing all possible retrosynthetic disconnections and, as such, reaction templates that can be applied to any given product.

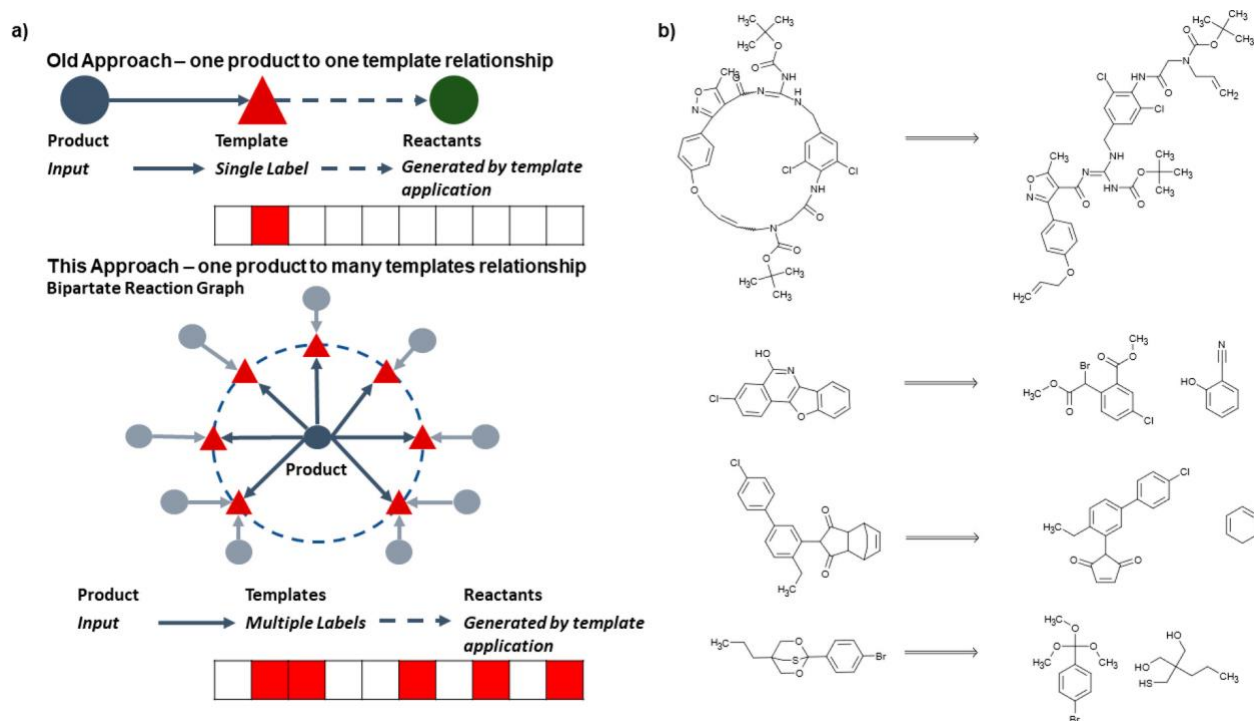


Figure 23: (a) Schematic of multilabel generation. Previously machine learning approaches to retrosynthetic planning have been trained considering a one product to one template relationship. However, as multiple templates/reactions may be used on a given compound, it is desirable to train the model considering 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 multilabel 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 multilabel dataset (Figure 23). Compared to the entirety of the datasets from which the ring formations were extracted, we found that ring-forming reactions constitute an upper limit of 4.5% and 5.8% of the USPTO and Reaxys datasets, respectively (Table 5). The number of ring forming reactions is likely lower than that reported as our crude measure of ring change (number of rings in product minus number of rings in reactants) includes a change in the number of rings resulting from protections and deprotections. An even smaller percentage of all the templates extracted from these datasets correspond to ring formations (Table 5). Therefore, an all-encompassing classifier that considers all extracted 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 5: 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 1 between Products and Reactants.^{a)}

Dataset	Ring Formations	Percentage of Ring-forming Reactions in Dataset	Ring Formation Templates Extracted	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 %

^{a)} Compared to the entirety of each respective dataset, the percentage of reactions corresponding to ring formations is estimated to be 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, considering 12% of the reactions in patents correspond to protections and deprotections.⁽³⁰⁾ The numbers shown are after filtering for templates occurring a minimum of 3 times and compared to all templates and reactions obtained from the dataset.

4.4.3 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 as explained in our previous work.¹³⁶ Products were represented as extended connectivity fingerprints (ECFP) with a radius of 2, using the Morgan algorithm in RDKit,³⁵ whereas templates were represented as binarized labels in a one-vs-all fashion using the scikit-learn library using the “LabelBinarizer”.¹⁸⁶ 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.¹⁸⁶

The network framed as a supervised multiclass classification problem was trained using Keras¹⁸⁷ with Tensorflow¹⁸⁸ as the back end, the Adam optimizer with an initial learning rate of 0.001,¹⁸⁹ and categorical cross entropy as the loss function (Figure 24). 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.¹³⁶ The multilabel approach enables faster training relative to the single-label approach while achieving similar accuracy and loss values (Figure 25).

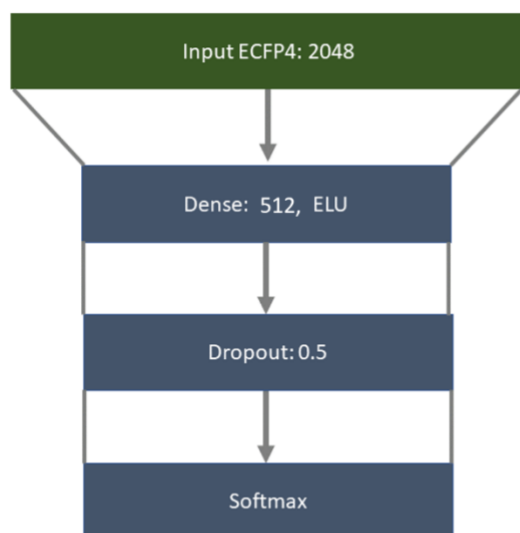


Figure 24: Architecture used to train the “rollout” policy taking molecules represented as ECFP4 as input, through a fully connected layer of 512 nodes, and ELU as the activation function followed by a dropout of 0.5 and softmax output layer.

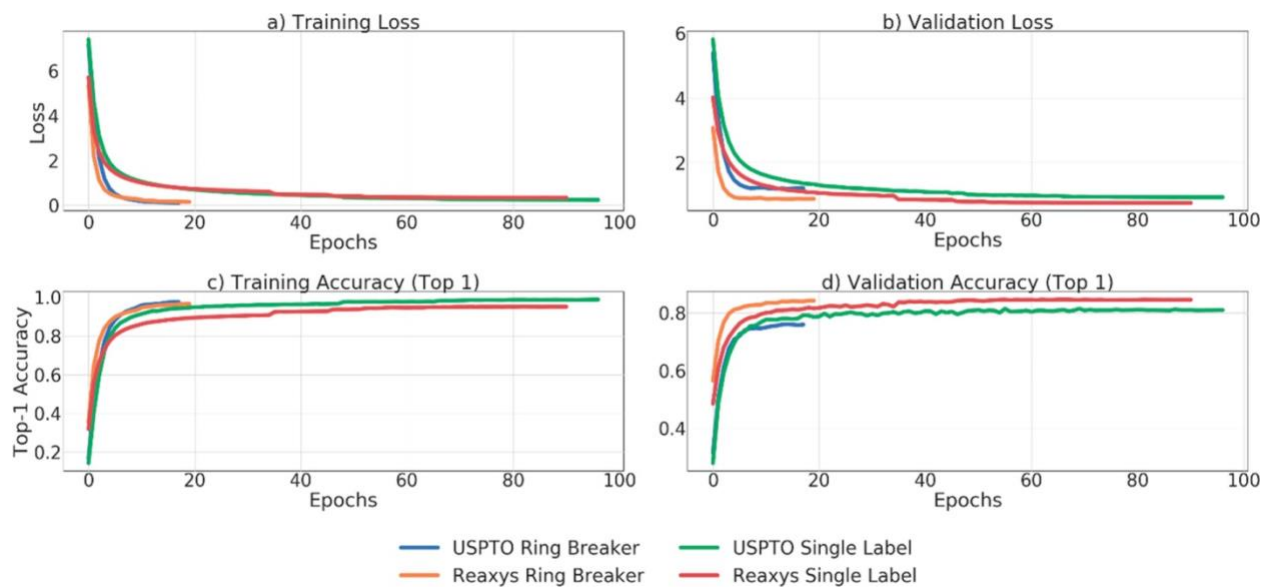


Figure 25: Training and validation curves for accuracy and loss. The RingBreaker network uses multilabel training which enables faster training times relative to the single-label approach, as can be seen by the relatively lower number of epochs required for training. The top-1 accuracy and loss are comparable across both approaches.

4.4.4 Brute Force Application, Filtering, and Prioritization

The “Ring Breaker” and standard models were used to make predictions for each substrate. In each case the templates were applied sequentially until all templates in the template library for each model were exhausted. The first applicable template, the first applicable template that corresponds to a ring formation, the number of templates applicable in the top 50 that correspond to a ring formation, the maximum number of applicable templates, the maximum number of applicable templates that correspond to a ring formation were recorded.

The filtered model was obtained by identifying all templates in the standard model that were used in ring formations (the subset used to train “Ring Breaker”). All templates that were not in the “Ring Breaker” set of templates were set to zero in the output vector. The output vector was then sorted by the probability associated with the remaining templates. This vector was then used as the predictions for the filtered model. The predictions were then applied sequentially as previously described.

4.4.5 Prediction of ZINC Fragments and DrugBank

The ring subsets described were obtained from the ZINC database and used as is.²⁰⁵ The curves plotted are a result of counting the number of heavy atoms in each compound for the ZINC subset and plotting the number of times a fragment with the corresponding heavy atom count occurs. Predictions were then made using each model and the dataset filtered for those fragments for which templates were predicted and successfully applied. From this we can determine how many compounds in the subset we can predict a one-step retrosynthesis for and how this varies with ring size for which we have used heavy atom count as a proxy measure. The overlap of each fragment set with the training sets is available in Overlap of ZINC fragment sets with the training sets.

4.5 Availability of Data and Materials.

Reaxys datasets were used under a license agreement. Precedents for reactions on the Reaxys Web server may not be present in our licensed dataset. The USPTO, DrugBank, and ZINC datasets are freely available. Filbert, NameRxn, and HazelNut were used for atom-mapping and classification under license Drugfrom NextMove software. All code used in the production of this work will be made available under an MIT license at <https://github.com/reymond-group/RingBreaker>.

4.6 Author Contributions

A.T. designed, conducted the research, and wrote the manuscript. A.T. and E.J.B. designed the concept. A.T. and N.S. designed and selected the compound sets. E.J.B., O.E., and J.-L.R. supervised the project.

5 Retrosynthetic accessibility score (RAscore) – rapid machine learned synthesizability classification from AI driven retrosynthetic planning

Computer aided synthesis planning (CASP) is part of a suite of artificial intelligence (AI) based tools that are able to propose synthesis routes to a wide range of compounds. However, at present they are too slow to be used to screen the synthetic feasibility of millions of generated or enumerated compounds before identification of potential bioactivity by virtual screening (VS) workflows. Herein we report a machine learning (ML) based method capable of classifying whether a synthetic route can be identified for a particular compound or not by the CASP tool AiZynthFinder. The resulting ML models return a retrosynthetic accessibility score (RAscore) of any molecule of interest, and computes at least 4500 times faster than retrosynthetic analysis performed by the underlying CASP tool. The RAscore should be useful for pre-screening millions of virtual molecules from enumerated databases or generative models for synthetic accessibility and produce higher quality databases for virtual screening of biological activity.

This chapter has previously appeared as a scientific article in Chemical Science. This section has been reproduced from the following article with permission from the Royal Society of Chemistry.

A. Thakkar, V. Chadimová, E. J. Bjerrum, O. Engkvist, J.-L. Reymond,
*Retrosynthetic accessibility score (RAscore) – rapid machine learned
synthesizability classification from AI driven retrosynthetic planning. Chem.
Sci. 2021, 12, 3339–3349. DOI: 10.1039/D0SC05401A*

5.1 Introduction

Artificial intelligence (AI) in chemical discovery has been driving improvements in the tools available to the chemical community. This has occurred primarily in the areas of *de novo* generation of new chemical entities (NCE),^{61,211} toxicology/bioactivity,²¹² and computer aided synthesis planning (CASP).^{9,138} The question as to which molecule to make and how to make it, is at the center of chemical discovery programs across academia and a range of industries, ranging from agrochemical to pharmaceutical.¹ Typically virtual screening (VS) workflows have been used to decide which compounds to make, starting from generated, enumerated, commercial, or public datasets which are then filtered using a variety of statistical and physics based modelling techniques until the search space is refined (Figure 26).^{2,213–215} The question and decision of which and how to make a given set of compounds is left to a team of chemists at the end of the VS workflow, prior to synthesis in the laboratory. To aid this filtering process a variety of computational tools which take synthesizability considerations into account have been employed over the last two decades.^{116,216,217}

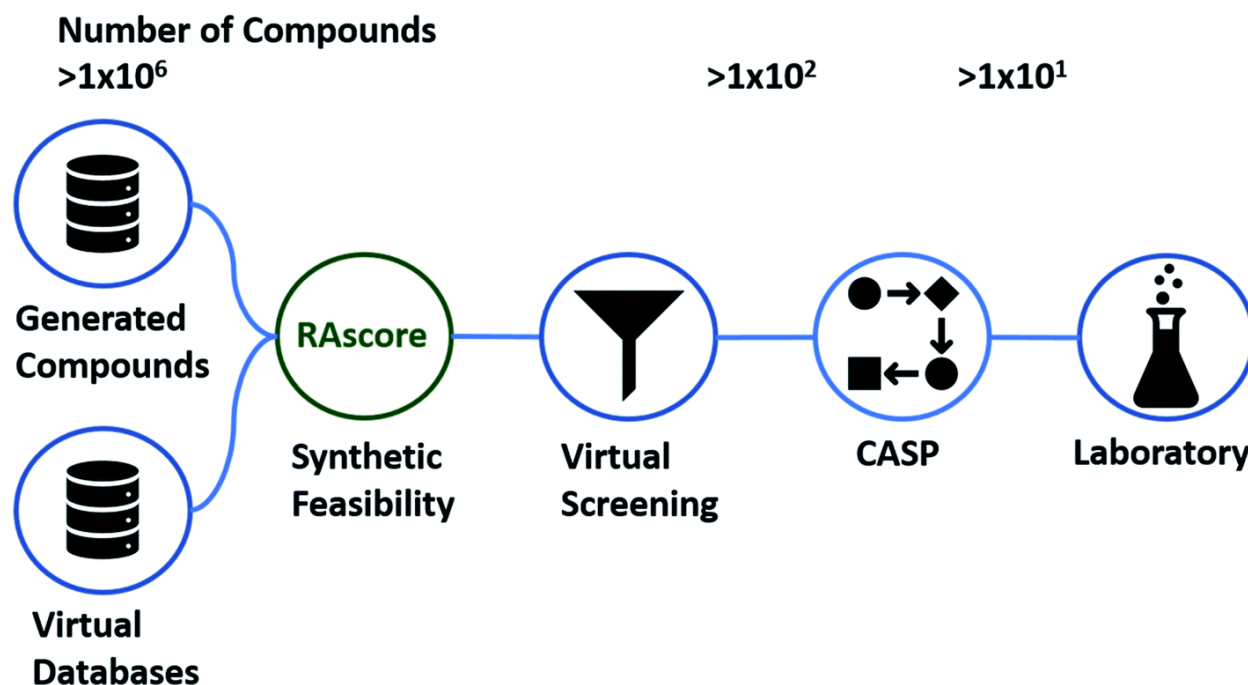


Figure 26: Example of a virtual screening (VS) workflow. The synthesis of compounds is typically considered at the end of the workflow as a final selection criteria, and it is at this point CASP is also used to filter compound libraries to synthesizable compounds. RAscore allows for pre-screening of compounds that may be synthetically accessible by CASP enabling use earlier in the VS workflow (green).

CASP has emerged as a method by which compounds can be filtered in the VS workflow, and during optimization cycles throughout the generative modelling process. Several recent CASP tools have been developed which may be used for these purposes, including but not limited to: Synthia (formerly Chematica),¹²⁶ ICSYNTH,¹³³ ASKCOS,¹³⁸ AiZynthFinder,²¹⁸ and IBM RXN.¹⁵¹ These can be used at two potential stages of the generation process, either to bias the

generation process or as a *post hoc* filter after the molecules have been generated.²¹⁹ Given a target compound, CASP can predict each step of the synthesis pathway towards commercially available building blocks. This makes it suitable for the *in silico* filtering of large compound libraries, and has been demonstrated by Gao and Coley for the case of generated compounds.²¹⁹ However, despite the vast amount of progress that has contributed to making the prediction of full synthetic routes computationally tractable,^{1,104,116,220} to the extent that some predictions may be made within a minute.^{1,138} The scale at which predictions must be conducted for large compound libraries consisting of several million or even billions of compounds can still be limiting.

To tackle the challenge of screening large compound libraries with synthesizability considerations, existing scores include the synthetic accessibility score (SAscore), synthetic complexity score (SCscore), and synthetic Bayesian accessibility (SYBA).^{218,221–223} The SAscore and SYBA are estimations of synthetic feasibility based on the occurrence of molecular fragments in public databases, whereas SCscore is learned from a reaction corpus, with the underlying assumption that products are more complex than their constituent reactants.

Herein, we propose the retrosynthetic accessibility score (RAscore) that enables rapid estimation of synthetic feasibility as determined from the predictions of CASP, in this case AiZynthFinder.²¹⁸ We investigate a machine learning classifier for retrosynthetic accessibility (RA) assessment called RAscore, trained on the outcomes generated from AiZynthFinder, which we have shown can increase the speed at which synthetic accessibility can be estimated, and separate compounds for which retrosynthetic routes can be found by AiZynthFinder. This is an improvement that adds value to existing synthesis scores, and when used in combination with the previous scores, the RAscore should enable pre-screening of compounds that can be later subjected to full retrosynthetic analysis. Thus, this enables CASP to be used at earlier stages of a VS workflow or during the generative modelling process.

We further emphasize that the RAscore may be retrained on data generated from any CASP tool. Therefore, the score will serve to reflect improvements in the continuously changing synthesis planning technology landscape, thereby overcoming current limitations, and can be customized to the specific needs of a project or user. The models and training protocols have therefore been made available for public use: <https://github.com/reymond-group/RAscore>.

5.2 Methods

5.2.1 AiZynthFinder – a tool for computer aided synthesis planning

AiZynthFinder is a template-based retrosynthetic planning tool based on the methodology of Segler and Waller.^{9,218} It consists of a neural network policy, which determines which reaction to use at a given retrosynthetic step, with Monte-Carlo tree search, as reported in our previous studies.¹³⁶ The code, data, and models are open source and available to the public: <https://github.com/MolecularAI/AiZynthFinder>. The reaction transforms have been

extracted from the US patent office extracts (USPTO) and used to train the model by which retrosynthetic expansion was conducted.⁹² Models on Reaxys and proprietary datasets have been examined in our previous studies, but have been omitted in this study due to their proprietary nature.¹³⁶ These can equally be used in place of the USPTO policy for those who have access to the data, and the extraction and training protocols can be found in the repository linked above.

AiZynthFinder considers retrosynthetic routes to be solved if the precursors or building blocks are commercially available. Therefore, as stopping criteria we use the ACD catalogue,¹⁹³ Enamine building block set,²²⁴ and AstraZeneca internal database. These are available from the respective vendors except for the AstraZeneca internal catalogue. In place of the vendors mentioned here, the AiZynthFinder GitHub repository contains a set of compounds extracted from the ZINC database,²²⁵ as highlighted in our previous work.²¹⁸

The score is inherently limited by the underlying CASP tool, however retraining of the RAscore is possible following the procedures outlined herein. Thus, the score can be customized for individual projects and users, as well as kept up to date with developments in synthesis planning technology. We emphasize that any synthesis planning tool should be able to be used for these purposes.

5.2.2 Retrosynthesis prediction for training set generation

Training and test datasets were generated by randomly sampling 200 000 compounds from ChEMBL,¹⁹⁰ as a reference set, and 100 000 compounds each from GDBChEMBL and GDBMedChem, to resemble compounds that would usually be out with the applicability domain of CASP.^{48,49} The compounds were subsequently subjected to retrosynthetic analysis using AiZynthFinder, and labelled as solved or unsolved. The time limit to search for retrosynthetic routes was set as 3 minutes per target compound, with a maximum of seven steps, a maximum of two hundred iterations, and expansion of fifty actions at each stage of the search as determined by the policy network up to a cumulative cutoff threshold of 0.995.

5.2.3 Machine learning classifiers for estimation of retrosynthetic accessibility

Estimation of retrosynthetic accessibility (RA) was framed as a binary classification problem, as the goal of the study was not to score complexity but rather identify with rapid approximation whether a compound could be synthesized or not by CASP, for which we use AiZynthFinder in this study. We trained a series of classifiers on the retrosynthetic predictions of AiZynthFinder using the label generation method stated previously. The trained classifier predicts whether or not a given compound is synthetically accessible as found by AiZynthFinder.

We examined the following classification algorithms: (a) a feed forward neural network classifier, (b) XGBoost classifier, and (c) random forest classifier. For each algorithm 2048 dimensional counted extended connectivity fingerprints were used with a radius set to 3 (ECFP6), and ECFP6 counts with features as generated by RDKit.^{20,35} In total six different models were trained for each

dataset, ChEMBL, GDBChEMBL, and GDBMedChem. SAScore, SCscore, and SYBA are continuous scores for complexity, thus we trained a classifier for each score for comparative purposes, where the score was used as the sole descriptor. For the score-based classifiers we used a feed forward neural network and logistic regression. The scores used as descriptors were calculated using RDKit and the models published by the authors of the corresponding publications.^{20,221–223}

Scikit-Learn was used to train the random forest model,¹⁸⁶ XGBoost for the XGB classifier, and Keras with Tensorflow for the feed forward neural networks.^{187,188} In each case the models were wrapped within an objective function using the Optuna framework for hyperparameter optimisation.²²⁶ All models with the exception of the feed neural network were optimized using a five-fold cross validation. The framework used to train the classifiers and models are available at <https://github.com/reymond-group/RAscore>, and can be used for any binary classification problem.

Each model was optimized with the Optuna hyperparameter optimization framework to find the optimal parameter set.²²⁶ In the case of the feed forward neural network, we treated the number of layers, the size of the layers, the activation function, the dropout rate, and the learning rate as hyperparameters, to find the optimal architecture within the bounds of the starting criterion as given in Example of the optimal architecture found by hyperparameter optimization for the ChEMBL dataset.

There was no overlap of compounds between training, validation, and test sets. This was determined by computing the InChI-keys of the compounds in the two sets and using the Python built-in set methods to find the intersection.²⁸ We did not check whether a compound was present in the training data used to train AiZynthFinder, however this is likely not to influence the performance of multi-step retrosynthesis, as the reaction datasets only consider single steps and is supported by our previous studies.¹³⁶

5.2.4 Average linkage as a method for evaluating machine learning based classifiers

We assessed model performance by computing how well solved and unsolved routes are separated using the concept of average linkage. Average linkage is a statistical method by which the distance between two clusters are treated as the average distance between all pairs of items, where each member of the pair belongs to one of the two clusters. In this instance, the two clusters are solved and unsolved compounds as determined by AiZynthFinder (other CASP tools may be used in place). The average linkage or separation between solved and unsolved compounds was determined by min–max scaling the values of each score such that they were normalized between 1 and 0 using the Scikit-Learn *MinMaxScaler*. The absolute pairwise distances were computed, and the average of the distances taken to yield a value that corresponds to the separation of the clusters as shown in (Figure 27).

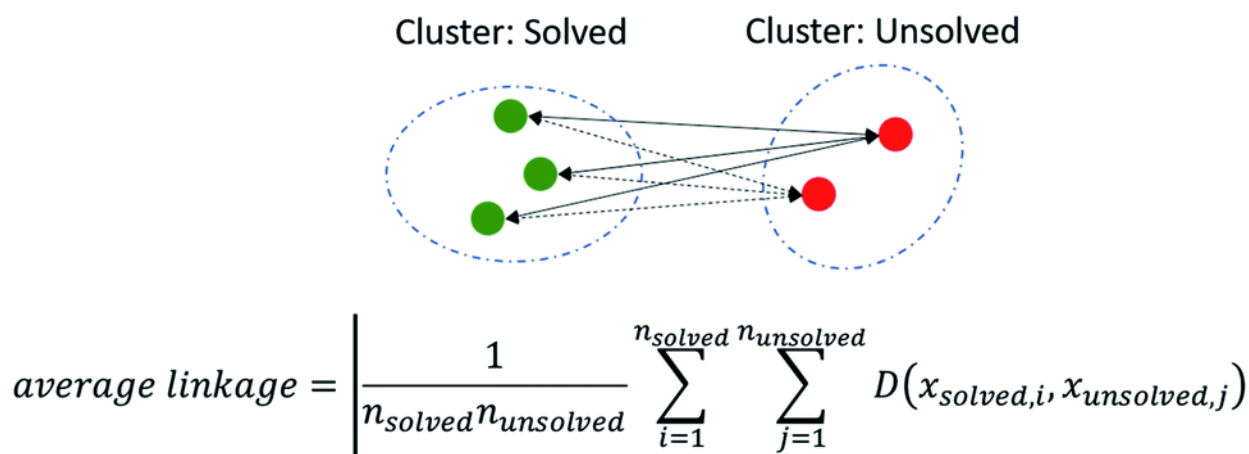


Figure 27: Illustrates the computation of the average linkage. The average linkage is a method by which the distance between two clusters are treated as the average distance between all pairs of items, where one member of the pair belongs to each cluster.

5.3 Results and discussion

5.3.1 Route statistics from the generation of labels for machine learning classifiers

Initially training and test datasets were generated by randomly sampling 200 000 compounds from ChEMBL,¹⁹⁰ as a reference set, and 100 000 compounds each from GDBChEMBL and GDBMedChem.^{48,49} The two are subsets of the GDB17 database.⁴⁴ ChEMBL was chosen to represent a selection of bioactive molecules and the GDB subsets chosen to be more challenging owing to their differing structural and physiochemical property distribution.⁴⁸ The compounds were subsequently subjected to retrosynthetic analysis using AiZynthFinder, and labelled as solved or unsolved.

Figure 28 shows statistics gathered for the predicted retrosynthetic routes during the label generation process. The percentage of solved routes increases monotonically, and the rate at which routes are solved decreases with the number of steps for each dataset. This is most noticeable for compounds requiring synthetic routes between 5 and 7 steps, where we observe a significant increase in the dataset coverage (Figure 28b), but no corresponding increase in the percentage of solved compounds (Figure 28a). ChEMBL has the highest percentage of solved compounds, whereas GDBMedChem and GDBChEMBL are consistently lower.

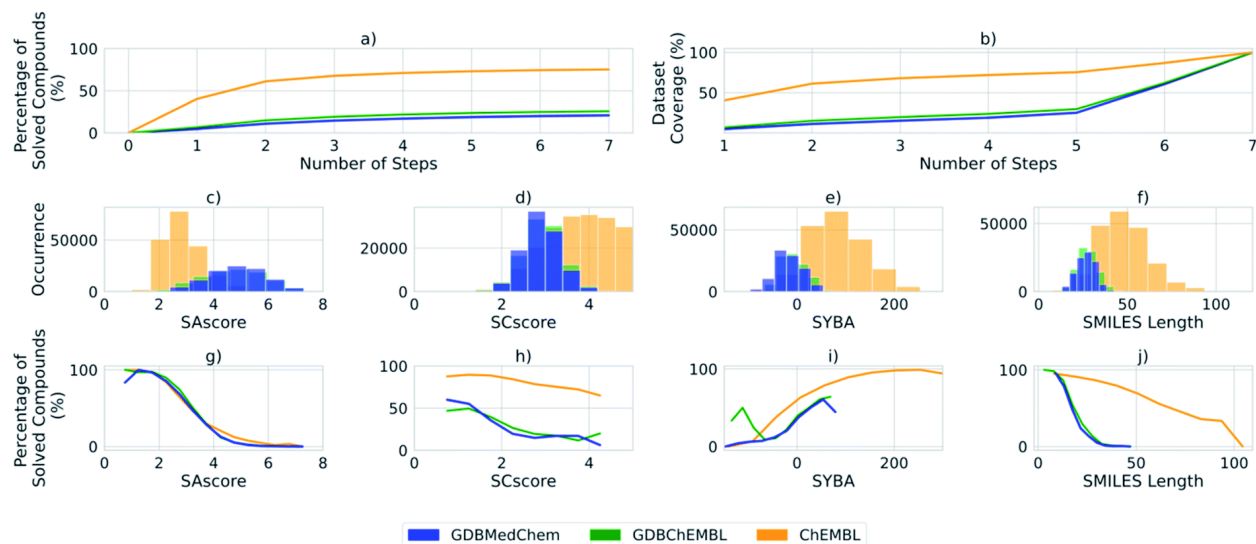


Figure 28: Statistics gathered for the retrosynthesis predicted during the label generation process for each dataset ChEMBL, GDBChEMBL, and GDBMedChem. The statistics are shown for all compounds sampled: 200 000 from ChEMBL, and 100 000 from each of the GDB subsets. (a) The percentage of solved compounds as a function of the number of steps, (b) the dataset coverage as a function of the number of steps. (c-f) Histograms depicting the distribution of the compounds in each dataset for each of the currently used scores. For SAscore and SCscore the lower the score the less complex and easier to synthesise a given compounds, whereas for SYBA positive values indicate easy to synthesise compounds and negative values hard to synthesise. (g-j) The percentage of solved compounds as a function of each of the currently used scores as computed for each bin in the histogram.

We observed a correlation between the percentage of solved compounds and the SAscore,²²³ SCscore,²²¹ and SYBA,²²² as well as SMILES length (Figure 28g-j), which are in agreement with the results obtained by Coley and Gao.²¹⁹ In the case of SAscore and SCscore, the lower the score the more likely it is that a synthetic route can be obtained for a compound, as found for all datasets. The ChEMBL sample exhibits a lower range of SAscore than the GDBMedChem and GDBChEMBL samples (Figure 28c), which may explain the higher percentage of solved compounds in ChEMBL as compared to GDBMedChem and GDBChEMBL (Figure 28a).

However, for SCscore (Figure 28d) the GDB subsets exhibit a lower range of scores in comparison to the ChEMBL sample. Thus, the inverse of the distribution we obtain for SAscore and can be rationalized by considering the assumptions made in the SCscore model. The SCscore is based on reactions rather than molecular fragments and assumes that the products of a reaction are more complex than the reactants. In most cases the products are also larger than the reactants, thus the assumption for SCscore falters for the GDB subsets because of their restricted size as shown by the difference in SMILES length between the ChEMBL and GDB subsets (Figure 28f-j). This is further supported by the lower percentage of solved routes for the GDB subsets (Figure 28f-j).

In the case of SYBA, the higher the score the more likely it is that a route can be found, negative values indicate hard to synthesize compounds. The distribution shown for SMILES length reflects the fact that the GDB subsets are skewed towards smaller molecules, and with lower heavy atom counts than those found in ChEMBL by virtue of the rules used in their

enumeration.^{48,49} The Figure 28j reveals that the rate at which compounds can be solved falls off much more rapidly with SMILES length for the GDB subsets than for ChEMBL.

5.3.2 Attempts at using SAScore, SCscore, and SYBA

We assessed the existing scores SAScore, SCscore, and SYBA for their ability to distinguish between compounds that could be solved by AiZynthFinder and those that could not (Figure 29). These scores have often been used to filter or estimate the synthetic accessibility of large datasets of virtual compounds.^{44,227,228} However, we have found that there is no threshold value at which the SA, SC, and SYBA scores can be set that clearly separates compounds that can and cannot be solved by AiZynthFinder, as shown by the overlapping histograms. This was observed for all datasets examined in this study (Attempts at using SAScore, SCscore and SYBA). Thus, there is potential for them to be misused when filtering large virtual libraries. To resolve this issue, we propose that the existing scores be used alongside the classifiers trained in this study to determine whether a synthetic route can be found, and how difficult it may be to realize the route in the wet lab.

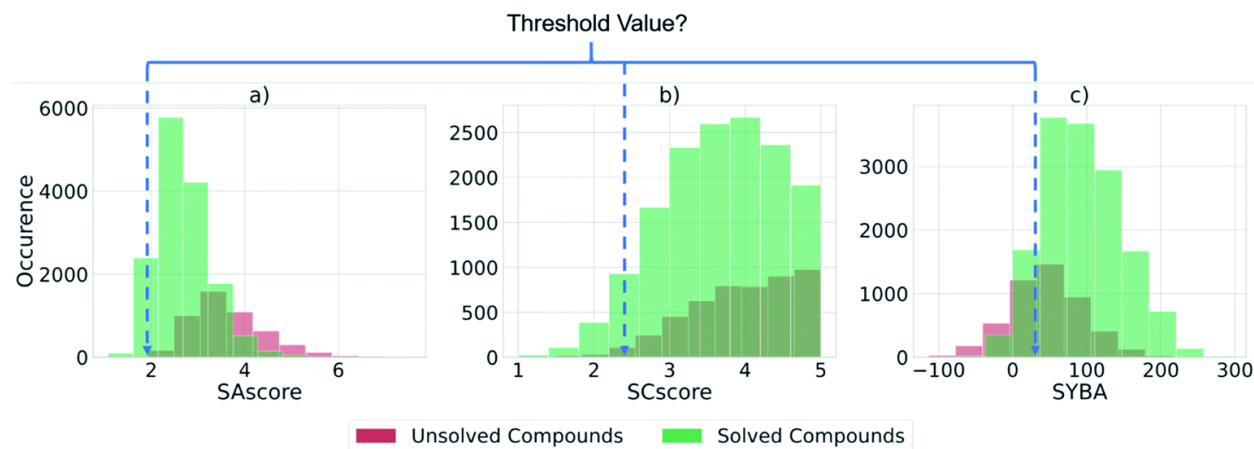


Figure 29: Histograms computed for the test set of ca. 20 000 ChEMBL compounds showing whether a retrosynthetic route could be found by AiZynthFinder for a given compound (green) or not (red), and their distributions across each of the scores in current use. There is no threshold value at which the current scores are able to separate compounds that can be solved by AiZynthFinder (green) from those that cannot (red). This highlights how the scores have potential for misuse in generative modelling and filtering sets of compounds.

5.3.3 Machine learning classifiers for estimation of retrosynthetic accessibility

The overlaps shown in Figure 29, demonstrate the need to be able to differentiate between compounds that can and cannot be synthesized by AiZynthFinder. Therefore, we trained a series of ML based classifiers to determine whether a given compound could be solved by AiZynthFinder. A selection of the results obtained for the trained classifiers are shown in Table 6 (refer to Machine Learning Classifiers for Estimation of Retrosynthetic Accessibility for all trained models). In each case the classifiers outperform the existing scores which were used as a baseline (SAScore, SCscore, and SYBA) both in terms of the AUC (area under the curve) and

average linkage with respect to their ability to classify compounds as solved or unsolved. When using the existing scores as descriptors to train the classifiers, we observed a marginal improvement in comparison to the score itself. This is because the existing scores are complexity based scores, thus have not been developed with the separation of compounds found synthetically accessible by CASP in mind. A more significant improvement in classifier performance was obtained when using ECFP6 counted vectors as molecular descriptors, both with and without features. The feed forward neural network (NN) based models consistently outperformed random forest and showed comparable performance to gradient boosting methods (XGB).

Table 6: Outlines the top 3 classifiers trained for each dataset alongside their corresponding metrics^{a)}

Dataset	Model	Descriptor	AUC	Accuracy	Precision	Recall	Average Linkage
ChEMBL	NN (RAScore)	ECFP6 counts with features	0.93	0.90	0.92	0.95	0.69
	NN	ECFP6 counts	0.94	0.90	0.92	0.95	0.68
	XGB	ECFP6 counts	0.95	0.91	0.92	0.96	0.65
	NN	SAScore	0.85	0.81	0.84	0.92	0.37
	NN	SCscore	0.61	0.75	0.61	1.00	0.27
	NN	SYBA score	0.74	0.78	0.78	0.97	0.21
	Baseline	SAScore	0.15	-	-	-	0.17
	Baseline	SCscore	0.39	-	-	-	0.22
	Baseline	SYBA	0.74	-	-	-	0.17
GDBChEMBL	NN (GDBscore)	ECFP6 counts	0.93	0.87	0.76	0.73	0.64
	NN	ECFP6 counts with features	0.94	0.88	0.78	0.74	0.63
	XGB	ECFP6 counts	0.94	0.89	0.81	0.73	0.61
	Baseline	SAScore	0.11	-	-	-	0.26
	Baseline	SCscore	0.38	-	-	-	0.14
	Baseline	SYBA	0.72	-	-	-	0.17
GDBMedChem	NN	ECFP6 counts	0.93	0.88	0.75	0.64	0.64
	NN	ECFP6 counts with features	0.94	0.89	0.77	0.66	0.63
	XGB	ECFP6 counts	0.94	0.89	0.78	0.64	0.61
	Baseline	SAScore	0.13	-	-	-	0.22
	Baseline	SCscore	0.39	-	-	-	0.14
	Baseline	SYBA	0.70	-	-	-	0.17

^{a)} For comparative purposes a baseline has been included which are the SAScore, SCscore, and SYBA. The metrics for these have been computed using Scikit-Learn and the average linkage computed as described in the methods. Classifiers were trained using each of the respective scores as descriptors to enable a direct comparison of classifier performance. These marginally outperform the baseline models in terms of AUC and average linkage. The top 3 classifiers for each dataset using ECFP6 variants consistently outperform the baseline models and their classifiers. For RAScore the top performing classifier on the ChEMBL dataset was chosen, and a separate GDB specific model chosen termed GDBscore which was the top performing classifier on the GDBChEMBL dataset.

We identified that the following classifiers were consistently the top three models across each of the datasets: feed forward neural networks using ECFP6 counts, feed forward neural networks using ECFP6 counts with features, and XGBoost using ECFP6 counts. For the RAScore we chose the top performing classifier for separating the compounds as determined by the average linkage. We also identified a GDB specific classifier which we term GDBscore in the same manner. The

GDBscore classifier was trained on the GDBChEMBL dataset, the classifier trained on GDBMedChem was found to have equivalent performance.

5.3.4 Prediction time

The importance of training ML based classifiers rather than simply predicting the full retrosynthetic pathway becomes clear when examining Table 7. Full retrosynthetic route prediction of the ChEMBL sample of 200 000 compounds to a set of commercially available building blocks took approximately 239 CPU days on a single machine with 8 CPUs and 64 GB of RAM, using AiZynthFinder. Parallelization of full synthetic route prediction is not possible on a single machine under the current implementation of AiZynthFinder, however, it is possible to split the compounds over several cores and distribute the workload over several machines as has been done in this study. In comparison 77 minutes were required for classifying retrosynthetic accessibility using RAscore. The increase in prediction speed by *ca.* 4500 times opens up the possibility of estimating the retrosynthetic accessibility of virtual compounds, for instance in drug discovery projects, and for the scoring of compounds resulting from generative models earlier in the virtual screening workflow. Similar increases in prediction time are also observed for the GDB subsets (Table 6).

Table 7: Percentage of solved compounds for each dataset and the run time^a required using AiZynthFinder.

	ChEMBL	GDBChEMBL	GDBMedChem
Percentage Solved	75.21	25.54	20.79
Size	200,000	100,000	100,000
AiZynthFinder Run Time (days)	239	149	151
Score Run Time (mins)	79 ^{b)}	30 ^{c)}	30 ^{c)}

^{a)} Expressed in days taken on a single machine with 8 CPUs and 64 GB of RAM, rounded to the nearest day.

The time taken in minutes for the neural network classifier with ECFP6 counted fingerprints is also given for comparative purposes. The neural network classifier, RAscore, is able to reproduce the results obtained from AiZynthFinder in a fraction of the time taken to predict full retrosynthetic routes.

^{b)} RAscore

^{c)} GDBscore

5.3.5 Applicability domain

Gao and Coley previously published the results of running retrosynthetic analysis with ASKCOS for a series of datasets consisting of both published and generated compounds.²¹⁹ We tested our trained classifier on the dataset used by Gao and Coley to determine the applicability domain of the classifier and gauge how well the ASKCOS predictions could be reproduced. We also used AiZynthFinder to predict retrosynthetic routes to the same set of compounds to establish whether the classifier could reproduce the underlying CASP tool.

For each dataset AiZynthFinder marginally outperforms ASKCOS, and is most striking for the GDB17 sample (Figure 30).⁴⁴ This is because AiZynthFinder only considers retrosynthetic analysis, whereas ASKCOS additionally factors in reaction prediction which enables pruning of unfeasible or low probability retrosynthetic pathways. Furthermore, as the reaction prediction

models are trained on published chemistry, and the majority of GDB17 compounds are unpublished or dissimilar to published compounds,⁴⁴ the pathways suggested are likely to be pruned resulting in the lower percentage of solved compounds for ASKCOS. Another difference that should be considered when comparing the two models is the building blocks available to each respective model. This can affect the ability of the CASP tool to find retrosynthetic routes and influences whether or not a compound is labelled as solved.

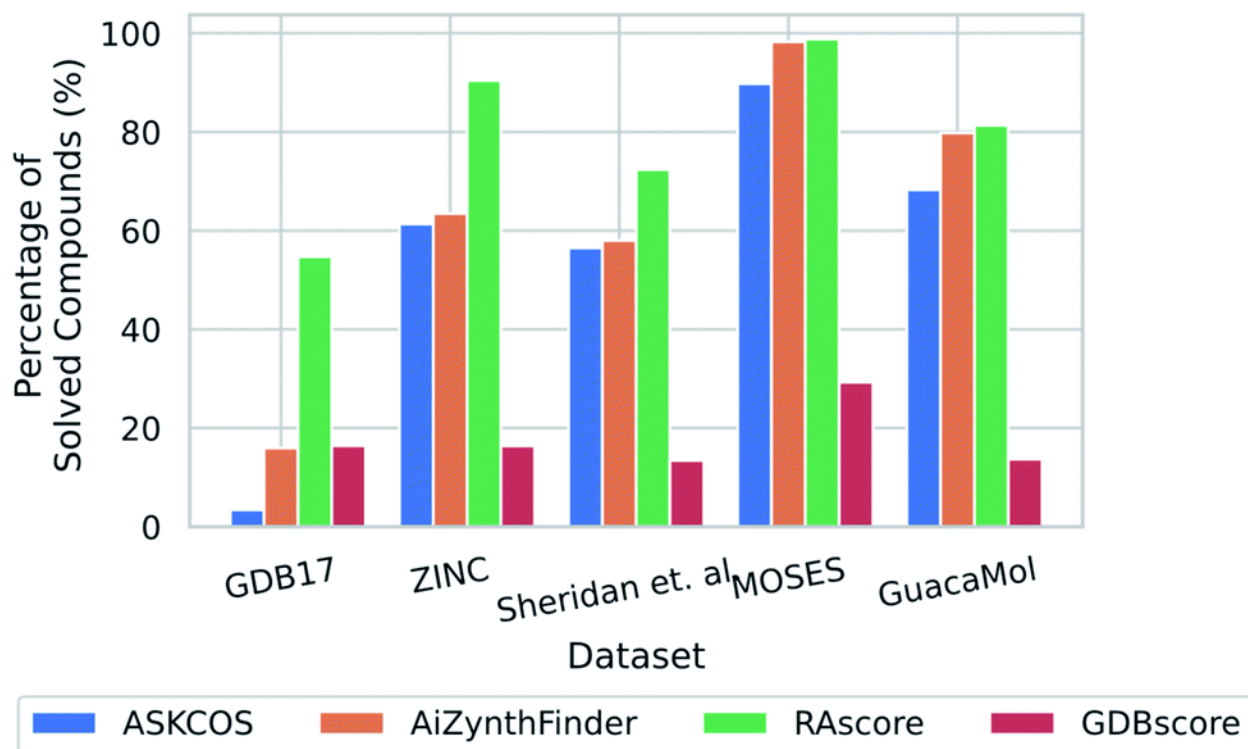


Figure 30: Applicability domain as determined by application to a set of compounds published by Gao and Coley in a previous study, full details of each dataset can be found in the referenced manuscript.²¹⁹

We found that the feed forward neural network classifier trained on ChEMBL that we term RAscore, overestimates the synthetic accessibility of GDB17 in comparison to ASKCOS and AiZynthFinder. This is also observed for the other datasets examined, however the extent to which RAscore overpredicts is less striking. To replicate the GDB17 dataset, we use GDBscore, which is a classifier trained on GDBChEMBL and find we can better reproduce the underlying AiZynthFinder synthesis planning tool. The MOSES dataset is based on the ZINC Clean Leads collection and GuacaMol is based on the ChEMBL database, both are used for evaluating distribution learning algorithms for drug discovery.^{229,230} The overprediction on both ZINC and the prediction in line with the MOSES dataset is surprising considering the compounds originate from the same database. However, this may be rationalized considering the samples differ in their distribution, and have been obtained from different collections within the ZINC database.^{227,230,231}

The overprediction on the Sheridan *et al.* dataset can be seen as positive as all compounds in the dataset were previously synthesized at Merck.²³² In addition, the prediction in line with the

GuacaMol set, implies that the classifier performs well on ChEMBL like compounds by virtue of the underlying training data.

5.3.6 Examples – Limitations of RAscore arising from CASP

We examined the test set from our ChEMBL sample for compounds within a Tanimoto similarity of 0.8 or greater. Some examples of pairs of compounds are shown in Figure 31. In the pairs shown one compound was unsolved by our retrosynthetic tool and the other labelled as solved. For each example we show that the topology is largely unchanged and only small edits have been made to the functionality of the molecule. The change in outcome with minor changes in functionality highlight a limitation of AiZynthFinder and likely other template based CASP tools. This can originate from: the representation of the input molecules, the way the templates are specified, and the distribution of similar samples in the dataset from which the reactions originate. The templates suggested for disconnections are unable to account for subtle changes in the reaction center, thus the appropriate precursors were not able to be enumerated. This arises because the molecular graph underlying the template does not match that of the substrate, thus there is no substructure match. These examples are not ‘true’ negatives in the sense that they cannot be experimentally realised in the wet-lab and are only negative in relation to the ability of the AiZynthFinder to conduct a retrosynthetic analysis. Some examples of such compounds which have led to poor separation of solved/unsolved compounds are shown in Figure 31.

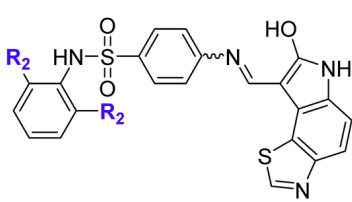
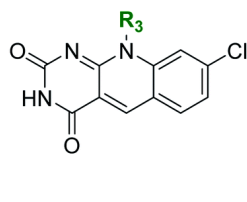
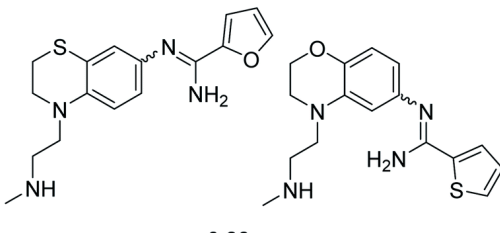
	a)		b)		c)	
						
Tanimoto	0.81		0.93		0.86	
	$R_2 = \text{H}$	$R_2 = \text{Me}$	$R_3 = \text{cyclohexyl}$	$R_3 = \text{cyclopropyl}$		
Solved	No	Yes	No	Yes	No	Yes
RAscore	1.00	1.00	0.99	0.98	0.99	0.99
SAscore	2.83	2.97	2.49	2.51	2.97	2.74
SCscore	4.43	4.48	3.51	3.53	4.55	4.53
SYBAscore	112.42	115.22	42.20	29.97	18.96	38.10

Figure 31: Examples of pairs of compounds from the test set that are similar to each other (Tanimoto > 0.8), where a retrosynthetic route could be found for one example in the pair but not the other. In each case only a slight modification of the compound leads to a change in the outcome from the CASP tool, consider (a) addition of two ortho-methyl groups on the terminal phenyl ring, (b) substitution of a cyclohexane moiety for a cyclopropane, and (c) a change in substitution pattern and ring morphology, leads to a change in outcome from solved to unsolved.

To understand why the solved/unsolved test cases were not easily separable, consider the examples in Figure 31. In the case of similar compounds, both solved and unsolved compounds are scored as synthetically feasible with values tending towards 1.0, despite AiZynthFinder not having found

a synthetic route. The example in Figure 31a, is a case for which RAscore predicts the compounds as synthetically accessible by AiZynthFinder despite a synthetic route having been found for only the compound with two *ortho*-methyl groups on the terminal phenyl ring. The RAscore learns that such minor changes to functionality are feasible by virtue of the machine learning approach, which does not take into consideration the inner workings of AiZynthFinder, but rather learns a mapping between inputs (compounds) and outputs (synthesizable by AiZynthFinder/unsynthesizable by AiZynthFinder). This behavior is an artefact of the subset of compounds from ChEMBL the model was trained on, and examples in which the model misclassifies compounds as synthesizable can also be found. Similar substitutions are shown in Figure 31b-c, whereby AiZynthFinder failed to suggest retrosynthetic disconnections leading to commercially building blocks.

In most cases the most similar molecule in the training set was below a Tanimoto value of 0.8 (Appendix C.5.1 Example Compound Similarity to Training Set), and potentially requires a different synthetic strategy as compared to the compounds shown for the test set (Figure 31). This raises another limitation of AiZynthFinder and potentially other CASP tools, which can be overcome by RAscore. The performance of a CASP tool is limited by the number and type of building blocks available. In some cases, it may be that the building blocks necessary are not included in the database underlying the CASP tool but are in fact available from other vendors. Furthermore, it can also be the case that similar building blocks are available that a medicinal chemist may consider for functionalization. In these cases, the RAscore is able to learn that it is likely that two analogues are synthetically accessible despite a retrosynthetic route having not been found. This is because RAscore is not based on a library of building blocks and has been trained with the compound as input and label (synthesizable by AiZynthFinder/unsynthesizable by AiZynthFinder) as output, thus has no knowledge of building blocks explicitly. The RAscore model learns similarity between compounds internally, and by doing so learns where to place a decision boundary between datapoints belonging to each cluster. This is the basis on which most machine learning techniques enable the models to extrapolate to similar compounds.

To exemplify the aforementioned arguments, consider the routes predicted by AiZynthFinder shown in Figure 32. If we again take the case of the phenyl moiety both with and without the *ortho*-methyl groups of the compound shown in Figure 31a, and examine the routes predicted for each, Figure 32a-b respectively, we observe differences in the predicted route in terms of the synthetic strategy used, thus step count. Therefore, similar compounds with largely unchanged topology can have considerably different synthetic routes predicted for them. One of the reasons this occurs is because each step in the route prediction is treated independently from the others. Thus, the neural network used in AiZynthFinder does not learn that similar compounds have the potential to be synthesized *via* similar routes as it has not been fed information about the route. Whilst a chemist may consider first synthesizing the scaffold, and subsequently functionalising it to yield the desired analogues, AiZynthFinder is currently unable to take into account such considerations. This is further exemplified in Figure 32a-b, whereby different synthetic routes necessitate different starting materials. The synthetic route proposed in Figure 32a can be used to

a

b

Regioselectivity issue during aminothiazole formation

Building Block Commercially Available

The RAScore has potential to overcome some of these limitations as it does not take into account route information explicitly. Rather the RAScore is based on the predictions of AiZynthFinder, and equally the predictions of any CASP tool should be able to be used in their place. This has the advantage that the RAScore is then able to approximate whether a synthetic route can be found using CASP for any given molecule, without having to compute the synthetic route each time.

5.4 Conclusions

Herein we have built on the improvements in AI driven CASP in recent years by combining the predictions made with our CASP tool, AiZynthFinder, with ML, to train a classifier returning a retrosynthetic accessibility score (RAscore). RAscore addresses the challenge of classifying compounds as synthetically feasible and is orders of magnitude faster than full retrosynthetic analysis by CASP, and with comparable performance. The RAscore demonstrates potential for rapid pre-screening of compounds for synthetic accessibility, enabling enrichment of synthetically feasible chemical space. Whereas previous synthetic accessibility and complexity based scores have potential for misuse when filtering large virtual libraries, as a result of being unable to determine a threshold value (Figure 29), we resolve this issue by proposing that the existing scores be used alongside the RAscore to determine whether a synthetic route can be found, and how difficult it may be to realize the route in the wet lab.

In addition, we highlight inherent limitations to be aware of in the RAscore arising from the performance and applicability of the underlying CASP tool, namely: (1) availability of building blocks, (2) different synthetic strategies towards the same scaffold, and (3) route predictions are treated independently to each other. The concept presented herein can be extended to any CASP tool and the predictions it generates, and the score retrained. The score will be made available under an MIT license at: <https://github.com/reymond-group/RAscore>.

5.5 Availability of data and materials

The score was made available under an MIT license at, as well as instructions on how to access the datasets and the framework for training the classifiers: <https://github.com/reymond-group/RAscore>.

AiZynthFinder is open source and is available under an MIT license at: <https://github.com/MolecularAI/AiZynthFinder>.

The dataset used to assess the applicability domain can be found at: https://github.com/wenhao-gao/askcos_synthesizability/tree/master/results/dataset.csv.

5.6 Author Contributions

A. Thakkar designed, conducted the research, and wrote the manuscript. V. Chadimová performed preliminary modelling studies. E. J. Bjerrum contributed ideas to the project. O. Engkvist, and J. L. Reymond supervised the project and assisted in writing the manuscript. All authors read and approved the final manuscript.

6 Linking Navigation of Chemical Libraries to Synthetic Route Prediction using Browser Based Visualisation Tools

The GDBRouteBrowser combines chemical library visualization with synthetic route prediction, to enable chemists to obtain a deeper understanding of the chemical library they are interested in. This is accomplished by overlaying properties that can be calculated or predicted, such as physiochemical, bioactivity, or those derived from synthesis prediction approaches on a graphical representation of the chemical space called TMAP. The GDBRouteBrowser can be tailored to the needs of a particular project or individual to facilitate compound prioritization. I demonstrate a workflow by which the GDBRouteBrowser can be used to test hypotheses relating to the prediction of synthetic routes. This enables rapid feedback to be obtained from experimental chemists within one tool, thus facilitating the identification of areas for further development of the underlying synthesis planning toolkit. Furthermore, I show that the GDBRouteBrowser can be used to identify series of compounds that may have related syntheses, thus can aid chemists in the ideation and compound prioritization process during a project. While I have used AiZynthFinder, the concept can be applied to the predictions of any CASP tool providing they are uploaded to the underlying MongoDB instance following the appropriate schema.

This chapter is unpublished research

6.1 Introduction

The development of computational techniques used in chemical discovery and synthesis have received increased interest in recent years. As computational tools and libraries have become more accessible, methods for generating molecules have given unprecedented access to virtual chemical space,^{44,48,49,55,61,63,65,175,230} there remains a bottle neck in translating the insights brought by computer aided drug design (CADD) to the wet lab. In part, this bottle neck arises from the need to estimate the synthetic accessibility of the generated virtual space,^{217,221–223,233} for which computer aided synthesis planning (CASP) tools have been developed to address the issue.^{9,126,132,138,218,219,234,235} Secondly, the way compounds are prioritized and selected from the generated space does not lend itself well to user interaction. Namely, spreadsheets or databases of compound data are given to end users to interact with. To overcome this, visualization techniques such as PCA, UMAP, t-SNE, and TMAP have been used to gain a "bigger picture" view of the chemical space and have been used interactively to examine and select compounds that may be shortlisted for synthesis.^{45,75,87,236} Once compounds are selected for synthesis, retrosynthetic analysis must be conducted either by a human or CASP. These predictions require a user to switch tools to generate synthesis plans, where the data is often stored inside the respective toolkit. Switching and learning different tools for compound prioritization, visualization, and synthesis prediction is undesirable owing to the high barrier of entry and necessity for a user to manually transfer data between the tools. Therefore, in this study I propose steps towards creating one tool combining compound library visualization with synthetic route prediction. This has the additional benefit that compounds can be prioritized based on synthetic route information, along with traditional physio-chemical descriptors and bioactivity data.

6.2 Methods and Implementation

6.2.1 AiZynthFinder – A Tool for Computer Aided Synthesis Planning

AiZynthFinder is a template-based retrosynthetic planning tool based on the methodology proposed Segler and Waller.^{9,218} It consists of a neural network policy, which determines which reaction to use at a given retrosynthetic step, with Monte-Carlo tree search, as reported in our previous studies.^{136,218} The code, and a set of pre-trained models based on publicly available data have been open sourced and are available to the public: <https://github.com/MolecularAI/AiZynthFinder>.

In this study a modified version of the AiZynthFinder expansion policy was used as trained in a previous study.¹⁴⁴ The reaction transforms were extracted from the US patent office extracts (USPTO),⁹³ Pistachio,⁹¹ Reaxys,¹⁷⁸ and AstraZeneca internal electronic laboratory notebook (ELN) and used to train the model termed prioritization network in Figure 33A.⁹² The model was further augmented using artificial labels determining the applicability of templates at a given retrosynthetic step following an approach used in our previous studies as shown in Figure 33B,

where I found an improvement in predictive performance and applicability for the topN predicted templates, and overall synthetic route finding capability.¹⁴⁴ For this reason, the modified model was deemed to be more appropriate for the task of finding retrosynthetic routes towards GDB molecules. Further details regarding the architecture and training of the model are given in our previous publication, although have not been open-sourced owing to the use of proprietary datasets.¹⁴⁴

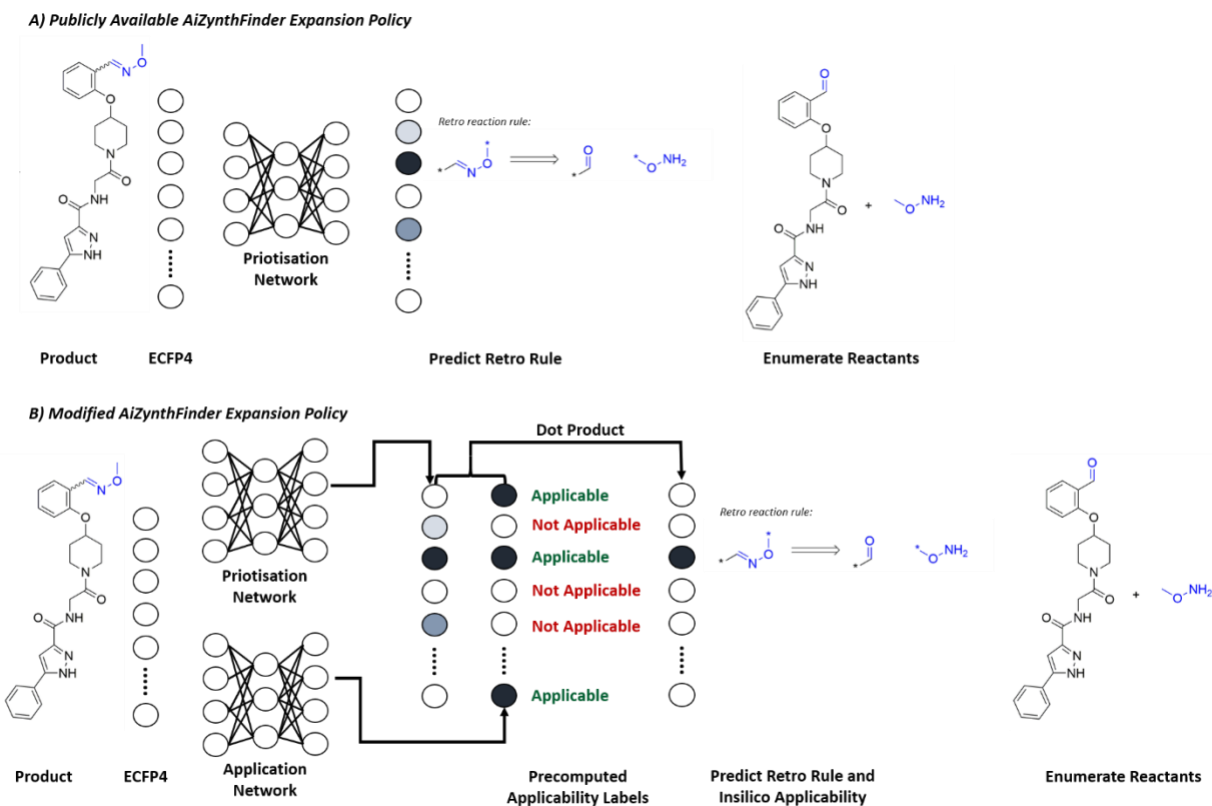


Figure 33: Schematic detailing differences between the architectures for a) the standard publicly available AiZynthFinder expansion policy and b) a modified version of the expansion policy based on proprietary data and accounting for the in-silico applicability of templates. The modified version of the AiZynthFinder expansion policy shown in (b) was used as trained in a previous study. The modified model showed improved performance with retrosynthetic route finding as demonstrated in our previous studies, and for this reason was deemed to be more suitable towards the task of finding retrosynthetic routes to GDB molecules.

AiZynthFinder considers retrosynthetic routes to be solved if the precursors or building blocks are commercially available. Therefore, as stopping criteria I use the MolPort catalogue, and Enamine building block set,²²⁴ considering only 'in-stock' compounds, which results in a stock catalogue of 653,397 building blocks. These are available from the respective vendors. In place of the vendors mentioned here, the publicly available AiZynthFinder GitHub repository contains a set of compounds extracted from the ZINC database,^{225,237} as highlighted in our previous work.²¹⁸

6.2.2 Precomputing Synthetic Routes

Synthetic routes were pre-computed using the AiZynthFinder package outlined previously. Route predictions can be considered an embarrassingly parallel problem, thus given a file containing SMILES, the chosen compound library was split into multiple batch jobs and submitted to a HPC cluster running SLURM as SLURM job arrays in an automated manner. The resulting output files were uploaded to a MongoDB instance and properties calculated using RDKit for later processing into TMAPs.^{20,87} This process can be repeated for any file corresponding to a project or compound library containing a SMILES on each line as necessary.

6.2.3 GDBRouteBrowser – Linking Chemical Library Visualisation to CASP

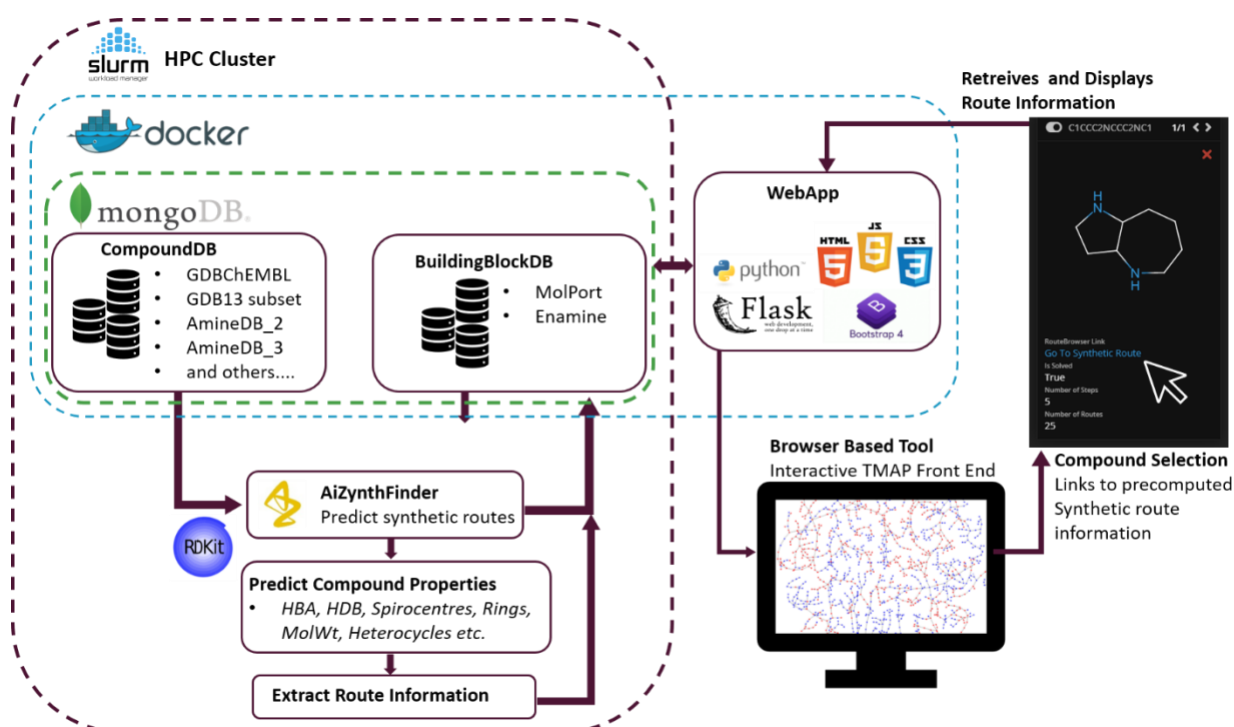


Figure 34: Architecture for the GDBRouteBrowser. Compounds from various chemical libraries are stored inside a MongoDB collection, called the CompoundDB. Likewise, the stock building blocks used for the CASP tool, in this case AiZynthFinder are stored in a separate collection, called BuildingBlockDB. Synthetic routes are precomputed using AiZynthFinder and distributed over a HPC cluster using SLURM arrays in an embarrassingly parallel manner. Computation of properties and extraction of route information from the the AiZynthFinder output is also conducted in parallel and uploaded to a MongoDB hosted on the server. TMAPs are then precomputed based on the MongoDB data and served via a Flask web application. The web application also serves requested route information and enables shortlisting of user chosen compounds. Two instances of MongoDB are used, one for development and HPC upload and a second MongoDB deployed on the production server, which is updated periodically by transfer of data from the development instance.

The GDBRouteBrowser is a browser-based tool developed in this study with the objective of unifying visualization and examination of chemical space, including the associated experimental data, computed properties, and synthetic route prediction. This is motivated by a need to reduce

barriers between available computational tools and facilitate experimental engagement. The architecture is shown in Figure 34. The GDBRouteBrowser's core functionalities are containerized using Docker. This facilitates hosting a MongoDB instance, which is connected to Flask web application using the Python programming language. The MongoDB serves as the backend for the web application and stores compound data, as well as precomputed synthetic routes using AiZynthFinder and property data computed by RDKit.²⁰ Non-RDKit properties such as those from experimental data, or those calculated by other models such as AiZynthFinder may also be stored within the schema. The compound database is used to compute the co-ordinates of a TMAP,⁸⁷ which uses the Faerun library for interactive visualization and creation of HTML pages embedding the property information.⁸⁰ The precomputed HTML is served by the Flask web application, and acts as the front end for visualizing the chemical library in an interactive manner. The TMAP displays the pre-computed properties by coloring the space according to the range of values available for a particular property, thereby enabling visualization of property distributions and their corresponding datapoints (Figure 35).

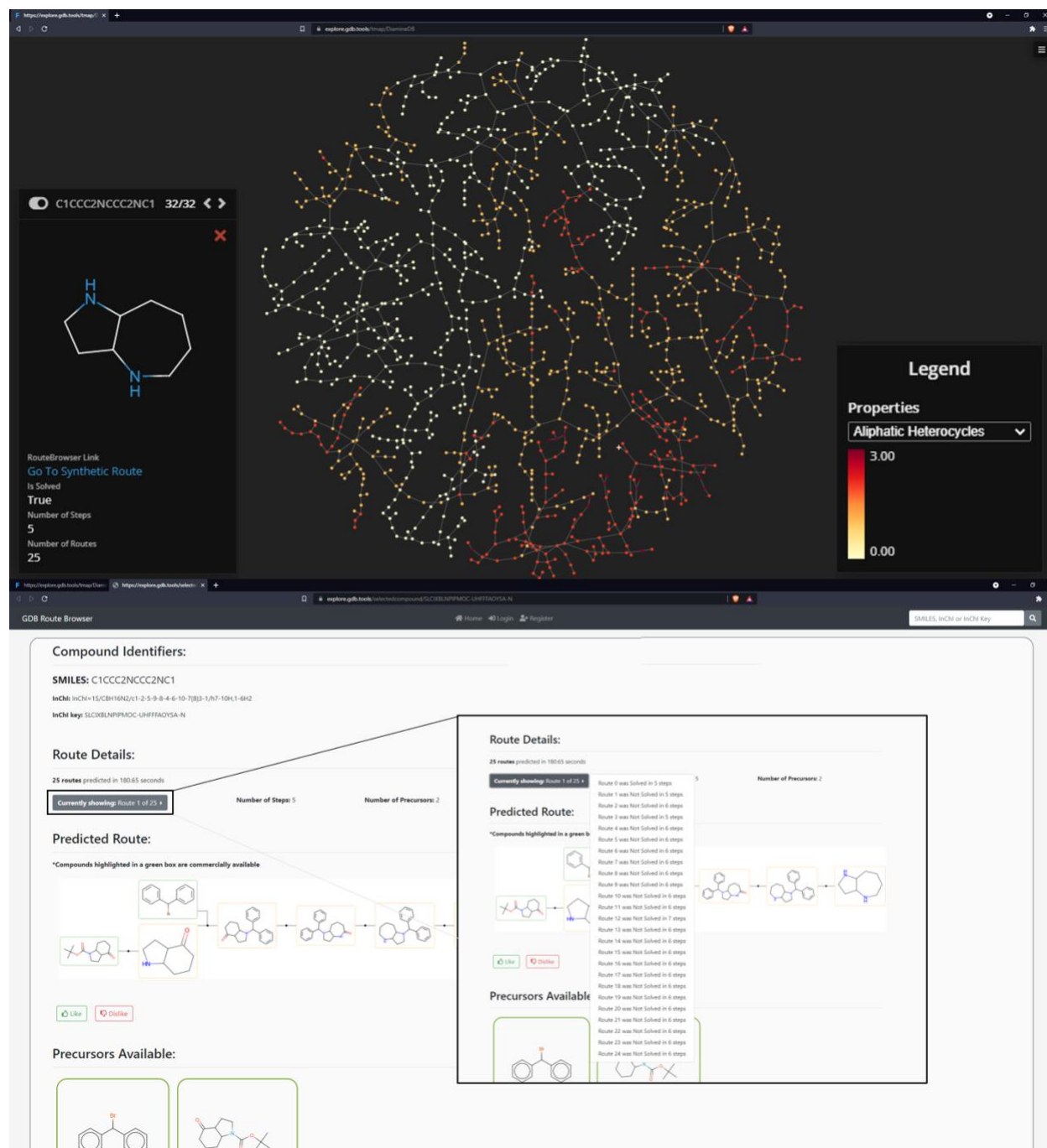


Figure 35: Left – TMAP visualization of the AmineDB_2 colored by the number of aliphatic heterocycles present (darker colors signify a higher number, and lighter colors lower numbers) as shown in the legend. Each point represents a compound, that when selected displays a card containing the compounds structure and chosen associated information. In this case there is a link to the synthetic route, whether a synthetic route to commercial precursors could be found using AiZynthfinder, and some information about the number of steps and routes predicted. Clicking the link opens another tab with route information (right). Right – Navigation to synthetic route information is provided through the TMAP or alternatively through search functionality. Options to view each predicted route and whether or not it was solved by AiZynthFinder is shown and the appropriate route chosen for display. The route is shown and may be expanded for ease of viewing. The precursors are also shown along with their SMILES to enable search in vendor catalogues.

Each compound in the visualization is represented as a spherical point. Clicking on any given point displays a card containing a summary of the compound's properties and a link to the synthetic route. The displayed summary statistics can be modified during TMAP generation, and links to the synthetic routes are embedded by addition of HTML during TMAP generation. The links provide a connection to an interface connected to the underlying MongoDB instance, which facilitates a richer display of information relating to the compound and pre-computed synthetic routes. Requests can be made to the server to display alternative routes. The new route selection is retrieved from the MongoDB and processed using AiZynthFinder's utilities to process the route into a temporary image file which is then served to the end user through the interface. During this process available precursors and their corresponding SMILES are extracted from the route, such that they can be displayed in the interface to facilitate easier searching in commercial catalogues for end users. In this study precomputation was used in lieu of dedicated compute facilities on the host server.

In place of AiZynthFinder, alternative CASP tools and the properties they compute can be included into the overall workflow by modification of the synthetic route prediction and route information extraction workflow. This would additionally require scripts to handle the output of alternate CASP tools and processing the predictions into images to facilitate viewing for the end user.

6.2.4 Compound Selection

Compounds were selected from the GDB databases investigated previously in our group. The subsets were chosen based on ongoing projects of interest within our group.

AmineDB_2: The database was derived from GDB4c containing up to 4 rings with a maximum of 14 atoms per ring.¹⁹⁸ AmineDB_2 was obtained by filtering and decorating the hydrocarbon rings present in GDB4c such that the maximum number of rings was limited to 2, with ring sizes between 5 and 7, and up to 2 amines in both exo- and endo- cyclic configurations. This resulted in a dataset of 1,323 molecules as obtained by an in-house algorithm designed by Josep Arús-Pous.

AmineDB_3: The database was derived from GDB4c containing up to 4 rings with a maximum of 14 atoms per ring.¹⁹⁸ AmineDB_3 was obtained by filtering and decorating the hydrocarbon rings present in GDB4c such that the maximum number of rings was limited to 3. This resulted in a dataset of 44,929 molecules as obtained by an in-house algorithm designed by Josep Arús-Pous.

GDB13_ABCDEFGH: The subset was obtained from a previous study concerned with the filtering of GDB-13. All filters used in the study were applied and resulted in a dataset of 994,840 molecules.²³⁸

GDBChEMBL_X: The GDBChEMBL database consists of ChEMBL like molecules as computed by the ChEMBL likeness score.⁴⁸ 100 M compounds were sampled from the GDBChEMBL database and filtered according to the following criteria. The number of hydrogen bond donors less than equal to 3 and the number of hydrogen bond acceptors also less than equal to three. The number of rotatable bonds less than equal to 3 and greater than or equal to 2 rings. The cLogP

must be in the range greater than equal to 1 and less than equal to 3. There must exactly one aliphatic or aromatic carbocycle in the molecule. The ring size had to be between 4 and 8, and no double bonds or quaternary centers were allowed. 1,490,508 compounds remained after filtering.

All compounds were subsequently subjected to retrosynthetic analysis using AiZynthFinder using a modified model as described previously. The AmineDB_2 was also predicted using the standard USPTO model.

6.3 Results and Discussion

To demonstrate the utility of linking synthetic route information to the navigation of chemical libraries I selected GDB subsets that were of ongoing interest to our lab. Specifically, I chose two amine databases, AmineDB_2 and AmineDB_3 consisting of a maximum of two and three rings respectively. These have previously been exploited by our group for the discovery of a novel potent Janus Kinase inhibitor,²³⁹ thus demonstrating the potential for GDB molecules to act as a source of building blocks for medicinal chemistry and drug discovery.²⁴⁰ Furthermore, our choice is governed by the potential for cyclic amines to act as the core for potent gamma secretase modulators as shown by Ratni and co-workers.²⁴¹

The databases were subjected to retrosynthetic analysis using AiZynthFinder (Table 8). AmineDB_2 has the highest number of compounds which could be solved by AiZynthFinder (Table 8), and there is a noticeable increase of 11.7 % when using the modified model considering pre-computed applicability labels compared to the standard USPTO model. AmineDB_3 has considerably less routes to commercially available precursors as determined by AiZynthFinder.

Table 8: Generalised overview of the results obtained by running AiZynthFinder on GDB subsets varying in their size and the types of compounds they contain, from cyclic amines to ChEMBL like compounds. The AmineDB_2 had the highest number of compounds for which routes could be found, and the AmineDB_3 and GDB_ABCDEFGH were comparable. The CPU wall time is expressed as the time taken for a computation on a single core with 4GB of RAM using an Intel® Xeon® E5-2360 v2 CPU, where all predictions were carried out on the CPU. The number of individual reaction steps predicted for GDB13_ABCDEFGH and GDBChEMBL_X is comparable to the number of reactions reported in the Reaxys® database.

Dataset	Compounds	% Solved by AiZynthFinder	Number of Routes	Number of Steps	CPU wall time (dd-hh:mm:ss)
AmineDB_2	1,323	56.2	12,207	59,987	02-08:30:29
AmineDB_2_USPTO	1,323	44.5	15,525	83,048	01-13:33:48
AmineDB_3	44,929	17.7	427,493	2,829,675	91-14:35:04
GDB13_ABCDEFGH	994,840	19.7	8,501,323	54,933,389	1362-17:27:44
GDBChEMBL_X	1,490,508	33.8	12,826,692	75,029,188	2025-03:59:35

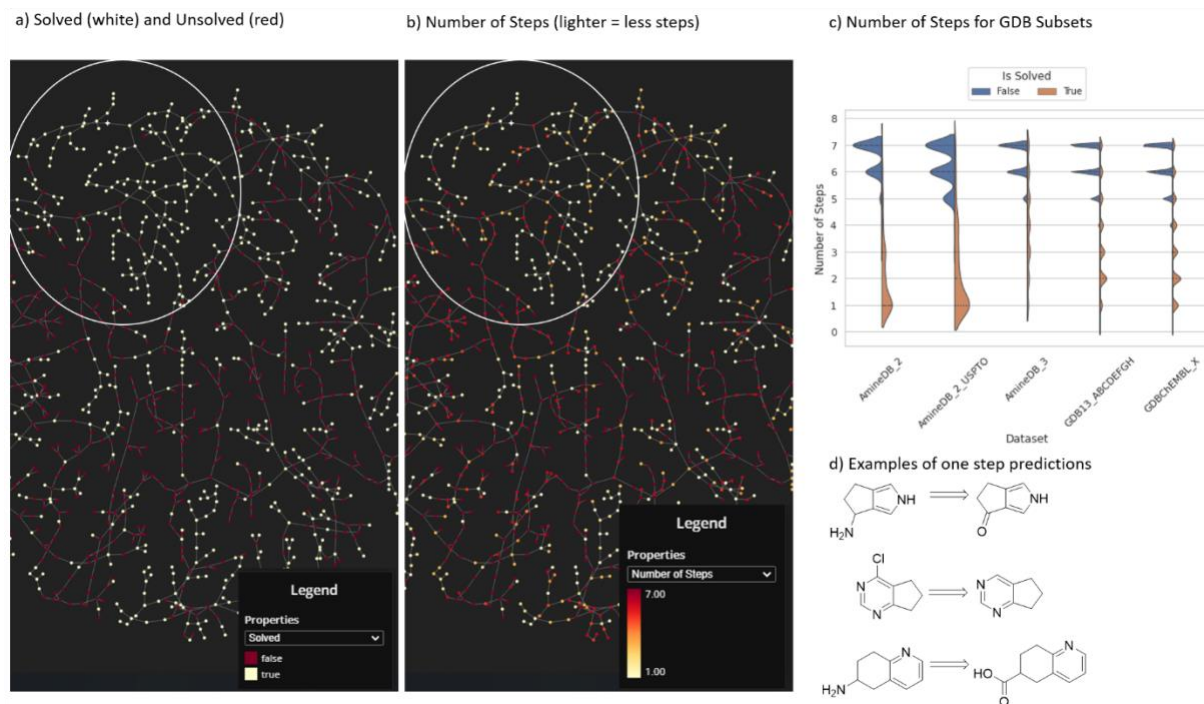


Figure 36: a) TMAP for the AmineDB_2 showing points representing compounds that were solved (white) and not solved (red) using a modified version of AiZynthFinder. b) TMAP for the AmineDB_2 showing the number of steps required to solve each compound, only considering the highest scoring route. Qualitatively we can observe that predicted routes were solved in either one step or the algorithm ran exhaustively, as shown by the white and red points for AmineDB_2. c) Distributions of the number of steps required for both solved and unsolved routes across all GDB subsets examined, and a reference curve for AmineDB_2 using the standard USPTO model. d)

Beyond these simple statistics the GDBRouteBrowser enables further investigation of the factors determining the success of synthetic route prediction. Consider AmineDB_2 (Figure 36), the TMAPs qualitatively show that the solved compounds require fewer steps to solve, as can be seen in the top left cluster of Figure 36a and Figure 36b. This is supported by quantitative assessment by examining the distributions of solved and unsolved routes for each dataset (Figure 36c). For each dataset, it is evident that when a successful retrosynthetic route is predicted, a shorter synthetic route is required in comparison to the unsolved compounds. This is governed by the stopping criteria, that a compound must be in stock for a route to be considered solved, thus the algorithm favors shorter routes leading to privileged scaffolds. The final compound can then be obtained via simple functional group interconversions as shown in Figure 36d.

Where privileged scaffolds are not available in the stock database, or the chemistry resulting in the privileged scaffold is not predicted, the algorithm must attempt to assemble the molecule. In the case of AmineDB_2 and AmineDB_3 the molecules consist of only carbon skeletons decorated by nitrogen in exo- and endo- cyclic positions. This poses a synthetic challenge as exo-cyclic nitrogen's as shown in Figure 36d create limited opportunities for ring assembly and functional group interconversions are often predicted. In this case the algorithm struggles to break down the carbon scaffold in most cases.

I further investigated the claim, that AmineDB_2 compounds containing solely exo-cyclic nitrogen's create limited opportunities for ring assembly and functional group interconversions are often predicted towards privileged scaffolds using the GDBRouteBrowser. Figure 37 shows a branch of a TMAP calculated for a subset of AmineDB_2 which was filtered such that it contained only compounds containing exo-cyclic nitrogen's, thus leaving behind a purely carbon core. Visual inspection of the TMAP led to the discovery of the branch shown in Figure 37, containing one exo-cyclic nitrogen at the ring junction. As can be seen in the examples in Figure 37, the exo-cyclic nitrogen is used strategically by AiZynthFinder in a Grignard reaction to alkylate the ring system prior to a ring closing metathesis step and reduction to form the final carbon scaffold. This was predicted for species *a* and *b* in Figure 37, however related species *c*, *d*, and *e* were predicted to be obtained through functional group interconversion of privileged scaffolds. Although synthetic routes breaking apart the aliphatic carbocycle were successfully predicted, selectivity issues may arise in the first two steps of the synthesis. Furthermore, on inspecting other branches of the TMAP only one further series was identified leading to disconnection of the aliphatic ring systems. The workflow of building a TMAP containing structures relevant to testing a hypothesis, and visually inspecting the routes and compounds could allow for rapid feedback to be obtained on synthesis prediction tools and enable related series and synthesis to be discovered.

I further examined molecules containing only endo-cyclic nitrogens to determine whether added functionality to the core enabled assembly of the ring system to be predicted. The TMAP generated for endo-cyclic nitrogen containing compounds was visually inspected and a branch containing a series of interest to our group identified (Figure 38). I again observe that where privileged scaffolds are available, the ring system may be obtained by functional group interconversions or deprotection strategies as is the case for compounds *h*, *i*, and *j* in Figure 38. Notably compound *g* is predicted to be obtained via a Beckmann rearrangement, a type of reaction commonly thought to be a weakness of template-based synthesis prediction methods. The choice of protecting group is non-standard, however this may be modified according to a synthetic chemist's experience, and the Beckmann rearrangement may be conducted from the commercially available N-Boc protected precursor. The route for compound *g* was unable to be predicted by the standard USPTO model, IBM RXN, and ASKCOS.^{138,242}

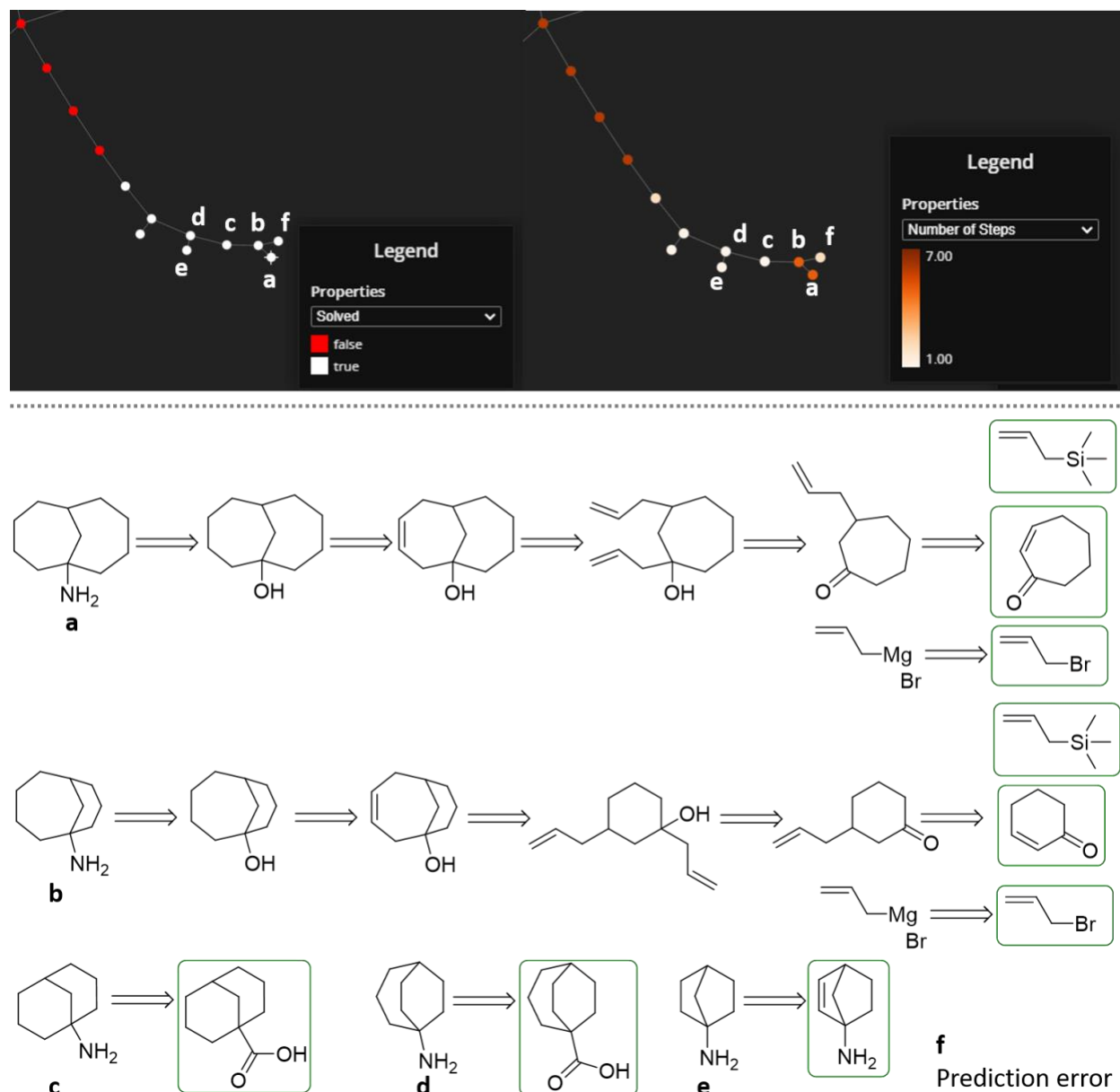


Figure 37: Top – TMAP calculated for a subset of AmineDB_2 containing only exocyclic nitrogens, colored by whether a synthetic route could be found by AiZynthFinder (left) and the number of steps (right). The TMAP is zoomed in on a branch detailing related compounds for which synthetic routes breaking down the aliphatic carbocycle core were predicted. Bottom – Examples of synthetic routes to related compounds in a series. AiZynthFinder predicts the same strategy to obtain compounds **a** and **b**, which breaks down the aliphatic carbocycle. However, for compounds **c**, **d**, and **e** AiZynthFinder predicts routes to privileged scaffolds where the final compound is obtained through a functional group interconversion. Compound **f** is a prediction error arising from the underlying templates and dataset.

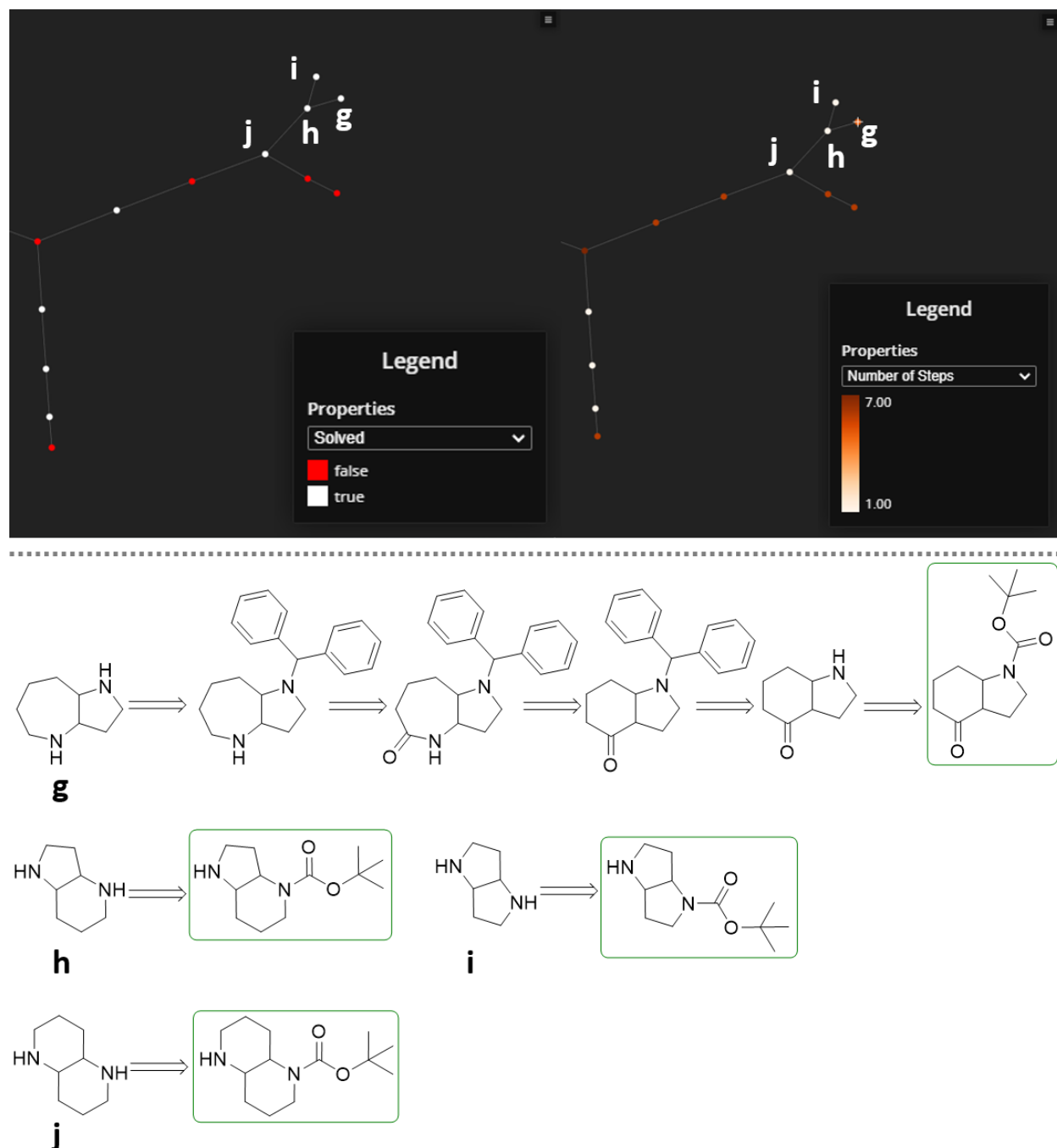


Figure 38: Top - TMAP calculated for a subset of AmineDB_2 containing only endocyclic nitrogens, colored by whether a synthetic route could be found by AiZynthFinder (left) and the number of steps (right). The TMAP is zoomed in on a branch detailing related compounds for which synthetic routes assembling the ring system core were predicted. Bottom – Examples of synthetic routes to related compounds in a series. AiZynthFinder predicts a Beckmann rearrangement to assemble the ring system from commercially available precursors. The choice of protecting group can be modified by the chemist. However, for compounds **h**, **i**, and **j** AiZynthFinder predicts routes to privileged scaffolds where the final compound is obtained through a functional group interconversion.

6.4 Conclusion

The GDBRouteBrowser is a browser-based tool linking the navigation of chemical libraries to synthetic route prediction. In this study, I have described the development and implementation of the GDBRouteBrowser, a tool combining compound library visualization with synthetic route information to enable exploration of chemical space to further experimental engagement.

In addition, I have demonstrated its utility as applied to GDB subsets to facilitate the exploration of chemical libraries, by both technical and non-technical users owing to its interactive display. I demonstrate a workflow by which the GDBRouteBrowser can be used to test hypotheses relating to the prediction of synthetic routes. This enables rapid feedback to be obtained from experimental chemists within one tool, thus facilitating the identification of areas for further development of the underlying CASP toolkit. Furthermore, I show that the GDBRouteBrowser can be used to identify series of compounds that may have related syntheses, thus can aid chemists in the ideation and compound prioritization process during a project.

The tool can be adapted to suit a project demands by overlaying synthetic route information, bioactivity data, or bespoke properties of interest as demonstrated in our case for exo- and endocyclic amines from AmineDB_2 and AmineDB_3. This proves particularly useful for the case of GDB molecules as their low molecular weight and structures are often shared with building blocks. Therefore, often syntheses are predicted in one-step from privileged scaffolds. Thus, using the GDBRouteBrowser, can enable identification of synthetically interesting and not yet explored structures.

Herein, I have presented a tool combining compound library visualization with synthetic route information to enable exploration of chemical space to further experimental engagement. While I have used AiZynthFinder, the concept can be applied to the predictions of any CASP tool providing they are uploaded to the underlying MongoDB instance following the appropriate schema.

6.5 Availability of data and materials

The code for the GDBRouteBrowser will be made available under an MIT license, as well as instructions on how to access the datasets: <https://github.com/reymond-group/GDBRouteBrowser>.

AiZynthFinder is open source and is available under an MIT license at: <https://github.com/MolecularAI/AiZynthFinder>.

Unfortunately, the models underlying data used for the modified AiZynthFinder model is not available to the public as they were obtained from proprietary sources.

6.6 Author Contributions

A. Thakkar and J.L. Reymond designed the concept. A. Thakkar designed and coded the workflow, conducted the research, and wrote the manuscript. J.L. Reymond supervised the project and contributed to interface usability, compound choice, and evaluation of synthetic routes.

7 Conclusion and Outlook

7.1 Summary

In this thesis I have examined the topic of computer aided synthesis planning as a means for augmenting chemical discovery and exploring chemical space. To do this, I co-led the development of a template-based synthesis planning methodology inspired from the works of Segler and Waller, that we call AiZynthFinder.

AiZynthFinder was first described in chapter 3 and uses automatically extracted templates, combined with neural networks and Monte-Carlo tree search to predict retrosynthetic routes to compounds ranging from drug-like molecules, to those obtained from low-molecular weight databases such as the GDB. The templates encode the reaction center, thus capture the atoms and bonds that have changed because of the reaction. Where previous approaches at the time relied on manual encoding of chemical reactions as templates, the methodology I have utilized, uses an automatic template extraction procedure negating the need for manual encoding. In chapter 3 I improved the template extraction algorithm by encoding key functional and protecting groups commonly used in organic synthesis. This was an improvement over the existing algorithms, which considered atoms and bonds from a given radius from the reaction centre and had limited functionality for recognizing where atoms and bonds contributed to functional or protecting groups. To assess the quality of the extracted templates, I developed a metric based on the ability to reproduce the data from which the templates were originally extracted. Using this metric, I was able to improve the template extraction process.

Subsequently, I examined how the extracted templates radius affected the downstream task of retrosynthetic planning and determined that owing to the specificity of the templates extracted at larger radii, the lower their performance when conducting full retrosynthetic analysis. This means that although a reaction may be present in the training set it cannot be applied *in silico* due to a sub-structure mismatch, a theme that repeatedly occurs through this thesis. Along with this finding I also found that accuracy is a misleading metric for determining the performance of a model to predict full retrosynthetic pathways. The reason for this is because accuracy does not account for *in silico* template applicability, nor does it account for the validity of each of the top N predictions. Thus, I propose two alternate metrics for template-based retrosynthesis prediction. Firstly, the number of applicable templates in the top N, as this must be maximized to ensure efficiency in the subsequent tree search and informs us as to how well the neural network is prioritizing applicable chemistry for the input product. Secondly, the overall performance towards a specified objective function during full synthetic route prediction, in our case the availability of a precursor in the commercial stock database. The latter metric can also be used for assessing template-free

retrosynthetic analysis, as it is methodology agnostic and requires assessment of whether a synthetic route was found towards a given objective.

In chapter 3 I further established that 2 % of the templates were shared when examining public and proprietary datasets. Whilst this represents a relatively small portion of the substructures that have been used historically in reactive chemistry, I found that using as low as 4.8 % of the available templates was sufficient for synthetic route prediction in the case of drug-like molecules. This occurs due to a bias or frequent use of certain reaction types, an observation which is also confirmed by other studies in the field. I have further investigated this finding by conducting retrosynthetic analysis on virtual libraries obtained from AstraZeneca, where I found a smaller set of templates was needed for compounds originating from combinatorial libraries, compared to those designed by a medicinal chemistry team. chapter 3 ends with the first description of AiZynthFinder, an open-source retrosynthetic planning software that was later published as an independent article after code refactoring and is currently deployed for use in AstraZeneca.

During field testing of AiZynthFinder within the AstraZeneca chemistry community, I determined a key bottleneck in synthesis prediction was predicting the formation of ring systems. To address this issue, I developed a domain specific model in the low data regime called RingBreaker which is outlined in chapter 4. While AiZynthFinder can be used to predict full retrosynthetic pathways to molecules of interest, the RingBreaker can be used interactively to predict single step ring forming reactions, as deemed appropriate by a chemist. As such it is currently deployed alongside AiZynthFinder in an interactive mode for use within AstraZeneca, however the code and a model based on publicly available data is available for community use.

A key technical development outlined in chapter 4, is the shift to a multi-label approach for single step retrosynthetic prediction. The reason for this is intuitive. There are several disconnections that are possible for a given product, thus several reactions and precursors that could be predicted. The single-label approach assumes that one product has one reaction that can be predicted for its disconnection, however on examination of the reaction datasets I found that there are multiple reactions leading to a given product. These can be encoded into the same label vector for a given product, and this has two consequences. The training no longer scales with the number of reactions in a dataset, but scales with the number of products, thus improving training times as the number of training examples for each epoch decreases while still retaining all the information as the output vectors become less sparse. Using the multi-label approach, I found that using public and proprietary datasets were complementary in the case of predicting ring formations, notably, in cases where one fails, the other is often able to suggest a suitable disconnection. As the training examples from which the templates were extracted come from substituted ring systems, it follows that attempting to use RingBreaker on unsubstituted ring systems does not always work. The reason for this is a sub-structure mismatch which prevents *in silico* applicability. Nevertheless, the RingBreaker was able to predict disconnections in line with the literature as shown for the “rings of the future” and is currently used to aid synthetic chemists in situations where AiZynthFinder fails, or where retrosynthetic ideas are needed from a starting point containing a ring system.

As an artefact of the RingBreaker study I observed that several templates were being applied outside of their original context, as described by the reaction database. These out of context templates led to the successful prediction of literature disconnections in the case of RingBreaker. Based on our findings of brute force *in silico* template application in chapter 4 I determined that this approach could be applied more generally to the overarching AiZynthFinder. This led to a study on generating artificial applicability labels for improving retrosynthetic prediction,¹⁴⁴ the model for which I use in chapter 6 to predict a Beckmann rearrangement which is not possible using the standard single label AiZynthFinder model, nor with other state of the art CASP tools such as IBM RXN and ASKCOS.^{138,242} We found training a model on pre-computed applicability labels,¹⁴⁴ enabled more efficient prioritization of *in silico* applicable templates and improved overall retrosynthetic route finding using the metrics described in Chapter 3.

Despite several algorithmic improvements compared to previous approaches, the prediction of full retrosynthetic pathways can be computationally expensive, thus limited by resource availability. This prohibits the application of retrosynthetic analysis to large molecular databases such as the GDB, and generative models that optimize for synthetic feasibility. To overcome this issue, I have developed a proxy model called the Retrosynthetic Accessibility score (RAScore) for the scoring of molecules as synthesizable or not, based on the predictions of AiZynthFinder. In addition, I showed that existing scores for synthetic accessibility do not resolve solved and unsolved compounds, i.e., those that can and cannot be synthesized from available building blocks. Thus, I introduce a metric called average linkage, which determines the average distance between all pairs of items in a cluster, in our case solved and unsolved compounds. I used the average linkage to assess a variety of machine learning based models along with an automatic machine learning (auto-ML) toolkit built specifically for this study and were able to identify a classifier that enables retrosynthetic accessibility to be estimated at least 4,500 times faster than running full retrosynthetic analysis. Therefore, opening the possibility to score large datasets of molecules in a reasonable time frame. The RAScore was restricted in applicability domain, thus for usage alongside the GDB dataset, the GDBscore based on a subset of GDB molecules was developed instead. The RAScore is currently deployed within AstraZeneca and the code and models are publicly available. It is recommended that the classifiers be retrained on samples of the compounds of interest to the user to enable more accurate predictions.

Finally, to ensure accessibility and ease of use of the developed tools for synthetic chemists, and to aid in the prioritization of GDB compounds, I developed the GDBRouteBrowser in chapter 6. The GDBRouteBrowser combines chemical library visualization with synthetic route prediction, to enable chemists to obtain a deeper understanding of the chemical library they are interested in. This is accomplished by overlaying properties on a graphical representation of the chemical space called TMAP, that can be calculated or predicted, such as physiochemical, bioactivity, or those derived from synthesis prediction approaches. The GDBRouteBrowser can be tailored to the needs of a particular project or individual to facilitate compound prioritization. I demonstrate a workflow by which the GDBRouteBrowser can be used to test hypotheses relating to the prediction of

synthetic routes. This enables rapid feedback to be obtained from experimental chemists within one tool, thus facilitating the identification of areas for further development of the underlying CASP toolkit. Furthermore, I show that the GDBRouteBrowser can be used to identify series of compounds that may have related syntheses, thus can aid chemists in the ideation and compound prioritization process during a project. While I have used AiZynthFinder, the concept can be applied to the predictions of any CASP tool providing they are uploaded to the underlying MongoDB instance following the appropriate schema.

Each of the AiZynthFinder, RingBreaker, and RAscore tools that I have developed and contributed to through this thesis are deployed internally for use within AstraZeneca, where they can be used as a standalone tool for retrosynthesis prediction, used interactively to explore single steps by a chemist, or used to score generated molecules. In addition, an open-source version of each tool based on public data is available for general use and development by the community.

7.2 Outlook

Computer aided synthesis prediction has seen a resurgence of interest since 2017, with the application of neural networks to chemical reaction prediction and retrosynthesis.^{135,243} Unlike historical attempts the current environment marks a moment where tools are becoming more widely available both programmatically and through user interfaces deployed for the public and internally within industry. This has been made possible through developments and accessibility of machine learning algorithms, improving digital literacy and education, improved hardware, open-source data, and the widespread usage of the internet as a means for distribution. Despite our current progress to deliver functional synthesis prediction software, there is still much work to be done to meet the demands of practicing chemists. Furthermore, the implications of predicting chemical synthesis are far reaching, and reach far beyond the scope of aiding laboratory chemists. This last section will outline some of the future developments that may be required for further adoption of synthesis prediction technologies.

7.2.1 Data

Chapter 3 discussed the importance of datasets in chemical synthesis prediction. Chemical data forms the bedrock on which machine learning or other algorithmic techniques can be applied. Unfortunately, reaction data is not always consistent in annotations and is biased towards the most frequently used and positive outcome reactions.⁹⁵ Notably, negative outcome reaction data are not often recorded, and their classification may be unreliable. This is because there can be multiple reasons for a failed or negative reaction, not least including the decision of the researcher to stop the reaction due to changing project demands, and cases where a researcher may accidentally lose material, resulting in unreliable data entry. Whilst these problems are bound to occur, methods and standards for reporting chemical data through electronic laboratory notebooks should be improved. One such framework that could improve reporting methods are the FAIR principles for data

management.²⁴⁴ Currently, electronic laboratory notebooks can be quite restrictive regarding the type of experiment conducted, often favouring single experiment entry over high throughput screening and successive design of experiments (DOE) for the identification of an optimal set of conditions. Furthermore, it is difficult for ELNs under their current implementation to capture multi-step synthesis. Therefore, it is not possible to easily extract full synthetic routes. This is also a problem in commercially available reaction datasets where single experiments are reported, and the complete synthetic sequence is not easily known. Whilst modifications of ELNs can be made to facilitate high throughput data capture, additional data that is generated during an experiment is difficult to capture or is captured on an independent platform. Consider the recording of heat and mass transfer, reaction quenching and work-up, or particle size during recrystallisation, which is often captured on proprietary software provided by the hardware vendor.

In addition, reporting reaction data in ELNs is a mixture of free-text and specific fields for data entry. Increasing the number of fields imposes a rigid reporting structure at the expense of adaptability, whereas free text is often not parsed during ELN export for use by informatics teams. Therefore, a degree of flexibility is also required in reaction reporting and data capture. To improve existing reporting schemes, methods for extracting and obtaining chemical data are being examined, ranging from natural language processing (NLP) which can tackle the free-text problem, to high throughput experimentation (HTE) and continuous processing which can yield negative examples on which to train subsequent models.^{100,188,245} Whilst current ELNs may not necessarily be suited for HTE data, the role of HTE as a method of collecting large amounts of data, faster, has been recognized.²⁴⁶ Furthermore, efforts to create a public repository for reaction data are underway in the community, as well as discussions on schema and the redevelopment of laboratory notebooks to improve data capture.^{96,247–250}

During early and late-stage chemical development there are several criteria to take into consideration aside from the feasibility of a particular reaction. The cost of materials, the purge of impurities, knowledge transfer protocols for scale up, purification and crystallization, stability of active pharmaceutical ingredients, and kinetic modelling to name a few examples may be stored in separate documents. The outcomes of route-finding campaigns are stored as individual entries in ELNs or documented in PowerPoint presentations and pdf documents. Thus, there is a need for a repository for this information which is machine readable. While it may be the case that not all the data will be used for modelling reasons, there is a need to improve infrastructure surrounding the storage and analysis of critical process data.

Improvements in publishing practices could also prove beneficial to structured data repositories of reaction information. For instance, consider the compendium of synthetic routes published every year for approved drugs.^{179,251} These could be used as benchmarking sets for quality of synthetic routes, as well as providing a baseline to improve upon. Yet they are not available in a format which is easily machine-readable, despite being a source of valuable information, they remain locked in the literature. The same can be said for specialized areas of chemistry, in which there are several reviews outlining structure–reactivity relationships.²⁵² These reviews can provide a wealth

of information to which physics-based modelling can compare to experiment, and data-driven modelling learn relationships between structure and function. The subject of open source publishing and reproducibility has also been subject to ongoing debate.^{253–255} There have been several improvements in this regard during recent years, however there remain several opportunities to facilitate this process including the use of code and data sharing platforms.

In principle, one can imagine a future of shared data infrastructure that is privacy preserving, has community governance, is accessible and easy to use by both modelers and experimentalists. The MELLODDY (Machine Learning Ledger Orchestration for Drug Discovery) has shown that privacy preserved data sharing and pre-competitive model building can be achieved, and the Open Reaction Database, has shown that there is community interest in shared data infrastructure.^{96,256}

7.2.2 Modelling Chemical Reactions and Pathways

For a model to learn from data, the data must be represented in a way that the model can understand. At present the three representations commonly used are, fingerprints, SMILES, and graph representations. Of the three, graph representations are better able to capture three-dimensional information and have shown to be useful for studying protein folding.²⁵⁷ While fingerprint and SMILES based approaches have been used to model regio-, chemo-, and stereo-selectivity,^{258–260} introduction of 3D structure may improve predictive performance through consideration of a molecule's conformation. However, consideration of a molecule's conformation is non-trivial and varies according to the solvent, and other species present in the environment and is governed by molecular interactions. While approximations may be possible using force-fields, and geometry optimizations through quantum chemistry packages, the latter can be computationally expensive. Although research is underway into the paradigm of quantum mechanics – machine learning (QM-ML) which promises to improve the speed at which reactivity may be modelled.^{261–263}

Coupling existing machine learning approaches for synthesis planning with physics-based models could allow for fast computation of synthesis pathways augmented by physics-based evaluation to score their validity.¹²⁶ This could take the form of evaluating competing reactivity pathways based on reaction conditions, to determine potential impurities and the ratios in which products are formed. Furthermore, current synthesis prediction approaches are restricted to one transformation per prediction and do not necessarily account for global reactivity. For instance, accounting for multiple reactivity sites, in the case of protection and deprotection, or oxidation and reduction reactions, which may affect more than one functional group.¹²⁶ The combination of physics-based approaches with machine-learning could additionally be used for improving catalyst design, for more sustainable, cheaper, and higher performing catalysts.^{264–266}

With sustainability and catalysis in mind, recognition, and incorporation of bio-catalyzed reactions where appropriate, in place of chemical transformations could enable environmentally friendly processes.^{267–270} Furthermore, building a network of chemical reactivity, as in the case of the Lapkin and Grzybowski groups has shown promise for determining strategic building blocks, as

well as navigating intellectual property issues.^{127,161} The implications of this technology, can be extended to the navigation of chemical supply chains, predicting future supply, and identifying commonalities in chemical manufacturing workflows.^{271–275} Both single step and multi-step pathways can be extracted from such reaction networks, and cost-benefit analysis conducted. Accurate pricing models could enable better decision making both for academic and industrial labs, enabling cost and supply based route scoring.

For single step reaction and retrosynthesis prediction, the field is still lacking a full comparison of the strengths and weaknesses of the various models. Template-based models require an extraction step in between the data and model building, whereas template-free models can be trained directly on the data. Therefore, while template-free models appear more scalable in the first instance, as they learn directly from data, this may not necessarily be the case as highlighted in the conclusion and chapter 4. However, where template-based models were thought to be more explainable, as templates can be linked directly back to reaction precedents, this can also be achieved for template-free models using a similarity-based approach and reaction fingerprints.^{276,277} Additionally, the wide use of natural language processing-based approaches, and an effort to understand their predictions has led to the development of tools capable of looking into the predictive model, so that it is no longer a black box.²⁷⁸ Template-free models also have the advantage that they do not rely on substructure matching, thus may be better able to learn patterns in complex chemical reactions such as rearrangements. However, the applicability domain of template-free models is potentially more restrictive than template-based models which can apply templates outside of the initial training domain.^{144,145} The diversity of molecules that is seen by template-based models can thus be extended by application of templates to hypothetical structures as an extension to our previous work on augmentation with artificial labels, which we show increases the performance of retrosynthetic prediction.^{144,145} Implementation and investigation of alternate tree search strategies, along with the development of scoring schemes to suit the area of application is another aspect which remains to be studied more widely.²⁷⁹

Whilst the above is not a comprehensive summary of the improvements that can be made in the modelling of synthesis planning, future developments should aim to become interoperable. This can be achieved through standardized formats for reaction routes and synthetic steps. At present research addressing each aspect of synthesis planning is siloed, and improving reporting standards and working towards interoperability, would enable the creation of modular open-source synthesis planning tools.

7.2.3 Education and Collaboration

Educating computational researchers regarding the problems faced by experimentalists is not enough. For effective, translatable models to be built, computational researchers must be able to understand the requirements of experimentalists. This means that each party needs to be familiar with basic vocabulary from the others domain such that they can define the features and metrics

required to push the field forward. Metrics do not have to be computational, the propensity to reduce everything to a number via a computational method is all too appealing for a computational researcher, as this enables cheap simulation of the problem and its optimisation. Experimentation is not cheap and is multi-parametric, thus suitable definitions of experimental successes need to be defined. Digital literacy among experimentalists and experimental literacy among computational researchers is key. An understanding of each other's domain may also help to ease the tension that computers will replace the chemists. The goal is to augment and enable the chemist and not replacement.

Several consortia and pre-competitive collaborations have been formed to train the next generation of scientists, and to educate industrial counterparts. This thesis is the result of one such collaboration for big data in chemistry, BigChem.^{280,281} Other consortia deploying synthesis planning models across a broad range of chemical companies are the MIT-MLPDS and C-CAS consortia based in the United States.^{282,283} SynTech, AI3SD (Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery), and AIDD are European based initiatives aiming to incorporate the latest computational advances into scientific discovery.^{284–286}

Despite the formation of consortia and public funding mandates for open-source data and code, current research is not always interoperable. Therefore, at present the wheel is often reinvented albeit with a blueprint, to incorporate research findings into an academic or industrial lab. Moving towards interoperable and reusable frameworks would allow researchers in the field to focus on their overall objective and build on ideas. In this regard we can look towards the field of computer science for inspiration. The open libraries that we, as computational researchers use routinely to build useful tools are a network of smaller interoperable projects. From the field of cheminformatics, the RDKit is one such example.²⁰

7.2.4 A Personal Vision for the Future

This is the beginning of a journey. To accelerate the way chemical discovery is conducted. An interdisciplinary approach bringing together experts from different fields is required. In addition, emphasis should be placed on the ease of use, interoperability, and accessibility of the tools that are developed. Successful approaches may be characterized as those with a shallow learning curve for the user, a rich data source for the developer, and tight-knit integration throughout the community. The approach should also be scalable, adaptable, reliable, interoperable, and most importantly, meet the needs of the end user.

The overarching goal is to create an improved ecosystem that is: open, community governed, privacy preserved, has a low barrier to entry, makes data, insights, and automation accessible, while allowing for the continuous changing of incentives to fuel a community-based effort to solving societal and scientific problems.

*"You can't go back and change the beginning, but you can start where you are
and change the ending."*

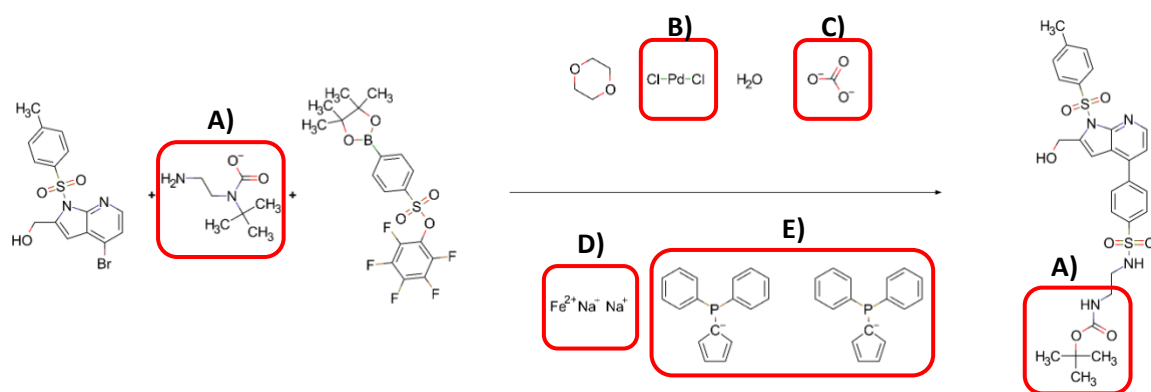
- C.S. Lewis

8 Appendix

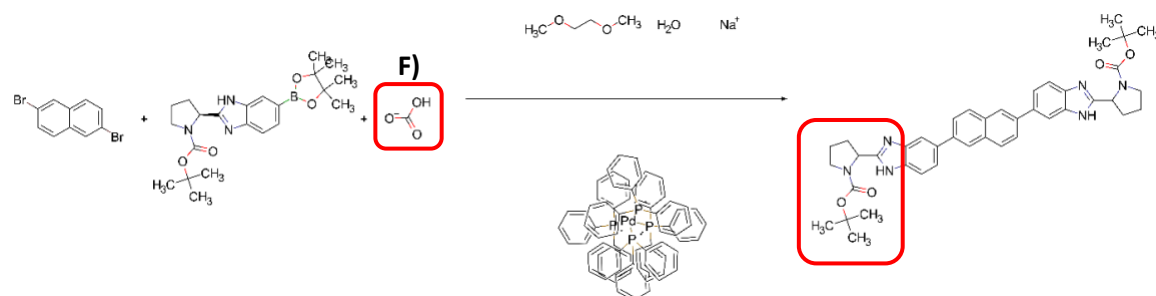
Appendix A Datasets and their influence on the development of computer assisted synthesis planning tools in the pharmaceutical domain

Appendix A.1 Data Inconsistencies

Data inconsistencies in the USPTO dataset which highlight a wider problem with reaction data. These have been filtered out in our approach; however this is not exhaustive.



Source: US20080146606A1



Source: US20130317213A1 [0753]

- A) Incorrect recording of the Boc protecting group, frequently used in organic synthesis. In addition, the charge is not balanced in the reactants.
- B) Palladium dichloride is often used to form the active catalyst in situ, however it is not clear which ligands are to be associated with the metal. A chemist can infer the active species;

however, the computer must be informed which species are grouped together. This is possible using ChemAxon extended SMILES (CXSMILES), which contain information regarding the grouping of constituent parts in the reaction. For our task these do not correspond to the changing molecular environment during the transformation, therefore are not included in the templates. As such, for the task of retrosynthesis, catalyst representation can be ignored, however is a key factor in reaction and condition prediction, so cannot be overlooked with respect to the wider field.

- C) Unidentified salt lacking annotations for its utility in the shown reaction
- D) Iron salt without corresponding ligands. Can be used to form a catalyst in situ or may come from a pre-formed/commercially available catalyst. It is not clear what role the species plays from a computational perspective.
- E) Phosphorus based ligands which do not contribute to the changing atoms and bonds in the reaction. It is not clear to which metal the ligands bind nor their role owing to missing annotations.
- F) Incorrect atom mapping arising from unbalanced reaction stoichiometry. Atoms in a species sharing substructure with the reactive species can have mislabeled atoms, thus the algorithmic extraction produces an incorrect template.

Appendix A.2 Top 10 Templates Across All Datasets

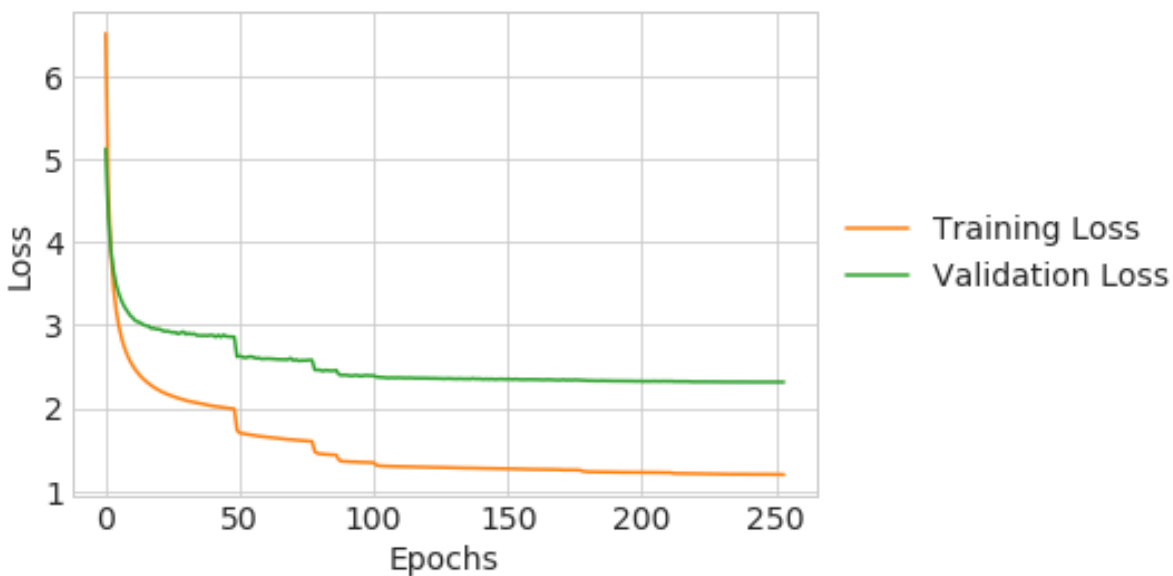
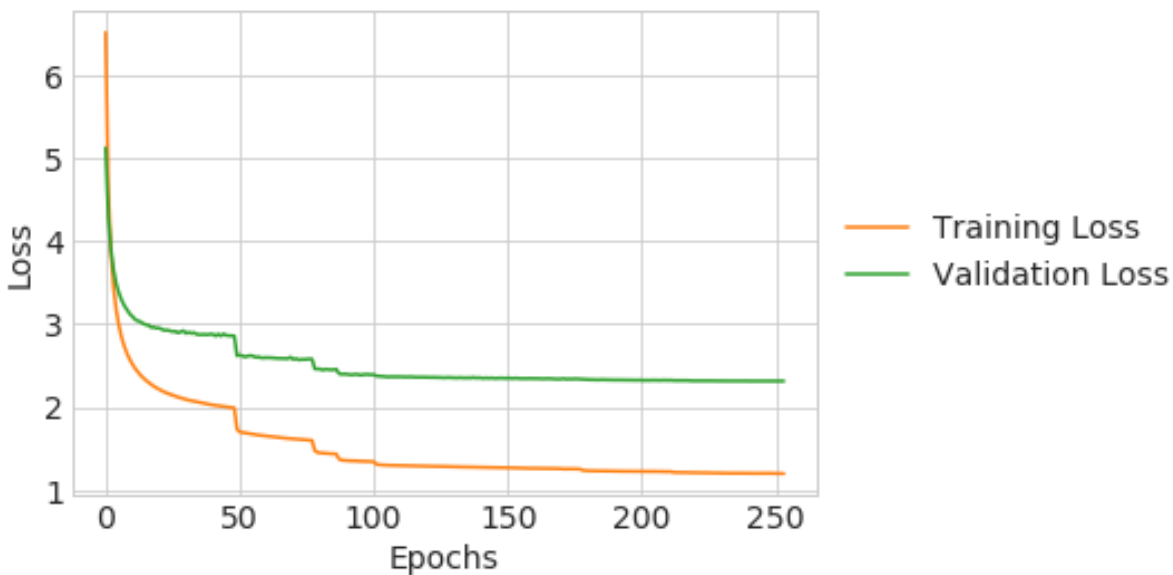
Top 10 templates across all datasets – csv file attached to publication

Appendix A.3 SMARTS Encoding Protecting and Functional groups

.txt file containing SMARTS patterns of the ca. 70 functional/protecting groups used

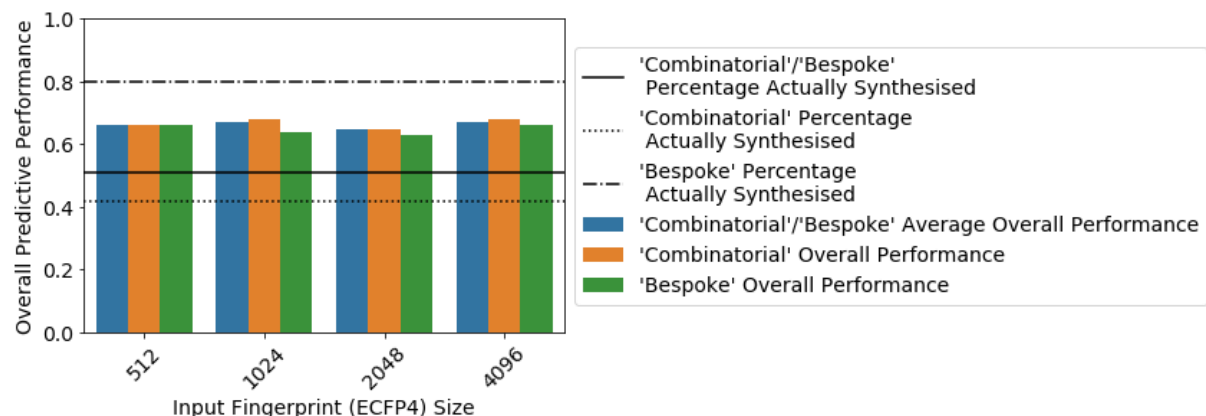
Appendix A.4 Example accuracy and loss curves

Accuracy and Loss curves for the model trained on the USPTO dataset, filtering for templates that occurred a minimum of three times.



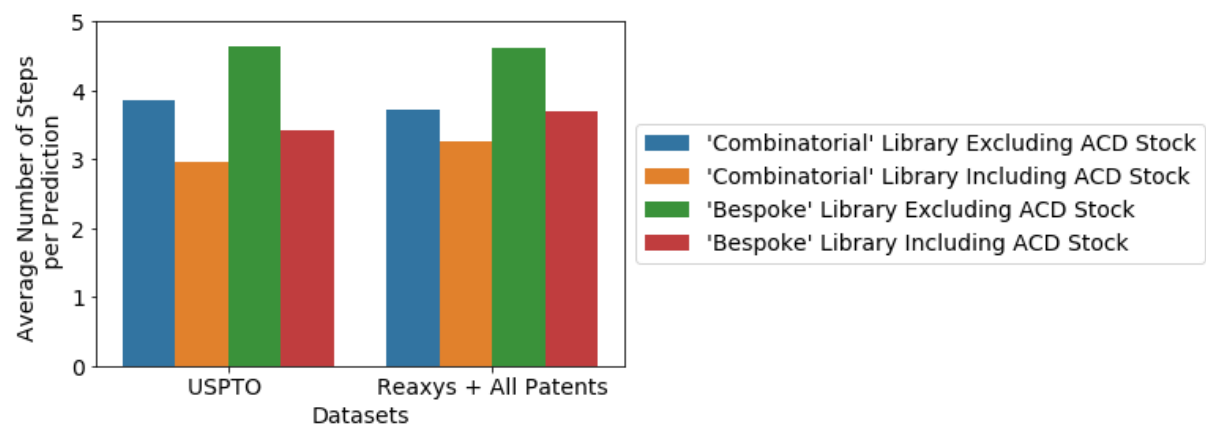
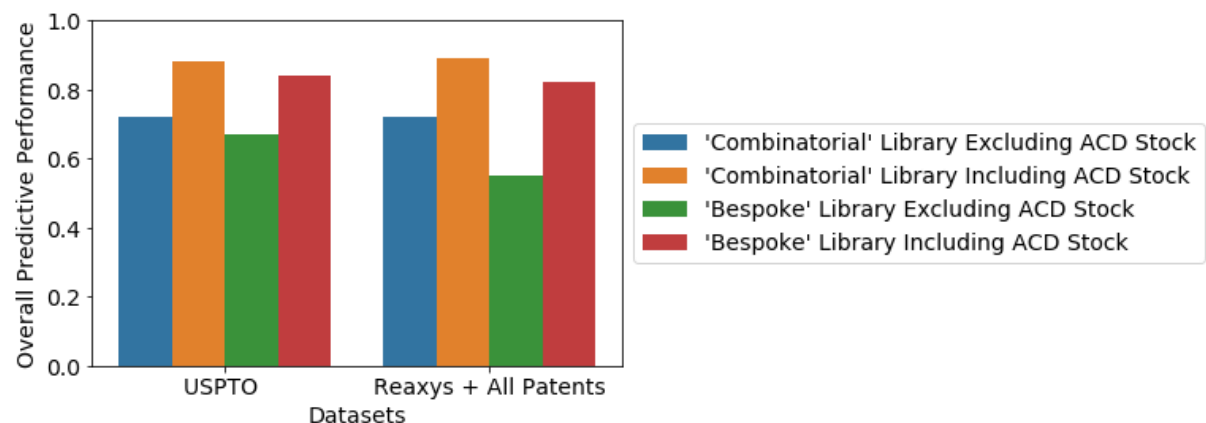
Appendix A.5 Input ECFP4 fingerprint size and performance

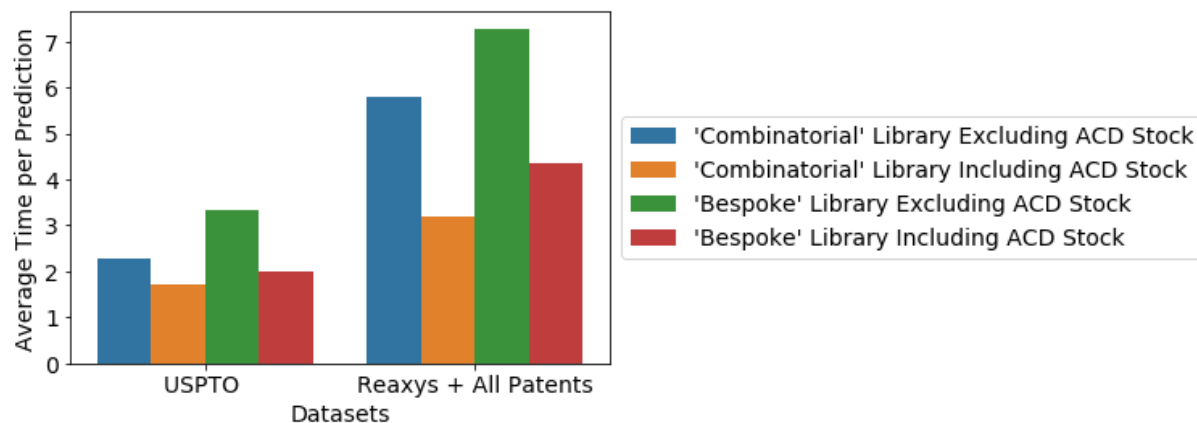
Comparison of the accuracy of models trained on different fingerprint sizes for the USPTO dataset to the iTrax virtual library dataset.



Appendix A.6 Performance and stock set of compounds

The performance of the model increases regardless of the dataset used when a larger stock set of compounds is used as an end point. Of note is the time taken to find full synthetic routes to the target compounds, where a larger stock set performs better.



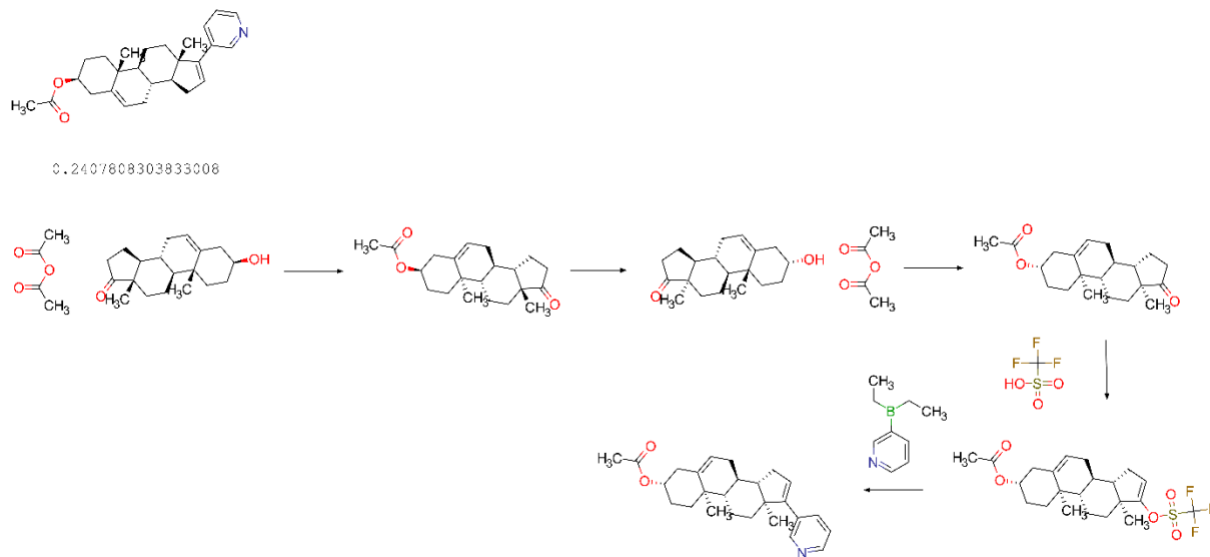


Appendix A.7 Exemplary Synthetic Routes

The following synthetic routes were found by the model trained on the USPTO dataset filtering for templates that occurred a minimum of three times. The AstraZeneca internal stock and Enamine building blocks were used as the stock set of compounds. The syntheses shown are for compounds in the top 125 small molecule therapies of 2018, where 47 % (59 of 125) compounds were present in the USPTO dataset. While, in some cases these have been solved faster than those that are not present, it is evident that there are selectivity conflicts and synthetic strategies employed by the algorithm that a trained chemist would avoid. These have not been filtered and the raw output of the top ranked route is given to illustrate the drawbacks of such a naïve route search yet exemplify that a naïve algorithm may in some cases still propose plausible routes.

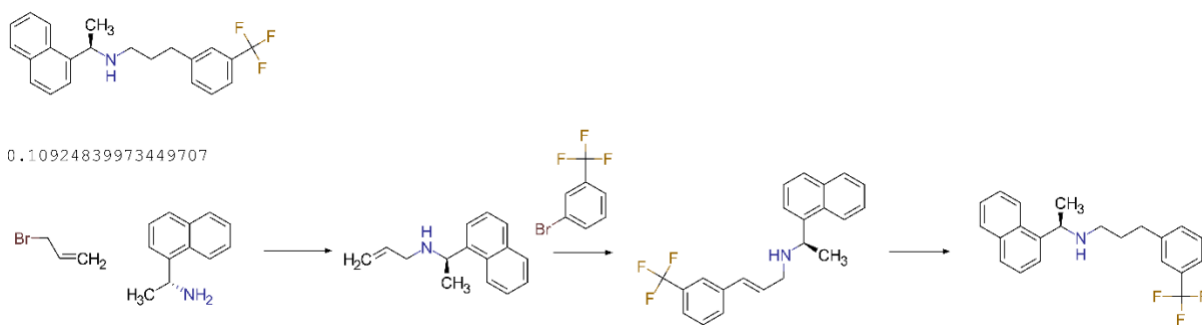
Compound: Abiraterone Acetate

Time to solved: 0.24 seconds



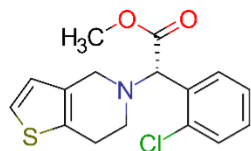
Compound: Cinacalcet

Time to solved: 0.12 seconds

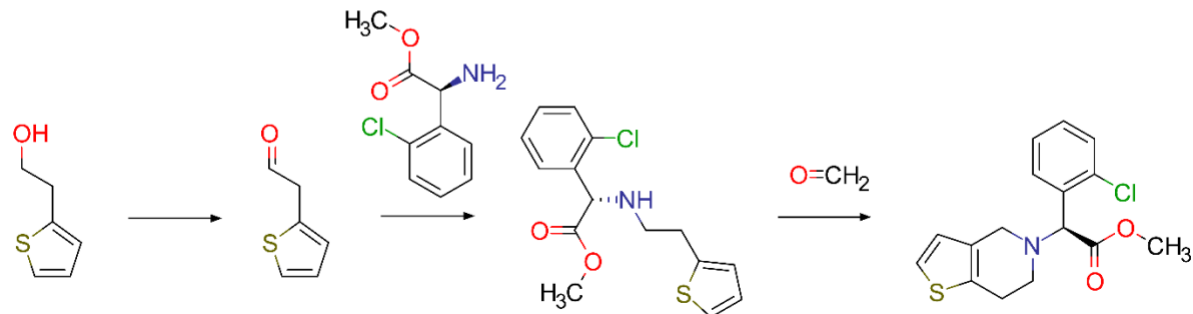


Compound: Clopidogrel

Time to solved: 7.59 seconds

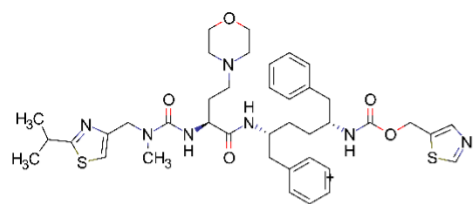


7.594855070114137

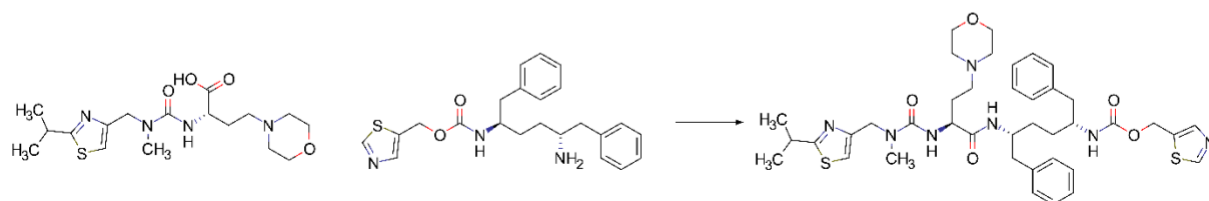


Compound: Cobicistat

Time to solved: 0.58 seconds

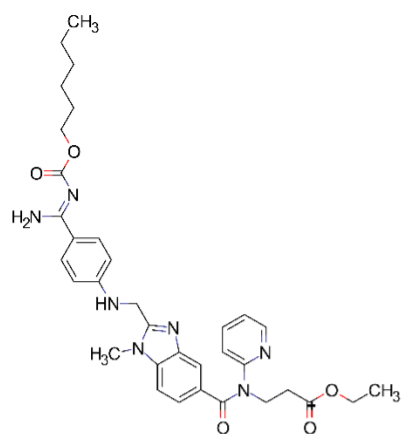


0.5831310749053955

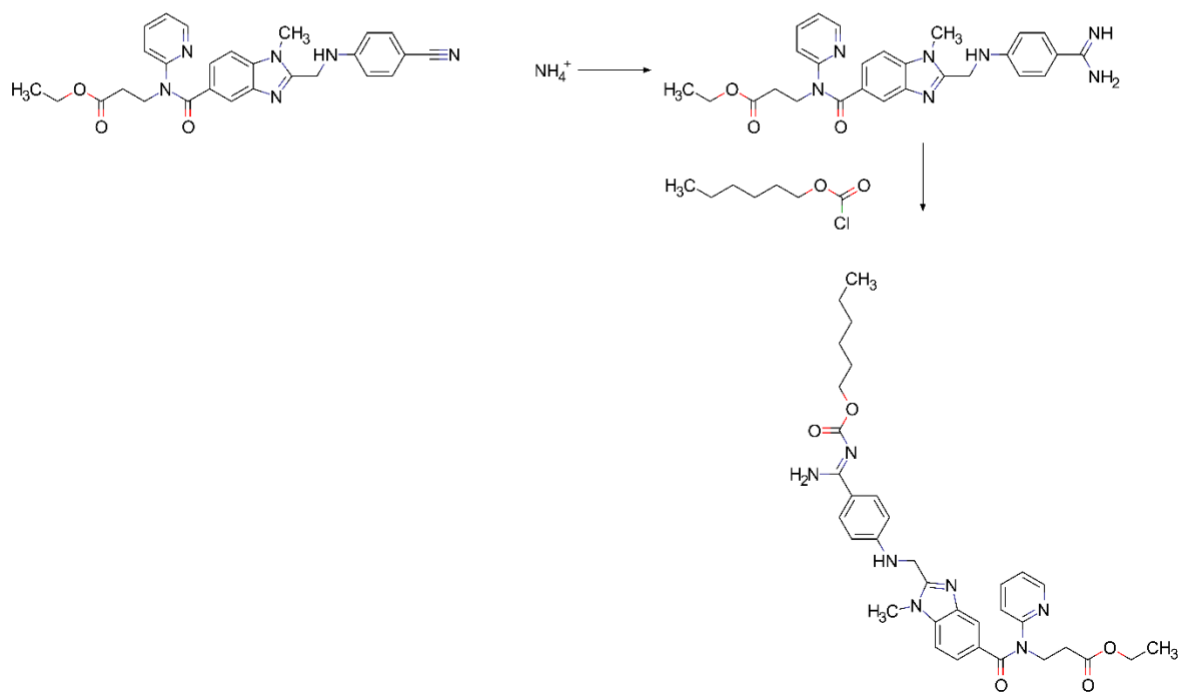


Compound: Dabigatran

Time to solved: 0.07 seconds

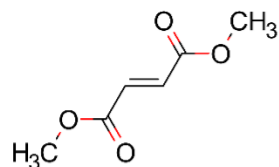


0.07490253448486328

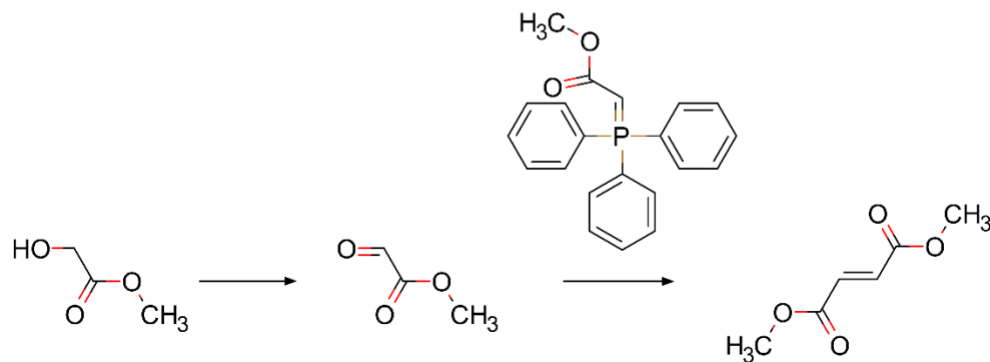


Compound: Dimethyl Fumarate

Time to solved: 1.25 seconds

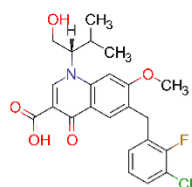


1.2493224143981934

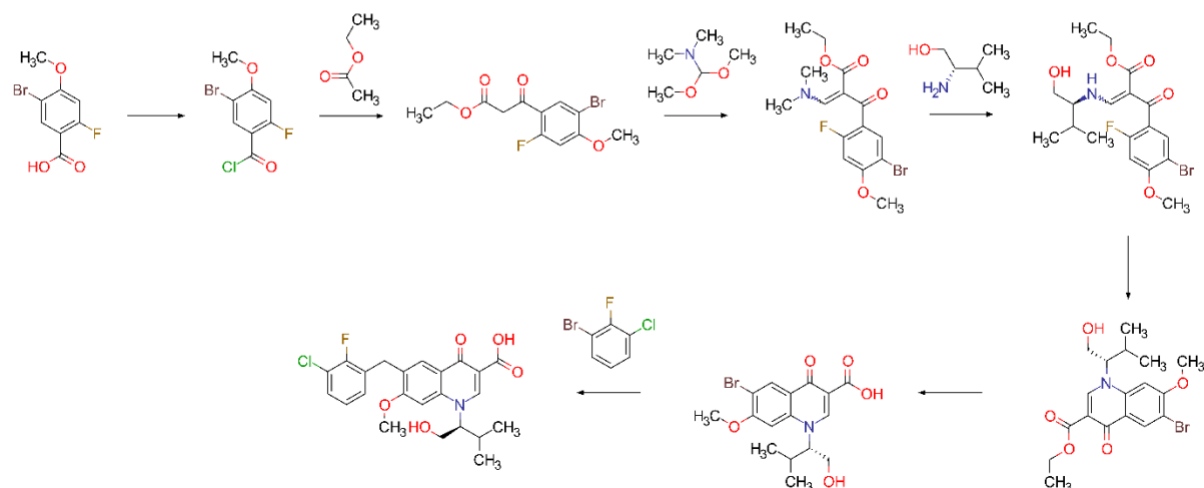


Compound: Elvitegravir

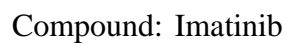
Time to solved: 13.64 seconds



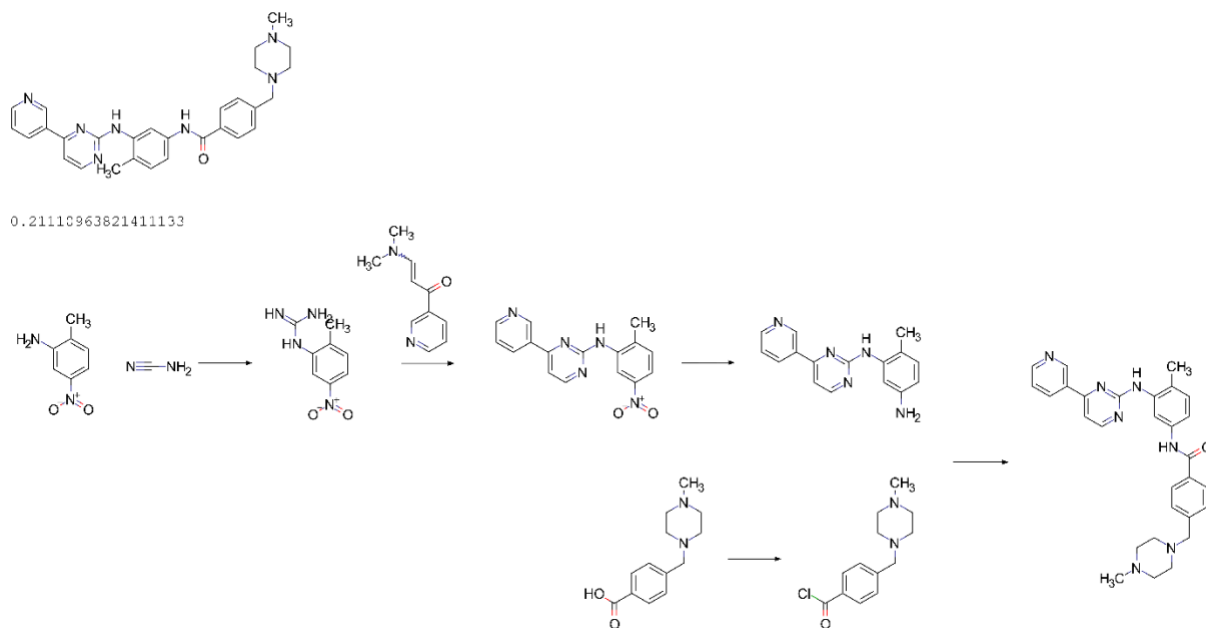
13.63359/850/9956



Time to solved: 0.49 seconds

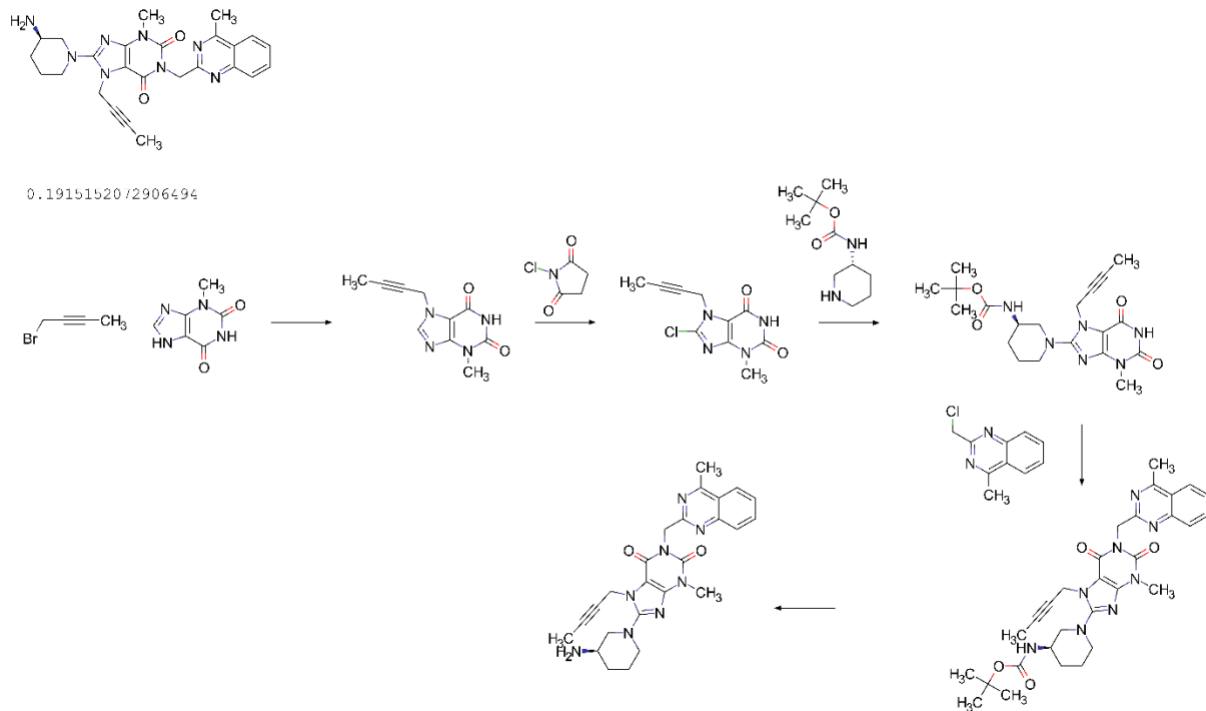


Time to solved: 0.21 seconds



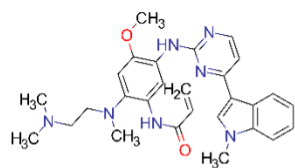
Compound: Linagliptin

Time to solved: 0.19 seconds

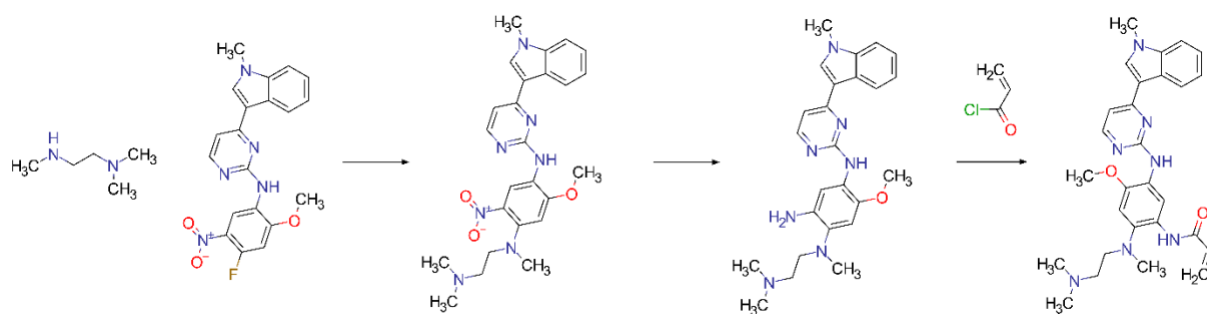


Compound: Osimertinib

Time to solved: 0.11 seconds

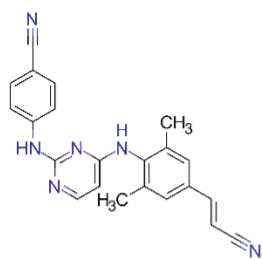


0.11084532737731934

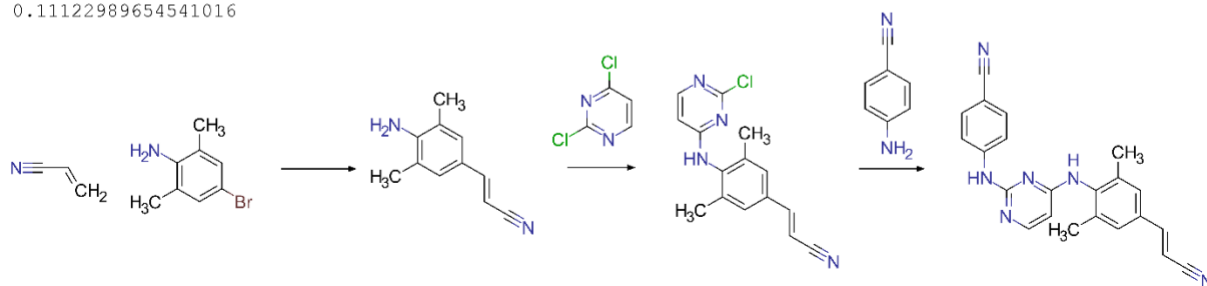


Compound: Rilpivirine

Time to solved: 0.11 seconds

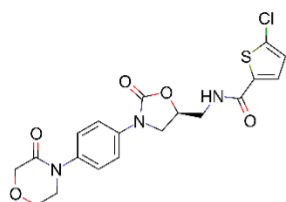


0.11122989654541016

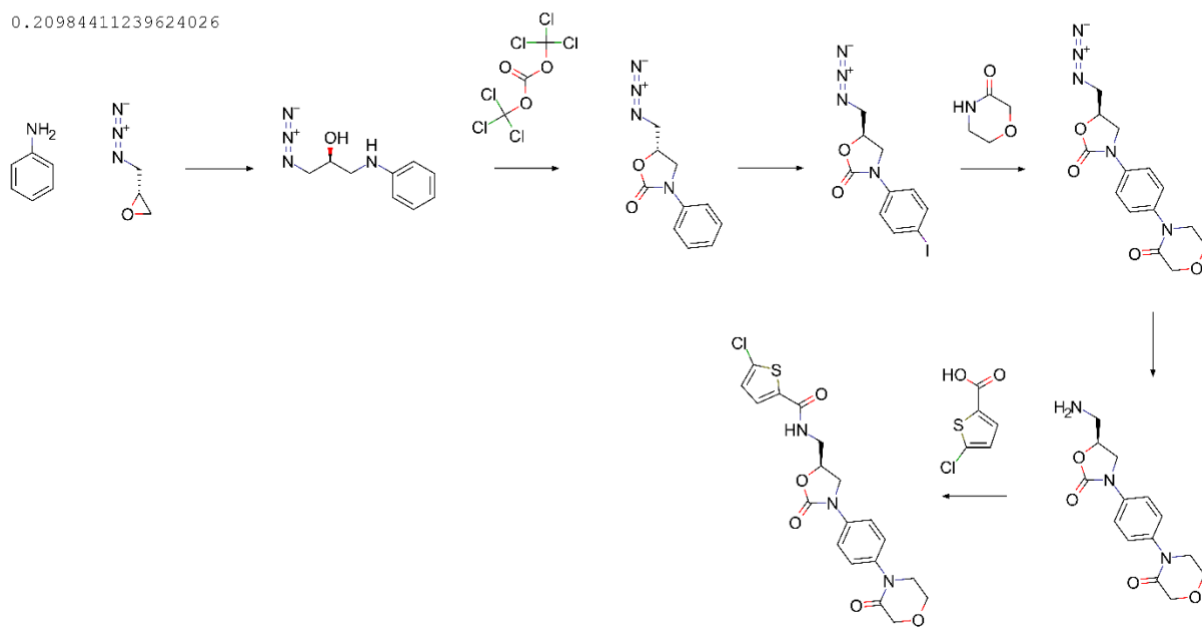


Compound: Rivaroxaban

Time to solved: 0.21 seconds

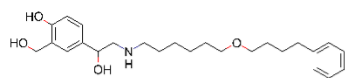


0.20984411239624026

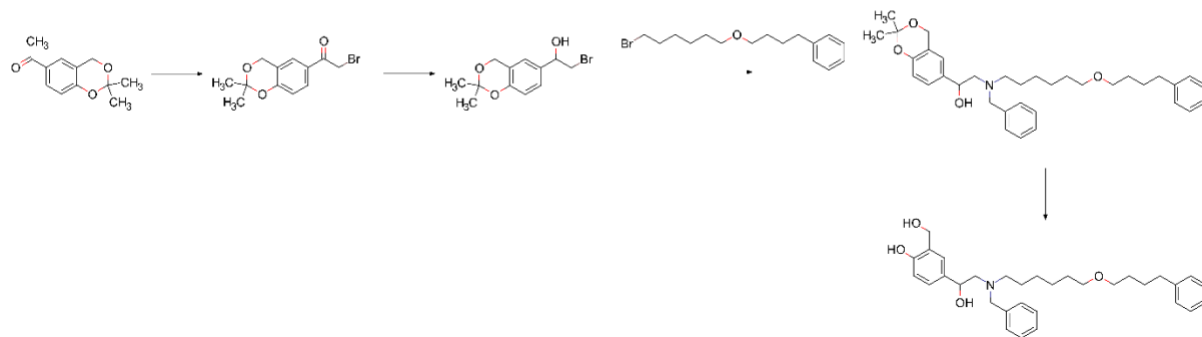


Compound: Salmeterol

Time to solved: 3.38 seconds



3.3805201053619385

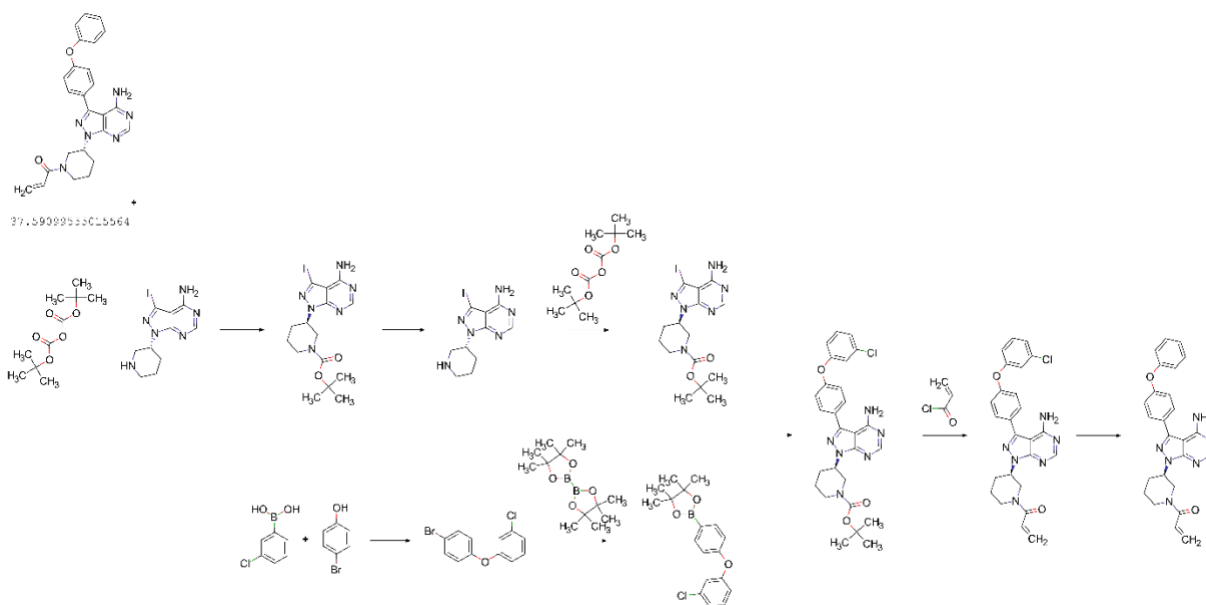


Compound: Apixaban

Unsolved: 27.10 seconds

Reason: Precursor not in stock

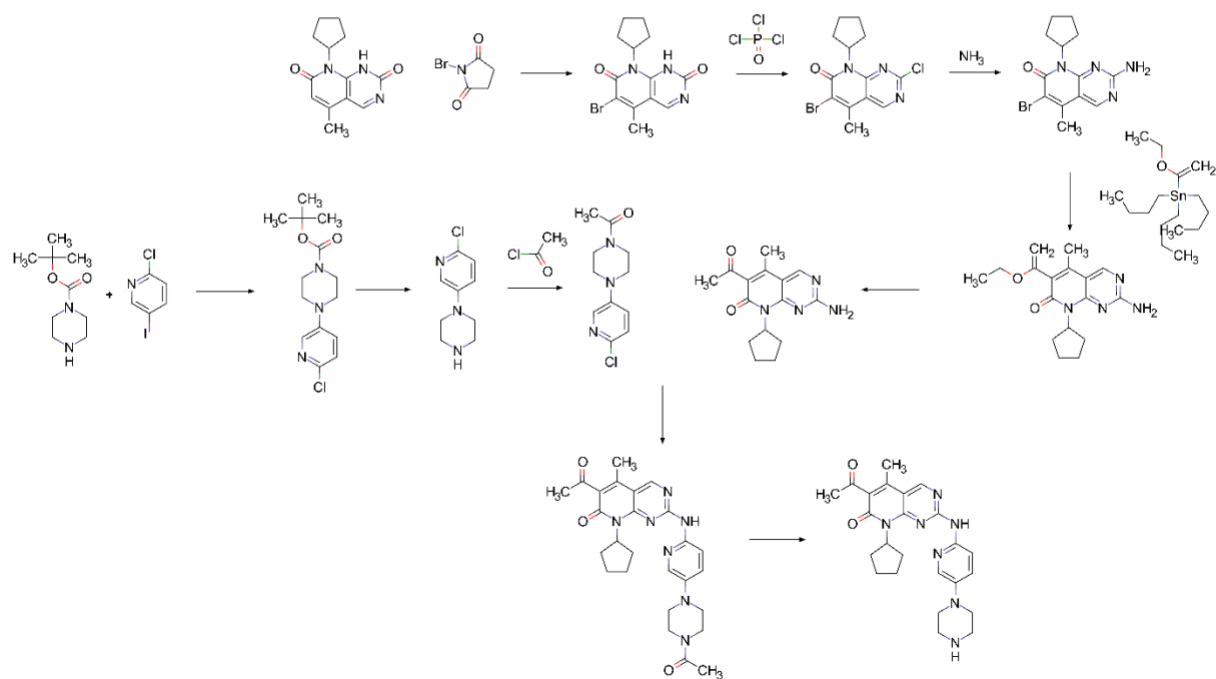
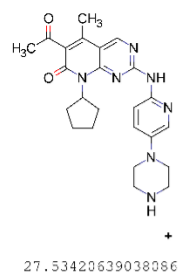




Compound: Palbociclib

Unsolved: 27.53 seconds

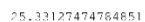
Reason: Precursor not in stock



Compound: Tenofovir Alafenamide

Unsolved: 25.33 seconds

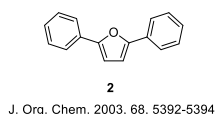
Reason: Precursor not in stock



Appendix B “Ring Breaker”: Neural Network Driven Synthesis Prediction of the Ring System Chemical Space

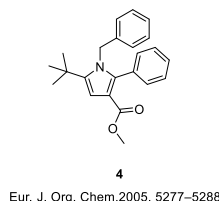
Appendix B.1 Brute force application, Filtering, and Prioritization

Paal–Knorr-furan synthesis



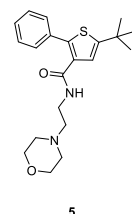
J. Org. Chem. 2003, 68, 5392-5394

Paal–Knorr-pyrrole synthesis



Eur. J. Org. Chem. 2005, 5277–5288

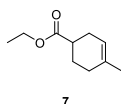
Paal–Knorr-thiophene synthesis



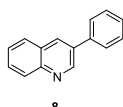
Eur. J. Org. Chem. 2005, 5277–5288



Diels–Alder reaction

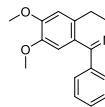


Org. Lett. 2006, 8, 12, 2487-2489



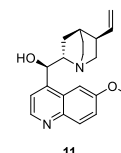
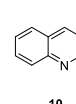
J. Org. Chem. 2016, 81, 656376572

Bischler–Napieralski reaction

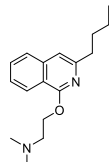


Org. Lett. 2008, 10, 16, 3485-3488

Skraup reaction



Pomeranz–Fritsch reaction

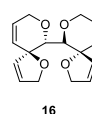


13

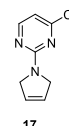
Robinson annulation



Ring-closing metathesis

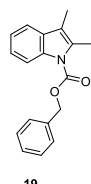


Chem. Rev. 2004, 104, 5, 2199-2238



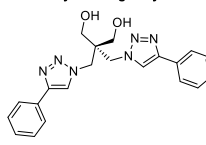
J. Org. Chem. 2008, 73, 7417–7419

Fischer indole synthesis



Org. Lett. 2009, 11, 23, 5454-5456

Azide-alkyne Huisgen cycloaddition



Angew. Chem. Int. Ed. 2002, 41, 2596-2599

The results for applying the standard model, it's subsequent filtering, and the ‘Ring Breaker’ model are contained in the .zip folder for ease of reading.

These contain the following information: target, first_applicable_rank, first_applicable_ringformation_rank, top_50_of_which_ringformation, top_100_of_which_ringformation, top_150_of_which_ringformation, top_200_of_which_ringformation, top_250_of_which_ringformation, top_300_of_which_ringformation, top_350_of_which

_ringformation,top_400_of_which_ringformation,top_450_of_which_ringformation,top_500_of_which_ringformation,max_applicable_exhaustive,max_applicable_ringformation_exhaustive

Appendix B.2 Selected Ring Formations and Training Set Overlap

Table 9: Indicates whether an exact structural match for each substrate has been found in the training sets used for the respective model. 1 indicates the presence of an exact match, and 0 indicates no exact match was found. The matches were found by comparing the InChI keys of the substrate against all InChI keys of the products in the respective training sets.

Substrate	USPTO Standard	USPTO Ring Breaker	Reaxys Standard	Reaxys Ring Breaker
1	0	0	1	1
2	0	0	1	1
3	0	0	1	1
4	0	0	1	1
5	0	0	1	0
6	0	0	1	0
7	1	1	1	1
8	1	0	1	0
9	0	0	1	1
10	1	0	1	1
11	0	0	0	0
12	1	0	1	0
13	0	0	1	0
14	1	0	1	1
15	0	0	1	1
16	0	0	0	0
17	0	0	1	1
18	1	0	1	1
19	0	0	0	0
20	0	0	1	1

Table 10: Indicates whether an substructure match for each substrate has been found in the training sets used for the respective model. 1 indicates the presence of an exact match, and 0 indicates no exact match was found.

Substrate	USPTO Standard	USPTO Ring Breaker	Reaxys Standard	Reaxys Ring Breaker
1	1	1	1	1
2	1	1	1	1
3	1	1	1	1
4	1	1	1	1
5	1	1	1	1
6	1	1	1	1
7	1	1	1	1
8	1	1	1	1
9	1	1	1	1
10	1	1	1	1

11	1	1	1	1
12	1	1	1	1
13	1	1	1	1
14	1	1	1	1
15	1	1	1	1
16	0	0	0	0
17	1	1	1	1
18	1	1	1	1
19	1	1	1	1
20	1	1	1	1

Appendix B.3 Discrepancy between predictive and exhaustive search

Figure 13 shows a discrepancy between the number of templates predicted in the top 50 and those that can be applied as determined by exhaustive application. The discrepancy exists for both the standard and ‘Ring Breaker’ models. To understand why this is the case, consider the case of two structurally related compounds, 8 and 10 in the set of substrates examined in the manuscript (Figure 12). Compound 10 is a substructure of compound 8, therefore it might be expected that the templates that can be applied to compound 10 could also be applied to compound 8 to form either ring of the bicyclic system.

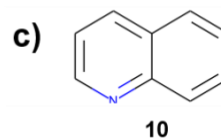
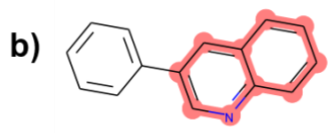
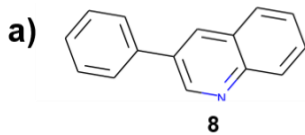
However, when all templates available to the ‘Ring Breaker’ model were applied exhaustively, we found that the subset of templates that were applicable differed for the two compounds. We observed that templates that could be applied to compound 10 could not necessarily be applied to compound 8, differing only by substitution at the position meta to the nitrogen.

For a template to be applied there must first be a substructure match between the template and queried compound. In the case of Figure 39e, the substructure shown appears to match both compounds 8 and 10 and has been highlighted in the two structures. However, the SMARTS encoding the substructure imposes restrictions that are not immediately obvious on a visual inspection of the structure. The position marked in blue (Figure 39e) is described by an aromatic carbon atom in the SMARTS which is equivalent in both structures. Whereas the atom marked in pink (Figure 39e) corresponds to an aromatic carbon bound to a hydrogen atom, additionally it has been specified that the carbon has a degree of 2 and no charge. This criterion matches the environment in compound 10, but substitution at the meta position with a phenyl group to form compound 8 violates the specification, therefore the template cannot be applied in this case. This serves to highlight that although two compounds can be structurally related, and a chemist may expect that the same reaction template may be applied to both, this is not necessarily the case, and depends on the generality of the way the templates are encoded, therefore the template extraction algorithm.

The crude measure used in this study to identify ring forming reactions from which templates were extracted is the difference in the number of rings between the products and reactants. Whilst this

measure does not reassure that other reactions beside those considered as ring formations seep into the training set and template space, the network does not prioritize such reactions. This is shown in Figure 39a, where none of the templates that can be applied are predicted in the top 50. Out of the five templates that can be applied, the one that corresponds to ring formation is also not prioritized by the network in this case. This leads to the discrepancy between all the templates that can be applied and the subset of which are predicted. Furthermore, Figure 2 in the manuscript shows the number of templates that can be exhaustively applied is below 50 for each substrate examined, thus it is not expected that all 50 templates predicted are applicable.

Templates that do not explicitly encode ring formations are present in the dataset as the use of the template is context dependent. For instance, consider Figure 39d below, in which a coupling reaction has been predicted. Whilst the template does not formally encode a ring formation, examination of the reaction from which it was originally extracted reveals that the reaction was used to couple two fragments of the same compound, thus forming a ring. This also goes to show that the template-based approach attempts to apply reactions in a context in which they may not have been used before, sometimes erroneously as in the case of exhaustive enumeration of compound 8, Figure 39a.



d) Template context dependency

The diagram illustrates template context dependency through three rows of chemical transformations. Each row shows a reactant structure on the left, a transformation arrow, and a product structure on the right. The reactant structures are based on a common template: a central C1-C4 bond with C2 and C3 branching from C1, and C5 and C6 branching from C4. The product structures show the result of a transformation that adds a substituent to the C4 position, while the C1-C2 and C3-C5 bonds remain.

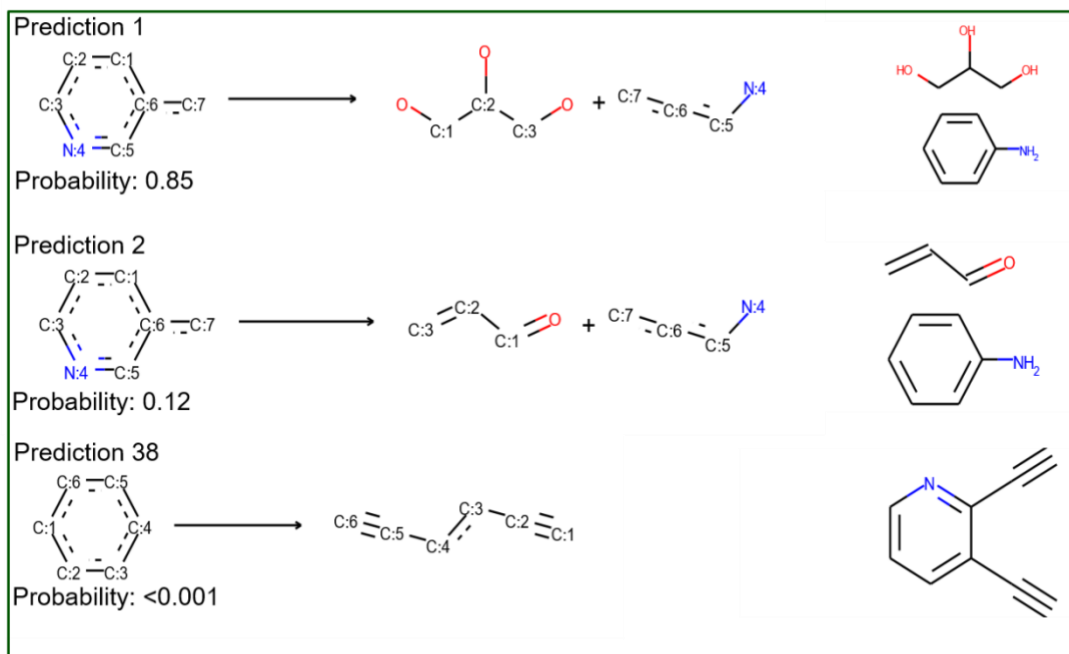
- Row 1:** The reactant is a hexagon with vertices labeled C:1 through C:6. The transformation arrow points to a product where the C4 position is substituted with a C≡C group, resulting in a structure with a C≡C-C≡C-C≡C chain.
- Row 2:** The reactant is a hexagon with vertices labeled C:1 through C:6. The transformation arrow points to a product where the C4 position is substituted with a C≡C group, resulting in a structure with a C≡C-C≡C-C≡C chain.
- Row 3:** The reactant is a hexagon with vertices labeled C:1 through C:6. The transformation arrow points to a product where the C4 position is substituted with a C≡C group, resulting in a structure with a C≡C-C≡C-C≡C chain.

The product structures on the right are chemical structures that correspond to the reactant structures. They are:

- Top: A benzene ring with a C≡C group at the C4 position.
- Middle: A benzene ring with a C≡C group at the C4 position.
- Bottom: A benzene ring with a C≡C group at the C4 position.



All applicable templates and outcomes, c)



- Not in top 50 predictions (Ring Breaker)

- in top 50 predictions (Ring Breaker)

e) Template substructure matching issue

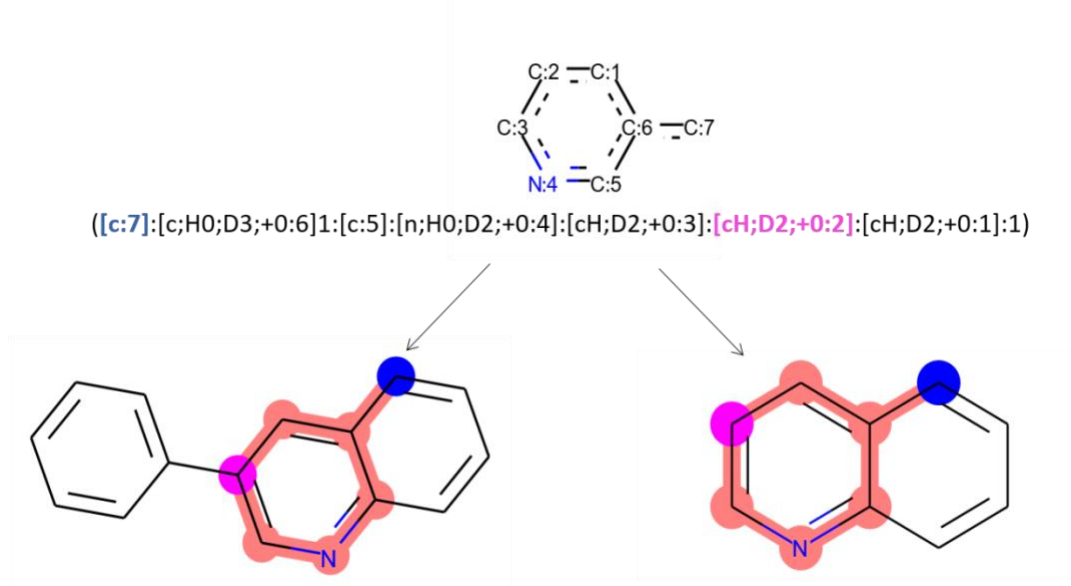
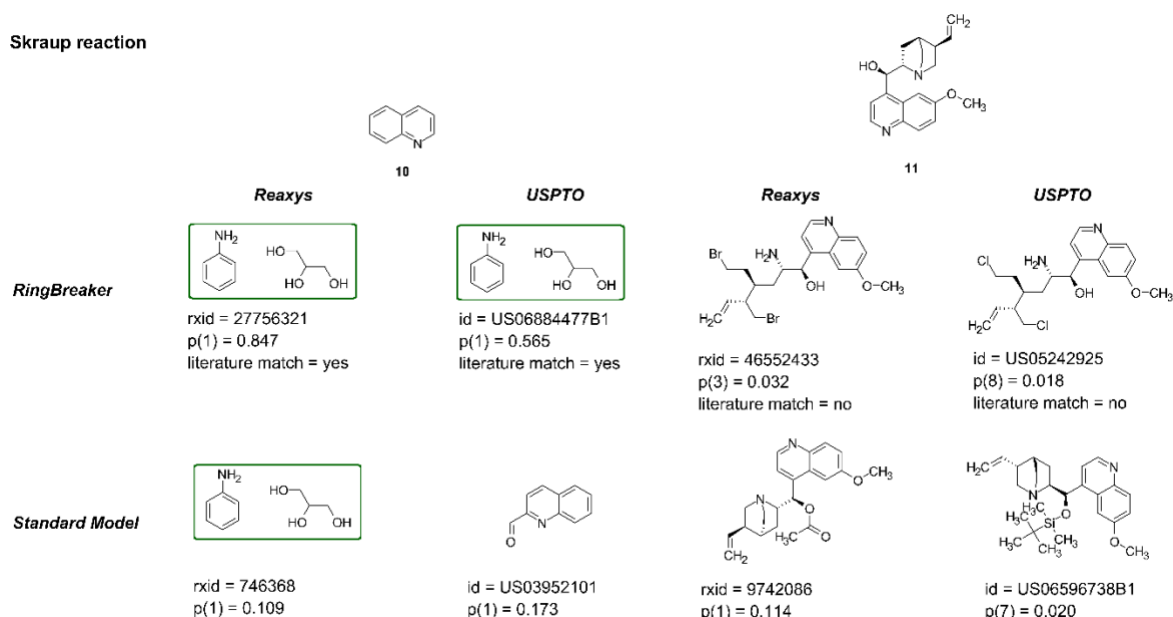


Figure 39 a) Compound 8 and c) compound 10, alongside the applicable templates found through an exhaustive search of the Reaxys ring breaker set through the filtered dataset and whether they are present in the top 50 predicted templates by the Ring Breaker model. b) highlights that both compound 8 (a) and compound 10 (b) share a substructure. d) Shows the application of templates for ring formations is context dependent. In some cases, reactions that are not formally considered ring forming may be used to form ring fragments. This comes originates from their use in the dataset as shown for the coupling reaction. e) Illustrates why templates which seemingly share the same substructure on the diagram/image level are not the same when the SMARTS pattern corresponding to the product of the template are considered. The template encodes the same chemical environment around the atom highlighted in blue, yet the environment around that in pink differs. The hydrogen and degree of the atom highlighted in pink is explicitly defined in the SMARTS pattern, and this only matches that of compound 10, not 8, due to a difference in substitution pattern. The sets of templates that can be applied to each of compound 8 and 10 differ as a result of the specific environments encoded by the SMARTS patterns.

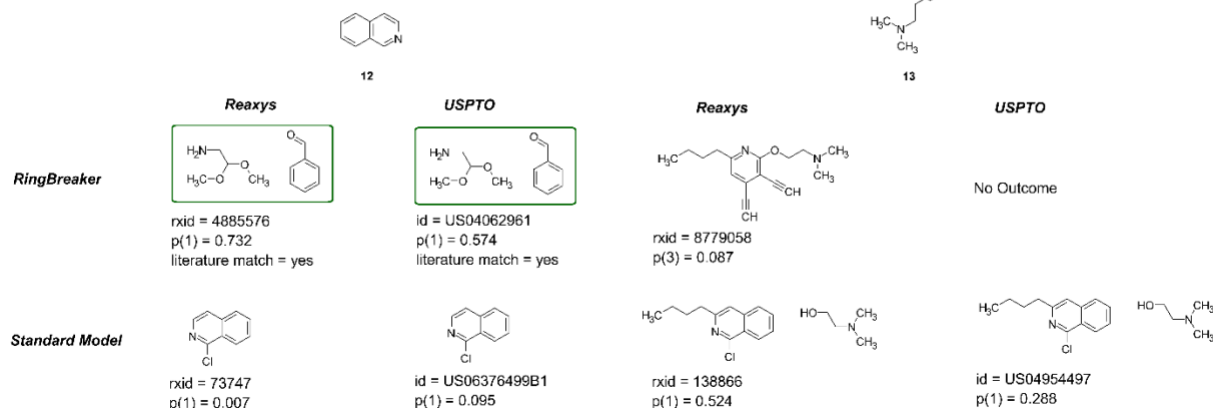
Appendix B.4 Prediction of well-known ring formations

The following predictions were made 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.

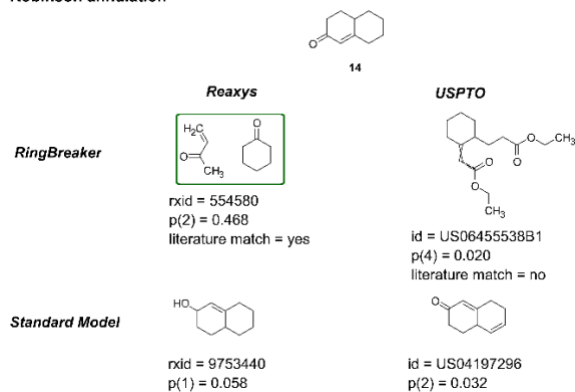
Skraup reaction



Pomeranz-Fritsch reaction



Robinson annulation



Ring-closing metathesis

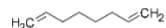


15

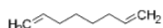
Reaxys

USPTO

RingBreaker



rxid = 9426547
p(1) = 0.845

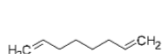


id = US07608590B2
p(1) = 0.996

Standard Model



rxid = 84879
p(1) = 0.388



id = US07608590B2
p(1) = 0.286



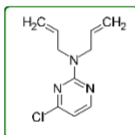
17

J. Org. Chem. 2008, 73, 7417-7419

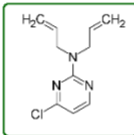
Reaxys

USPTO

RingBreaker

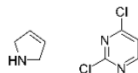


rxid = 9426547
p(1) = 0.998
literature match = yes

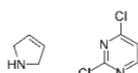


id = US07608590B2
p(1) = 0.576
literature match = yes

Standard Model



rxid = 44295261
p(1) = 0.427



id = US04518596
p(1) = 0.743



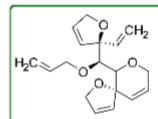
16

Chem. Rev. 2004, 104, 5, 2199-2238

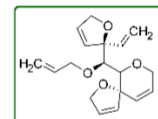
Reaxys

USPTO

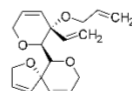
USPTO



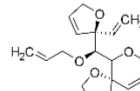
rxid = 29734072
p(1) = 0.910
literature match = yes



id = US20090264445A1
p(5) = 0.023
literature match = yes

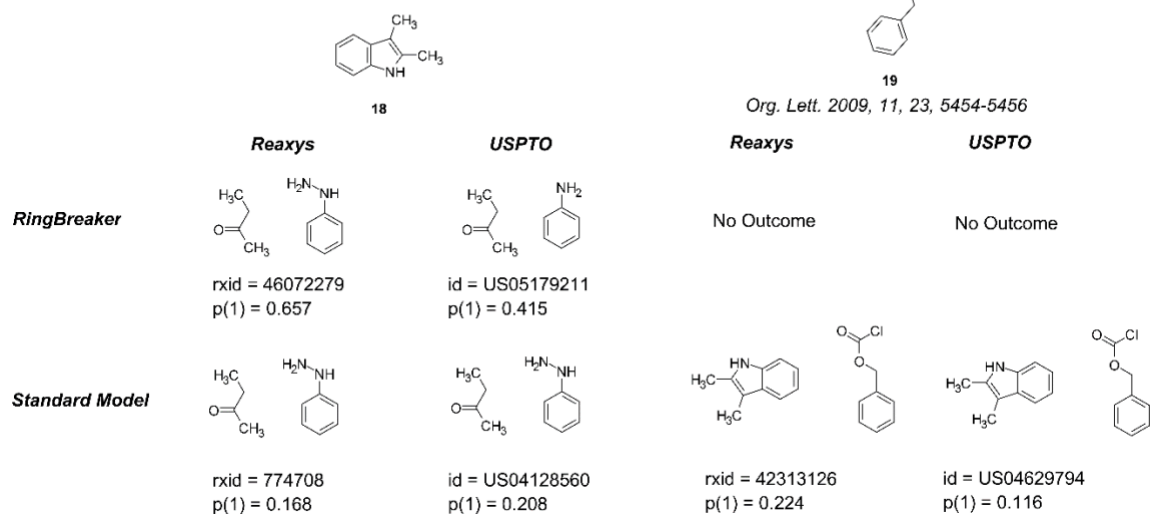


rxid = 9428477
p(1) = 0.048
literature match = yes

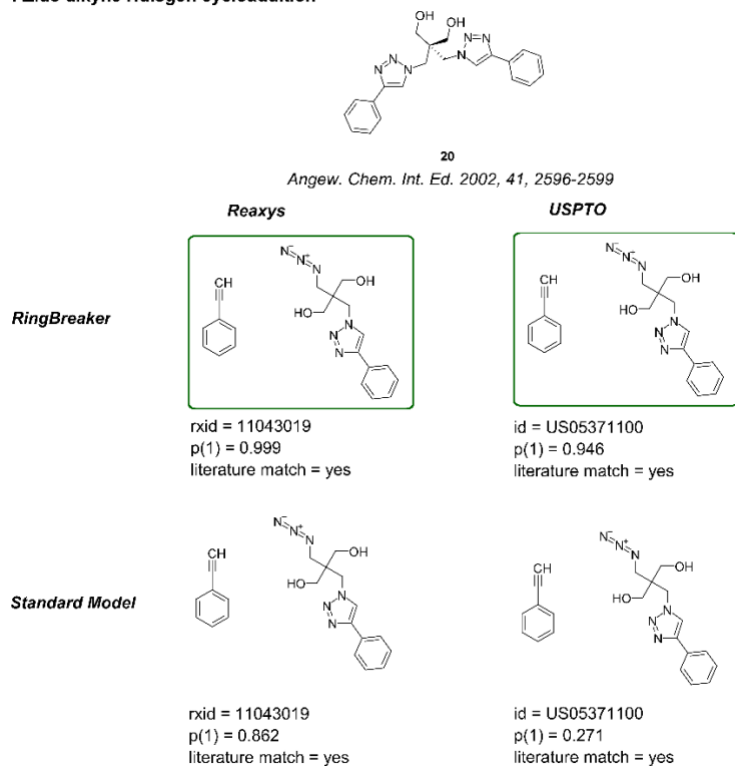


id = US20090264445A1
p(3) = 0.036

Fischer indole synthesis

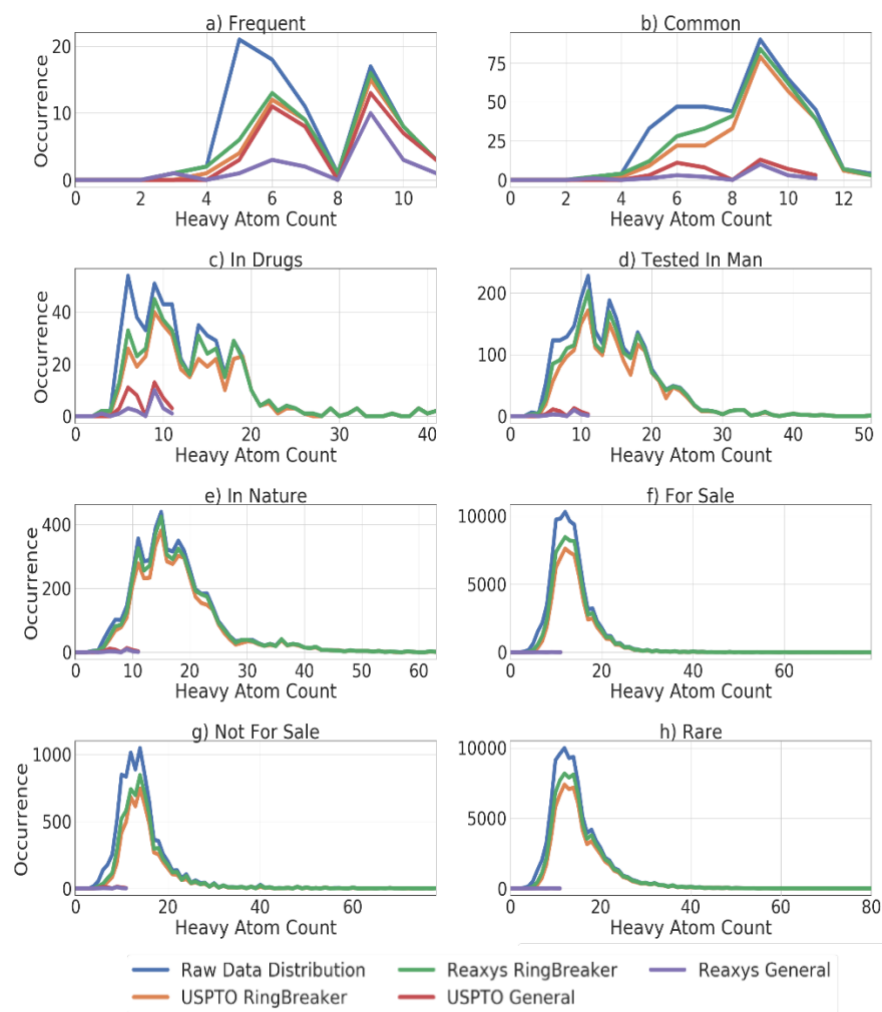


Azide-alkyne Huisgen cycloaddition



Appendix B.5 Prediction of Fragments

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) frequent (occurring in greater than 1 M substances) b) common (occurring in greater than 100 K substances) c) present in drugs d) present in substances tested in man, an extended set of those present in drugs e) present in nature f) are purchasable g) are not purchasable according to ZINC h) 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.



Appendix B.6 Overlap of ZINC fragment sets with the training sets

Table 11: Percentage of fragments from the ZINC ring datasets with an exact structural match to those in the USPTO and Reaxys datasets. The values are that of an exact match, and so are a lower bound estimate of the overlap with the USPTO and Reaxys datasets. This is because some fragments from the ZINC sets may be substructures of compounds in the training set, and so will have been seen during training the model. The greatest overlap is in the ring systems that frequently occur followed by those that commonly occur for the standard models. For the Ring Breaker models the pattern observed for the standard models is observed for the Reaxys dataset and the inverse is true for the USPTO dataset. The model performs surprisingly well on the rare fragments for which there is a low percentage overlap.

ZINC Ring Set	USPTO Standard (%)	USPTO Ring Breaker (%)	Reaxys Standard (%)	Reaxys Ring Breaker (%)	Zinc Set Size
Common	31	9	64	29	388
For Sale	2	< 1	6	2	91,695
Frequent	33	7	90	51	82
In Drugs	33	7	90	51	541
In Man	21	5	41	18	2,472
In Nature	11	3	23	9	5,433
Not For Sale	4	1	10	4	9,340
Rare	1	< 1	4	1	100,900

Table 12: Percentage of fragments from the ZINC ring datasets with a sub-structure match to those in the USPTO and Reaxys datasets.

ZINC Ring Set	USPTO Standard (%)	USPTO Ring Breaker (%)	Reaxys Standard (%)	Reaxys Ring Breaker (%)	Zinc Set Size
Common	96	83	99	93	388
For Sale	-	-	-	-	91,695
Frequent	100	100	100	100	82
In Drugs	82	54	92	76	541
In Man	53	28	81	53	2,472
In Nature	-	10	-	28	5,433
Not For Sale	-	< 1	-	< 1	9,340
Rare	-	-	-	-	100,900

Appendix B.7 Overlap of approved drugs in Drug Bank with the training sets

Table 13: Percentage of compounds from the approved drugs in DrugBank with an exact structural match to those in the USPTO and Reaxys datasets. The matches were found by comparing the InChI keys of the substrate against all InChI keys of the products in the respective training sets.

Dataset	USPTO Standard (%)	USPTO Breaker (%)	Ring	Reaxys Standard (%)	Reaxys Breaker (%)	Ring	Dataset size
DrugBank Approved	31		2	52		5	2039

Appendix C Retrosynthetic accessibility score (RAscore) – rapid machine learned synthesizability classification from AI driven retrosynthetic planning

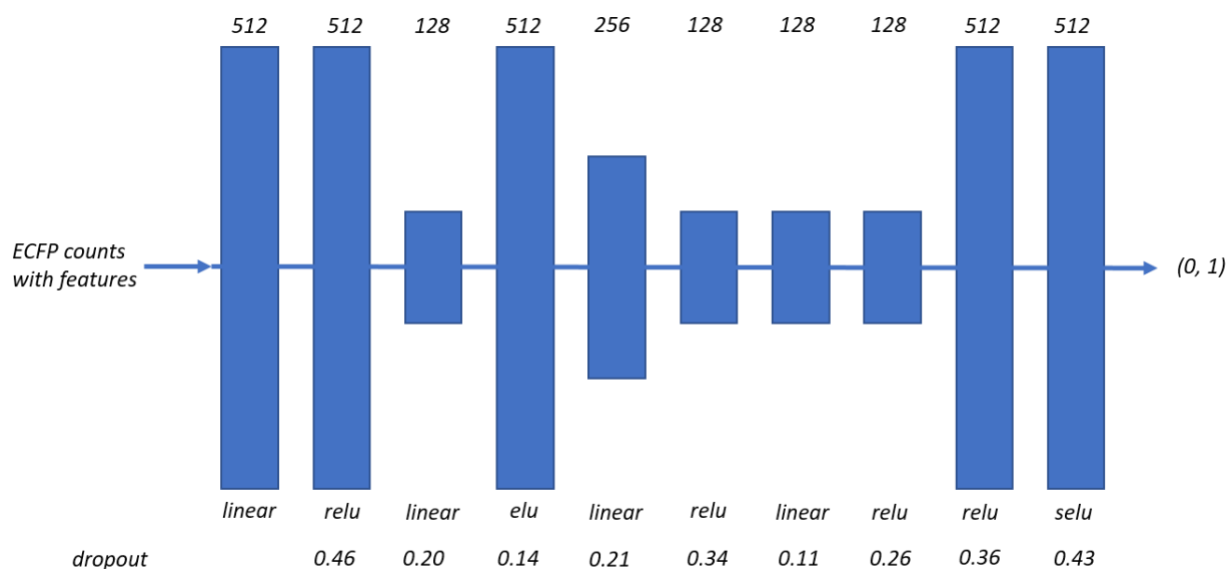
Appendix C.1 Retrosynthesis Prediction for Training Set Generation

Refer to the attached .csv files for the training and test datasets.

Refer to the configuration file found in our GitHub repository for the settings used during the label generation process: <https://github.com/reymond-group/RAscore>

Appendix C.2 Machine Learning Classifiers for Estimation of Retrosynthetic Accessibility

Appendix C.2.1 Example of the optimal architecture found by hyperparameter optimization for the ChEMBL dataset.



Optimal architecture found for the ChEMBL dataset using Optuna for hyperparameter optimization. The hyperparameters to optimize were, the number of layers, the size of the layers, the activation function, dropout rate, and the learning rate. Full details of the parameters found are given in the SI. As input for the model 2048 dimensional counted ECFPs with features and a radius of 3 were used, and as output the class label, solved or unsolved as obtained via retrosynthetic analysis using our CASP tool.

Appendix C.2.2 Optimal Model Hyperparameters

The following boundaries were used to train the classifiers:

Appendix C.2.2.1: Logistic Regression

```
"algorithm": {"LogisticRegression": {  
  "solver": ["newton-cg", "lbfgs", "sag", "saga"],  
  "C": {  
    "low": 0.1,  
    "high": 1.5 }  
}
```

Appendix C.2.2.2: Random Forest

```
"algorithm": {"RandomForestClassifier": {  
  "max_depth": {  
    "low": 10,  
    "high": 20  
  },  
  "n_estimators": {  
    "low": 10,  
    "high": 100  
  },  
  "max_features": ["sqrt", "log2"]  
}
```

Appendix C.2.2.3 XGB Classifier

```
"algorithm": {"XGBClassifier": {  
  "max_depth": {  
    "low": 10,  
    "high": 20  
  },  
  "n_estimators": {  
    "low": 10,  
    "high": 100  
  }  
}
```

```

    },
    "learning_rate": {
        "low": 0.05,
        "high": 0.2
    }
}

```

Appendix C.2.2.4 Neural Network

```

"algorithm": {
    "DNNClassifier": {
        "layer_1": [128, 256, 512],
        "activation_1": ["relu", "elu", "selu", "linear"],
        "dropout_1": 0.1,
        "max_layers": 10,
        "layer_size": [128, 256, 512],
        "layer_activations": ["relu", "elu", "selu", "linear"],
        "layer_dropout": {"low": 0,
                          "high": 0.5},
        "learning_rate": {"low": 1e-5,
                          "high": 1e-1}
    }
}

```

Appendix C.2.2.5 Parameters for NN ecfp counts with features

```

{"layer_1": 512, "activation_1": "linear", "num_layers": 10, "units_2": 512, "activation_2": "relu",
 "dropout_2": 0.45834579304621176, "units_3": 128, "activation_3": "linear", "dropout_3":
0.20214636121010582, "units_4": 512, "activation_4": "elu", "dropout_4":
0.13847113009081813, "units_5": 256, "activation_5": "linear", "dropout_5":
0.21312873496871235, "units_6": 128, "activation_6": "relu", "dropout_6":
0.33530504087548707, "units_7": 128, "activation_7": "linear", "dropout_7":
0.11559123444807062, "units_8": 128, "activation_8": "relu", "dropout_8":

```

0.2618908919792556, "units_9": 512, "activation_9": "relu", "dropout_9": 0.3587291059530903, "units_10": 512, "activation_10": "selu", "dropout_10": 0.43377277017943133, "learning_rate": 1.5691774834712003e-05}

Appendix C.2.2.6: Parameters for NN ecfp counts

{"layer_1": 256, "activation_1": "selu", "num_layers": 2, "units_2": 128, "activation_2": "relu", "dropout_2": 0.15578695546915372, "learning_rate": 2.632240761263429e-05}

Appendix C.2.2.7: Parameters for XGBoost ecfp counts

{"max_depth": 19, "n_estimators": 97, "learning_rate": 0.19984033197055842}

Appendix C.2.2.8 Parameters for SA Score Logistic Regression

{"C": 0.18582521970918675, "solver": "lbfgs"}

Appendix C.2.2.9 Parameters for SC Score Logistic Regression

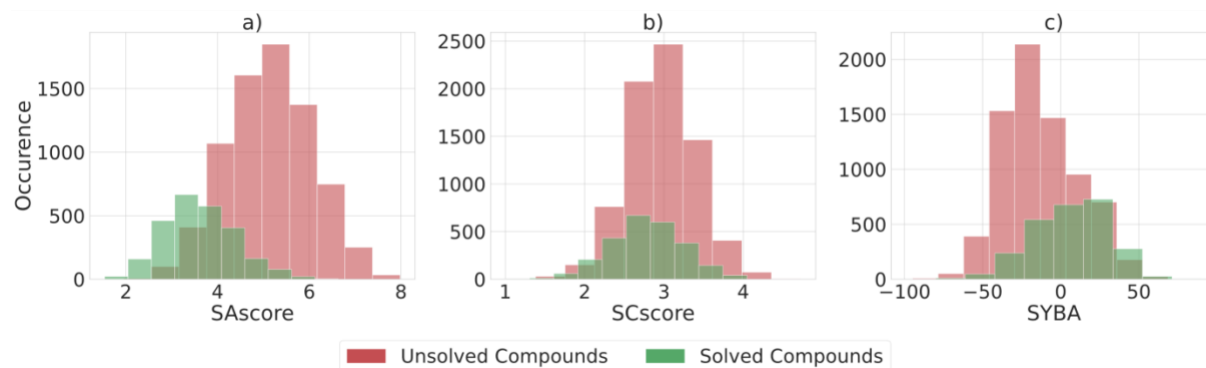
{"C": 0.22237611805770982, "solver": "saga"}

Appendix C.2.2.10 Parameters for SYBA Logistic Regression

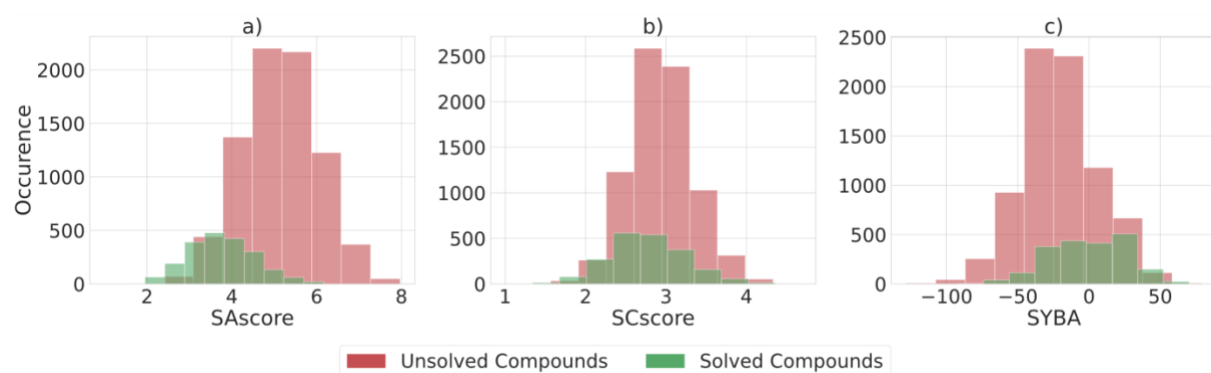
{"C": 0.39649336266446344, "solver": "newton-cg"}

Appendix C.2.3 Attempts at using SAScore, SCscore and SYBA

Appendix C.3.1 GDBChEMBL



Appendix C.3.2 GDBMedChem

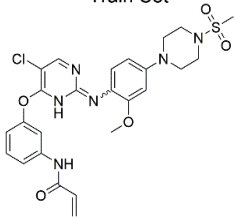
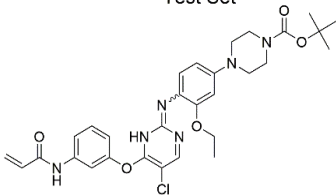
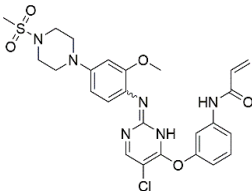
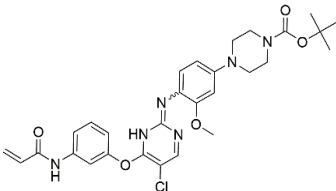
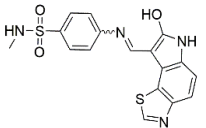
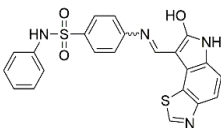
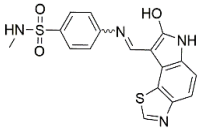
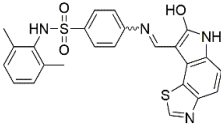
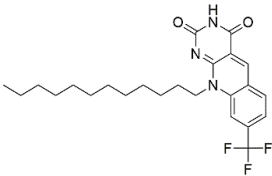
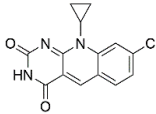
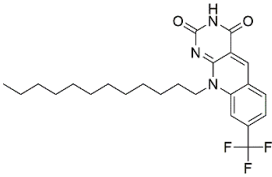
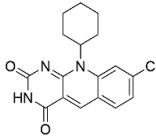
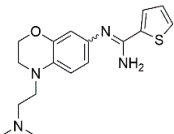
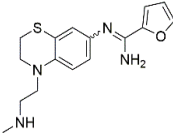
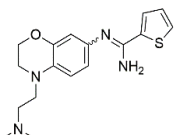
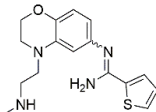


Appendix C.2.4 Machine Learning Classifiers for Estimation of Retrosynthetic Accessibility

Dataset	Model	Descriptor	ROC-AUC	Accuracy	Precision	Recall	Average Linkage
ChEMBL	NN (RAscore)	ECFP6 counts with features	0.93	0.90	0.92	0.95	0.69
ChEMBL	NN	ECFP6 counts	0.94	0.90	0.92	0.95	0.68
ChEMBL	XGB	ECFP6 counts	0.95	0.91	0.92	0.96	0.65
ChEMBL	XGB	ECFP6 counts with features	0.95	0.90	0.91	0.96	0.62
ChEMBL	RF	ECFP6 counts	0.90	0.83	0.82	0.99	0.30
ChEMBL	RF	ECFP6 counts with features	0.89	0.82	0.82	0.99	0.28
ChEMBL	NN	SA score	0.85	0.81	0.84	0.92	0.37
ChEMBL	Logistic	SA score	0.85	0.81	0.83	0.94	0.36
ChEMBL	Logistic	SC score	0.61	0.75	0.75	1.00	0.27
ChEMBL	NN	SC score	0.61	0.75	0.61	1.00	0.27
ChEMBL	Logistic	SYBA score	0.74	0.78	0.79	0.96	0.25
ChEMBL	NN	SYBA score	0.74	0.78	0.78	0.97	0.21
ChEMBL	-	SAscore	0.15	-	-	-	0.17
ChEMBL	-	SCscore	0.39	-	-	-	0.22
ChEMBL	-	SYBA	0.74	-	-	-	0.17
GDBChEMBL	NN (GDBscore)	ECFP6 counts	0.93	0.87	0.76	0.73	0.64
GDBChEMBL	NN	ECFP6 counts with features	0.94	0.88	0.78	0.74	0.63
GDBChEMBL	XGB	ECFP6 counts	0.94	0.89	0.81	0.73	0.61
GDBChEMBL	XGB	ECFP6 counts with features	0.94	0.88	0.80	0.71	0.61
GDBChEMBL	RF	ECFP6 counts with features	0.89	0.81	0.76	0.40	0.36
GDBChEMBL	RF	ECFP6 counts	0.88	0.81	0.81	0.31	0.32
GDBChEMBL	-	SAscore	0.11	-	-	-	0.26
GDBChEMBL	-	SCscore	0.38	-	-	-	0.14
GDBChEMBL	-	SYBA	0.72	-	-	-	0.17
GDBMedChem	NN	ECFP6 counts	0.93	0.88	0.75	0.64	0.64
GDBMedChem	NN	ECFP6 counts with features	0.94	0.89	0.77	0.66	0.63
GDBMedChem	XGB	ECFP6 counts	0.94	0.89	0.78	0.64	0.61
GDBMedChem	XGB	ECFP6 counts with features	0.94	0.89	0.79	0.64	0.61
GDBMedChem	RF	ECFP6 counts with features	0.89	0.83	0.80	0.27	0.36
GDBMedChem	RF	ECFP6 counts	0.88	0.81	0.91	0.10	0.32
GDBMedChem	-	SAscore	0.13	-	-	-	0.22
GDBMedChem	-	SCscore	0.39	-	-	-	0.14
GDBMedChem	-	SYBA	0.70	-	-	-	0.17

Appendix C.2.5 Limitations of Template Based CASP Tools

Appendix C.5.1 Example Compound Similarity to Training Set

Train Set	Solved	Tanimoto	Test Set
	1	0.77	
	1	0.83	
	1	0.74	
	1	0.73	
	0	0.46	
	0	0.45	
	0	0.70	
	0	0.61	

9 Bibliography

1. Struble, T. J. *et al.* Current and Future Roles of Artificial Intelligence in Medicinal Chemistry Synthesis. *J. Med. Chem.* (2020) doi:10.1021/acs.jmedchem.9b02120.
2. Schneider, P. *et al.* Rethinking drug design in the artificial intelligence era. *Nat. Rev. Drug Discov.* **19**, 353–364 (2020).
3. Jordan, A. M. Artificial Intelligence in Drug Design—The Storm Before the Calm? *ACS Med. Chem. Lett.* **9**, 1150–1152 (2018).
4. Feigenbaum, E. A. Some challenges and grand challenges for computational intelligence. *J. Assoc. Comput. Mach.* **50**, 32–40 (2003).
5. Turing, A. M. I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind* **LIX**, 433–460 (1950).
6. Battaglia, P. W. *et al.* Relational inductive biases, deep learning, and graph networks. 1–40 (2018).
7. Howarth, A., Ermanis, K. & Goodman, J. M. DP4-AI automated NMR data analysis: straight from spectrometer to structure. *Chem. Sci.* **11**, 4351–4359 (2020).
8. Chen, H., Engkvist, O., Wang, Y., Olivecrona, M. & Blaschke, T. The rise of deep learning in drug discovery. *Drug Discov. Today* (2018) doi:10.1016/J.DRUDIS.2018.01.039.
9. Segler, M. H. S., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555**, 604–610 (2018).
10. Muratov, E. N. *et al.* QSAR without borders. *Chem. Soc. Rev.* **49**, 3525–3564 (2020).
11. Unterthiner, T., Mayr, A., Klambauer, G. & Hochreiter, S. Toxicity Prediction using Deep Learning. *arXiv e-prints* arXiv:1503.01445 (2015).

12. Lindsay, R. K., Buchanan, B. G., Feigenbaum, E. A. & Lederberg, J. *Applications of Artificial Intelligence for Organic Chemistry: The DENDRAL Project. Artificial Intelligence Series* (McGraw-Hill Book Company, 1980).
13. Griffen, E. J., Dossetter, A. G. & Leach, A. G. Chemists: AI Is Here; Unite To Get the Benefits. *J. Med. Chem.* (2020) doi:10.1021/acs.jmedchem.0c00163.
14. Bender, A. Will Robotics, AI and Cloud Computing In Chemical Synthesis Save Drug Discovery? A Closer Look. <http://www.drugdiscovery.net/2020/09/09/will-robotics-ai-and-cloud-computing-in-chemical-synthesis-save-drug-discovery-a-closer-look/> (2020).
15. Mak, K.-K. & Pichika, M. R. Artificial intelligence in drug development: present status and future prospects. *Drug Discov. Today* **24**, 773–780 (2019).
16. McKinney, S. M. *et al.* International evaluation of an AI system for breast cancer screening. *Nature* **577**, 89–94 (2020).
17. Dalby, A. *et al.* Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **32**, 244–255 (1992).
18. David, L., Thakkar, A., Mercado, R. & Engkvist, O. Molecular representations in AI-driven drug discovery: a review and practical guide. *J. Cheminformatics* **12**, 56 (2020).
19. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Model.* **28**, 31–36 (1988).
20. RDKit: Open-Source Cheminformatics; <http://www.rdkit.org>.
21. O'Boyle, N. M. Towards a Universal SMILES representation - A standard method to generate canonical SMILES based on the InChI. *J. Cheminformatics* **4**, 22 (2012).

22. Schneider, N., Sayle, R. A. & Landrum, G. A. Get Your Atoms in Order-An Open-Source Implementation of a Novel and Robust Molecular Canonicalization Algorithm. *J. Chem. Inf. Model.* **55**, 2111–2120 (2015).
23. Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **5**, 107–113 (1965).
24. ChemAxon Extended SMILES and SMARTS - CXSMILES and CXSMARTS - Documentation. https://docs.chemaxon.com/display/docs/chemaxon-extended-smiles-and-smarts-cxsmiles-and-cxsmarts.md#src-1806633_ChemAxonExtendedSMILESandSMARTS-CXSMILESandCXSMARTS-Fragmentgrouping.
25. Daylight Theory: SMILES. <https://daylight.com/dayhtml/doc/theory/theory.smiles.html#RTFrnx5>.
26. Daylight Theory: SMARTS - A Language for Describing Molecular Patterns. <https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html>.
27. Nugmanov, R. I. *et al.* CGRtools: Python Library for Molecule, Reaction, and Condensed Graph of Reaction Processing. *J. Chem. Inf. Model.* **59**, 2516–2521 (2019).
28. Heller, S. R., McNaught, A., Pletnev, I., Stein, S. & Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *J. Cheminformatics* **7**, 23 (2015).
29. Pletnev, I. *et al.* InChIKey collision resistance: An experimental testing. *J. Cheminformatics* **4**, (2012).
30. Grethe, G., Blanke, G., Kraut, H. & Goodman, J. M. International chemical identifier for reactions (RInChI). *J. Cheminformatics* **10**, 22 (2018).

31. Grethe, G., Goodman, J. M. & Allen, C. H. International chemical identifier for reactions (RInChI). *J. Cheminformatics* **5**, 45 (2013).
32. Jacob, P. M., Lan, T., Goodman, J. M. & Lapkin, A. A. A possible extension to the RInChI as a means of providing machine readable process data. *J. Cheminformatics* **9**, (2017).
33. Varnek, A., Fourches, D., Hoonakker, F. & Solov'ev, V. P. Substructural fragments: An universal language to encode reactions, molecular and supramolecular structures. *J. Comput. Aided Mol. Des.* **19**, 693–703 (2005).
34. Fujita, S. Description of Organic Reactions Based on Imaginary Transition Structures. 1. Introduction of New Concepts. *J. Chem. Inf. Comput. Sci.* **26**, 205–212 (1986).
35. Rogers, D. & Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).
36. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **23**, 3–25 (1997).
37. Bohacek, R. S., McMartin, C. & Guida, W. C. The art and practice of structure-based drug design: A molecular modeling perspective. *Med. Res. Rev.* **16**, 3–50 (1996).
38. Valeur, E. *et al.* New Modalities for Challenging Targets in Drug Discovery. *Angew. Chem. Int. Ed.* **56**, 10294–10323 (2017).
39. Capecchi, A., Awale, M., Probst, D. & Reymond, J.-L. PubChem and ChEMBL beyond Lipinski. *Mol. Inform.* **38**, 1900016 (2019).
40. McKay, B. D. Practical Graph Isomorphism Congressus Numerantium. in 45–87 (1981).
41. Fink, T. & Reymond, J.-L. Virtual Exploration of the Chemical Universe up to 11 Atoms of C, N, O, F: Assembly of 26.4 Million Structures (110.9 Million Stereoisomers) and Analysis

- for New Ring Systems, Stereochemistry, Physicochemical Properties, Compound Classes, and Drug Discove. *J. Chem. Inf. Model.* **47**, 342–353 (2007).
42. Fink, T., Bruggesser, H. & Reymond, J.-L. Virtual Exploration of the Small-Molecule Chemical Universe below 160 Daltons. *Angew. Chem. Int. Ed.* **44**, 1504–1508 (2005).
43. Blum, L. C. & Reymond, J.-L. 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13. *J. Am. Chem. Soc.* **131**, 8732–8733 (2009).
44. Ruddigkeit, L., van Deursen, R., Blum, L. C. & Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **52**, 2864–2875 (2012).
45. Ruddigkeit, L., Blum, L. C. & Reymond, J.-L. Visualization and Virtual Screening of the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **53**, 56–65 (2013).
46. Coley, C. W., Eyke, N. S. & Jensen, K. F. Autonomous discovery in the chemical sciences part II: Outlook. *Angew. Chem. Int. Ed.* **59**, 23414–23436 (2019).
47. Visini, R., Awale, M. & Reymond, J.-L. Fragment Database FDB-17. *J. Chem. Inf. Model.* **57**, 700–709 (2017).
48. Bühlmann, S. & Reymond, J.-L. ChEMBL-Likeness Score and Database GDBChEMBL. *Front. Chem.* **8**, (2020).
49. Awale, M., Sirockin, F., Stiefl, N. & Reymond, J.-L. Medicinal Chemistry Aware Database GDBMedChem. **38**, 1900031 (2019).
50. Kim, S. *et al.* PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Res.* **49**, D1388–D1395 (2021).
51. Mendez, D. *et al.* ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* **47**, D930–D940 (2019).

52. CAS Content. <https://www.cas.org/about/cas-content>.
53. Elsevier. Reaxys. <https://www.elsevier.com/solutions/reaxys> (2021).
54. Enamine REAL Space. <https://enamine.net/compound-collections/real-compounds>.
55. Patel, H. *et al.* SAVI, in silico generation of billions of easily synthesizable compounds through expert-system type rules. *Sci. Data* **7**, 384 (2020).
56. Gillet, V. J. *et al.* SPROUT: Recent developments in the de novo design of molecules. *J. Chem. Inf. Comput. Sci.* **34**, 207–217 (1994).
57. Jensen, J. H. A graph-based genetic algorithm and generative model/Monte Carlo tree search for the exploration of chemical space. *Chem. Sci.* **10**, 3567–3572 (2019).
58. Lewell, X. Q., Judd, D. B., Watson, S. P. & Hann, M. M. RECAPRetrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.* **38**, 511–522 (1998).
59. Leach, A. R. & Hann, M. M. The in silico world of virtual libraries. *Drug Discov. Today* **5**, 326–336 (2000).
60. Hu, Q., Peng, Z., Kostrowicki, J. & Kuki, A. LEAP into the Pfizer Global Virtual Library (PGVL) Space: Creation of Readily Synthesizable Design Ideas Automatically BT - Chemical Library Design. in (ed. Zhou, J. Z.) 253–276 (Humana Press, 2011). doi:10.1007/978-1-60761-931-4_13.
61. Segler, M. H. S., Kogej, T., Tyrchan, C. & Waller, M. P. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Cent. Sci.* **4**, 120–131 (2018).

62. Olivecrona, M., Blaschke, T., Engkvist, O. & Chen, H. Molecular de-novo design through deep reinforcement learning. *J. Cheminformatics* **9**, 48 (2017).
63. Arús-Pous, J. *et al.* Exploring the GDB-13 chemical space using deep generative models. *J. Cheminformatics* **11**, 20 (2019).
64. Winter, R. *et al.* Efficient multi-objective molecular optimization in a continuous latent space. *Chem. Sci.* **10**, 8016–8024 (2019).
65. Kotsias, P.-C. *et al.* Direct steering of de novo molecular generation with descriptor conditional recurrent neural networks. *Nat. Mach. Intell.* **2**, 254–265 (2020).
66. Krishna Gottipati, S. *et al.* Learning To Navigate The Synthetically Accessible Chemical Space Using Reinforcement Learning. *arXiv* arXiv:2004.12485 (2020).
67. Bradshaw, J., Paige, B., Kusner, M. J., Segler, M. H. S. & Hernández-Lobato, J. M. A Model to Search for Synthesizable Molecules. *arXiv e-prints* arXiv:1906.05221 (2019).
68. Ghosh, S. *et al.* Structure-based virtual screening of chemical libraries for drug discovery This review comes from a themed issue on Combinatorial chemistry and molecular diversity Edited. *Curr. Opin. Chem. Biol.* **10**, 194–202 (2006).
69. Schneider, G. Virtual screening: an endless staircase? *Nat. Rev. Drug Discov.* **9**, 273–276 (2010).
70. Pereira, J. C., Caffarena, E. R. & Dos Santos, C. N. Boosting Docking-Based Virtual Screening with Deep Learning. *J. Chem. Inf. Model.* (2016) doi:10.1021/acs.jcim.6b00355.
71. Muegge, I. & Oloff, S. Advances in virtual screening. *Drug Discov. Today Technol.* doi:10.1016/j.ddtec.2006.12.002.
72. Horvath, D. A Virtual Screening Approach Applied to the Search for Trypanothione Reductase Inhibitors. *J. Med. Chem.* **40**, 2412–2423 (1997).

73. Maia, E. H. B., Assis, L. C., de Oliveira, T. A., da Silva, A. M. & Taranto, A. G. Structure-Based Virtual Screening: From Classical to Artificial Intelligence . *Frontiers in Chemistry* vol. 8 343 (2020).
74. Oprea, T. I. & Gottfries, J. Chemography: The Art of Navigating in Chemical Space. *J. Comb. Chem.* **3**, 157–166 (2001).
75. Awale, M., Van Deursen, R. & Reymond, J. L. MQN-mapplet: Visualization of chemical space with interactive maps of DrugBank, ChEMBL, PubChem, GDB-11, and GDB-13. *J. Chem. Inf. Model.* (2013) doi:10.1021/ci300513m.
76. Awale, M. & Reymond, J.-L. Similarity Mapplet: Interactive Visualization of the Directory of Useful Decoys and ChEMBL in High Dimensional Chemical Spaces. *J. Chem. Inf. Model.* **55**, 1509–1516 (2015).
77. Jin, X. *et al.* PDB-Explorer: a web-based interactive map of the protein data bank in shape space. *BMC Bioinformatics* **16**, 339 (2015).
78. Awale, M. & Reymond, J.-L. Web-based 3D-visualization of the DrugBank chemical space. *J. Cheminformatics* **8**, 25 (2016).
79. Awale, M., Probst, D. & Reymond, J.-L. WebMolCS: A Web-Based Interface for Visualizing Molecules in Three-Dimensional Chemical Spaces. *J. Chem. Inf. Model.* **57**, 643–649 (2017).
80. Probst, D. & Reymond, J.-L. FUn: a framework for interactive visualizations of large, high-dimensional datasets on the web. *Bioinformatics* **34**, 1433–1435 (2018).
81. Pearson, K. LIII. On lines and planes of closest fit to systems of points in space. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **2**, 559–572 (1901).
82. Hotelling, H. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **24**, 417–441 (1933).

83. Van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, (2008).
84. McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *ArXiv Prepr. ArXiv180203426* (2018).
85. T. Kohonen. Exploration of very large databases by self-organizing maps. in *Proceedings of International Conference on Neural Networks (ICNN'97)* vol. 1 PL1-PL6 vol.1 (1997).
86. Bishop, C. M., Svensén, M. & Williams, C. K. I. GTM: The Generative Topographic Mapping. *Neural Comput.* **10**, 215–234 (1998).
87. Probst, D. & Reymond, J.-L. Visualization of very large high-dimensional data sets as minimum spanning trees. *J. Cheminformatics* **12**, 12 (2020).
88. Beilstein, F. K. *Beilstein Handbook of Organic Chemistry: Collective Indexes. 5. Supplementary Series: Covering the Literature from 1960 Through 1979. Compound-name Index for Volumes 23-25: E-Px.* (Beilstein Information, 1994).
89. van Hilten, L. G. *How data scientists are uncovering chemical compounds hidden in patents.* (Elsevier Connect, November, 2019).
90. InfoChem. SPRESI. <https://www.infochem.de/about/spresi> (2019).
91. John Mayfield, Daniel Lowe, & Roger Sayle. *Pistachio*. (NextMove Software, 2018).
92. Lowe, D. Chemical Reactions from US Patents (1976-Sep2016). https://figshare.com/articles/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873 (2018).
93. Lowe, D. M. Extraction of chemical structures and reactions from the literature (Doctoral thesis). (2012). doi:10.17863/CAM.16293.

94. Jensen, K. F., Coley, C. W. & Eyke, N. S. Autonomous discovery in the chemical sciences part I: Progress. *Angew. Chem. Int. Ed.* **59**, 22858–22893 (2019).
95. Jia, X. *et al.* Anthropogenic Biases in Chemical Reaction Data Hinder Exploratory Inorganic Synthesis. *Nature* **573**, 251–255 (2019).
96. Kearnes, S. M. *et al.* The Open Reaction Database. *J. Am. Chem. Soc.* (2021) doi:10.1021/jacs.1c09820.
97. Buitrago Santanilla, A. *et al.* Nanomole-scale high-throughput chemistry for the synthesis of complex molecules. **347**, 49–53 (2015).
98. Lin, S. *et al.* Mapping the dark space of chemical reactions with extended nanomole synthesis and MALDI-TOF MS. *Science* **361**, eaar6236–eaar6236 (2018).
99. Steiner, S. *et al.* Organic synthesis in a modular robotic system driven by a chemical programming language. *Science* **363**, eaav2211 (2019).
100. Vaucher, A. C. *et al.* Automated extraction of chemical synthesis actions from experimental procedures. *Nat. Commun.* **11**, 3601 (2020).
101. UDM. (Pistoia Alliance, 2021).
102. Robinson, R. LXIII.—A synthesis of tropinone. *J. Chem. Soc. Trans.* **111**, 762–768 (1917).
103. Corey, E. J. & Wipke, W. T. Computer-assisted design of complex organic syntheses. *Science* **166**, 178–92 (1969).
104. Corey, E. J., Long, A. K. & Rubenstein, S. D. Computer-Assisted Analysis in Organic Synthesis. *Science* **228**, 408–418 (1985).
105. Corey, E. J. General methods for the construction of complex molecules. in *The Chemistry of Natural Products* 19–37 (Butterworth-Heinemann, 1967). doi:https://doi.org/10.1016/B978-0-08-020741-4.50004-X.

106. Pensak, D. A. & Corey, E. J. LHASA—Logic and Heuristics Applied to Synthetic Analysis. in *Computer-Assisted Organic Synthesis* vol. 61 1–32 (American Chemical Society, 1977).
107. Warren, S. & Wyatt, P. *Organic Synthesis: The Disconnection Approach, 2nd Edition*. (Wiley, 2008).
108. Wipke, W. T., Ouchi, G. I. & Krishnan, S. Simulation and evaluation of chemical synthesis—SECS: An application of artificial intelligence techniques. *Artif. Intell.* **11**, 173–193 (1978).
109. Gelernter, H. L. *et al.* Empirical Explorations of SYNCHEM. *Science* **197**, 1041 LP – 1049 (1977).
110. Wiswesser, W. J. *A Line-Formula Chemical Notation*. (Thomas Crowell Company publishers, 1954).
111. Gasteiger, J. & Jochum, C. EROS A computer program for generating sequences of reactions. in 93–126 (Springer Berlin Heidelberg, 1978).
112. Gasteiger, J. *et al.* A new treatment of chemical reactivity: Development of EROS, an expert system for reaction prediction and synthesis design BT - Organic Synthesis, Reactions and Mechanisms. in 19–73 (Springer Berlin Heidelberg, 1987).
113. Gasteiger, J. *et al.* Models for the representation of knowledge about chemical reactions. *J. Chem. Inf. Comput. Sci.* **30**, 467–476 (1990).
114. Gasteiger, J., Ihlenfeldt, W. D. & Röse, P. A collection of computer methods for synthesis design and reaction prediction. *Recl. Trav. Chim. Pays-Bas* **111**, 270–290 (1992).
115. Jorgensen, W. L. *et al.* CAMEO: a program for the logical prediction of the products of organic reactions. *Pure Appl. Chem.* **62**, 1921 (1990).
116. Ihlenfeldt, W.-D. & Gasteiger, J. Computer-Assisted Planning of Organic Syntheses: The Second Generation of Programs. *Angew. Chem. Int. Ed. Engl.* **34**, 2613–2633 (1996).

117. Ugi, I. *et al.* Computer-Assisted Solution of Chemical Problems—The Historical Development and the Present State of the Art of a New Discipline of Chemistry. *Angew. Chem. Int. Ed. Engl.* **32**, 201–227 (1993).
118. Ugi, I. *et al.* Models, concepts, theories, and formal languages in chemistry and their use as a basis for computer assistance in chemistry. *J. Chem. Inf. Model.* **34**, 3–16 (1994).
119. Ugi, I. *et al.* New Applications of Computers in Chemistry. *Angew. Chem. Int. Ed. Engl.* **18**, 111–123 (1979).
120. Dugundji, J. & Ugi, I. An algebraic model of constitutional chemistry as a basis for chemical computer programs. in *Computers in Chemistry* 19–64 (Springer Berlin Heidelberg, 1973).
121. Satoh, H. & Funatsu, K. SOPHIA, a Knowledge Base-Guided Reaction Prediction System - Utilization of a Knowledge Base Derived from a Reaction Database. *J. Chem. Inf. Comput. Sci.* **35**, 34–44 (1995).
122. Szymkuć, S. *et al.* Computer-Assisted Synthetic Planning: The End of the Beginning. *Angew. Chem. Int. Ed.* **55**, 5904–5937 (2016).
123. Klucznik, T. *et al.* Efficient Syntheses of Diverse, Medicinally Relevant Targets Planned by Computer and Executed in the Laboratory. *Chem* **4**, 522–532 (2018).
124. Molga, K., Gajewska, E. P., Szymkuć, S. & Grzybowski, B. A. The logic of translating chemical knowledge into machine-processable forms: a modern playground for physical-organic chemistry. *React. Chem. Eng.* **4**, 1506–1521 (2019).
125. Skoraczyński, G. *et al.* Predicting the outcomes of organic reactions via machine learning: Are current descriptors sufficient? *Sci. Rep.* (2017) doi:10.1038/s41598-017-02303-0.
126. Mikulak-Klucznik, B. *et al.* Computational planning of the synthesis of complex natural products. *Nature* (2020) doi:10.1038/s41586-020-2855-y.

127. Molga, K., Dittwald, P. & Grzybowski, B. A. Navigating around Patented Routes by Preserving Specific Motifs along Computer-Planned Retrosynthetic Pathways. *Chem* **5**, 460–473 (2019).
128. Jaworski, W. *et al.* Automatic mapping of atoms across both simple and complex chemical reactions. *Nat. Commun.* **10**, 1434 (2019).
129. Warr, W. A. A Short Review of Chemical Reaction Database Systems, Computer-Aided Synthesis Design, Reaction Prediction and Synthetic Feasibility. *Mol. Inform.* **33**, 469–476 (2014).
130. Todd, M. H. Computer-aided organic synthesis. *Chem. Soc. Rev.* **34**, 247 (2005).
131. Jensen, K. F., Coley, C. W. & Eyke, N. S. Autonomous discovery in the chemical sciences part I: Progress. *Angew. Chem. Int. Ed.* (2019) doi:10.1002/anie.201909987.
132. Law, J. *et al.* Route Designer: A Retrosynthetic Analysis Tool Utilizing Automated Retrosynthetic Rule Generation. *J. Chem. Inf. Model.* **49**, 593–602 (2009).
133. Bøgevig, A. *et al.* Route Design in the 21st Century: The IC *SYNTH* Software Tool as an Idea Generator for Synthesis Prediction. *Org. Process Res. Dev.* **19**, 357–368 (2015).
134. Coley, C. W., Barzilay, R., Jaakkola, T. S., Green, W. H. & Jensen, K. F. Prediction of Organic Reaction Outcomes Using Machine Learning. *ACS Cent. Sci.* **3**, 434–443 (2017).
135. Segler, M. H. S. & Waller, M. P. Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction. *Chem. - Eur. J.* **23**, 5966–5971 (2017).
136. 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. *Chem. Sci.* **11**, 154–168 (2020).

137. Baylon, J. L., Cilfone, N. A., Gulcher, J. R. & Chittenden, T. W. Enhancing Retrosynthetic Reaction Prediction with Deep Learning Using Multiscale Reaction Classification. *J. Chem. Inf. Model.* **59**, 673–688 (2019).
138. Coley, C. W. *et al.* A Robotic Platform for Flow Synthesis of Organic Compounds Informed by AI Planning. *Science* **365**, eaax1566 (2019).
139. Wei, J. N., Duvenaud, D. & Aspuru-Guzik, A. Neural Networks for the Prediction of Organic Chemistry Reactions. *ACS Cent. Sci.* **2**, 725–732 (2016).
140. Silver, D. *et al.* Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484 (2016).
141. Gao, H. *et al.* Using Machine Learning To Predict Suitable Conditions for Organic Reactions. *ACS Cent. Sci.* **4**, 1465–1476 (2018).
142. Coley, C. W. *et al.* A graph-convolutional neural network model for the prediction of chemical reactivity. *Chem. Sci.* **10**, 370–377 (2019).
143. 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.* **63**, 8791–8808 (2020).
144. Bjerrum, E. J., Thakkar, A. & Engkvist, O. Artificial Applicability Labels for Improving Policies in Retrosynthesis Prediction. *Mach. Learn. Sci. Technol.* **2**, (2020).
145. Fortunato, M. E., Coley, C. W., Barnes, B. C. & Jensen, K. F. Data Augmentation and Pretraining for Template-Based Retrosynthetic Prediction in Computer-Aided Synthesis Planning. *J. Chem. Inf. Model.* (2020) doi:10.1021/acs.jcim.0c00403.
146. Liu, B. *et al.* Retrosynthetic Reaction Prediction Using Neural Sequence-to-Sequence Models. *ACS Cent. Sci.* **3**, 1103–1113 (2017).

147. Cadeddu, A., Wylie, E. K., Jurczak, J., Wampler-Doty, M. & Grzybowski, B. A. Organic chemistry as a language and the implications of chemical linguistics for structural and retrosynthetic analyses. *Angew. Chem. - Int. Ed.* (2014) doi:10.1002/anie.201403708.
148. Nam, J. & Kim, J. Linking the Neural Machine Translation and the Prediction of Organic Chemistry Reactions. *arXiv:1612.09529* (2016) doi:arXiv:1612.09529.
149. 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. *Chem. Sci.* **9**, 6091–6098 (2018).
150. Öztürk, H., Özgür, A., Schwaller, P., Laino, T. & Ozkirimli, E. Exploring chemical space using natural language processing methodologies for drug discovery. *Drug Discov. Today* (2020) doi:https://doi.org/10.1016/j.drudis.2020.01.020.
151. Schwaller, P. *et al.* Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Cent. Sci.* **5**, 1572–1583 (2019).
152. Karpov, P., Godin, G. & Tetko, I. V. A Transformer Model for Retrosynthesis. in *Artificial Neural Networks and Machine Learning – ICANN 2019: Workshop and Special Sessions* (eds. Tetko, I. V, Kůrková, V., Karpov, P. & Theis, F.) 817–830 (Springer International Publishing, 2019).
153. Schwaller, P., Hoover, B., Reymond, J.-L., Strobelt, H. & Laino, T. Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. *Sci. Adv.* **7**, eabe4166 (2021).
154. Bort, W. *et al.* Discovery of novel chemical reactions by deep generative recurrent neural network. *Sci. Rep.* **11**, 3178 (2021).

155. Kayala, M. A., Azencott, C.-A., Chen, J. H. & Baldi, P. Learning to Predict Chemical Reactions. *J. Chem. Inf. Model.* **51**, 2209–2222 (2011).
156. Fooshee, D. *et al.* Deep learning for chemical reaction prediction. *Mol. Syst. Des. Eng.* **3**, 442–452 (2018).
157. Ishida, S., Terayama, K., Kojima, R., Takasu, K. & Okuno, Y. Prediction and Interpretable Visualization of Retrosynthetic Reactions Using Graph Convolutional Networks. *J. Chem. Inf. Model.* **59**, 5026–5033 (2019).
158. Shi, C., Xu, M., Guo, H., Zhang, M. & Tang, J. A Graph to Graphs Framework for Retrosynthesis Prediction. *arXiv e-prints* arXiv:2003.12725 (2020).
159. Somnath, V. R., Bunne, C., Coley, C. W., Krause, A. & Barzilay, R. Learning Graph Models for Template-Free Retrosynthesis. *arXiv e-prints* arXiv:2006.07038 (2020).
160. Jacob, P.-M. & Lapkin, A. Statistics of the network of organic chemistry. *React. Chem. Eng.* **3**, 102–118 (2018).
161. Weber, J. M., Lió, P. & Lapkin, A. A. Identification of strategic molecules for future circular supply chains using large reaction networks. *React. Chem. Eng.* **4**, 1969–1981 (2019).
162. Grzybowski, B. A., Bishop, K. J. M., Kowalczyk, B. & Wilmer, C. E. The Wired Universe of Organic Chemistry. *Nat. Chem.* **1**, 31 (2009).
163. Jacob, P.-M. & Lapkin, A. Prediction of Chemical Reactions Using Statistical Models of Chemical Knowledge. *ChemRxiv* (2018) doi:10.26434/chemrxiv.6954908.v1.
164. Segler, M. H. S. & Waller, M. P. Modelling Chemical Reasoning to Predict and Invent Reactions. *Chem. – Eur. J.* **23**, 6118–6128 (2017).
165. Engkvist, O. *et al.* Computational prediction of chemical reactions: current status and outlook. *Drug Discov. Today* (2018) doi:10.1016/j.drudis.2018.02.014.

166. Schreck, J. S., Coley, C. W. & Bishop, K. J. M. Learning Retrosynthetic Planning through Simulated Experience. *ACS Cent. Sci.* **5**, 970–981 (2019).
167. Christ, C. D., Zentgraf, M. & Kriegl, J. M. Mining Electronic Laboratory Notebooks: Analysis, Retrosynthesis, and Reaction Based Enumeration. *J. Chem. Inf. Model.* **52**, 1745–1756 (2012).
168. Coley, C. W., Green, W. H. & Jensen, K. F. Machine Learning in Computer-Aided Synthesis Planning. *Acc. Chem. Res.* **51**, 1281–1289 (2018).
169. Schneider, N., Lowe, D. M., Sayle, R. A., Tarselli, M. A. & Landrum, G. A. Big Data from Pharmaceutical Patents: A Computational Analysis of Medicinal Chemists' Bread and Butter. *J. Med. Chem.* **59**, 4385–4402 (2016).
170. Surrey, A. R. *Name Reactions in Organic Chemistry (Second Edition)*. (Academic Press, 1961). doi:10.1016/B978-1-4832-3227-0.50004-3.
171. Baskin, I. I., Madzhidov, T. I., Antipin, I. S. & Varnek, A. A. Artificial Intelligence in Synthetic Chemistry: Achievements and Prospects. *Russ. Chem. Rev.* **86**, 1127–1156 (2017).
172. Yadav, M. K. On the synthesis of machine learning and automated reasoning for an artificial synthetic organic chemist. *New J. Chem.* **41**, 1411–1416 (2017).
173. Ravitz, O. Data-driven computer aided synthesis design. *Drug Discov. Today Technol.* **10**, e443–e449 (2013).
174. Ley, S. V, Fitzpatrick, D. E., Ingham, R. J. & Myers, R. M. Organic Synthesis: March of the Machines. *Angew. Chem. Int. Ed.* **54**, 3449–3464 (2015).
175. Olivecrona, M., Blaschke, T., Engkvist, O. & Chen, H. Molecular de-novo design through deep reinforcement learning. *J. Cheminformatics* **9**, 48 (2017).

176. Plowright, A. T. *et al.* Hypothesis driven drug design: improving quality and effectiveness of the design-make-test-analyse cycle. *Drug Discov. Today* **17**, 56–62 (2012).
177. Green, C. P., Engkvist, O. & Pairaudeau, G. The Convergence of Artificial Intelligence and Chemistry for Improved Drug Discovery. *Future Med. Chem.* **10**, 2573–2576 (2018).
178. Reaxys© , Copyright © 2019 Elsevier Limited except certain content provided by third parties. Reaxys is a trademark of Elsevier Limited.
179. Flick, A. C. *et al.* Synthetic Approaches to the New Drugs Approved During 2017. *J. Med. Chem.* (2019) doi:10.1021/acs.jmedchem.9b00196.
180. NextMove Software | HazELNut. <https://www.nextmovesoftware.com/hazelnut.html>.
181. 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.* **59**, 2529–2537 (2019).
182. NextMove Software | NameRxn. <https://www.nextmovesoftware.com/hazelnut.html>.
183. Daylight Theory Manual, Daylight Version 4.9, Release Date 08/01/11. <https://daylight.com/dayhtml/doc/theory/theory.smiles.html#RTFrnx5>.
184. Peter G. M. Wuts, T. W. G. *Greene's Protective Groups in Organic Synthesis*. (John Wiley & Sons, Inc, 2006). doi:10.1002/0470053488.
185. <https://www.organic-chemistry.org/protectivegroups/>.
186. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in {P}ython. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
187. Chollet, F. (2015) Keras, GitHub. <https://github.com/fchollet/keras>. (1AD).
188. Abadi, M. *et al.* *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*.

189. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *ArXiv arXiv:1412*, (2014).
190. Gaulton, A. *et al.* The ChEMBL database in 2017. *Nucleic Acids Res.* **45**, D945–D954 (2017).
191. Browne, C. B. *et al.* A Survey of Monte Carlo Tree Search Methods. *IEEE Trans. Comput. Intell. AI Games* **4**, 1–43 (2012).
192. Hagberg, A. A., Schult, D. A. & Swart, P. J. Exploring Network Structure, Dynamics, and Function using NetworkX. in *Proceedings of the 7th Python in Science Conference (SciPy2008)* (eds. Varoquaux, G., Vaught, T. & Millman, J.) 11–15 (2008).
193. <https://www.acdlabs.com/index.php>.
194. Njardarson, J. T. Top 200 Brand Name Drugs by Retail Sales in 2018. https://njardarson.lab.arizona.edu/sites/njardarson.lab.arizona.edu/files/2018Top200PharmaceuticalRetailSalesPosterLowResFinal_0.pdf.
195. Taylor, R. D., MacCoss, M. & Lawson, A. D. G. Rings in Drugs. *J. Med. Chem.* **57**, 5845–5859 (2014).
196. 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.* **54**, 3451–3479 (2011).
197. Boström, J., Brown, D. G., Young, R. J. & Keserü, G. M. Expanding the Medicinal Chemistry Synthetic Toolbox. *Nat. Rev. Drug Discov.* **17**, 709 (2018).
198. Visini, R., Arús-Pous, J., Awale, M. & Reymond, J.-L. Virtual Exploration of the Ring Systems Chemical Universe. *J. Chem. Inf. Model.* **57**, 2707–2718 (2017).
199. Liu, Y. & Gray, N. S. Rational Design of Inhibitors that Bind to Inactive Kinase Conformations. *Nat. Chem. Biol.* **2**, 358–364 (2006).

200. Pitt, W. R., Parry, D. M., Perry, B. G. & Groom, C. R. Heteroaromatic Rings of the Future. *J. Med. Chem.* **52**, 2952–2963 (2009).
201. Mok, N. Y. & Brown, N. Applications of Systematic Molecular Scaffold Enumeration to Enrich Structure–Activity Relationship Information. *J. Chem. Inf. Model.* **57**, 27–35 (2017).
202. Tyagarajan, S. *et al.* Heterocyclic Regioisomer Enumeration (HREMS): A Cheminformatics Design Tool. *J. Chem. Inf. Model.* **55**, 1130–1135 (2015).
203. Ward, R. A. & Kettle, J. G. Systematic Enumeration of Heteroaromatic Ring Systems as Reagents for Use in Medicinal Chemistry. *J. Med. Chem.* **54**, 4670–4677 (2011).
204. Silva Júnior, P. E. *et al.* Synthesis of Two ‘Heteroaromatic Rings of the Future’ for Applications in Medicinal Chemistry. *RSC Adv.* **6**, 22777–22780 (2016).
205. ZINC. <http://zinc.docking.org/rings/subsets/>.
206. Wishart, D. S. *et al.* DrugBank 5.0: A Major Update to the DrugBank Database for 2018. *Nucleic Acids Res.* **46**, D1074–D1082 (2017).
207. 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. *J. Org. Chem.* **81**, 6563–6572 (2016).
208. Movassaghi, M. & Hill, M. D. A Versatile Cyclodehydration Reaction for the Synthesis of Isoquinoline and β -Carboline Derivatives. *Org. Lett.* **10**, 3485–3488 (2008).
209. 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. *J. Org. Chem.* **68**, 5392–5394 (2003).

210. Minetto, G., Raveglia, L. F., Sega, A. & Taddei, M. Microwave-Assisted Paal–Knorr Reaction – Three-Step Regiocontrolled Synthesis of Polysubstituted Furans, Pyrroles and Thiophenes. *Eur. J. Org. Chem.* 5277–5288 (2005) doi:10.1002/ejoc.200500387.
211. Blaschke, T. *et al.* REINVENT 2.0 – an AI Tool for De Novo Drug Design. *ChemRxiv* (2020) doi:10.26434/chemrxiv.12058026.v2.
212. Mayr, A., Klambauer, G., Unterthiner, T. & Hochreiter, S. DeepTox: Toxicity Prediction using Deep Learning. *Front. Environ. Sci.* **3**, 80 (2016).
213. Walters, W. P. Virtual Chemical Libraries. *J. Med. Chem.* **62**, 1116–1124 (2019).
214. Chevillard, F. & Kolb, P. SCUBIDOO: A Large yet Screenable and Easily Searchable Database of Computationally Created Chemical Compounds Optimized toward High Likelihood of Synthetic Tractability. *J. Chem. Inf. Model.* **55**, 1824–1835 (2015).
215. Zhavoronkov, A. *et al.* Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat. Biotechnol.* **37**, 1038–1040 (2019).
216. Baber, J. C. & Feher, M. Predicting Synthetic Accessibility: Application in Drug Discovery and Development. *Mini-Rev. Med. Chem.* **4**, 681–692 (2004).
217. Gillet, V. J., Myatt, G., Zsoldos, Z. & Johnson, A. P. SPROUT, HIPPO and CAESA: Tools for de novo structure generation and estimation of synthetic accessibility. *Perspect. Drug Discov. Des.* **3**, 34–50 (1995).
218. Genheden, S. *et al.* AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning. *J. Cheminformatics* **12**, 70 (2020).
219. Gao, W. & Coley, C. W. The Synthesizability of Molecules Proposed by Generative Models. *J. Chem. Inf. Model.* (2020) doi:10.1021/acs.jcim.0c00174.

220. Cook, A. *et al.* Computer-aided synthesis design: 40 years on. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2**, 79–107 (2012).
221. Coley, C. W., Rogers, L., Green, W. H. & Jensen, K. F. SCScore: Synthetic Complexity Learned from a Reaction Corpus. *J. Chem. Inf. Model.* **58**, 252–261 (2018).
222. Voršilák, M., Kolář, M., Čmelo, I. & Svozil, D. SYBA: Bayesian estimation of synthetic accessibility of organic compounds. *J. Cheminformatics* **12**, 35 (2020).
223. Ertl, P. & Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminformatics* **1**, 8 (2009).
224. Enamine. <https://enamine.net/building-blocks>.
225. Sterling, T. & Irwin, J. J. ZINC 15 - Ligand Discovery for Everyone. *J. Chem. Inf. Model.* (2015) doi:10.1021/acs.jcim.5b00559.
226. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: A Next-generation Hyperparameter Optimization Framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* 2623–2631 (2019) doi:10.1145/3292500.3330701.
227. Gómez-Bombarelli, R. *et al.* Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **4**, 268–276 (2018).
228. Méndez-Lucio, O., Baillif, B., Clevert, D.-A., Rouquié, D. & Wichard, J. De novo generation of hit-like molecules from gene expression signatures using artificial intelligence. *Nat. Commun.* **11**, 10 (2020).
229. Brown, N., Fiscato, M., Segler, M. H. S. & Vaucher, A. C. GuacaMol: Benchmarking Models for de Novo Molecular Design. *J. Chem. Inf. Model.* **59**, 1096–1108 (2019).

230. Polykovskiy, D. *et al.* Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. arXiv:1811.12823 (2018).
231. Irwin, J. J., Sterling, T., Mysinger, M. M., Bolstad, E. S. & Coleman, R. G. ZINC: A Free Tool to Discover Chemistry for Biology. *J. Chem. Inf. Model.* **52**, 1757–1768 (2012).
232. Sheridan, R. P. *et al.* Modeling a Crowdsourced Definition of Molecular Complexity. *J. Chem. Inf. Model.* **54**, 1604–1616 (2014).
233. Thakkar, A., Chadimová, V., Bjerrum, E. J., Engkvist, O. & Reymond, J.-L. Retrosynthetic accessibility score (RAscore) – rapid machine learned synthesizability classification from AI driven retrosynthetic planning. *Chem. Sci.* **12**, 3339–3349 (2021).
234. Bradshaw, J., Paige, B., Kusner, M. J., Segler, M. H. S. & Hernández-Lobato, J. M. A Model to Search for Synthesizable Molecules. *ArXiv E-Prints* arXiv:1906.05221 (2019).
235. Krishna Gottipati, S. *et al.* Learning To Navigate The Synthetically Accessible Chemical Space Using Reinforcement Learning. *arXiv* arXiv:2004.12485 (2020).
236. Delalande, C. *et al.* Optimizing TRPM4 inhibitors in the MHFP6 chemical space. *Eur. J. Med. Chem.* **166**, 167–177 (2019).
237. ZINC. <http://zinc.docking.org/>.
238. Blum, L. C., van Deursen, R. & Reymond, J.-L. Visualisation and subsets of the chemical universe database GDB-13 for virtual screening. *J. Comput. Aided Mol. Des.* **25**, 637–647 (2011).
239. Meier, K., Arús-Pous, J. & Reymond, J.-L. A Potent and Selective Janus Kinase Inhibitor with a Chiral 3D-Shaped Triquinazine Ring System from Chemical Space. *Angew. Chem. Int. Ed.* **60**, 2074–2077 (2021).

240. Meier, K., Bühlmann, S., Arús-Pous, J. & Reymond, J.-L. The Generated Databases (GDBs) as a Source of 3D-shaped Building Blocks for Use in Medicinal Chemistry and Drug Discovery. *Chim. Int. J. Chem.* **74**, 241–246 (2020).
241. Ratni, H. *et al.* Discovery of RO7185876, a Highly Potent γ -Secretase Modulator (GSM) as a Potential Treatment for Alzheimer's Disease. *ACS Med. Chem. Lett.* **11**, 1257–1268 (2020).
242. Schwaller, P. *et al.* Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chem. Sci.* **11**, 3316–3325 (2020).
243. Liu, B. *et al.* Retrosynthetic Reaction Prediction Using Neural Sequence-to-Sequence Models. *ACS Cent. Sci.* **3**, 1103–1113 (2017).
244. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
245. Ahneman, D. T., Estrada, J. G., Lin, S., Dreher, S. D. & Doyle, A. G. Predicting reaction performance in C–N cross-coupling using machine learning. *Science* **360**, 186–190 (2018).
246. Eyke, N. S., Green, W. H. & Jensen, K. F. Iterative experimental design based on active machine learning reduces the experimental burden associated with reaction screening. *React. Chem. Eng.* **5**, 1963–1972 (2020).
247. RMG Database. <https://rmg.mit.edu/database/>.
248. iochem. <https://www.iochem-bd.org/>.
249. Smith, D. G. A. *et al.* The MolSSI QCArchive project: An open-source platform to compute, organize, and share quantum chemistry data. *WIREs Comput. Mol. Sci.* **n/a**, e1491 (2020).
250. The MolSSI Quantum Chemistry Archive. <https://qcarchive.molssi.org/>.
251. Flick, A. C. *et al.* Synthetic Approaches to New Drugs Approved during 2018. *J. Med. Chem.* (2020) doi:10.1021/acs.jmedchem.0c00345.

252. Rohrbach, S. *et al.* Concerted Nucleophilic Aromatic Substitution Reactions. **58**, 16368–16388 (2019).
253. Walters, W. P. Modeling, Informatics, and the Quest for Reproducibility. *J. Chem. Inf. Model.* **53**, 1529–1530 (2013).
254. Clark, R. D. A path to next-generation reproducibility in cheminformatics. *J. Cheminformatics* **11**, 62 (2019).
255. Landrum, G. A. Reproducibility in cheminformatics and computational chemistry research: certainly we can do better than this. *J. Cheminformatics* **5**, O4 (2013).
256. MELLODDY. <https://www.melloddy.eu/>.
257. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
258. Struble, T. J., Coley, C. W. & Jensen, K. F. Multitask prediction of site selectivity in aromatic C–H functionalization reactions. *React. Chem. Eng.* **5**, 896–902 (2020).
259. 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.* **11**, 4874 (2020).
260. Kromann, J. C., Jensen, J. H., Kruszyk, M., Jessing, M. & Jørgensen, M. Fast and accurate prediction of the regioselectivity of electrophilic aromatic substitution reactions. *Chem. Sci.* **9**, 660–665 (2018).
261. Smith, J. S., Isayev, O. & Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* **8**, 3192–3203 (2017).
262. Noé, F., Tkatchenko, A., Müller, K.-R. & Clementi, C. Machine Learning for Molecular Simulation. *Annu. Rev. Phys. Chem.* **71**, 361–390 (2020).

263. Ramakrishnan, R., Dral, P. O., Rupp, M. & Anatole Von Lilienfeld, O. Big Data Meets Quantum Chemistry Approximations: The Δ -Machine Learning Approach. *J Chem Theory Comput* **13**, 25–25 (2015).
264. Gensch, T. *et al.* A Comprehensive Discovery Platform for Organophosphorus Ligands for Catalysis. (2021) doi:10.26434/chemrxiv.12996665.v1.
265. Santiago, C. B., Guo, J.-Y. & Sigman, M. S. Predictive and mechanistic multivariate linear regression models for reaction development. *Chem. Sci.* **9**, 2398–2412 (2018).
266. Shields, B. J. *et al.* Bayesian reaction optimization as a tool for chemical synthesis. *Nature* **590**, 89–96 (2021).
267. Finnigan, W., Hepworth, L. J., Flitsch, S. L. & Turner, N. J. RetroBioCat as a computer-aided synthesis planning tool for biocatalytic reactions and cascades. *Nat. Catal.* **4**, 98–104 (2021).
268. Turner, N. J. & O'Reilly, E. Biocatalytic retrosynthesis. *Nat. Chem. Biol.* **9**, 285–288 (2013).
269. Probst, D. *et al.* Molecular Transformer-aided Biocatalysed Synthesis Planning. (2021) doi:10.26434/chemrxiv.14639007.v1.
270. Kreutter, D., Schwaller, P. & Reymond, J.-L. Predicting enzymatic reactions with a molecular transformer. *Chem. Sci.* **12**, 8648–8659 (2021).
271. Gao, H. *et al.* Combining retrosynthesis and mixed-integer optimization for minimizing the chemical inventory needed to realize a WHO essential medicines list. *React. Chem. Eng.* **5**, 367–376 (2020).
272. Weber, J. M., Guo, Z., Zhang, C., Schweidtmann, A. M. & Lapkin, A. A. Chemical data intelligence for sustainable chemistry. *Chem. Soc. Rev.* **50**, 12013–12036 (2021).
273. Weber, J. M., Lió, P. & Lapkin, A. A. Identification of strategic molecules for future circular supply chains using large reaction networks. *React. Chem. Eng.* **4**, 1969–1981 (2019).

274. Bishop, K. J. M., Klajn, R. & Grzybowski, B. A. The Core and Most Useful Molecules in Organic Chemistry. *Angew. Chem. Int. Ed.* **45**, 5348–5354 (2006).
275. Szymkuć, S. *et al.* Computer-generated “synthetic contingency” plans at times of logistics and supply problems: scenarios for hydroxychloroquine and remdesivir. *Chem. Sci.* **11**, 6736–6744 (2020).
276. Schneider, N., Lowe, D. M., Sayle, R. A. & Landrum, G. A. Development of a Novel Fingerprint for Chemical Reactions and Its Application to Large-Scale Reaction Classification and Similarity. *J. Chem. Inf. Model.* **55**, 39–53 (2015).
277. Schwaller, P. *et al.* Mapping the space of chemical reactions using attention-based neural networks. *Nat. Mach. Intell.* **3**, 144–152 (2021).
278. Schwaller, P., Hoover, B., Reymond, J.-L., Strobelt, H. & Laino, T. Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. *Sci. Adv.* **7**, eabe4166 (2021).
279. Chen, B., Li, C., Dai, H. & Song, L. Retro*: learning retrosynthetic planning with neural guided A* search. in *International Conference on Machine Learning* 1608–1616 (PMLR, 2020).
280. Tetko, I. V., Engkvist, O., Koch, U., Reymond, J.-L. & Chen, H. BIGCHEM: Challenges and Opportunities for Big Data Analysis in Chemistry. *Mol. Inform.* **35**, 615–621 (2016).
281. Big Data in Chemistry | BIGCHEM Project | Fact Sheet | H2020. *CORDIS / European Commission* <https://cordis.europa.eu/project/id/676434>.
282. MLPDS – Machine Learning for Pharmaceutical Discovery and Synthesis Consortium. <https://mlpds.mit.edu/>.
283. CCAS. <https://ccas.nd.edu/>.

284. Home. *SynTech CDT* <https://www.syntechcdt.com>.

285. Home. *AI 4 Scientific Discovery* <https://www.ai3sd.org/>.

286. Advanced machine learning for Innovative Drug Discovery | AIDD Project | Fact Sheet |
H2020. *CORDIS / European Commission* <https://cordis.europa.eu/project/id/956832>.

10 Declaration of Consent

Declaration of consent

on the basis of Article 18 of the PromR Phil.-nat. 19

Name/First Name: Thakkar, Amol

Registration Number: 17-144-122

Study program: Chemistry and Molecular Sciences

Bachelor ☐

Master ☐

Dissertation ☒

Title of the thesis: Computer Aided Synthesis Prediction to Enable Augmented Chemical Discovery and Chemical Space Exploration


Supervisor: Professor Dr. Jean-Louis Reymond

I declare herewith that this thesis is my own work and that I have not used any sources other than those stated. I have indicated the adoption of quotations as well as thoughts taken from other authors as such in the thesis. I am aware that the Senate pursuant to Article 36 paragraph 1 litera r of the University Act of September 5th, 1996 and Article 69 of the University Statute of June 7th, 2011 is authorized to revoke the doctoral degree awarded on the basis of this thesis.

For the purposes of evaluation and verification of compliance with the declaration of originality and the regulations governing plagiarism, I hereby grant the University of Bern the right to process my personal data and to perform the acts of use this requires, in particular, to reproduce the written thesis and to store it permanently in a database, and to use said database, or to make said database available, to enable comparison with theses submitted by others.

Bern 03-12-21

Place/Date


Signature