Learning the Language of Chemical Reactions – Atom by Atom

Linguistics-Inspired Machine Learning Methods for Chemical Reaction Tasks

Inaugural dissertation
of the Faculty of Science,
University of Bern

presented by

Philippe Schwaller

from Luterbach, SO

Supervisors of the doctoral thesis:
Prof. Dr. Jean-Louis Reymond
Department of Chemistry and Biochemistry, University of Bern

Dr. Teodoro Laino
IBM Research – Europe

Learning the Language of Chemical Reactions – Atom by Atom

Linguistics-Inspired Machine Learning Methods for Chemical Reaction Tasks

Inaugural dissertation
of the Faculty of Science,
University of Bern

presented by

Philippe Schwaller

from Luterbach, SO

Supervisors of the doctoral thesis:
Prof. Dr. Jean-Louis Reymond
Department of Chemistry and Biochemistry, University of Bern

Dr. Teodoro Laino
IBM Research – Europe

Accepted by the Faculty of Science.

Bern, 22.03.2021

The Dean
Prof. Dr. Zoltan Balogh

"If I have seen further, it is by standing on the shoulders of giants."

- Isaac Newton, 1675

Dedicated to my grandmother Cécile Tâche,
and in loving memory of my grandfather Jean Tâche …

Acknowledgements

The work presented in my thesis would not have been possible without the support of numerous people who endorsed me, supported me, and contributed to the individual projects. I count myself lucky to have been part of many thriving collaborations.

First of all, I wish to express my gratitude to my supervisors Prof. Jean-Louis Reymond and Dr. Teodoro Laino, for the guidance, trust, and freedom they have given to me to explore my ideas.

Jean-Louis, thank you for welcoming me in the group, allowing me to work on exciting projects in close collaboration with synthetic chemists and supporting me throughout my thesis. I love your openness to new approaches, your vision for impactful research, and the unique setup you have in your lab with a well-balanced mixture of synthetic chemists, biochemists, and cheminformaticians.

Many thanks, Teo, for giving me the opportunity to join IBM Research, for constantly believing in what I was doing, and for encouraging me to push the boundaries. I also enjoyed our 1-to-1 conversations, after which I always felt a bit wiser, not only scientifically. Your advice helped me to make important decisions. I much appreciate the exposure you allowed me to have within IBM and externally. I am proud of being part of the outstanding RXN for Chemistry team that you have built over the last years.

Théo, living together and developing the prototypes IBM RXN for Chemistry was great fun – it felt almost like solving Kaggle challenges in a team or building up a small start-up. Thank you very much for making the initial contact, without you I probably would not have joined IBM in the first place. Alain, I much enjoyed our fruitful discussions and interactions. Your ability to analyse and dissect complex problems into smaller, more approachable ones impresses me. Alessandra, your positive energy and motivation is a great plus for the group. Fascinating to see how many concurrent projects you manage to deal with. Working with you is always a great pleasure! Daniel, it was excellent to collaborate with you first in Reymond group and now at IBM Research. Matteo, Joppe, Alec, what would I do without you. I appreciate all the work you do behind the scenes, on the cloud platforms, the codebase and the cluster. Thanks a lot to the other team members (Antonio, Federico, Heiko, Leonid, Dimitrios, Alessandro and Oliver), and the people from the IBM Research Zurich lab. Back when times were normal, I enjoyed daily coffee breaks and table tennis games with numerous people, including Roxana, Florian, Matthieu, Bojana, Ignacio, Igor, Riccardo (2x), Rico, Vishnu, Aris and Anton. Michael, after our time at the University of Manchester, it was a pleasure to meet you again in the Zurich lab. Let's see where our paths cross next. Thanks to the IBM Research communications team with Angela, Chris and Leonid. Let me also mention Linda and Anne-Marie who proof-read most of our publications and Claudia from the human resources, who took good care of me, thank you. During my PhD, I had the chance to work with people outside the Zurich lab. Ben and Hendrik, you are just awesome people, thanks for all. Our collaboration was one of the highlights of my PhD.

Next, I would like to thank all members of the Reymond group. I enjoyed coming to the University once a week. It was fantastic to be part of the group! Giorgio, thanks a lot for so being open and interested in novel methods. Working and discussing ideas with you was inspiring. I would have loved to spend more time with you in the lab. Amol, Alice, David, Sven and Josep, thank you for the exciting discussions, the beers at Sattler, and the Italian pasta. Sandra, thank you for taking care of the administrative work so that we as PhD students can focus on the science.

Ryan, I am looking forward to times when we can meet again and socialise at conferences. Bingqing, the final title of my thesis was inspired by one of our conversations. Thank you for that and for giving me the opportunity to be participate in your machine learning group meetings. I would also like to thank Michael, better known as Moret. I am excited to see what you all come up with in the future.

I really appreciate the time Prof. Alán Aspuru-Guzik and Prof. Philippe Renaud, who were part of my thesis examination committee, have taken and the valuable feedback they have given to me.

Without building on top of existing open-source frameworks and using published machine-accessible data, the work presented in this thesis would not be feasible. Many thanks to Greg, Nadine, Daniel, Roger, and all RDKit contributors for enabling such exciting research and great discussions. Thanks to Bastian for his great mimosis latex template, which I adapted to write this thesis.

I want to thank my friends from Düdingen - Matthias, Jan, Patrick, Claudio, Simon, Christophe, Jonathan, Ruben, and Benjamin. It is awesome how, after all those years and the different paths we have taken, we still meet up regularly.

I am deeply grateful to my family, mum, dad, and Julie, who always support me, independent of where I am and where I plan to go next.

And finally, a special thank you goes to Larissa – my fiancée and the most wonderful person in my life – for her constant love, support, and encouragement over the last ten years. I am looking forward to many more exciting adventures. There is nothing better than exploring the world together!

Abstract

Over the last hundred years, not much has changed how organic chemistry is conducted. In most laboratories, the current state is still trial-and-error experiments guided by human expertise acquired over decades. What if, given all the knowledge published, we could develop an artificial intelligence-based assistant to accelerate the discovery of novel molecules? Although many approaches were recently developed to generate novel molecules in silico, only a few studies complete the full design-make-test cycle, including the synthesis and the experimental assessment. One reason is that the synthesis part can be tedious, time-consuming, and requires years of experience to perform successfully. Hence, the synthesis is one of the critical limiting factors in molecular discovery.

In this thesis, I take advantage of similarities between human language and organic chemistry to apply linguistic methods to chemical reactions, and develop artificial intelligence-based tools for accelerating chemical synthesis. First, I investigate reaction prediction models focusing on small data sets of challenging stereo- and regioselective carbohydrate reactions. Second, I develop a multi-step synthesis planning tool predicting reactants and suitable reagents (e.g. catalysts and solvents). Both forward prediction and retrosynthesis approaches use black-box models. Hence, I then study methods to provide more information about the models' predictions. I develop a reaction classification model that labels chemical reaction and facilitates the communication of reaction concepts. As a side product of the classification models, I obtain reaction fingerprints that enable efficient similarity searches in chemical reaction space. Moreover, I study approaches for predicting reaction yields. Lastly, after I approached all chemical reaction tasks with atom-mapping independent models, I demonstrate the generation of accurate atom-mapping from the patterns my models have learned while being trained self-supervised on chemical reactions.

My PhD thesis's leitmotif is the use of the attention-based Transformer architecture to molecules and reactions represented with a text notation. It is like atoms are my letters, molecules my words, and reactions my sentences. With this analogy, I teach my neural network models the language of chemical reactions - atom by atom. While exploring the link between organic chemistry and language, I make an essential step towards the automation of chemical synthesis, which could significantly reduce the costs and time required to discover and create new molecules and materials.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Aims and objectives

Molecules and materials are all around us, and discovering new ones is one of the main drivers of technological progress. Although for discovery, computational studies can simulate molecular properties in silico, actual experiments have to be performed to synthesise and test the molecules. Those design-make-test cycles are long and costly. Chemical synthesis is currently a key limiting factor in the discovery process [1]. This thesis aims to improve artificial intelligence (AI)-assisted synthesis planning systems and accelerate chemical discovery by investigating the link between human language and organic chemistry, and modelling chemical reactivity using linguistics-inspired approaches.

Motivated by recent breakthroughs in natural language processing (NLP) [2, 3], I develop transformer-based methods for reaction prediction, multi-step synthesis planning and other related chemical reaction tasks. Like humans, my models learn by repeatedly seeing examples of successful reactions and extracting the underlying patterns. One crucial difference is that the models can learn from large collections of millions of reactions in a few days, while it would take more than a lifetime to do the same for a human. Given the extracted knowledge, my models can then assist chemists in deciding what reactions to perform, how to design their synthesis routes, or select experiments by predicting reaction yields. The objectives of this thesis are the following:

- Develop atom-mapping independent chemical reaction models.
- Collaborate with synthetic chemists and get their feedback on the models.
- Analyse transformer-based reaction prediction models for low-data regime reaction classes.
- Examine transformer-based reaction prediction models for regio- and stereoselective reactions.
- Construct a multi-step synthesis planning tool.
- Formulate evaluation metrics for single-step retrosynthesis models that are better suited than top-N accuracy.
- Develop transformer-based chemical reaction classification models to make the predictions of other models more explainable.
- Design chemical reaction fingerprints without requiring a reactant-reagent separation and reaction centre information.
- Investigate physics-agnostic chemical reaction yield prediction models.
- Understand why transformer-based models, like transformers, work well on chemical reactions.
- Develop a reaction atom-mapping tool, which is not based on heuristics or trained on atom-map data generated using heuristics, but guided by a signal learned from unmapped chemical reactions.

- Keep the models in a reasonable size range so that the results can be reproduced in a few days on affordable hardware.

1.2 Thesis outline

The thesis is outlined as follows:

- Chapter 2 introduces computer-assisted synthesis planning and discusses the essential elements to develop data-driven chemical reaction models. First, machine-readable molecular and reaction representations and their formats are presented. Second, the chemical reaction data sets, particularly the ones available in open-source, are described. Then, the basics of deep-learning models, including modern language models like Transformers [2, 3], are introduced and chemical reaction tasks are outlined. Lastly, the analogy between human language and organic chemistry is illustrated.

- Chapter 3 focuses on chemical reaction prediction for challenging reaction classes. How small specific and large generic data sets can be leveraged to increase the prediction accuracy of reaction prediction models like Molecular Transformers [4] is studied using transfer learning. The two investigated transfer learning strategies are applied to carbohydrate reactions. Stereochemistry, one of the weaknesses of previous reaction prediction approaches [4, 5, 6], is critical for carbohydrate reactions and governs the reactivity. It is a direct collaboration with synthetic chemists. One of the three test sets, on which the models are evaluated is a 14-step synthesis of a lipid-linked oligosaccharide performed by Giorgio Pesciullesi.

- Chapter 4 describes the methods behind the IBM RXN for Chemistry multi-step synthesis planning tool, where two Molecular Transformer [4] models are coupled. One suggests precursor molecules sets given a product molecule and the other scores chemical reactions given precursors-product combinations. A hyper-graph beam search is used to find the most promising routes. The developed retrosynthesis models not only predict reactants but simultaneously also suitable reagents for the reactions without distinguishing between them. This reaction representation allows the approach to be atom-mapping independent. Moreover, metrics to evaluate single-step retrosynthesis models more appropriately than top-N accuracy are presented. Those metrics also better capture human expert evaluations.

- Chapter 5 introduces the application of encoder transformer models to chemical reactions. The models are trained on chemical reaction classifications tasks, where, given the chemical reaction simplified molecular-input line-entry system (SMILES) as input, the aim is to predict the corresponding reaction class. Synthetic chemists commonly use reaction classes or name reactions to communicate complex reactivity concepts in simple terms. Being able to assign classes to reactions, directly provides additional information for chemists. For example, the reactions predicted by the models in chapter 4 can be classified and therefore, better be understood by chemists. Moreover, chapter 5 describes the usage of the outputs of the reaction encoder as reaction fingerprint. The reaction fingerprints can be used to perform efficient similarity searches and create chemical reaction maps. Those fingerprints provide a link from predicted reactions to similar reactions in reaction data sets. For example in

training sets of reaction prediction models, the retrieved reactions can then be analysed to understand why the models made their predictions and increase model explainability.

- Chapter 6 uses the work reaction encoder transformer models presented in chapter 5 and describes how they can be fine-tuned on chemical reaction yield prediction, a regression task. First, the predictions on two small high-throughput experiment data sets containing Buchwald-Hartwig and Suzuki-Miyaura cross-coupling reactions are analysed. Then, the extracted yield information from the open-source United States Patent and Trademark Office (USPTO) data set is studied. As expected, the models work better on the high-throughput experiment data than on the noisy patent data stemming from multiple sources.

- Chapter 7 extends chapter 6 and explores data augmentation strategies for chemical reaction yield prediction. Molecule order permutations, SMILES randomisations, and a mixture of both are investigated on the Buchwald-Hartwig high-throughput experiment data set. The data-augmented reaction transformers perform better than previous approaches, including physics-based descriptors plus random forest models, even in the low data regime with less than 100 training points. Morevover, an approach for epistemic uncertainty estimation using test-time augmentation is introduced. The uncertainty estimates correlate with the error of the predictions including the out-of-distribution test sets.

- Chapter 8 introduces an unsupervised attention-guided approach to compute atom-mapping in chemical reactions. By opening the black-box transformer models presented in chapter 5 and visualising their inner workings, attention heads were found that consistently produce an atom-mapping signal in the attention weights. In those heads, product atoms attend the corresponding reactant atom and vice versa. This observation means that the models were able to capture the grammar of chemical reactions without explicitly being taught. Using the atom-mapping signal, an atom-mapping tool is developed that outperforms existing tools.

- Chapter 9 concludes the thesis, summarises the main contributions and provides an outlook on future challenges and opportunities.

1.3 Publications

The thesis consists of first author and equal contribution publications presented as separate chapters (equal contribution is indicated by •):

- Chapter 2: P Schwaller, T Laino. Data-Driven Learning Systems for Chemical Reaction Prediction: An Analysis of Recent Approaches. in Machine Learning in Chemistry: Data-Driven Algorithms, Learning Systems, and Predictions. *ACS Symp. Ser.*, 2019, 61–79.

- Chapter 3: G Pesciullesi•, P Schwaller•, T Laino, JL Reymond. Transfer learning enables the molecular transformer to predict regio- and stereoselective reactions on carbohydrates. *Nat. Commun.*, 2020, 11, 4874.

- Chapter 4: P Schwaller, R Petraglia, V Zullo, V H Nair, R A Haeuselmann, R Pisoni, C Bekas, A Iuliano, T Laino. Predicting retrosynthetic pathways using a combined linguistic model and hyper-graph exploration strategy. *Chem. Sci.*, 2020,11, 3316–3325.
- Chapter 5: P Schwaller, D Probst, AC Vaucher, VH Nair, D Kreutter, T Laino, JL Reymond. Mapping the Space of Chemical Reactions using Attention-Based Neural Networks. *Nat. Mach. Intell.*, 2021, 3, 144–152.
- Chapter 6: P Schwaller, AC Vaucher, T Laino, JL Reymond. Prediction of Chemical Reaction Yields using Deep Learning. *Mach. Learn.: Sci. Technol.*, in press, 2021.
- Chapter 7: P Schwaller, AC Vaucher, T Laino, JL Reymond. Data augmentation strategies to improve reaction yield predictions and estimate uncertainty. *NeurIPS Workshop on Machine Learning for Molecules*. 2020. DOI:10.26434/chemrxiv.13286741
- Chapter 8: P Schwaller, B Hoover, JL Reymond, H Strobelt, T Laino. Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. *Sci. Adv.* in press, 2021.

The following is a list of publications and preprints that were co-authored during but not incorporated into the thesis:

- AC Vaucher, F Zipoli, J Geluykens, VH Nair, P Schwaller, T Laino. Automated extraction of chemical synthesis actions from experimental procedures. *Nat. Commun.*, 2020, 11 (1), 2041–1723
- H Öztürk, A Özgür, P Schwaller, T Laino, E Ozkirimli. Exploring chemical space using natural language processing methodologies for drug discovery. *Drug Discovery Today*, 2020, 25 (4), 689–705
- VH Nair, P Schwaller, T Laino. Data-driven Chemical Reaction Prediction and Retrosynthesis. *Chimia*, 2020, 73 (12), 997–1000
- A Toniato, P Schwaller, A Cardinale, J Geluykens, T Laino. Unassisted Noise-Reduction of Chemical Reactions Data Sets. *Nat. Mach. Intell.*, in press, 2021.
- AC Vaucher, P Schwaller, J Geluykens, VH Nair, A Iuliano, T Laino. Inferring experimental procedures from text-based representations of chemical reactions. ChemRxiv preprint, 2020. DOI:10.26434/chemrxiv.13286741
- AC Vaucher, P Schwaller, T Laino. Completion of partial reaction equations. *NeurIPS Workshop on Machine Learning for Molecules*, 2020. DOI:10.26434/chemrxiv.13273310
- D Kreutter, P Schwaller, JL Reymond. Predicting Enzymatic Reactions with a Molecular Transformer. *ChemRxiv preprint*, 2020. DOI:10.26434/chemrxiv.13161359

2 Recent Data-Driven Learning Systems for Chemical Reactions

One of the critical challenges in efficient synthesis route design is the accurate prediction of chemical reactivity. Unlocking it, could significantly facilitate chemical synthesis and hence, accelerate the discovery of novel molecules and materials. With the current rise of AI algorithms, access to cheap computing power and the wide availability of chemical data, it became possible to develop entirely data-driven mathematical models able to predict chemical reactivity. Similar to how a human chemist would learn chemical reactions, those models learn the underlying patterns in the data by repeatedly looking at examples. In this chapter, I introduce the state-of-the-art data-driven learning systems for forward chemical reaction prediction and retrosynthesis, analyse different reaction representations, the available data sets and the model architectures. I discuss the advantages and limitations of the different AI models' strategies. The intention is to provide a critical assessment of the different data-driven approaches developed in the last years not only for the cheminformatics community but also for the AI models end-users, the organic chemists, for early adoption of such technologies.

Parts of this chapter have been published as a book chapter in ACS Symposium Series:

Reprinted with permission from P Schwaller, T Laino. Data-Driven Learning Systems for Chemical Reaction Prediction: An Analysis of Recent Approaches. in Machine Learning in Chemistry: Data-Driven Algorithms, Learning Systems, and Predictions. *ACS Symp. Ser.*, 2019. 61-79.

2.1 Dream of computer assisted synthesis planning

Approaching the universe of organic chemistry can be an ordeal for beginner students, who typically experience difficulty in predicting the products of chemical reactions. It takes a certain amount of practice and knowledge to make the process more successful and efficient. Problems that may appear great challenges for an undergraduate student may be embarrassingly simple for a synthetic organic chemist with more than 30 years of experience. However, the complexity of the molecular space is such that the prediction of chemical reaction may become a difficult task even for expert synthetic organic chemists. Just like humans created computer programs to confront expert player at Chess [7], Jeopardy [8] and Go [9, 10] it happened that chemists encoded the vast collection of instructions for making molecules, available in the rich chemistry literature, into computer software with the purpose of creating an expert system to assist chemists in designing efficient routes to target molecules for organic synthesis. At the origin of this revolution was the

pioneering work of Corey [11, 12]. Around 1967 three groups started to address this problem constructing computer programs – LHASA by Corey et al. [13], SECS by Wipke and Dyott [14], and SYNCHEM by Gelernter et al. [15, 16] – that were searching synthetic strategies in synthesising known and unknown compounds, using a chemical knowledge base, rather than performing a reaction retrieval from a database of literature examples.

The idea at the base of the reaction prediction or synthesis planning was that by analysing an input molecule with a catalogue of retro-reactions (or transforms) encoded in memory, one could retrieve the descriptions of all the possible changes which will occur in the course of a particular reaction. It is inherent in such an approach that the planned syntheses will be based only on a combination of encoded transforms. EROS [17, 18] was the first attempt to use a large chemical data set to cast the problem of reaction prediction into a mathematical framework. Molecules and reactions were represented by specific matrices (bond-electron matrix and reaction matrix, respectively [19]) and the synthesis planning was cast as a pure matrix-matrix multiplication problem. This mathematical model was used as a basis for a variety of deductive computer programs for the solution of chemical problems, and EROS [17, 18] can be considered the first attempt to use AI for the reaction prediction problem. Since the mid-nineties, we witnessed an increased interest in the development of different approaches based on data with CAMEO [20], WODCA [18] and SOPHIA [21], being the pioneering technologies in this field exploiting advanced mathematical frameworks. Similar to LHASA [13] and SYNCHEM [15] but with a bigger commitment of resources, Chematica [22] few years later, used human experts to extract chemical reactions from the literature and to encode them with rules. The project [22] started at the beginning of the year 2000 and went on for more than a decade before it was publicly announced. Albeit the decision to encode the broad knowledge of organic chemistry with rules was not new [17, 20], Chematica [22, 23] was the first to achieve a high level of accuracy in forward and retrosynthetic reaction prediction. This competitive advantage was explainable with the multi-year efforts to codify the most extensive set of rules ever, including reaction core, reactivity conflicts, substituents and groups requiring protection during multi-step synthesis. Despite the recent scientific and business successes [22, 24], the approach is not sustainable in the long-term: manually extracting rules from literature is a tedious work and prone to human error. Rules tend to be very brittle, as for every new reaction outside the scope of the current rules a new rule, which does not contradict the existing 100k thousand rules has to be added. Finally, the involvement of humans in the entire curation process makes the maintenance and development of the software unscalable due to the ever-growing amount of data produced and published. For a more extensive review of the history of computer-assisted synthesis programs I refer the reader to [25, 26, 27, 28].

Starting from 2010 on, thanks to advances in machine learning algorithms, more powerful computational resources and to the availability of a vast amount of open-source chemical data, we witnessed the development of a multitude of different types of mathematical models that tried to offer a valid alternative to the rule-based approaches. The advantage of these mathematical models is that once trained on a data set, they can infer the patterns hidden in the data in a few hundreds of milliseconds. Similar to what a human chemist would do, data-driven models learn from examples, ideally without having humans encoding domain specific knowledge, such as reaction rules in organic synthetic chemistry. The main difference is that a mathematical model can analyse and incorporate the whole literature, millions of distinct chemical reactions, in a matter of days, which would take more than a lifetime for a human.

Organic chemistry syntheses are still mainly designed by human experts, who relying on their personal experience, intuition and years of training try to come up with reasonable steps. Along the way, the route is improved, typically, by trial-and-error. If one step fails, and no alternative route is found to circumvent the failing step, the whole route is rethought with different initial steps. The later the failing step, the higher the costs. Data-driven chemical reaction models could be used to validate individual steps in a multi-step synthesis. One goal is to estimate the risk of a specific reaction and place the reactions that are more likely to fail at the beginning of the synthesis route. Data-driven chemical models could also be used to predict side products and impurities and as inexpensive cross-validation of outcomes generated by time-consuming and computation-intensive simulations. Therefore, it is no surprise that such models are believed to profoundly change the way chemists will design synthesis in the near future. Similar to what happened after Deep Blue beat Gary Kasparov with computers assisting human players in chess matches (centaur chess), we envisage scientific assistants, supporting human chemists by giving them access to the knowledge hidden in a much wider variety of chemical reactions. While the recent mathematical approaches [4, 5, 6, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38] are based on data, their architecture can be very different with unique responses to specific data sets. For non-experts, it can be a real hurdle to rationalise the subtleties of the different implementations and critically assess which models are more business-ready to use in daily life applications than others. For the rest of this chapter, I focus on data-driven approaches for the problem of forward reaction prediction and retrosynthesis, describing artificial neural network-based models that myself and others developed in the last years and that can be trained on previously published experimental data. I also discuss the training data. While recently large reaction corpora (including Reaxys [39] and SciFinder [40] became wider available in the last 15 years, their usage in model training is still hindered by their limited access for data analysis and model training purposes. Innovation in designing new data-driven models requires unconditional data availability: for organic chemistry reaction prediction, the experienced acceleration was strongly correlated to the possibility of accessing a large set of chemical reactions consisting of millions of tabulated examples, extracted from the USPTO patents [41, 42]. Therefore, my discussion about machine learning models will focus more in details on all those approaches that trained and tested on the USPTO data set, comparing the performance and analyse the details of the respective implementation. The discussion includes the models, behind the platform known as IBM RXN for Chemistry, made freely available in 2018 [4, 43, 44], using an NLP approach for reaction prediction in organic chemistry.

2.2 Chemical representation and formats

An essential aspect of data-driven models is the representation of the data used during the training process. To uncover and better understand the highly non-linear patterns in organic chemistry and reaction prediction using machine learning, data should be made available in a machine-readable format and as accurate and clean as possible. Still today, those highly complex chemical reactions are simplified to quite abstract reaction diagrams, challenging to interpret with the use of a computer program. Reaction diagrams consist mainly of four main parts: in the centre is the arrow, which points in the direction the reaction proceeds; to the left are the starting materials; above and below the arrow the additional reagents, agents and spectator molecules (e.g. catalysts

and solvents) and finally, the products on the right. As simple as this basic scheme seems, there are several non-obvious challenges. Firstly, the distinction between what is a starting material and what an agent is vague. What one chemist would call a reactant, would be a reagent for another, and the placement in one of the two categories would give more indication on what the chemist, who draw the diagram in the first instance, focused on his attention, instead of the actual role of the compound. Hence, having a molecule above or below the arrow does not necessarily mean that it does not transfer part of its atoms to the final product. Secondly, usually, only the major target is reported and not the whole product distribution. Trivial products like water or alcohol are often left out to simplify the diagram representation, which can become even more cryptic in case there is the need to report enantiomers and racemic mixtures. Lastly, depending on the reaction conditions, the outcome of a reaction can be different. Ideally, a chemical representation would contain information on the reaction conditions (e.g. temperature, time, pH), the reaction yield and the enantiomeric excess. As they are not always added to reaction diagrams, the corresponding text has to be consulted to get a full picture. While a human expert can easily make the connections between the diagram and additional information found in the supporting text, no reliable methods exist to date to extract all the information from reaction diagram and combine it with the textual information to generate a machine-readable representation. Even then, it happens that some of the crucial details are not disclosed. An effort was made in the last decades to create different standards to codify reaction information into machine-readable format to efficiently store, compare and analyse chemical reactions. RXNfiles [45] and RDFiles [45] are quite similar, with RXNfiles containing the molecular information of a single reaction and RDfiles containing multiple reactions with additional information on the reaction conditions, atom-mapping and reaction centre. Reaction SMILES or SMIRKS [46, 47] contain reactants, agents and products, the last being separated by a '>' symbol. Although the format supports atom-mapping, there are no extra fields for reaction conditions and reaction centre information. SMILES arbitrary target specification (SMARTS), describing the molecular pattern, are extended with the '>' symbol to encode reaction rules, also called reaction templates. Chemical Markup Language (CML) [48, 49] is the equivalent to Extensible Markup Language (XML) for chemical information. As this format is very flexible, it allows for the most complete description of chemical reactions. However, no clear standards exist, which makes the data exchange and comparison between research groups difficult. RInChI [50, 51], based on the IUPAC International Chemical Identifier [52], is a line notation describing groups of reactants, agents and products. As the aim of RInChI is to generate a unique and unambiguous reaction descriptor to link and find chemical reactions, atom-mapping is not supported [51]. While the RInChI only contains standardised structural information, RAuxInfo stores the conformation and orientation of the compounds used to generate the RInChI. Moreover, hashing algorithms allow to generate shorter keys for the reactions, which facilitate the search of reactions. To store reaction conditions, stoichiometry of reactants and agents, as well as yields and conversion ratios a RInChI extension called ProcAuxInfo has been proposed [53]. The information on the individual molecules involved in a chemical reaction is represented either as fingerprints (e.g. ECFP [54]), line notations (e.g. SMILES and InChI) or graphs. In contrast to the latter two, fingerprinting methods are non-invertible hashes. In molecular graphs, the nodes usually correspond to the atoms and the edges of the graph to the bonds. Molecular graphs are often hydrogen depleted. Line notations are text-based representations of molecular graphs. Recently, two novel line notations have emerged: DeepSMILES [55], an adapta-

tion of SMILES, and SELF-referencing Embedded Strings (SELFIES) [56]. Both aim to facilitate the construction of syntactically valid molecular graphs which could improve the performance of data-driven models. For a more extensive review of molecular descriptors, I point the reader to the work of Sanchez-Lengeling and Aspuru-Guzik [57] and of David et al. [58].

2.3 Chemical reaction data

The path chemical reaction information has to flow from the laboratory, where the reaction was conducted, through an article or patent publication and finally, extracted by whatever means to be stored in a database is extremely lossy and error prone. As suggested before [25, 53], there should be a standard on how to report chemical data, such that every data point supporting a publication is submitted in a machine-readable format together with the manuscript. Such a shortcut, where the reaction information would go from the author, through the use of standardised electronic lab notebooks (ELN), directly through an open database, would be ideal and allow the field to advance rapidly. To date, there are few options of reaction data sets collection, but most of them are commercial, close-access and come with terms and conditions that do not allow training of open-access AI models. Lowe [41, 42] generated the largest open-access reaction data set. Originally, the text-mining tool was developed at the University of Cambridge [41], it was later improved by NextMove Software [59] and takes advantage of the latest improvements and technologies in Natural Language Understanding and text-mining in the field of chemistry. The data set is available in two formats: SMILES and CML. The reaction SMILES ('.rsmi') file contain not only the reaction SMILES, but also the patent number, paragraph, year, text-mined and calculated yield. The CML files are more complete, containing the paragraph, from which the reaction was extracted, the names of the compounds, which were converted to SMILES and action lists, describing the steps taken during the procedure (heating, cooling, stirring, ...). Most of the data-driven models took into account the information in the easily readable '.rsmi' file. The reactions in the USPTO data set are atom-mapped using Epam's Indigo toolkit [60]. However, those atom-maps are wrong in many cases [42]. Although the most recent atom-mapping approach is based on heuristics [61], it is simple to draw reactions, where the atom-mapping is ambiguous, as seen in Figure 2.1. The work of Schneider et al. [62] has shown that between Indigo Toolkit [60] and NameRXN [63], two tools able to generate atom-mapping, only in 22% of the reactions on 50k random reactions from the USPTO data set the set of reactants matched. Therefore, because of the inherent difficulty in determining the precise mapping, all methods, which are based on atom-mapping, are fundamentally limited by the underlying software, which generates the atom-mapping. This observation motivated us to develop atom-mapping independent approaches, as described in chapters 3-7. In chapter 8, I will then introduce an atom-mapping tool that was build by analysing the inner workings of a neural network trained self-supervised on the USPTO data set [42].

While it is impressive how much information, could be extracted from the US patents, the USPTO data set is far from being perfect. It is not free from systematic extraction errors, contains partly incomplete reactions with a preponderant tendency to misinterpret organometallic compounds. In particular, the incomplete reactions are a severe problem for data-driven reaction prediction methods. Despite the usage of the atom-mapping to check whether all atoms

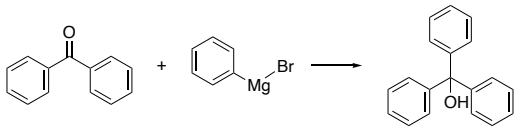

Figure 2.1: **Example chemical reaction**. A Bromo Grignard reaction with non-trivial atom-mapping, as any phenyl group in the product could correspond to any phenyl in the reactants. There is more than one correct atom-mapping. The reaction SMILES for this reaction is: "O=C(c1ccccc1)c1ccccc1.Br[Mg]c1ccccc1>>OC(c1ccccc1)(c1ccccc1)c1ccccc1".

on the product side were also present on the reactant side, there is no possibility to check if all the necessary solvents and catalysts were correctly extracted. One reason for these errors is the incorrect spellings of IUPAC names in patents. As a consequence, models are trained on similar reactions not always explicitly containing the catalysts and hence, infer that the catalyst is not important for the reaction to take place. For example, the models trained with such data perfectly predict a coupling reaction without seeing the metal catalyst. There is another problem with organometallic compounds when using SMILES. In fact, SMILES were designed to represent organic compounds only and there is no obvious way to treat bonds within organometallic systems in SMILES. Moreover, for data-driven reaction prediction models, it is not clear if the correct bond representation is crucial in attaining a higher prediction accuracy. Acting as catalysts their presence or absence is often more important, then the exact bonding description within the organometallic centre.

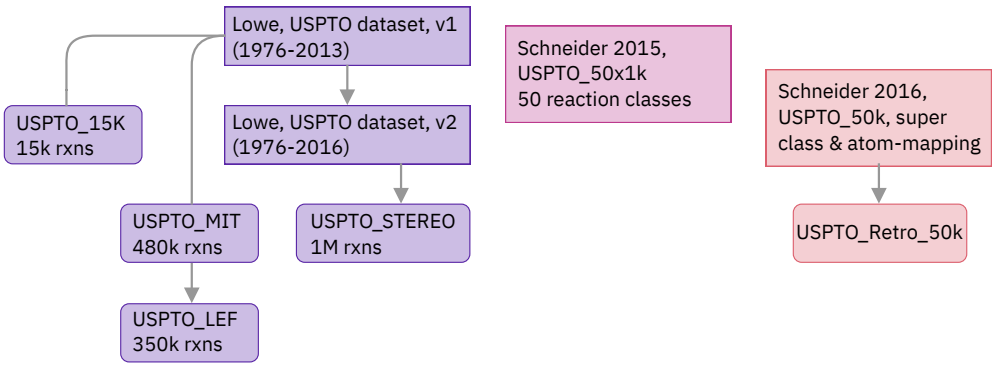

Figure 2.2: **USPTO data family tree.** The USPTO data set family tree with the four versions of the text-mined data set [41, 42, 62, 64] and the different subsets [5, 35, 36, 65] used to benchmark data-driven reaction prediction models.

Since the publication of the USPTO data set [42] an entire family of reaction prediction benchmark subsets with different flavours appeared, as shown in Figure 2.2. All these subsets were made available at publication time, including the correct splitting between the training, validation and test set. The publicly available data allows not only to reproduce the scientific outcome reported in a publication but also a direct and statistical meaningful comparison between the dif-

ferent approaches. The most used benchmark set is the USPTO_MIT set. However, during the filtering process Jin et al. [5] removed all reactions containing stereochemical information. As stereochemical information might be crucial for the functionality of molecules, Schwaller et al. [36] generated another subset of the USPTO data set keeping stereochemistry, therefore referred to as USPTO_STEREO. Bradshaw et al. [37] later removed all reactions without a so-called linear electron flow topology (excluding e.g. pericyclic reactions), hence, simplifying the data set. This subset of USPTO_MIT containing only 73% of the reactions is referred to as USPTO_LEF. Schneider et al. published two independent USPTO subsets with additional reaction meta data. The first contained 1k reaction example corresponding to 50 reaction classes (e.g. Thioether synthesis 1.8.5) [64]. The second contained reaction superclasses (e.g Heteroatom alkylation and arylation 1) and atom-mapping assigned by NameRXN [62] and was later used by Liu et al. [65] as single-step retrosynthesis benchmark data set.

2.4 Deep learning models

Besides data representation and the actual data, the third crucial ingredient for machine learning for chemical reactions are the models. This thesis focuses on deep learning, which is a subset of machine learning and artificial intelligence. In deep learning, models automatically extract useful pieces of information from raw data to inform future predictions. One motivation to use them, compared to traditional machine learning techniques, is that the input features are not hand-engineered but learned from the data. Hence, the features tend to be less brittle, and the algorithms scalable to larger data sets.

Deep learning methods are based on artificial neural networks. McCullogh and Pitts [66] introduced the basic concept inspired by biological neurons already in 1943. A single artificial neuron, also called perceptron by Rosentblatt [67], is mathematically described as follows:

$$f^{(i)}(\boldsymbol{x}) = \sigma(\boldsymbol{w}^T\boldsymbol{x} + \boldsymbol{b}), \tag{2.1}$$

where $\boldsymbol{w}$ are learnable weights with which the inputs $\boldsymbol{x}$ are multiplied, $\boldsymbol{b}$ is a learnable bias term and σ a non-linear activation function. Real-world data, particularly in chemistry, is often non-linear. The non-linearities introduced by the activation functions, such as a sigmoid function, a hyperbolic tangent, or a rectified linear unit (ReLu) [68], make it possible to model such data. By connecting the inputs $\boldsymbol{x}$ to multiples neurons, a dense layer can be created. Those fully-connected dense layers consisting of many perceptrons are one of the fundamental building blocks of neural networks as depicted in Figure 2.3 *a*. Deep neural networks are made by stacking multiple layers.

In dense layers, the weights are independent and no symmetry can be exploited as inductive bias as all inputs are connected to all outputs. Other neural network building blocks have stronger inductive biases [69]. Convolutional layers learn filters that extract local correlations from neighbouring inputs [70], as shown in Figure 2.3 *b*. The weights of a convolutional layer are shared across space, which makes them translation invariant. 1D convolutional layers can be applied to text or time series [71], 2D ones to pixels in images [70], and 3D ones to voxels [72]. Closely related and often used for molecular inputs are graph convolutional layers [73] depicted in Figure 2.3 *c*. Instead of computing the output on neighbouring pixels/voxels, graph convolutional layers compute the outputs based on the adjacent nodes in the graph. Other neural network building blocks

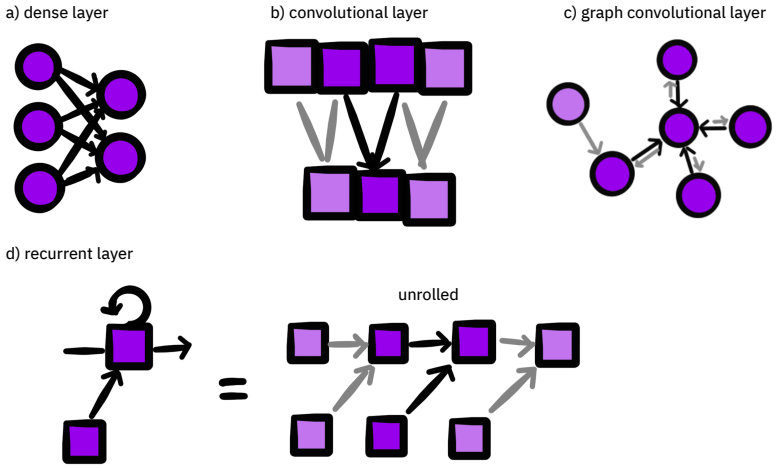

Figure 2.3: **Neural network building blocks.** Fundamental building blocks of neural networks: a) a fully-connected dense layer, b) a convolutional layer, c) a graph convolutional layer and d) a recurrent layer.

are recurrent layers, as shown in Figure 2.3 *d*. Recurrent layers often model sequential inputs in, for example, text or time series. They are similar to dense layers but with a feedback connection. Therefore, their output not only depends on the current input but also on the previous state of the layer. Recurrent layer weights are shared in time. Popular variations of recurrent neural networks are gated recurrent units (GRU) [74] and Long-Short Term Memory (LSTM) [75]. One advantage that convolutional and recurrent layers have compared to dense layers is that they do not require fixed-size inputs.

2.4.1 Neural network model training

Deep neural network models are made of stacked differentiable layers. Each of the layers contains weights, which have to be trained before the model can output reasonable predictions. During training, the weights are updated iteratively to minimise a predefined loss function. Given some inputs from the training set, the loss function compares the predicted values with the true expected values. As the neural networks are made of differentiable functions, the gradient of the loss with respect to the weights can be computed throughout the network using backpropagation [76]. The weights are then adapted using the gradient [77] to minimise the loss in the next iteration. This procedure is repeated until convergence or reaching another predefined stopping criterion, such as the maximum number of steps. How strong the weights are adapted at every step of the training depends on the learning rate, one of the most important hyperparameters to tune for a neural network. Models trained with too small learning rates can get stuck in a local minimum, and those trained with too large ones might diverge. In practice, adaptive learning rates which increase or decrease depending on the training are commonly used [78]. For detailed information and recommendations on how to best train neural networks, prevent overfitting and tune hyperparameters, I refer the reader to the Deep Learning book by Goodfellow et al. [79].

Deep learning recently became popular not only because of the wide availability of data and hardware but also because of the software. With frameworks like Tensorflow [80] and PyTorch [81], the differentiation of the networks and backpropagation are implemented and done automatically. Moreover, it is a common practice in machine learning to share code implementations with publications [82]. Researchers can focus on solving new tasks using neural network models instead of reimplementing everything from scratch.

2.4.2 Encoder-decoder architectures

Depending on the target task, the neural network building blocks are combined differently. In NLP, a common task is neural machine translation, in which the goal is to translate sentences from one language (e.g. English) to another language (e.g. French). It is usually a supervised task because the training data in one language is annotated with the other language's corresponding translations. The sentences can be represented as a series of words, sub-words, or characters [83]. Those sentence sub-parts are called tokens. Hence, the process to separate a sentence into its tokens is called tokenisation. To tackle the translation task, the most common model architecture is an encoder-decoder model. The first encoder-decoder sequence-to-sequence (seq-2-seq) models were introduced by Sutskever et al. [84] and Cho et al. [74]. Those models consist of an encoder that reads in the input token sequence in one language and converts the sequence into a context vector. The context vector is given as input to the decoder. The decoder then sequentially predicts the output sequences, token by token, based on the context vectors and all previously predicted tokens. A special END token signals the end of the predicted sequence. To make arbitrary token vocabulary sizes compatible with the models, tokens are converted to input feature vectors using trainable word embeddings [85], which after training could capture the meaning of the words they encoded [86].

In the work of Sutskever et al. [84] and Cho et al. [74] the encoder and decoder were built using recurrent neural networks. However, the performance of the early seq-2-seq models was limited by the fixed-sized context vector between encoder and decoder [74, 84]. For long and complex input sequences, the context vector was too small to give the decoder enough information to predict the target correctly.

To overcome this limitation, Bahdanau et al. [87] and Luong et al. [88] independently suggested a method called attention:

$$\text{attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \underbrace{\text{softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}})}_{\text{attention weights}} \boldsymbol{V} \tag{2.2}$$

where $\boldsymbol{Q}$ the query, $\boldsymbol{K}$ the key and $\boldsymbol{V}$ the value matrices. d_k is a scaling factor later introduced by Vaswani et al. [2]. In seq-2-seq models with attention [87, 88], instead of encoding the whole sequence into a fixed-size context vector, the encoder computes one vector for every input token resulting in a context matrix. In the attention equation above, the context matrix information is used for $\boldsymbol{K}$ and $\boldsymbol{V}$. At every decoding step, the decoder queries the context matrix using $\boldsymbol{Q}$. for the most relevant information to predict the next token. The attention function returns the values weighted by how aligned keys and queries are. The output of the softmax function, the

so-called attention weights, can be visualised to show what the decoder is focusing on to predict the output tokens.

2.4.3 Transformers

In 2017, in a ground-breaking study called "Attention is all you need", Vaswani et al. [2], introduced the encoder-decoder Transformer architecture (Figure 2.4 *a*). In contrast, to previous seq-2-seq models, the authors did not use recurrent layers in the Transformer but instead built the architecture on attention layers only. As attention layers have no sequential inductive bias, the token order information is given to the model through positional encodings. The encoder of the Transformer consists of stacks of self-attention layers with multiple attention heads. For self-attention layers, the queries, keys and values are the outputs of the previous attention layer. The encoder computes representations of every input token given all input tokens. This approach is useful as the same tokens might have different meanings depending on the sentence. Self-attention is shown in Figure 2.4 *d*. Two different attention mechanism are implemented in the decoder. The first is encoder-decoder attention (Figure 2.4 *f*), which resembles the attention used in previous seq-2-seq models [83, 87], where the decoder queries the encoder keys and values. The second is the masked self-attention (Figure 2.4 *e*), where the queries, keys and values are from the decoder. In contrast to self-attention, future keys are masked to prevent revealing information to the decoder about the tokens it has to predict. Another novelty of the work by Vaswani et al. [2] was the multi-head attention. Every attention layer in the Transformer has a defined number of attention heads. Every head can learn an independent function to attend the features and specialise on a particular pattern in the sequences. For instance, a head could learn to focus on the punctuation and another on the subject of the phrase. Using this novel architecture, Vaswani et al. [2] set new records in the English-to-German and English-to-French neural machine translation task.

The Transformer models presented by Vaswani et al. [2] already had between 65M and 213M trainable weights. The recent trends in NLP have been to make improvements by training larger and larger models. However, not all language tasks have enough data to train such large models efficiently. One approach shown to help achieve better results is pretraining [89, 90, 91]. The main idea of pretraining is simple: a model is first trained on an auxiliary task for which more data exist or data can be generated automatically. The pretrained model then starts with more favourable weights than randomly initialised ones to learn the actual task and achieves better results. This second stage, where the model is trained on the actual task, is called fine-tuning. Leveraging data from multiple data sets to achieve better model generalisation is known as transfer learning [92]. It is also possible to perform transfer learning by training on multiple data sets simultaneously [93], called multi-task learning. Inspired by successes of pretraining and the Transformer architecture, Radford et al. [94] introduced Generative Pre-trained Transformer (GPT, Figure 2.4 *b*), a decoder-only transformer pretrained on a large text corpus. Decoder-only because similar to the original Transformer decoder it was trained with a masked self-attention by predicting the next tokens in a sequence from left to right. GPT improved upon previous models on the General Language Understanding Evaluation (GLUE) benchmark [95]. In the meantime, the same group in OpenAI developed GPT-2 [94] and GPT-3 [96]. The improvements on NLP benchmarks and more human-like language generation were achieved by increasing the number of trainable weights from 100M in GPT, to 1.5B in GPT-2 to 175B in GPT-3 and using larger text corpora.

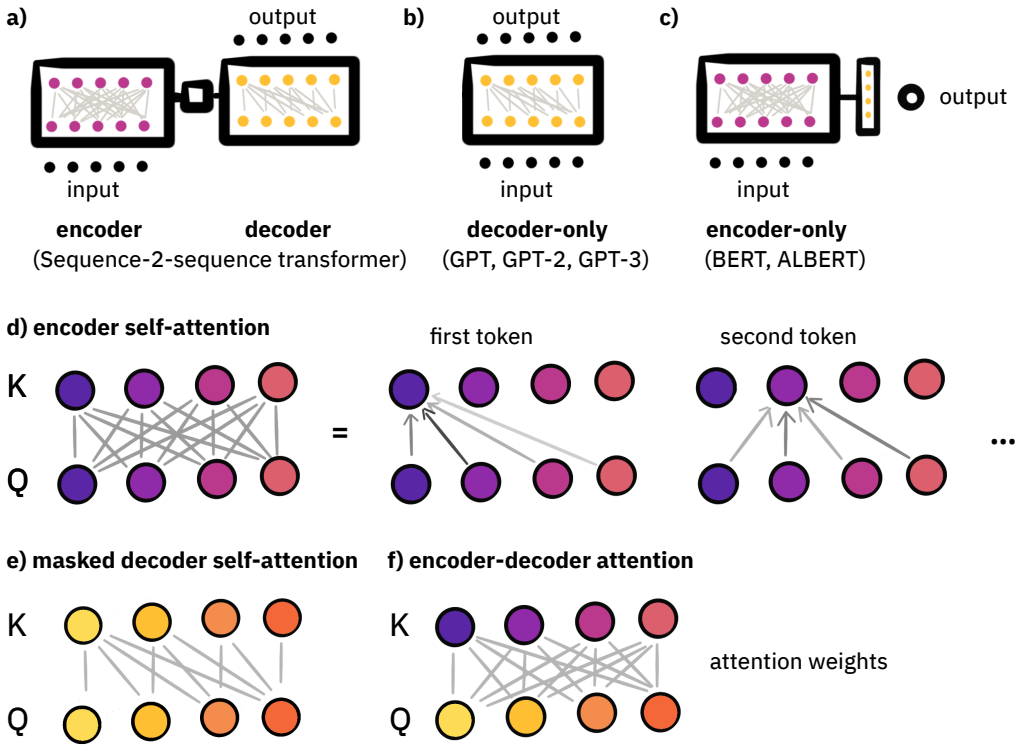

Figure 2.4: **Transformers.** a) Encoder-decoder model. b) Decoder-only model. c) Encoder-only model. d) Self-attention, where the inputs to key and query are the same used in Transformer encoders. On the right, are schematic drawing of the attention to the first and second token. e) Masked self-attention, where key and query are the same. To not give the decoder information about the future it is supposed to predict, the representations of those tokens are masked and not queried. f) Encoder-decoder attention, attention between encoder keys and decoder queries.

Devlin et al. [3] developed a language representation model called Bidirectional Encoder Representations from Transformers (BERT, Figure 2.4 *c*). In contrast to the autoregressive decoder-only approach by Radford et al. [94], BERT is an autoencoding model, which learns the token representations by correcting corrupted sequences. Through the unmasked encoder self-attention, the token representations are computed in the context of all sequence tokens. The two pretraining tasks that Devlin et al. [3] introduced are Masked Language Modelling (MLM) and next sequence prediction. Similar to GPT, the BERT models contained between 110M and 340M weights. Training such only feasible for big corporations with dedicated hardware, the cloud compute costs to train one GPT-2 model was estimated 43k USD by Strubell et al. [97]. Recently, several studies focused on achieving similar performance with smaller models. Lan et al. [98] developed a lite version of BERT (ALBERT), where the number of parameters is reduced by sharing the weights across the layers. Sanh et al. [99] used distillation techniques [100] to compress the BERT's knowledge into a smaller model. Other groups modified the attention algorithm to make more efficient, and hence, better scale to longer sequences [101, 102, 103, 104, 105].

2.5 Machine learning for chemical reactions

As seen above, the availability of open-data, software, and more powerful hardware enabled the development of novel approaches to tackle challenges in NLP that went beyond simple regression problems. Similarly, organic chemical reactions tasks worth solving to accelerate the organic synthesis and better understand model predictions can be defined. Those tasks can be supervised, where the inputs are annotated/labelled and the models are trained to predict the labels. Or unsupervised, where the models learn data representations from unlabelled input.

Figure 2.5 provides an overview of the different chemical reaction tasks addressed in this thesis. Reaction prediction in *a*, where products are predicted given precursors, is a task for which well-defined benchmark data sets exists. Similar benchmarks are less appropriate for single-step retrosynthesis in *b*, where precursor sets are predicted given a product, as multiple correct precursor sets might exist resulting in the same product. Forward reaction prediction and single-step retrosynthesis can be formulated as generative tasks, where the product is generated atom by atom given the precursors or vice versa. Multi-step synthesis planning in *c*, where the aim is to find routes from the desired product to commercially available molecules, is more challenging but is required for the synthesis of most molecules. Other synthesis relevant chemical reaction tasks can provide more information about the generative models' predictions and help to better understand the predictions. Reaction classification models in *d* can be used to label chemical reactions and communicate their underlying concepts. Chemical reactions are discrete, and it is difficult to search for similar reactions when the reactants and the reaction centre are not determined. Alternatively, chemical reactions can be encoded and represented in a continuous space as reaction fingerprints in *e*. Reaction fingerprints can then be used to query for similar reactions in a data set. Accurate predictions of reaction yields in *f* with uncertainty estimation could be used to guide synthesis planning tools and chemists in their choice of what experiment to perform. Finally, for atom-mapping independent models, the atom-mapping is not tracked during the prediction. Atom-mapping tools in *g* can tag the corresponding atoms on precursor and product sides from which the reaction centre, the reactant-reagent split and the grammar of chemical reactions can be derived.

2.5.1 Forward reaction prediction

From the different tasks on chemical reactions that can be approached with machine learning, the chemical reaction prediction task, where likely products are predicted given precursors, is potentially the most obvious one.

The idea of chemical reaction prediction models is not new. Pioneering examples are EROS [17], CAMEO [20], WODCA [18] and SOPHIA [21], all of which were built on top of either a rather small-scale reaction or knowledge database. Satoh and Funatsu [21] presented the first approach not requiring the reaction type or class as input for the prediction and recognised the potential of using reaction outcome prediction models for the validation of retrosynthesis steps in synthesis planning tools. Here, I focus on purely data-driven chemical reaction prediction methods taking advantage of novel machine-learning techniques based on artificial neural networks.

The recent data-driven approaches can be distinguished by analysing the model, the data, the input features and the outputs, as shown in Table A.1. There are several types of network archi-

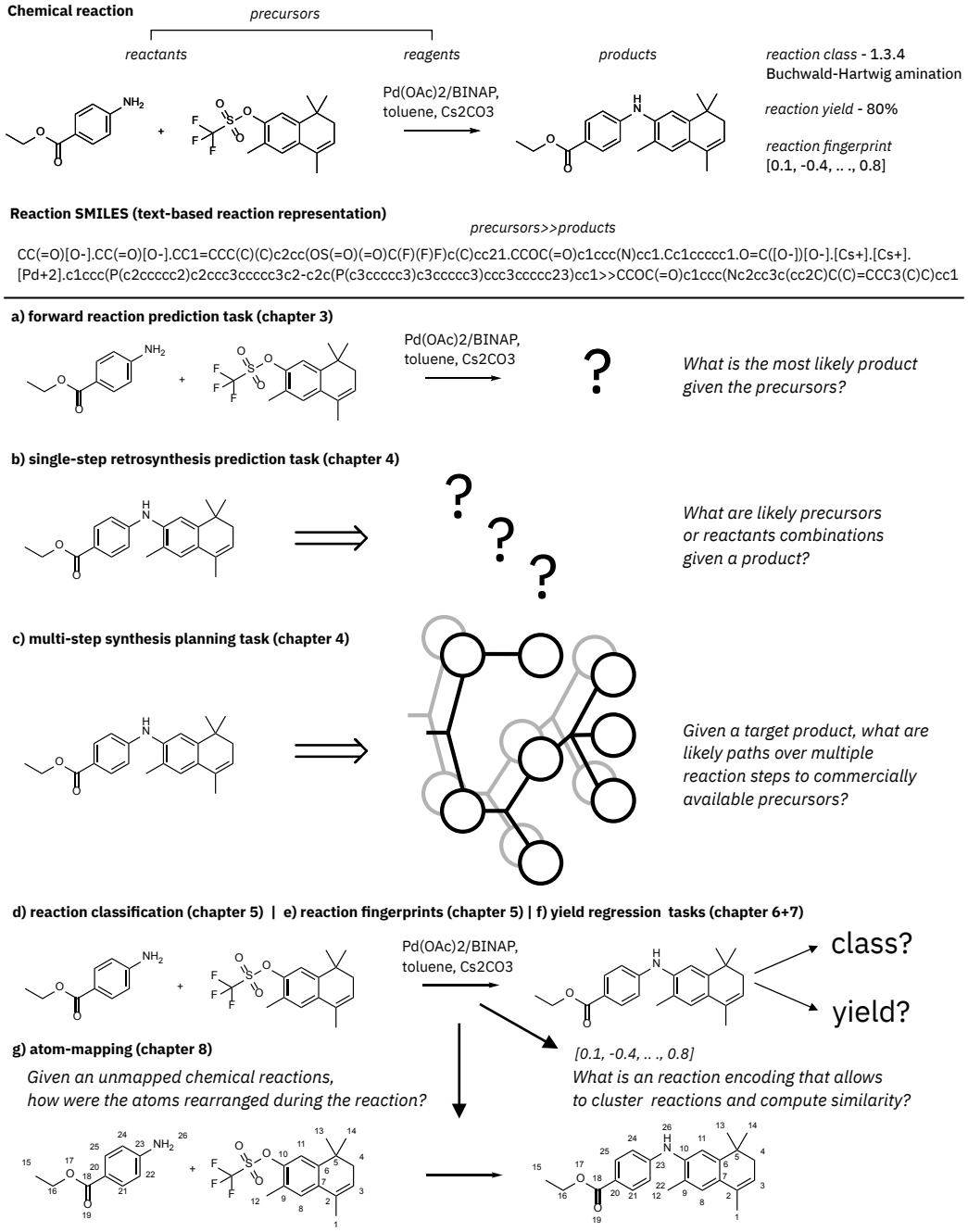

Figure 2.5: **Chemical reaction tasks.** A Buchwald-Hartwig amination reaction taken as example for different chemical reaction tasks.

tectures used for reaction prediction. Feed-forward neural networks learn a function, which maps a fixed-sized vector through the network to another fixed-sized vector. Seq-2-seq and transformer networks are auto-regressive encoder-decoder architectures and have the advantage that they can

handle inputs of varying lengths, as well as generate outputs of varying lengths. Graph neural networks learn a function applied to a node in a graph and its neighbours. Those neural network architectures have all different inductive biases [69]. Kayala et al. [29] used a neural network to predict mechanistic steps through the identification and ranking of electron sources and sinks. The inputs to the network contained a combination of the reaction conditions, hand-crafted molecular features and the local neighbourhood of the individual atoms. As a chemical reaction can consist of a sequence of mechanistic steps, multiple such predictions would be required to get the final product of a reaction. Building further on this idea, Kayala and Baldi [30] developed the ReactionPredictor, which ranked the atomic interactions based on the output of three separate feed-forward neural network, the first trained for polar, the second for pericyclic and last for radical reactions. The main drawback is that data on mechanistic steps is not readily available. Therefore, Kayala et al. [29] generated their own data using their rule-based expert system [106]. In a more recent work by Foshee et al. [34], the Baldi group extended their data set from 5.5k to 11k elementary reactions. Still applying a very similar approach for the prediction of mechanical steps, they showed that a bi-directional long short-term memory network using solely a SMILES string as input nearly matches the electron source/sink identification performance of their feed-forward neural network with more chemical inputs. Wei et al. [31] used feed-forward neural networks to identify, which SMARTS transformation out of 16 reaction templates to apply to a set of two reactants plus one reagent. Their approach was based on the concatenation of differentiable molecular fingerprints [73]. Therefore, their network could be trained end-to-end and did not require any hand-crafted features. In contrast, Segler and Waller [32] modelled the reaction prediction task with graph-reasoning model to find missing links in a knowledge graph made of binary reactions from the Reaxys [39] database. In another work, Segler and Waller [33], used a neural network to rank reaction templates, which were automatically extracted from the Reaxys [39] database. Reactions were represented using traditional fingerprints, which construct a fixed-sized vector based on the presence and absence of individual local motives in the molecules. Segler et al. [107] developed an in-scope filter to estimate the reaction feasibility based on their fingerprint. As also pointed out by Coley et al. [108], a reaction template might match different reactive sites in the reactants and therefore, generate more than one product. Hence, template ranking is not sufficient to predict the most likely product of a reaction. To overcome this problem, Coley et al. [35] proposed a different approach. Instead of ranking the templates, they applied all the templates matching the reactants in a first step to generate possible candidate products. The products were then ranked by a neural network. Recognising the drawbacks of hashing the reactant molecules to a fixed-sized fingerprint, Coley et al. [35] designed edit-based reaction representation based on the atoms that had a change in bond type and hydrogen count. The inputs to their model were augmented with structural information, as well as easily computable geometric and electronic information. The method was tested on a rather small subset of the USPTO data set [41] containing 20k reactions. In general, template-based methods are fundamentally limited by the set of templates they are based on and cannot predict anything outside the scope of this set. While automatically generated template sets scale well, it is still not straightforward to produce a good set of templates [33, 35, 109]. Usually, the number of neighbouring atoms or the distance around the reaction centre has to be specified. This leads to a trade-off between a large amount of very specific templates and a small amount of overly generic templates. Moreover, the local environment near the reaction centre might not be sufficient to describe the reaction. Another drawback

of automatic template extraction is that the reaction centre is typically identified using the atom-mapping, which depending on the source might not be correct. All in all, before the end of 2017 most of the data-driven reaction prediction approaches were either rule-based or small-scale. In the meantime, template-free large-scale approaches emerged, which can be categorised into two main classes, namely bond change predictions and product molecule generation, an overview is found in Figure 2.6.

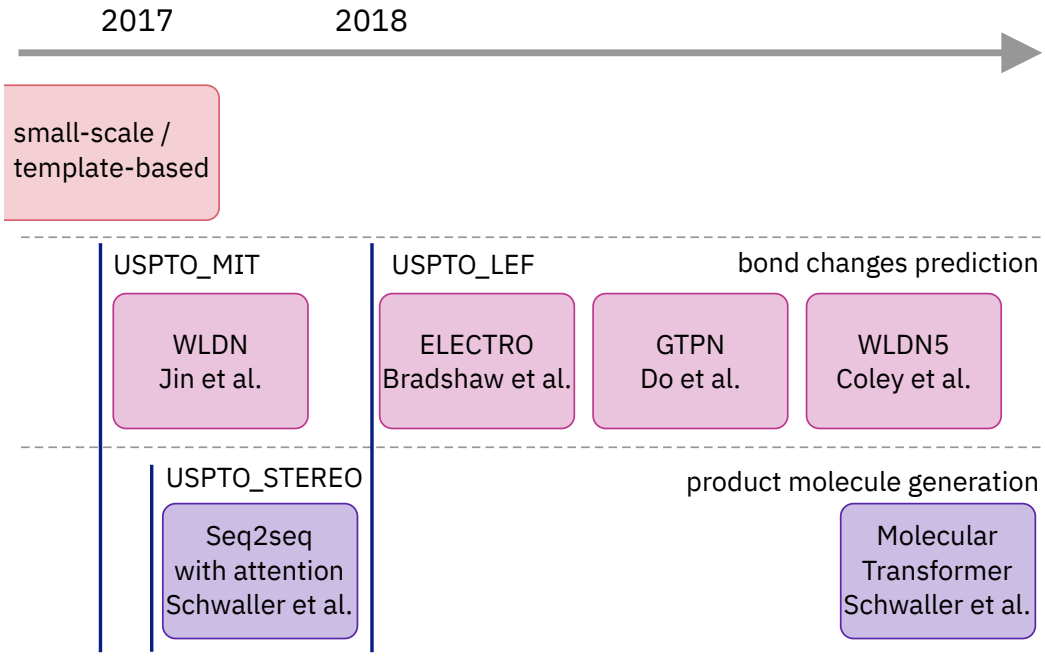

Figure 2.6: **Timeline of reaction prediction models.** Timeline of the recent developments of large-scale data-driven reaction prediction models that can be compared using the different USPTO reaction subsets. There are two main strategies, bond changes predictions and product molecule generation.

Jin et al. [5] presented the Weisfeiler-Lehman Network/Weisfeiler-Lehman Difference Network approach (WLN/WLDN), which uses a two-step process to predict bond changes within the reactants. In the first step, a graph-convolutional neural network calculates the pair-wise reactivity between atoms and identifies possible reaction centres. After the reaction centres are filtered, a Weisfeiler-Lehman Difference network ranks the bonds most likely reacting. The final product molecule is generated by applying the suggested bond changes to the reactants. Jin et al. [5] made their data set and training, validation and test split publicly available, from here on referred as USPTO_MIT. The data set contained no reactions with stereochemical information. Reaction SMILES containing stereoisomers were previously filtered out, as this would have required a more sophisticated approach, able to predict not only bond changes but also changes in atomic labels, for example, specifying 3-dimensional configuration at a tetrahedral carbon. The open-source USPTO_MIT data set made it possible to compare with alternative methods directly. In the same

year, Schwaller et al. [36] published a SMILES-2-SMILES approach using a seq-2-seq model with an attention layer. Seq-2-seq models generate product molecules, SMILES token by SMILES token, using a recurrent neural network [110]. While the usage of neural machine translation models for reaction prediction had already been proposed by Nam & Kim [111] and for retrosynthesis by Liu et al. [65], it was the first large-scale demonstration of a seq-2-seq model. Schwaller et al. [36] showed that representing reactants and reagents solely with SMILES attention-based seq-2-seq models, could compete with graph-based models where the node features were composed of more chemical information. The attention-weights could be visualised and revealed that the decoder focuses on one or more relevant atoms in the reactants while predicting each atom of the product. Compared to the bond change prediction approaches, SMILES-2-SMILES approaches construct the whole product molecule token by token. To solve the ambiguity of atomic order in SMILES, Schwaller et al. [36] used the canonical SMILES to specify an order in which the atoms have to be predicted. Besides predicting accuracies similar to the original work of Jin et al. [5] on the USPTO_MIT set, Schwaller et al. published the USPTO_STEREO data set to compare models able to predict stereoisomers (to the level they can be described in SMILES). Beyond the proof of scaling seq-2-seq models with large data sets, Schwaller et al. [36] introduced a new metric for measuring accuracy, by weakly separating reactants and reagents with a > token and representing only the most common reagents. This metric was unfortunately endorsed by other groups [5, 6, 37, 38] creating a measure of comparison that brings the development of such models in the wrong direction. Separating reactant and reagents leads to simplification of the reaction prediction problem, as one must already know the reacting molecules to do the separation, as pointed out by Griffiths et al. [112], The prediction problem is then reduced to the prediction of the correct reactive sites. This metric has been corrected for reaction prediction [4].

Similar to the Baldi group [106], Bradshaw et al. [37] followed an approach inspired by textbook organic chemistry and arrow pushing diagrams. They developed a model to predict electron paths. To do so, they analysed the graph-edits published by Jin et al. [5]. Their method could only be applied to USPTO_LEF, a subset of USPTO_MIT. In their paper, Bradshaw et al. [37] claim that they predict not only the product, but also the "mechanism". While they might get the mechanism of simple reactions, the underlying mechanistic steps often involve more electron movements then can be read out by comparing the final product with the starting material. Predicting the correct product does not mean that the predicted electron path is correct, as graph-edits cannot be taken as ground truth for mechanistic steps. For instance, a push to a catalyst in a coupling reaction could not be represented in their method as they add the reagents (solvents, catalysts) only as global features. The work of Bradshaw et al. [37] is interesting as they tackle the problem with new machine learning approaches. Similarly, Do [38] suggested a Graph Transformation Policy network, to learn the best policy to predict bond changes. The model did not have the restriction of only being able to predict the USPTO_LEF but could also be used on the USPTO_MIT data set, where it after invalid product removal achieved a top-1 accuracy of 83.2%. Late 2018, Coley et al. [6] improved their previous WLN/WLDN approach presented in Jin et al. [5] and called it a graph convolutional neural network (GCNN) approach. The main difference is that they changed the enumeration criterium in the first step. Instead of generating candidates using the top-6 atom pairs, they allow up to 5 simultaneous bond changes out of the top-16 bond changes for the enumeration. This change leads to higher coverage of products in the test set and hence, also an improvement in the overall accuracy, reaching considerable 85.6% top-1 on the

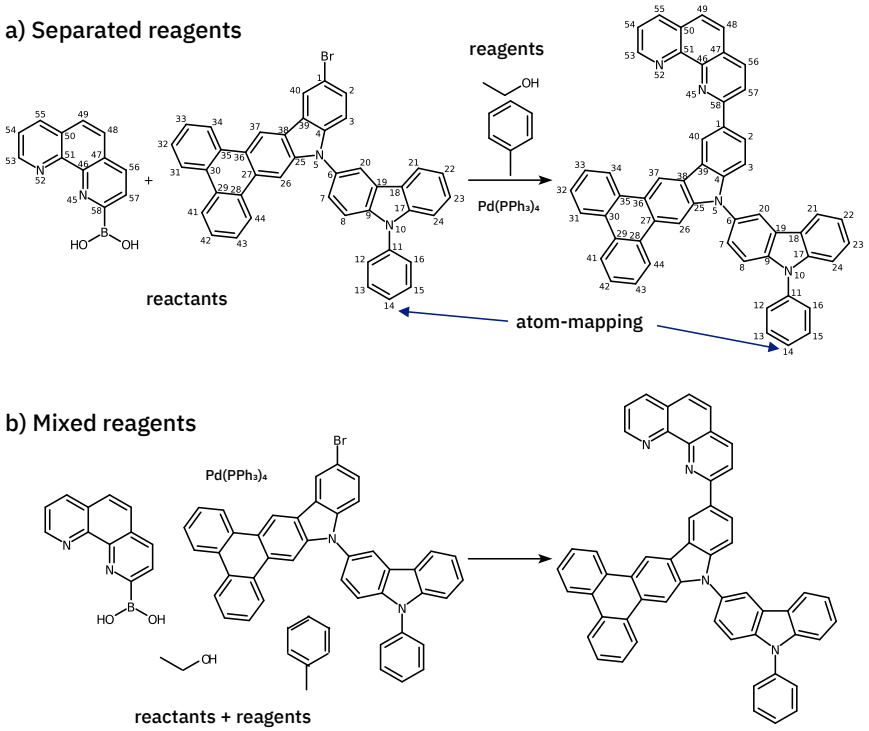

Figure 2.7: **Reactants-reagents separation.** Visualisation of the two chemical reaction representation settings. a) shows the separate reagents setting, where the information of which molecule contributes atoms to the product and which molecule does not is explicitly contained. Unfortunately, this requires knowing the atom-mapping and therefore, also knowing the product before making the prediction. b) in contrast, shows the mixed setting, where no distinction is made between reactants and reagents. The model has to figure out itself, which molecules are the most likely to react together. The mixed setting makes the reaction prediction problem more realistic, but also more challenging.

USPTO_MIT with separated reagents. The approach is still a two-step process and therefore, not end-to-end. Parameters like the maximum number of bond changes to take into account have to be determined empirically over the validation set and might change for another reaction data set. The coverage of the first step sets the upper bound for the accuracy of the second step. Schwaller et al. [4] demonstrated for the first time accuracies of over 90% on the USPTO_MIT data set. They called their model Molecular Transformer, as it was built on top of the Transformer architecture [2] introduced in Section 2.4.3. To prevent the model to learn only from the canonical representation, the training set inputs were augmented with non-canonical versions of the SMILES [46]. Schwaller et al. [4] not only show significant improvements in terms of top-1 accuracy on the USPTO_MIT data set but also on the USPTO_STEREO and a time-split Pistachio reaction test set containing stereochemical information. One major advantage of this approach is that the Molecular Transformer outperforms all previous approaches even when no distinction is made

between reactants and reagents in the input. Therefore, the approach is the first, which is completely template and atom-mapping independent. The difference between a separated reagent and the so-called mixed reagents reaction representation is visualised in Figure 2.7. It is also interesting to note as SMILES-based linguistic approaches have often been discredited because of the possibility to introduce syntactical errors during the SMILES inference process. Syntactical errors are the norm, and actually, the capacity for an underlying AI model to learn the grammar rules behind SMILES codification is very much depending on the architecture used. For instance, the work made by Schwaller et al. [4] using the Molecular Transformer clearly shows that less than 1% of the top-1 prediction is grammatically invalid. Remarkably, the underlying AI model learns not only the domain knowledge (organic chemistry) but also the SMILES grammar, to a level that can be considered close to perfection.

In 2020, Qian et al. [113] developed a GCNN and used probabilistic and symbolic inference to enforce chemical constraints and account for prior chemical knowledge. Similar to other graph neural networks for reaction prediction [5, 6, 37], the approach is limited to predicting reactions without stereochemical information. Recently, Tetko et al. [114] demonstrated that using the Molecular Transformer [4] results can be improved by applying extensive data augmentation and using computationally more expensive testing protocols. At test time, Tetko et al. [114] generated for every input reaction up to 100 data-augmented copies. The predicted product was then determined by taking the most frequent predicted product from the 100 inputs presented to the model. Using this test-time augmentation (TTA), they were able to achieve an accuracy of 92% on the standard separated USPTO_MIT data set. One of the advantages seq-2-seq models have during training compared to GCNN approaches is that the model gets feedback for every token in the sequence and not only for a few graph edits. Sacha et al. [115] represented the products of a reaction as a canonical sequence of graph-edits predicted by a GCNN. This idea makes it possible to train GCNN models similarly to seq-2-seq models.

Table 2.1 reports the top-1, top-2 and top-3 accuracies of the different approaches on the patent data sets set, where top-N accuracy means that the reported product could be found in the N most likely predictions of the model.

In Table 2.1, it becomes apparent that even recent work focused on predicting reactions without stereochemical information. However, stereochemistry, the 3-dimensional arrangement of atoms, affects chemical reactivity. While graph-edit-based approaches are currently unable to handle stereoisomers [5, 6, 113], predicting reactions, where stereochemistry plays a role, is a weakness of the Molecular Transformer. In Chapter 3, I present an approach to improve the Molecular Transformer predictions on challenging carbohydrate reactions using a small training data set using transfer learning [116]. This work also includes the first experimental validation of deep learning chemical reaction prediction models.

In the work of Coley et al. [6] and Schwaller et al.[4], the attention weights are used to enhance the explainability of their predictions and make the models more transparent, one of the major criticisms of those data-driven black-box models. Coley et al. [6] calculate pair-wise interactions between reactant and reagents atoms (source) during the first step of their approach. The most reactive sites can be identified by selecting one atom and highlighting those interactions with all the other source atoms. In the Molecular Transformer, instead, this would correspond to a visualisation of the self-attention in the encoder. Using the Molecular Transformer not only the encoder and decoder self-attentions can be visualised, but more interestingly, also the decoder-encoder

Table 2.1: **Standard benchmark reaction prediction results** Top-3 accuracies of the recent data-driven reaction prediction models on the different USPTO subsets. Currently, only product generation models are able to take into account stereochemical information and make predictions on the USPTO_STEREO data set. For all the models, a significant accuracy increase is observed between Top-1 and Top-2. TTA=Test-time augmentation.

Accuracy	Top-1 [%]	Top-2 [%]	Top-3 [%]
USPTO_MIT Separated reagents			
Jin et al. [5]	79.6		87.7
Schwaller et al. [36]	80.3	84.7	86.2
Do et al. [38]	83.2		86.0
Coley et al. [6]	85.6	90.5	92.8
Schwaller et al. [4]	90.4	93.7	94.6
Schwaller et al. (ensemble) [4]	91.0	94.2	95.2
Qian et al. [113]	90.4	93.2	94.1
Tetko et al. (TTA 100x) [114]	92	95.4	
Sacha et al. [115]	89.3	92.7	94.4
USPTO_MIT Mixed reagents			
Jin et al. [5]	74.0		86.7
Schwaller et al. [4]	88.6	92.4	93.5
Tetko et al. (TTA 100x) [114]	90.6	94.4	
Sacha et al. [115]	86.3	90.3	92.4
USPTO_STEREO Separated reagents			
Schwaller et al. [36]	65.4	71.8	74.1
Schwaller et al. [4]	78.1	84.0	85.8
USPTO STEREO Mixed reagents			
Schwaller et al. [4]	76.2	82.4	84.3

attention. The latter can be interpreted as how important source atoms are to predict a specific product atom. Empirical evaluations of those attention weight maps show that the model learned something similar to atom-mapping, as seen for a Bromo Suzuki coupling reaction in Figure 2.8.

As shown above, the two main reaction prediction approaches construct the major products of a reaction using product generation or bond changes prediction methods. While the recent product generation methods are completely atom-mapping independent [4], the atom-mapping is required to generate the ground-truth bond changes for the bond changes prediction methods [6,

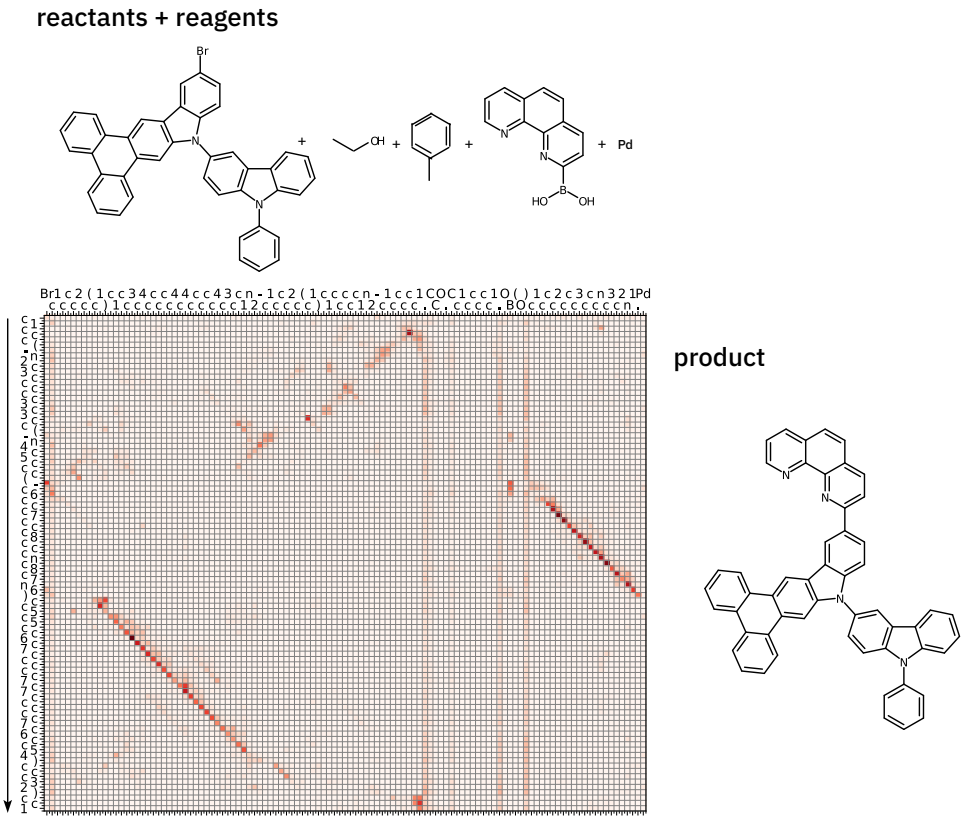

Figure 2.8: **Attention weights for a Bromo Suzuki coupling reaction.** Product-reactants attention generated by the Molecular Transformer [4] for a Bromo Suzuki coupling reaction. Attention weights show how important an input token (horizontal) was for the prediction of an output token (vertical). It can be seen that the model focused on the corresponding molecule parts in the reactants, while predicting the product.

113]. As atom-mapping is still typically generated by rule-based approaches, the bond changes prediction methods inherit the limitations of the underlying approach used for the atom-mapping.

2.5.2 Synthesis route planning tools

Similar approaches to the ones that can be used for forward chemical reaction prediction can be used for single-step retrosynthetic predictions. "Retro" because the aim is to predict precursor molecules from a given product molecule. Liu et al. [65] introduced a seq-2-seq approach for single-step retrosynthesis and evaluated the model with Top-N accuracy. Top-N accuracy means that the reported precursors are present in the first N predictions by the model. Although top-N accuracy is simple to compute, it is not well suited for the retrosynthesis task. Numerous precursor sets could lead to the desired product and not only the reported one.

Moreover, the single-step retrosynthesis task can be formulated in multiple ways, significantly impacting its difficulty. Tetko et al. [114], recently suggested that it is enough to predict the largest fragment of the precursors, as expert chemists can fill in the rest of the information (other reactants and reagents). The task, as originally described by Liu et al. [65], was to predict reactants without reagents. The difference between reactants and reagents is often subtle and subjective. I use a more challenging formulation of the task in this thesis predicting all precursor molecules including reagents. This formulation allows my approach to be fully atom-mapping independent. To date, most studies still focus on predicting the reactants only [23, 27, 33, 65, 107, 108, 117, 118, 119, 120, 121, 122, 123, 124, 125] and require atom-mapping to make the reactant-reagent distinction in the training data.

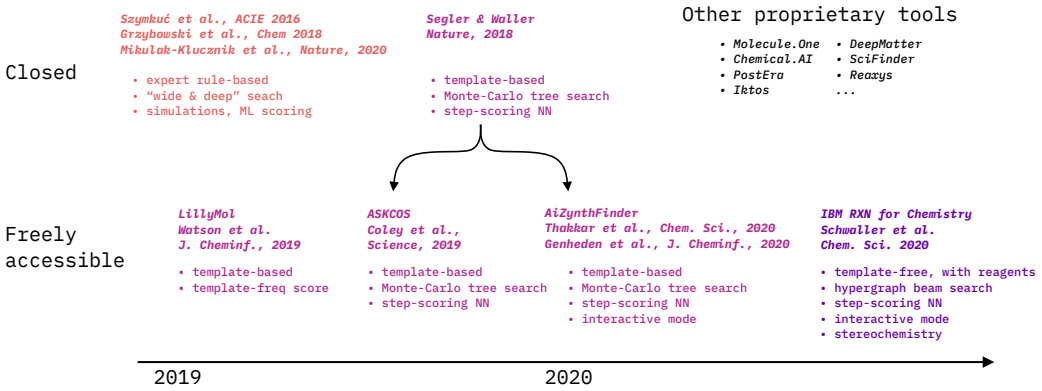

Figure 2.9: **AI-driven synthesis planning tools.** Overview of the most influential closed [22, 23, 24, 107] and openly accessible synthesis planning tools [43, 109, 118, 125, 126].

The usefulness of single-step retrosynthesis approaches is limited, as most of the synthesis routes require multiple reaction steps, to resolve the route and reach commercially available molecules. The real challenge is the multi-step synthesis task. It is unclear how well the metrics like top-N accuracy that are optimised in most of the single-step retrosynthesis studies translate to the multi-step task. In a ground-breaking work in 2018, Segler and Waller [107] introduced a template-based approach combined with a Monte-Carlo tree search (MCTS) to plan multi-step synthesis routes. They used a neural network to score individual reactions and prune the tree. As the approach was template-based, the predicted reaction steps included reactants only. Segler and Waller [107] performed a chemical Turing test by giving predicted and literature routes for nine molecules to 45 graduate-level organic chemists and letting them evaluate, which ones they like most without knowing the source. This test showed that on the small sample set of nine routes, the generated routes were on par with the literature routes. There was no significant preference for one of the two sources.

Unfortunately, similar to the work by the team behind Chematica, who recently predicted routes to natural products [24], the work by Segler and Waller [107] is not open-source or accessible through an application programming interface (API). Hence, it is not comparable to other approaches. Based on the work by Segler and Waller [107], Coley et al [125] and Thakkar et al. [109,

126] have implemented open-source algorithms for template-based and MCTS-based multi-step synthesis planning.

In Chapter 4, I will present my multi-step retrosynthesis planning approach [43]. It uses two Molecular Transformer models, one for precursor set suggestions and one for reaction scoring, and hypergraph beam-search to find optimal routes.

2.6 Learning the language of chemical reactions

Organic chemistry and written language have much in common [127]. Similar to letters and words, there is a well-defined set of atoms that can be used to construct molecules. Not every combination of letters makes a valid word and not every combination of atoms a stable molecule. If, in this analogy, atoms are letters and molecules are words, chemical reactions or sets of molecules can be seen as sentences. With text-based representations like SMILES [46, 47], the molecular graphs can be linearised using their spanning trees and encoded as a sequence of symbols. Note that in SMILES bond lengths and angles remain undefined. Hence, the molecules represented as SMILES do not contain conformer information. Moreover, hydrogen atoms are typically removed from the graphs before generating the SMILES and set to be implicit. Text-based molecular representations make it possible to apply NLP-inspired approaches, like Transformers [2, 3, 98] to molecules and chemical reactions.

Compared to graph neural network-based approaches operating on the molecular graphs, transformer-based approaches operating on SMILES might not be the immediate first choice. Still, one of their advantages is that in SMILES stereochemical information can be encoded to a certain extent. In contrast in graphs, it is harder to incorporate this information as it might depend not only on the nearest neighbours' order but also on further not directly connected neighbours [128]. Apart from enforcing a prior on connecting covalently bonded atoms, graph-neural networks are not that different from the transformer architecture. The Transformer architecture [2] can be seen as a graph-neural network where input tokens are nodes, and all of them are connected. The connections between the nodes are learned from examples through the attention mechanism. The attention mechanism is the common feature in all neural networks applied in this thesis. The initial token feature vectors in the Transformer model are computed with a context-independent token embedding layer [85], a neural networks layer that maps the token vocabulary to the input size of attention layers in the model. Those feature vectors might already carry some meaning. But they only take into account the single tokens and not the rest of the sentence. For instance, the word "bank" will be represented the same independent of the sequence talking about water or money. Similarly in chemical reactions, the token embedding layer will produce the same representation for any "C" token or "N". But the functionality and meaning of word and atom tokens much depend on their context.

A simple reweighing scheme to compute better feature vectors could be based on direct neighbours. However, in language and also chemical sequences, there are often long-range dependencies between tokens that refer to each other, or single tokens that alter the meaning of the whole sequence. In sentence like "The cat that jumped over the fence is black.", the token "black" at the end refers to the token "cat" in second position not to the seemingly closer "fence". Changing the sentence to "The cat that jumped over the fence is *not* black." by including one additional word

alters the meaning of the sentence. To perform well in language and chemical tasks, models have to capture long-range dependencies and fine-grained modifications in human language and chemical sequences alike. In chemical reactions, changing an electron-withdrawing functional group to an electron-donating group may significantly alter the reactivity and lead to a different reaction outcome. This change does not necessarily need to be close to the reaction centre. Hence, a better reweighing scheme than proximity to compute context-dependent feature vectors is required.

Self-attention can be seen as a method to reweigh individual inputs, in our case token feature vectors, based on all inputs. In self-attention, the key, query and value layers get the same feature vectors as input. First, the outputs matrices of the key and query layers are multiplied, then scaled and normalised. Those normalised scores, also called attention weights, are then multiplied with the outputs of the value layer. Figure 2.10 *a* visualises such a self-attention block. The key, query and value layers all contain trainable weights. By modifying those weights, the model learns to attend contextual information and produce more meaningful token representations. As the dimensions of the outputs of attention block are the same as the input dimensions, multiple such blocks can be stacked one after another.

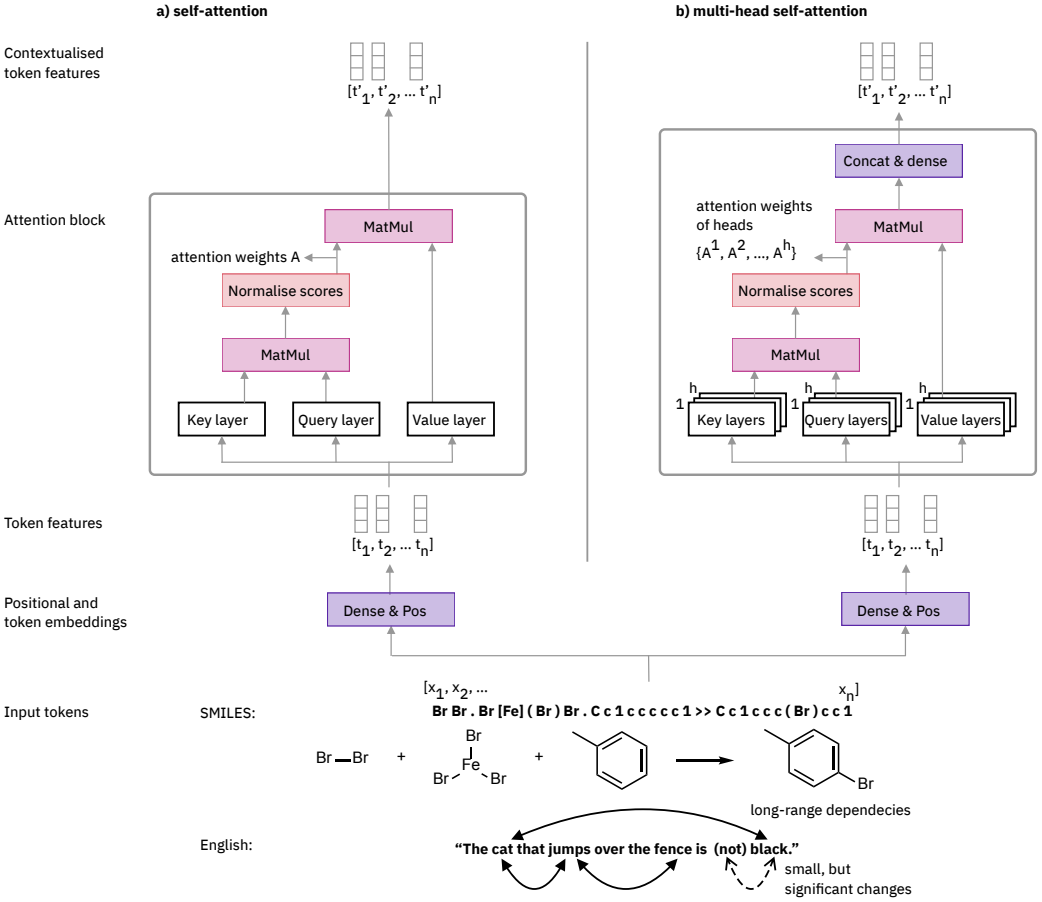

Figure 2.10: **Self-attention block.** a) single-head and b) multi-head attention

Looking at the example sentence and the word "jump", we could ask what or who jumped? The cat. And, over what or on what the subject jumped? Over the fence. Depending on the question, the attention mechanism would need to attend different tokens to generate an adequate answer. Multi-head attention [2] made it possible for models to attend the context using different attention functions simultaneously. As shown in Figure 2.10 *b*, multiple key, query and values layers are used in parallel instead of a single. In a multi-head attention block, the outputs of the last matrix multiplication are then concatenated and passed through an additional dense layer in order to return a feature matrix of the same dimensions as given as input. The parallel layers are the heads, and for each head an attention matrix is returned. Vig [129] and Hoover et al. [130] used visual inspection to demonstrate that after training the different heads attended different features in a sentence.

In a chemical reaction example, the first head could attend neighbouring atoms, the second atoms within the same molecule, the third important functional groups that might influence the reactivity. The multiple attention heads allow the model to focus on multiple tokens that are far apart and, consequently, generate better contextualised token representations.

The same concepts used for self-attention also apply to encoder-decoder attention, where the queries originate from the decoder part of the models, and masked self-attention, where future tokens are masked and their attention weights are set to zero.

Throughout this thesis, I will demonstrate how the analogy between human and chemical reaction language can be exploited to tackle chemical reaction tasks using Transformer models [2, 3, 4, 98].

3 Transfer learning enables the Molecular Transformer to predict regio- and stereoselective reactions on carbohydrates

Organic synthesis methodology enables the synthesis of complex molecules and materials used in all fields of science and technology and represents a vast body of accumulated knowledge optimally suited for deep learning. While most organic reactions involve distinct functional groups and can readily be learned by deep learning models and chemists alike, regio- and stereoselective transformations are more challenging because their outcome also depends on functional group surroundings. Here, we challenge the Molecular Transformer model to predict reactions on carbohydrates where regio- and stereoselectivity are notoriously difficult to predict. We show that transfer learning of the general patent reaction model with a small set of carbohydrate reactions produces a specialised model returning predictions for carbohydrate reactions with remarkable accuracy. We validate these predictions experimentally with the synthesis of a lipid-linked oligosaccharide involving regioselective protections and stereoselective glycosylations. The transfer learning approach should be applicable to any reaction class of interest.

This chapter has previously appeared as a scientific article in Nature Communications:

G Pesciullesi•, P Schwaller•, T Laino, J Reymond. Transfer learning enables the molecular transformer to predict regio- and stereoselective reactions on carbohydrates. *Nat. Commun.*, 2020, 11, 4874, (CC BY 4.0). The syntheses were performed and analysed by Giorgio Pesciullesi.

3.1 Introduction

Organic synthesis is a complex problem-solving task in which the vast knowledge accumulated in the field of organic chemistry is used to create new molecules, starting from simple commercially available building blocks [131]. Because of its complexity, organic synthesis is believed to be one of the main bottlenecks in pharmaceutical research and development [132], and having accurate models to predict reaction outcome could boost chemists' productivity by reducing the number of experiments to perform.

Machine learning has long been present in the chemical domain, tackling challenges such as Quantitative Structure-Activity Relationship predictions [133], virtual screening [134] and quantum chemistry [135, 136]. Enabled by algorithmic advances in deep learning [2, 73, 84, 88] and the

availability of large reaction data sets [41, 42], reaction prediction methods have emerged in recent years [4, 5, 6, 36, 37, 38, 43, 111, 113, 137]. Those reaction prediction methods can be divided into two categories [138], bond change prediction methods using graph neural networks [5, 6, 37, 38, 113] and product SMILES generation using sequence-2-sequence models [4, 36].

Reaction prediction tasks are typically evaluated on the USPTO_MIT benchmark [5], which does not contain molecules with defined stereocentres. Currently, the best prediction algorithm in terms of performance is the Molecular Transformer [2, 4]. The architecture is based on the ground-breaking work by Vaswani et al. [2], which revolutionised the field of neural machine translation, where sentences in one language are translated into another language. In contrast, for reaction prediction, the model learns to translate the precursors' Simplified molecular-input line-entry system (SMILES) [46] representation into the product SMILES.

The Molecular Transformer can be accessed for free through the IBM RXN for Chemistry platform [44]. Compared to other methods, such as graph neural networks-based ones, the advantages of the Molecular Transformer approaches are that they do not require mapping between the product and reactant atoms in the training [112] and inputs can contain stereochemistry. In fact, sequence-2-sequence approaches, like the Molecular Transformer [2, 4], are currently the only large-scale reaction prediction approaches capable of handling stereochemistry. Stereochemistry is systematically avoided in graph-based methods, as the connection table and adjacency matrix of two stereoisomers is identical. Although stereoselectivity can theoretically be predicted by the Molecular Transformers [4], it is one of their most significant weaknesses because of the lack of clean training data. To date, their performance on predicting specific stereochemical reactions has not been investigated.

In this work, we investigate the adaptation of the Molecular Transformer to correctly predict regio- and stereoselective reactions. As study case we focus on carbohydrates, a class of molecules for which the stereochemistry and the high degree of functionalisation are key reactivity factors. Carbohydrate chemistry is essential for accessing complex glycans that are used as tool compounds to investigate fundamental biological processes such as protein glycosylation[139, 140, 141], as well as for the preparation of synthetic vaccines [142, 143, 144]. Predicting the outcome of carbohydrate transformations, such as regioselective protection/deprotection of multiple hydroxyl groups or the stereospecificity of glycosylation reactions, is a very difficult task even for experienced carbohydrate chemists[145, 146], implying that this field of research might particularly benefit from computer-assisted reaction prediction tools.

First, we investigate transfer learning with a specialised subset of reactions as a means to adapt the Molecular Transformer to achieve high performance on carbohydrate reactions. Transfer learning, where a model is trained on a task with abundant data and either simultaneously trained or subsequently fine-tuned on another task with less data available [93], has recently let to significant advancements in Natural Language Processing [3, 91, 147, 148]. For instance, it has been used to improve translation performance in low-resource languages [147]. More recently, unsupervised pre-training transfer learning strategies have successfully been applied to sequence-2-sequence models [148, 149]. In the chemical domain, transfer learning has enabled the development of accurate neural network potential for quantum mechanical calculations [150] and shows great potential to solve other challenges [151]. For transfer learning we use a set of 20k carbohydrate reactions from the literature, comprising protection/deprotection and glycosylation sequences. We explore multi-task learning, as well as sequential transfer learning, and show that the adapted model, called

the Carbohydrate Transformer, performs significantly better than the general model on carbohydrate transformations and a model trained on carbohydrate reactions only.

Second, we perform a detailed experimental assessment of the deep learning reaction prediction model and test the Carbohydrate Transformer on unpublished reactions. Our assessment consists of a 14 step total synthesis of a modified substrate of a eukaryotic oligosaccharil transferase (OST). We also challenge our Carbohydrate Transformer to predict the reactions from the recently published total syntheses of the trisaccharide of Pseudomonas aeruginosa and Staphylococcus aureus [152] as a further assessment on more complex carbohydrate reactions. Those reactions would be considered challenging to predict, even for carbohydrate experts.

Overall, we observe a consistent top-1 prediction accuracy above 70%, which roughly means a 30% increase compared to the original Molecular Transformer baseline. We find that the confidence score is a good predictor of prediction reliability and that many wrong predictions have chemical reasons such as the lack of reagent stoichiometry in the training data. The approach we used to learn carbohydrate reactions could be applied to any reaction class. Hence, it is expected to have a significant impact on the field of organic synthesis, as models like the Molecular Transformer [4] can easily be specialised for the reaction sub-spaces that individual chemists are most interest in.

3.2 Results

3.2.1 Data availability scenarios

Besides the additional complexity, the main challenges for learning to predict stereochemical reactions is the data. In the largest open-source reaction data set by Lowe [41, 42], which fueled the recent advancements in machine learning for chemical reaction prediction, stereochemistry and specifically reactions involving carbohydrates are underrepresented and of poor quality. Hence, those reactions are problematic to learn.

In this work, we explore two real world scenarios, where there exist a large data set of generic chemical reactions and a small data set of complex and specific reactions. In our case, we use a data set derived from the US patent reactions by Lowe [42] as the large data set containing 1.1M reactions. We call this data set USPTO. For the specific reaction, we chose carbohydrates reactions, but the methods described could be applied to any reaction class of interest. We manually extracted reactions from the Reaxys [39] database, selected from papers of 26 authors in the field of carbohydrate chemistry. The small data set of 25k reactions will be referred to as CARBO for the remainder of the publication. We split the USPTO and the CARBO data set into train, validation and test sets. The reaction data was canonicalised using RDKit [153]. A more detailed description of the data is found in Supplementary Note 1.

If the access to the large and small sets is given, the two data sets can be used simultaneously for training. We call this first scenario multi-task. However, depending on the situation, direct access to the data of the generic data set may not be possible. For example, a company A may have proprietary reaction data precluded from external sharings. Company A could still train a model using their own data and share their model without revealing the exact data points. The trained model extracts some general chemical reactivity knowledge and could be shared without exposing

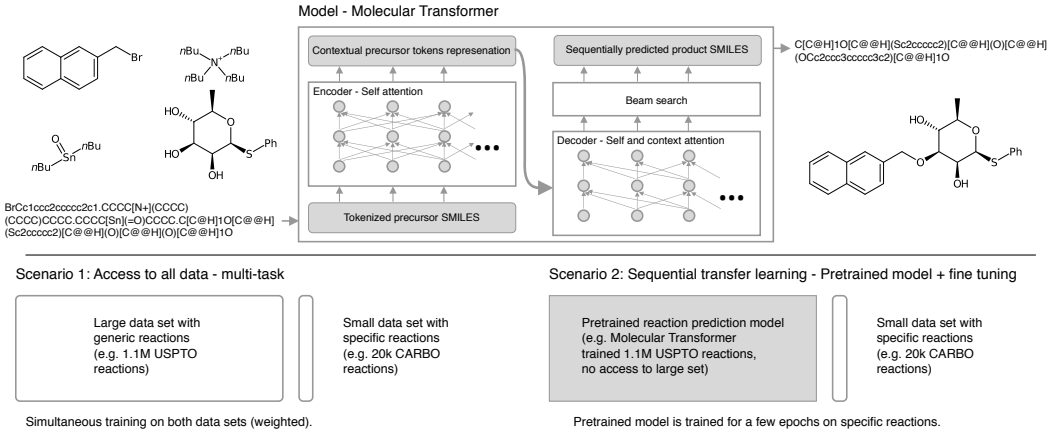

Figure 3.1: **Molecular Transformer model and data scenarios.** Sequence-2-sequence prediction of carbohydrate reactions and the two transfer learning scenarios, namely, multi-task and sequential training.

company proprietary information. This pre-trained model could then serve as a starting point to further train the model on another source of reactions. We call this scenario fine-tuning.

A visualisation of the model and the two scenarios can be found in Figure 3.1.

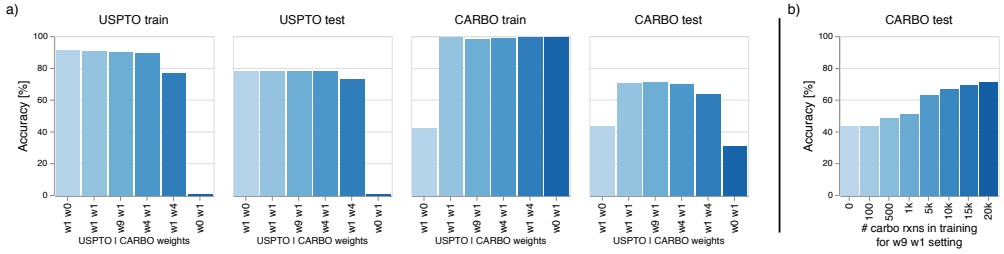

Figure 3.2: **Multi-task scenario results.** a) Top-1 accuracy of models trained with different weights on the USPTO and CARBO data set (the first number corresponds to the weight on the USPTO data set and the second to the weight on the CARBO data set). b) Top-1 accuracy for a model trained in the weight 9 weight 1 setting, where the number of reactions in the CARBO data set was reduced. Source data are provided as a Source Data file.

In the multi-task scenario, we investigated different reaction weighting schemes between the two sets. A comparison of the top-1 accuracies on the USPTO train, USPTO test, CARBO train and CARBO test sets for models trained with different weights for the USPTO train and CARBO train sets are shown in Figure 3.2 a). The weights describe in what proportion reactions from the two sets are shown per training batch. For example, weigtht 1 on USPTO and weight 1 on CARBO means that for one USPTO reaction one CARBO reaction is shown. As can be seen in the Figure, the highest accuracy on the CARBO test set (71.2 %) is obtained with weight 9 on the USPTO set and weight 1 on the CARBO set (w9w1). As expected, training only with

the CARBO train set leads to a poor CARBO test set accuracy (30.4 %). As 20k reactions are not enough for the model to learn predict organic chemistry. The accuracy reached by the model trained purely on the USPTO data reaches 43.3 %. It therefore performs better than the model trained purely on the CARBO reactions. In Figure 3.2 b), we assess the effect of the size of the CARBO train set. The accuracy continuously increases from 43.3 to 71.2 % with an increasing number of reactions in the train set.

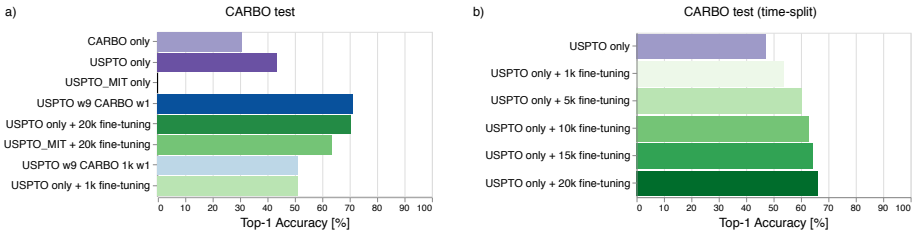

Figure 3.3: **Fine-tuning scenario results.** a) CARBO random split test set performance for different training strategies. In green are the top-1 accuracies of the models that were fine-tuned on either 1k or 20k CARBO reactions shown. For comparison, we included in purple the top-1 accuracies of the models trained on the single data sets (CARBO, USPTO and USPTO_MIT). Blue are the performances of models trained in the multi-task scenario. b) CARBO time split test set performance for different fine-tuning set sizes. Source data are provided as a Source Data file.

For the fine-tuning scenario, where access to the large generic data set is not given but a model, pre-trained on the large data set, is available instead, the results on the CARBO and USPTO test sets are shown in Figure 3.3 a). After training the model on the CARBO train set, the top-1 accuracy reaches a 70.3%, similar to the model that was trained on the two data sets simultaneously. The observed behavior is the same when less CARBO reactions are available. Also for 1k CARBO reactions, the fine-tuning model matched the accuracy of the corresponding multi-task model.

For this scenario, we analysed the effect of the train, validation and test split in more detail. We compared the random split described above to a time split, where we included CARBO reactions first published before 2016 into the train and validation sets and the reactions published from 2016 into the test set (2831 reactions). We investigated different fine-tune set sizes (1k, 5k, 10k, 15k and 20k). As seen in Figure 3.3 b), compared to the random split the top-1 accuracy with the 20k fine-tuning dropped slightly to 66% but it is still substantially larger than the accuracy that could be obtained with the generic USPTO training set only. Already with 5k CARBO reactions, an accuracy above 60% was reached. The larger the CARBO fine-tuning set, the better the performance of the fine-tuned model.

Besides the fact that the reactions in the large data set do not need to be revealed, another advantage is the short fine tuning training time. The fine tuning requires only 5k steps compared to 250k steps in the multi-task scenario. However, if time and access to both data sets are given, it is better to train simultaneously on all data for a longer time as the performance on the large data set does not decrease, as it does in the fine-tuning scenario. If the interest is only in a specific reaction class, short adaptation times or if generic data is not available, then fine-tuning a pre-trained model is better.

To further demonstrate the effectiveness of the fine-tuning approach, we performed an experiment where we pre-trained a model on a data set without stereochemical information. To do so, we used the USPTO_MIT data set by Jin et al. [5]. As seen in Figure 3.3 a), although the pre-trained model does not manage to predict any CARBO test set reactions, after fine-tuning for 6k steps the model reaches an accuracy of 63.3 %. The accuracy was not as high as with USPTO pre-training but a significant improvement over the 0.0 % correctly predicted reactions by the pre-trained model. The low accuracy after pre-training was expected as none of the chiral centre tokens (e.g. "[C@H]", "[C@@H]") were present in the training set. The fine-tuning result shows that the Molecular Transformer model is able to learn new concepts within a few thousands training steps on 20k data points.

In the next sections, we will compare the model trained only on the USPTO data, which was also used as pre-trained model (USPTO model) with the model that was then fine-tuned on the 20k CARBO reactions (CARBO model).

3.2.2 Experimental assessment

Although the accuracy of the transformer has been widely assessed [4], an experimental validation is still missing. Here, we decided to validate both the transformer and the augmented precision of the CARBO model on a recently realised synthetic sequence from our own laboratory, absent from the training data. This sequence is a 14 step synthesis of lipid linked oligosaccharide (LLO) **15** to be used as a substrate to study oligosaccharyl transferases (OST) [154, 155] (Figure 3.4). The sequence contains typical carbohydrate chemistry: protecting group manipulations (steps: b, h, i, l n, p), functional group manipulations (step c, d), regioselective protections (step e), a β-selective glycosylation (step g) and an α-selective phosphorylation (step m). The latter regio- and stereoselective transformations are of particular interest because their selectivity is generally difficult to control and to predict, even for experienced synthetic chemists.

We used both the general USPTO model and the fine-tuned CARBO model to predict 13 of the 14 steps in the sequence (step b was removed since it appeared in the training set). The USPTO only made four correct predictions (31%), which were either standard protecting group manipulations (step a, g, n) or functional group exchanges (step c). The CARBO model also correctly predicted these four simple reactions, but additionally, made another 6 correct predictions, including the regioselective benzoylation (**5** to **6**, step e) and the β-selective phosphorylation (**11** to **12**, step m), corresponding to a 77 % success rate and a 46 % improvement over the USPTO model, in line with the overall statistics presented above.

In detail, the CARBO model only made three mistakes. The first one concerns the reduction of the primary iodide **4** to a methyl group in **5** by hydrogenation, which is mistakenly predicted to also reduce the benzyl glycoside. The USPTO model makes the same mistake. Both models have not learned that carrying out the reaction in the presence of ammonia reduces the catalyst activity and avoids debenzylation, as no such reaction was present in the training sets. The second mistake concerns a similar reduction of the benzyl glycoside in **10** (step l), which is predicted to yield the β-lactol while the product **11** is in fact formed as an anomeric mixture. Again, the USPTO model makes the same mistake. Both models ignore that the intially formed β-lactol equilibrates spontaneously to the anomeric mixture via ring opening. Finally, the CARBO model predicts a shortened prenyl chain in the phosphate coupling reaction forming the protected LLO **14** (step

o), which does not make chemical sense. In this case it should be noted that the CARBO training set does not contain a single LLO molecule, and that the USPTO model performs worse since it returns an invalid SMILES for this reaction.

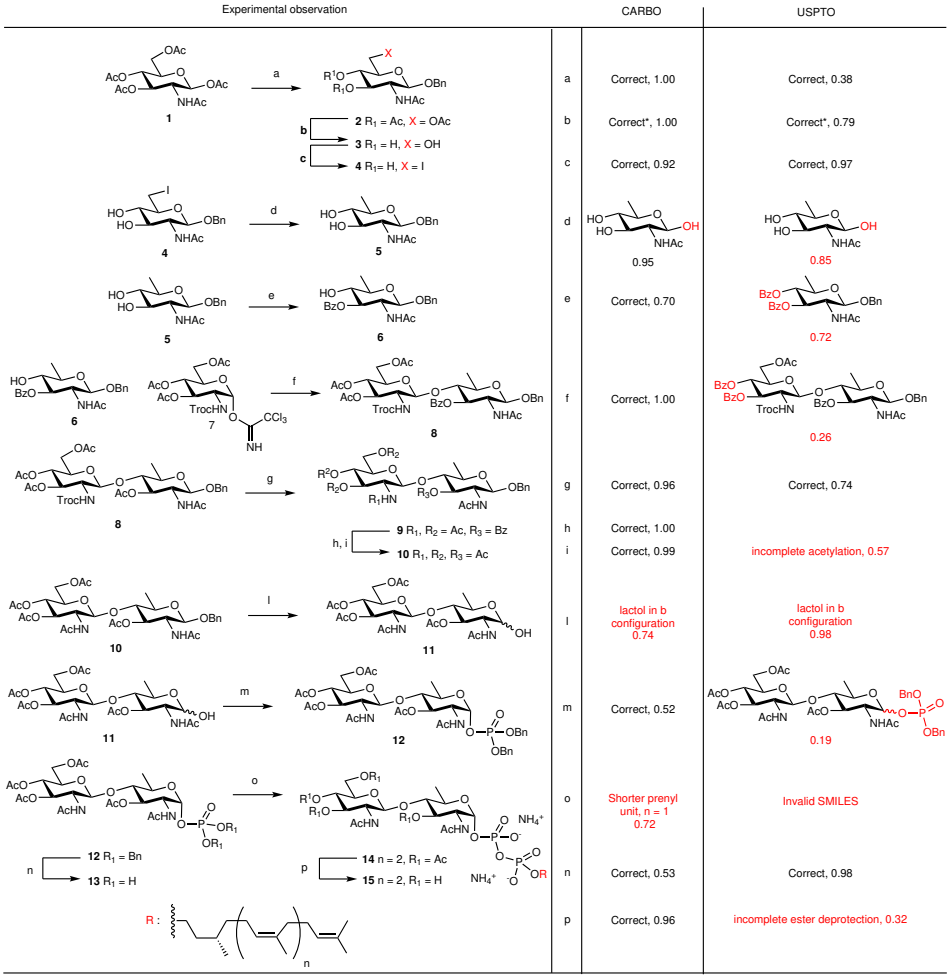

Figure 3.4: **Synthesis of lipid linked oligosaccharide (LLO).** Reaction conditions : a) BnOH, $Yb(OTf)_3$, DCE, 90° C, 2h, 78%. b) MeONa, MeOH, sonication, 30 min. c) PPh_3, I_2, imidazole, THF, 1h, reflux, 88% over two steps. d) Pd/C, NH_4OH, H_2, THF/H_2O, 30 min, 77%. e) BzCl, pyr, -35°, 70%. f) BF_3Et_2O, 4 Å MS, DCM, 26 h, 73%. g) Zn, Ac_2O, AcOH, DCE 50°, 3h, 96% h) MeONa, MeOH/DMF, 4 days. i) Ac_2O, 4-(Dimethylamino)pyridine, pyr, 76% over three steps. l) H_2, THF/H_2O, 10 bar, 16h m) LiHMDS, tetrabenzylpyrophosphate, 53%. n) H_2, THF/MeOH, 1h. o) farnesylnerol, CDI, DMF, then **11**, 5 days, 18%. p) MeOH, NH_4OH, 16h, qte. (*): reaction present in the training set.

We obtained similar prediction performances from both models when analysing a recently published total syntheses of the trisaccharide repeating unit of Pseudomonas aeruginosa and Staphylococcus aureus [152]. Those synthetic sequences comprises four difficult regio- and stereoselective glycosylation steps and five regioselective protection steps that are of particular interest. Out of

the 38 reactions that are absent from the training set in this sequence (Supplementary Figures 2-7), the USPTO model predicts only 15 reactions (39 %) correctly, and none of the difficult steps mentioned above. The CARBO model performs much better and correctly predicts 26 of the 38 reactions, corresponding to a 68% overall accuracy and a 29% gain over the USPTO model. In particular, the CARBO model correctly predicts the regioselectivity of the dimethyltinchloride mediated benzoylation of L-Rhamnopyranoside **16** (step no. 10, 3.5), the difficult regio- and stereoselective glycosylation at position 3 of the terminal fucosyl in disaccharide **18** (step no. 24) as well as the regioselective protection of the same disaccharide at position 3 (step no. 29), all of which are non-obvious even for synthetic chemists. Interestingly, the CARBO model predicts a double substitution of bis-triflate **19** instead of the correct single substitution at position 2, which the USPTO model correctly predicts. In this case it should be noted that the outcome of the reaction is dictated by stoichiometry (only one equivalent of the azide nucleophile), an information which is absent from the training data. In contrast to the USPTO training set, that contains only single azide substitutions, the CARBO training set contains single, as well as double substitutions. An analysis of the stereo centres in both data sets can be sound in Supplementary Table 1 and Supplementary Figure 1.

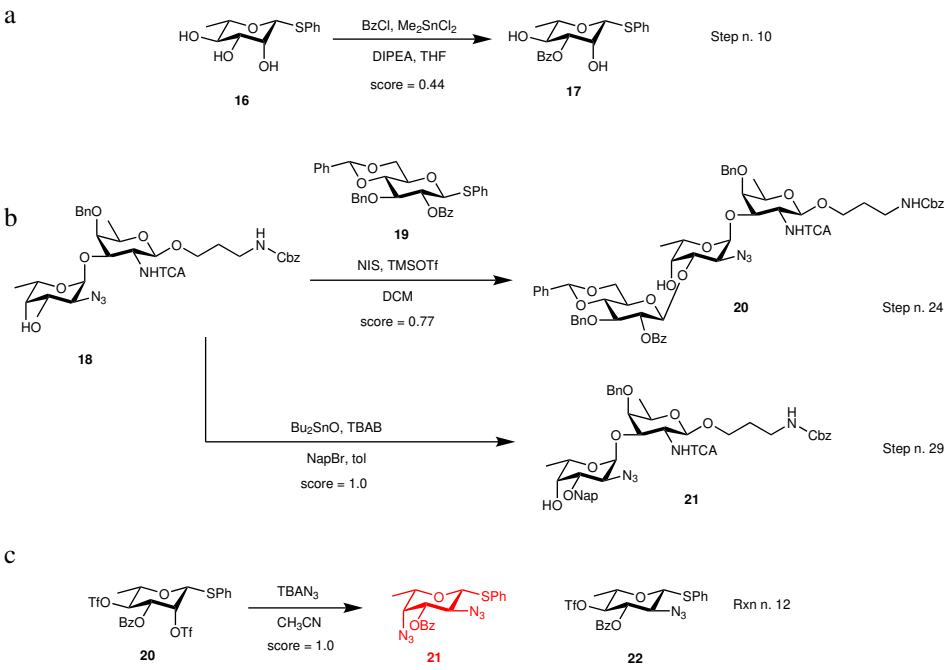

Figure 3.5: **Reactions predicted from recent literature.** (a) and (b): Reactions correctly predicted. (c) wrongly predicted reaction (red structure) due to missing reagent stoichiometry in the model: only one equivalent of NaN_3 was used resulting in single substitution, while the model predicts double substitution.

Every predicted reaction is associated with a confidence score [4], which is calculated from the product of the probabilities of the predicted product tokens. Interestingly, the confidence score

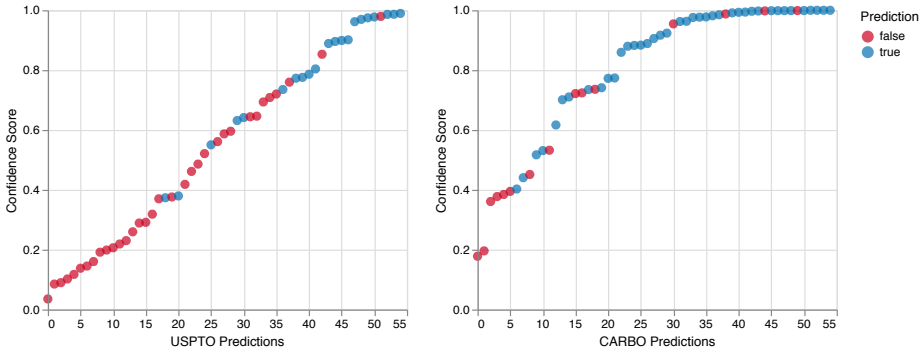

Figure 3.6: **Analysis of prediction confidence scores.** Predictions (ordered by confidence score) for the experimental assessment. Source data are provided as a Source Data file.

correlates with the correctness of the prediction (figure 3.6). For both models most of the correct predictions have a score higher than 0.8.

To have a closer look at the capabilities of the model to self-estimate it's own uncertainty, we analysed every reaction in detail. In some cases we observe epimerisation or rearrangments that have little chemical significance and are associated with low score values. This even occurs in more trivial transformations, such as amine acetylation of the trisaccharide in reaction 27 (scheme S3). Although the model is not able to predict the correct product, its low score seems to indicate that the model senses its own mistake. The second class are arguably wrong predictions that have high confidence for chemical reasons. Such an example is the previously discussed reaction 12 (Scheme 2, entry c) whose outcome is influenced by stoichiometry that together with other reaction conditions, is excluded from the training data, making this reactions extremely difficult to predict.

Similar to previous work [4], one of the limitations of current SMILES-2-SMILES models is that environmental reaction conditions like temperature and pressure are not taken into account. Those conditions are often missing in the data sets, and even if present, it would not be straightforward to codify temperature profiles applied during chemical reactions. Another limitation is the data coverage and quality. As pointed out above, most of the wrong predictions can be explained with the data that the models have seen during training.

The availability of large high-quality open-source reaction data set containing information detailed on amounts, stoichiometry and reaction conditions could substantially improve reaction prediction models.

3.3 Discussion

In this work, we demonstrated that transfer learning can be successfully applied to a generally trained transformer model using as few as 20k data points to derive a specific model that predicts reactions from a specific class with significantly improved performance. Transfer learning of the general molecular transformer model, trained on the USPTO data set to a specific set of reac-

tions, to obtain a high performance specialised model as demonstrated here should be generally applicable towards any subclass of specific reactions of interest.

Here we used transfer learning to improve predictions of regio- and stereoselectivity, a central aspect of synthetic chemistry that has not been systematically evaluated previously by reaction prediction models, in part due to the fact that the Molecular Transformer is currently the only model able to handle stereochemistry. As a test case we examined carbohydrates, a well-defined class of molecules for which reactions are difficult to predict even for experienced chemists, and subjected our model to experimental validation. We anticipate that the Carbohydrate Transformer will serve the practical purpose of improving the efficiency of complex carbohydrate syntheses. The model can guide chemists by predicting and scoring potential carbohydrates reactions before performing them experimentally. The fact that the confidence score correlates with prediction accuracy offers a simple metric to judge the quality of predictions. The shortcomings noted should be addressable by extending the training set with reactions that are not predicted well.

3.4 Methods

3.4.1 Reaction prediction model

All the experiments in this work were run with the Molecular Transformer model [4], which is illustrated in Figure 3.1. For details on the architecture we refer the reader to [2, 4]. We used Pytorch [81] and the OpenNMT [156] framework to build, train and test our models. Hyperparameters and a detailed description of the data sets can be found in the supplementary information. The investigated task is reaction prediction, where the aim is to predict the exact structural formula, including stereochemistry, of the products that are formed from a given a set of precursors as input. In the inputs, no difference is made between reactant and reagent molecules [4]. Following previous work [4, 36, 111], we use accuracy as the evaluation metric. The reported accuracies describe the percentage of correct reactions. A reaction is counted as correct only if the predicted products exactly matches the products reported in the literature after canonicalisation using RDKit [153]. The canonicalisation is required as multiple SMILES can represent the same molecule.

3.4.2 Chemical synthesis

All reagents were purchased from commercial sources and used without further purifications unless otherwise stated. All reactions were carried out in flame-dried round-bottomed-flask under an argon atmosphere, except if specified. Room temperature (rt) refers to ambient temperature. Temperatures of 0°C were maintained using an ice-water, -78°C with acetone/dry ice bath and the other temperatures using a cryostat. Dry solvents were obtained by passing commercially available pre-dried, oxygen-free formulations through activated alumina columns. Hydrogenation was performed at room pressure using H_2 filled balloon. Chromatographic purifications were performed with silica gel pore size 60 Å, 230-400 mesh particle size (sigma-aldrich). Thin layer chromatography (TLC) was performed using ALUGRAM Xtra Sil G/UV on pre-coated aluminium sheets, using UV light as a visualising, and an basic aqueous potassium permanganate solution and ceric ammonium molybdate (CAM) as developing agents. NMR spectra for ^{1}H, ^{13}C, DEPT, ^{31}P, COSY, HSQC, HMBC and NOE were recorded at room temperature with a

Bruker AV (400 MHz ^{1}H). Spectra were and processed using TopSpin 3.6.1 software. Chemical shifts are reported in δ (ppm) relative units to residual solvent peaks $CDCl_3$ (7.26 ppm for ^{1}H and 77.2 ppm for ^{13}C) and MeOD (3.31 ppm for ^{1}H and 49.00 ppm for ^{13}C). Splitting patterns are assigned as s (singlet), d (doublet), t (triplet), q (quartet), quint (quintet), multiplet (m), dd (doublet of doublets), and td (triplet of doublets). High resolution mass spectra (HRMS) was provided by the "Service of Mass Spectrometry" at the Department of Chemistry and Biochemistry in Bern and were obtained by electron spray ionisation (ESI) in positive or negative mode recorded on a Thermo Scientific LTQ OrbitrapXL. For the experimental procedures, NMR spectra and physical data of compounds 2-15, see Supplementary Note 3 of [116].

Data & code availability

The USPTO data set derived from Lowe [42] that we used for training and evaluation, our carbohydrate reactions, as well as the ones from the work of Behera et al. [152] are available from (`https://github.com/rxn4chemistry/OpenNMT-py/tree/carbohydrate_transformer`). Source data are provided with this paper.

The code and trained models are available from (`https://github.com/rxn4chemistry/OpenNMT-py/tree/carbohydrate_transformer`). The models are compatible with OpenNMT-py [156, 157], which was used for training and evaluation. The SMILES tokenisation function for preprocessing the inputs is found on the Molecular Transformer repository [4, 158]. The setup and hyperparameters can also be found in Supplementary Note 2.

4 Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy

We present an extension of our Molecular Transformer model combined with a hyper-graph exploration strategy for automatic retrosynthesis route planning without human intervention. The single-step retrosynthetic model sets a new state of the art for predicting reactants as well as reagents, solvents and catalysts for each retrosynthetic step. We introduce four metrics (coverage, class diversity, round-trip accuracy and Jensen-Shannon divergence) to evaluate the single-step retrosynthetic models, using the forward prediction and a reaction classification model always based on the transformer architecture. The hyper-graph is constructed on the fly, and the nodes are filtered and further expanded based on a Bayesian-like probability. We critically assessed the end-to-end framework with several retrosynthesis examples from literature and academic exams. Overall, the frameworks have an excellent performance with few weaknesses related to the training data. The use of the introduced metrics opens up the possibility to optimise entire retrosynthetic frameworks by focusing on the performance of the single-step model only.

This chapter has been published as a scientific article in Chemical Science:

P Schwaller, R Petraglia, V Zullo, V H Nair, R A Haeuselmann, R Pisoni, C Bekas, A Iuliano, T Laino. Predicting retrosynthetic pathways using a combined linguistic model and hyper-graph exploration strategy. *Chem. Sci.*, 2020, 11, 3316-3325 (CC BY-NC 3.0). Published by The Royal Society of Chemistry. The hyper-graph beam search was developed and implemented by Riccardo Petraglia. The predicted routes were analysed by Valerio Zullo and Anna Iuliano.

4.1 Introduction

The field of organic chemistry has been continuously evolving, moving its attention from the synthesis of complex natural products to the understanding of molecular functions and activities [159, 160, 161]. These advancements were made possible thanks to the vast chemical knowledge and intuition of human experts, acquired over several decades of practice. Among the different tasks involved, the design of efficient synthetic routes for a given target (retrosynthesis) is arguably one of the most complex problems. Key reasons include the need to identify a cascade of disconnections schemes, suitable building blocks and functional group protection strategies. Therefore,

it is not surprising that computers have been employed since the 1960s [131], giving rise to several computer-aided retrosynthetic tools.

Rule-based or similarity-based methods have been the most successful approach implemented in computer programs for many years. While they suggest very effective [23, 162] pathways to molecules of interest, these methods do not strictly learn chemistry from data but rather encode synthon generation rules. The main drawback of rule-based systems is the need for labourious manual encoding, which prevents scaling with increasing data set sizes. Moreover, the complexity in assessing the logical consistency among all existing rules and the new ones increases with the number of codified rules and may sooner or later reach a level where the problem becomes intractable.

4.1.1 The dawn of AI-driven chemistry.

While human chemical knowledge will keep fueling the organic chemistry research in the years to come, a careful analysis of current trends [23, 27, 33, 65, 107, 108, 117, 118, 119, 120, 121, 122, 123, 124, 125] and the application of basic extrapolation principles undeniably shows that there are growing expectations on the use of Artificial Intelligence (AI) architectures to mimic human chemical intuition and to provide research assistant services to all bench chemists worldwide.

Concurrently to rule-based systems, a wide range of AI approaches have been reported for retrosynthetic analysis [107, 108], prediction of reaction outcomes [4, 6, 30, 32, 35, 36] and optimisation of reaction conditions [163]. All these AI models superseded rule-based methods in their potential of mimicking the human brain by learning chemistry from large data sets without human intervention.

This extensive production of AI models for Organic chemistry was made possible by the availability of public data [41, 42]. However, the noise contained in this data generated by the text-mining extraction process heavily holds back their potential. In fact, while rule-based systems [164] demonstrated, through wet-lab experiments, the capability to design target molecules with less purification steps and hence, leading to savings in time and cost [165], the AI approaches [65, 107, 108, 109, 162, 166, 167, 168, 169, 170, 171] still have a long way to go.

Among the different AI approaches [172] those treating chemical reaction prediction as natural language processing (NLP) problems [127] are becoming increasingly popular. They are currently state of the art in the forward reaction prediction realm, scoring an undefeated accuracy of more than 90% [4]. In the NLP framework, chemical reactions are encoded as *sentences* using reaction SMILES [46] and the forward- or retro- reaction prediction is cast as a translation problem, using different types of neural machine translation architectures. One of the most significant advantages of representing synthetic chemistry as a language is the inherent scalability for larger data sets, as it avoids important caveats such as the need for humans to assign reaction centres [162, 164]. The Molecular Transformer architecture [158] is currently the most popular approach to treat chemistry as a language. Its trained models fuel the cloud-based IBM RXN [44] for Chemistry platform.

4.1.2 Transformer-based retrosynthesis: current status.

Inspired by the success of the Molecular Transformer [4, 44, 158] for forward reaction prediction, a few retrosynthetic models based on the same architecture were reported shortly after [166, 167,

169, 170, 171]. Zheng et al. [166] proposed a template-free self-corrected retrosynthesis predictor built on the Transformer architecture. The model achieves 43.7% top-1 accuracy on a small standardised (50k reactions) data set [173]. They were able to reduce the initial number of invalid candidate precursors from 12.1% to 0.7% using a coupled neural network-based syntax checker. Previous work reported less than 0.5% of invalid candidates in forward reaction prediction [4], without the need of any additional syntax checker. Karpov et al. [167] described a Transformer model for retrosynthetic reaction predictions trained on the same data set [173]. They were able to successfully predict the reactants with a top-1 accuracy of 42.7%. Lin et al. [169] combined a Monte-Carlo tree search, previously introduced for retrosynthesis in the ground-breaking work by Segler et al. [107], with a single retrosynthetic step Transformer architecture for predicting multi-step reactions. In a single-step setting, the model described by Lin et al. [169] achieved a top-1 prediction accuracy of over 43.1% and 54.1% when trained on the same small data set [173] and a ten times larger collection, respectively. Duan et al. [171] increased the batch size and the training time for their Transformer model and were able to achieve a top-1 accuracy of 54.1% on the 50k USPTO data set [173]. Later on, the same architecture was reported to have a top-1 accuracy of 43.8% [170], in line with the three previous transformer-based approaches [166, 167, 169] but significantly lower than the accuracy previously reported by Duan et al [171]. Interestingly, the transformer model was also trained on a proprietary data set [170], including only reactions with two reactants with a Tanimoto similarity distribution peaked at 0.75, characteristic of an excessive degree of similarity (roughly two times higher than the USPTO). Despite the high reported top-1 accuracy using the proprietary training and testing set, it is questionable how a model that overfits a particular ensemble of identical chemical transformations could be used in practice. Recently, a graph enhanced transformer model [174] and a mixture model [175] were proposed, achieving a top-1 accuracy of 44.9% and more diverse reactant suggestions, respectively, with no substantial improvements over previous works.

Except for the work of Lin et al. [169], all transformer-based retrosynthetic approaches were limited to a single step only. None of the previously reported works attempts the concurrent predictions of reagents, catalysts and solvent conditions but only reactants.

In this work, we present an extension of our Molecular Transformer architecture combined with a hyper-graph exploration strategy to design retrosynthetic pathways without human intervention. Compared to all other existing works using AI, we predict reactants as well as reagents for each retrosynthetic step, which significantly increases the difficulty of prediction[112]. Throughout the article, we will refer to reactants and reagents (e.g. solvents and catalysts) as precursors (see Figure 4.1). We criticise the use of the confidence level intrinsic to the retrosynthetic model (top-N accuracy) and introduce new metrics (coverage, class diversity, round-trip accuracy and Jensen-Shannon divergence) to evaluate the single-step retrosynthetic model, using the corresponding forward prediction and a reaction classification model. This provides a general assessment of each retrosynthetic step capturing the essential aspects a model should have to perform similarly to human experts in retrosynthetic analysis.

The optimal synthetic pathway is found through a beam search on the hyper-graph of the possible disconnection strategies. The hyper-graph is constructed on the fly, and the nodes are filtered and subject to further expansion based on a Bayesian-like probability that makes use of the forward prediction likelihood and the SCScore [176] to prioritise synthetic steps. This strategy allows circumventing potential selectivity traps, penalising non-selective reactions and precursors with

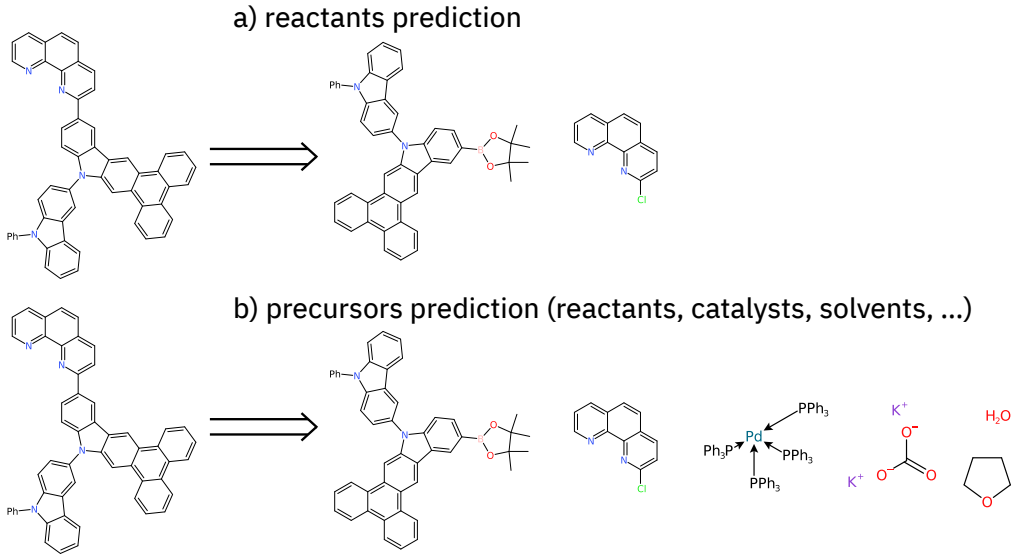

Figure 4.1: **Example precursor set suggestions.** Retrosynthesis step suggestion for 13-(1,10-phenanthrolin-2-yl)-10-(9-phenyl-9H-carbazol-3-yl)-10H-phenanthro[9,10-b]carbazole using a Chloro-Suzuki coupling reaction. In a) only the reactants are predicted. In b) all the precursors are predicted, which increases the overall difficulty of the single-step prediction task. While for a) two molecules consisting of a total of 68 atoms are predicted, the target of b) are six molecules consisting of 157 atoms.

higher complexity than the targets and leads to termination when commercially available building blocks are identified. We relate the quality of the retrosynthetic tree to the likelihood distributions of the forward prediction model and suggest the use of the Jensen-Shannon divergence to characterise the similarity of the distributions. This holistic analysis provides first the time a way to improve the quality of multi-step retrosynthetic tools systematically.

Finally, we critically assessed the entire AI framework by reviewing several retrosynthetic problems, some of them from literature data and others from academic exams. We show that reaching high performance on a subset of metrics for single-step retrosynthetic prediction is not beneficial in a multi-step framework. We also demonstrate that the use of all newly defined metrics provides an evaluation of end-to-end solutions, thereby focusing only on the quality of the single-step prediction model. The trained models and the entire architecture is freely available online [44]. The potential of the presented technology is high, augmenting the skills of less experienced chemists but also enabling chemists to design and protect the intellectual property of non-obvious synthetic routes for given targets.

4.2 Methods

4.2.1 Evaluation metrics for single-step retrosynthetic models

The evaluation of retrosynthetic routes is a task for human experts. Unfortunately, every evaluation is tedious and difficult to scale to a large number of examples. Therefore, it is challenging to generate statistically relevant results for more than a few different model settings. By using an analogy with human experts, we propose to use a forward prediction model [21, 107] and a reaction classification model to assess the quality [177] of the retrosynthetic predictions. The forward prediction model estimates the likelihood of the forward reaction of a single-step retrosynthesis and the classification model provides its corresponding class. Model scores have already been used as an alternative to human annotators to evaluate generative adversarial networks [178]. In our context, we define a retrosynthetic prediction as valid if the suggested set of precursors leads to the original product when processed by the forward chemical reaction prediction model (see Figure 4.2). More detail about the forward prediction and the reaction classification model can be found in the Supporting Information. Here we introduce four new metrics (round-trip accuracy, coverage, class diversity and the Jensen-Shannon divergence) to thoroughly evaluate retrosynthetic models.

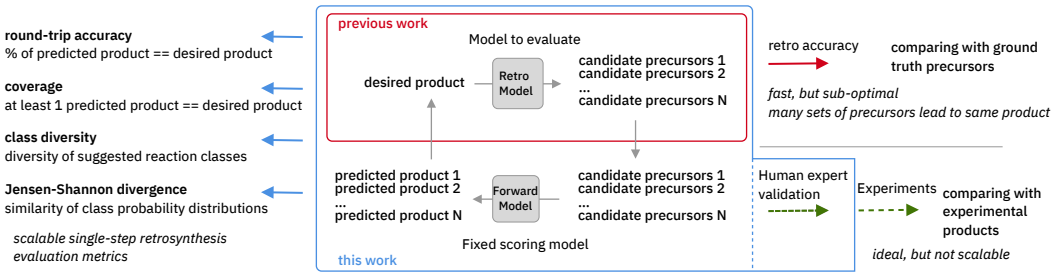

Figure 4.2: **Overview of single-step retrosynthesis evaluation metrics.**

The round-trip accuracy quantifies what percentage of the retrosynthetic suggestions is valid. This metric is an crucial evaluation as it is desirable to have as many valid suggestions as possible. This metric is highly dependent on the number of beams, as generating more outcomes through the use of a beam search might lead to a smaller percentage of valid suggestions due to lower quality suggestions in case of a higher number of beams.

The coverage quantifies the number of target molecules that produce at least one valid disconnection. With this metric, one wants to prevent rewarding models that produce many valid disconnections for only a few reactions, which would result in a small coverage. A retrosynthetic model should be able to produce valid suggestions for a wide variety of target molecules.

The class diversity is complementary to the coverage, as instead of relating to targets it counts the number of diverse reaction superclasses predicted by the retrosynthetic model, upon classification. A single-step retrosynthetic model should predict a wide diversity of disconnection strategies, which means generating precursors leading to the same product, with the corresponding reactions belonging to different reaction classes. Allowing a multitude of different disconnection

strategies is beneficial for an optimal route search and essential, precisely when the target molecule contains multiple functional groups.

Finally, the Jensen-Shannon divergence, which is used to compare the likelihood distributions of the suggested reactions belonging to different classes above a threshold of 0.5, is calculated as follows:

$$JSD(P_0, P_1, ..., P_{11}) = H\left(\sum_{i=0}^{11} \frac{1}{12} P_i\right) - \frac{1}{12} \sum_{i=0}^{11} H(P_i), \tag{4.1}$$

where P_i denote the probability distributions and $H(P)$ the Shannon entropy for the distribution P.

To calculate the Jensen-Shannon divergence, we split the single-step retrosynthetic reactions into superclasses and use the likelihoods predicted by the forward model to build a likelihood distribution within each class. This metric is crucial to assess the quality of a sequence of retrosynthetic steps. Having a model with a dissimilar likelihood distribution would be equivalent to having a human expert favour a few specific reaction classes over others. This would result in an introduction of bias favouring those classes with dominant likelihood distributions. While it is desirable to have a peaked distribution, as this is an evident sign of the model learning from the data, it is also desirable to have all the likelihood distributions equally peaked, with none of them exercising more influence than the others during the construction of a large number of retrosynthetic trees. The inverse of the Jensen-Shannon divergence ($1/JSD$) is a measure of the similarity of the likelihood distributions among the different superclasses and we use this parameter as an effective metric to guarantee uniform likelihood distributions among all possible predicted reaction classes. Uneven distributions are directly connected to the nature of the training data set. All these four metrics have been critically designed and assessed with the help of human domain experts. Their combined use paves the way for a systematic improvement of entire retrosynthetic frameworks, by adequately tuning data sets that optimise the different single-step performance indicators in a multi-objective fashion.

Additionally, we use the open-source chemoinformatics software RDKit [153] to evaluate the percentage of syntactically valid predicted molecules (grammatically correct SMILES).

4.2.2 Hyper-graph exploration

A retrosynthetic tree is equivalent to a directed acyclic hyper-graph, a mathematical object composed of hyper-arcs (A) that link nodes (N). The main difference compared to a typical graph is that a hyper-arc can link multiple nodes, similar to what happens in a retrosynthesis: if a node represents a target molecule, the hyper-arcs connecting to different nodes represent all possible reactions involving those corresponding molecules. Hyper-arcs have an intrinsic direction defining whether the reaction is forward or retro (see Figure 4.3).

A retrosynthetic route needs to be free of any loops, i.e. acyclic. This requirement renders the retrosynthetic route a hyper-tree [179], in which the root is the target molecule and the leaves are the commercially available starting materials (see Figure 4.4).

In cases where the hyper-graph of the entire chemical space is available, an exhaustive search may reveal all the possible synthetic pathways leading to a target molecule from defined starting

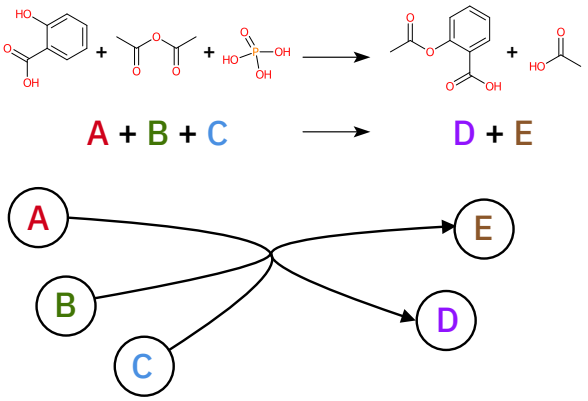

Figure 4.3: **Reaction hyper-graph.** A generic reaction (top of the picture) can be represented as a hyper-graph. Each molecule involved in the reaction becomes a node in the hyper-graph while the hyper-arc, connecting the reactants and reagents to the product, represents the reaction arrow.

materials. Instead, here we build the hyper-tree on the fly: only the nodes and arcs expanding in the direction of the most meaningful retrosynthesis are calculated and added to the existing tree. The retrosynthesis exploration uses a SCScore [176]-based Bayesian-like probability to decide the direction along which the graph is expanded, driving the tree towards more simple precursors. In Figure 4.5, we show a schematic representation of the multi-step retrosynthetic workflow. Given a target molecule, we use a single-step retrosynthetic model to generate a certain number of possible disconnections (i.e. precursors set). We canonicalise the predicted reaction smiles and determine their reaction class. We compute the SCScore as well as the reaction likelihood with the forward prediction model on the corresponding inchified entry. In order to discourage the use of non-selective reactions, we filter the single-step retrosynthetic predictions by using a threshold on the reaction likelihood returned by the forward model. The likelihood and SCScore of the filtered predictions are combined to compute a probability score to rank all the options. In case all the predicted precursors are commercially available the retrosynthetic analysis provides that option as a possible solution and the exploration of that tree branch is considered complete. If not, we repeat the entire cycle using the precursors as initial target molecules until we reach either commercially available molecules or the maximum number of specified retrosynthesis steps. The single-step forward and retrosynthetic predictive models, as well as the multi-step framework, do not contain explicitly encoded chemical knowledge: the only chemical knowledge embedded is the one learned from the data during the training processes. The algorithmic details and the path scoring function are detailed in the supplementary information.

4.3 Results

4.3.1 Single-step retrosynthesis

The Top-N accuracy score is the preferred method to evaluate the quality of single-step predictive models. While this is entirely justified for the evaluation of forward reaction prediction, its us-

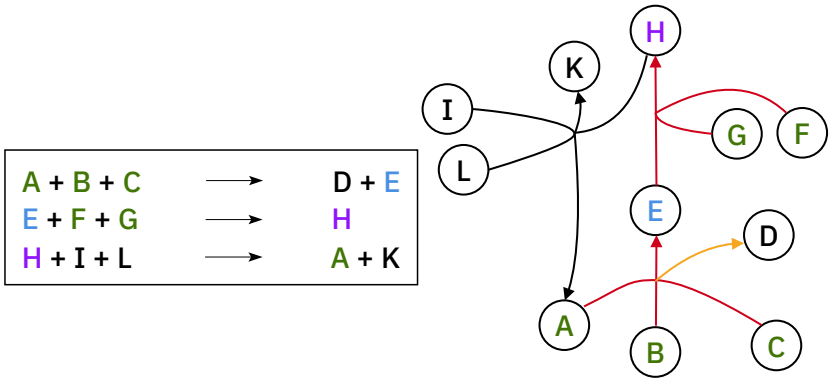

Figure 4.4: **Example of hyper-graph complexity.** The Molecule H is the target (purple label). The red lines represent the synthetic path from commercially available precursors (highlighted in green) to the target molecule. The yellow line, does not affect the retrosynthesis of H, neither does the last reaction with black lines.

age in the context of single-step retrosynthetic models is misleading, as recently suggested also by Thakkar et al. [109]. Top-N accuracy means that the ground truth precursors were found within the first N suggestions of the retrosynthetic model. In contrast to forward prediction models, a target molecule rarely originates from one set of precursors only. Often the presence of different functional groups allows a multitude of possible disconnection strategies to exist, leading to different sets of reactants, as well as possible solvents and catalysts.

The analysis of the USPTO stereo data set, derived from the text-mined open-source reaction data set by Lowe [41, 42], and of the Pistachio data set [180], shows that 6% of the products, and 14% respectively, have at least two different sets of precursors. While these numbers only reflect the organic chemistry represented in each data set, the total number of possible disconnections is undoubtedly larger. Considering the limited size of existing data sets, it is evident that, in the context of retrosynthesis, the top-N accuracy rewards the ability of a model to retrieve expected answers from a data set more than that to predict chemically meaningful precursors. Therefore, a top-N comparison with the ground truth is not an adequate metric for assessing retrosynthetic models.

Here, we dispute the previous use of top-N accuracy in single-step retrosynthetic models [65, 107, 108, 162, 166, 167, 168, 169, 170, 171] and propose four new different metrics (round-trip accuracy, coverage, class diversity and Jensen-Shannon divergence [181], see Section 4.2.1) for their evaluation.

During the development phase, we trained different retrosynthetic transformer-based models with two different data sets, one fully based on open-source data (*stereo*) and one on based commercially available data from Pistachio (*pistachio*). In some cases, the data set was inchified [52] (labelled with_*i*). Table 4.1 shows the results for the retrosynthetic models, evaluated using a fixed forward prediction model (*pistachio_i*) on two validation sets (*stereo* and *pistachio*). The coverage represents the percentage of desired products for which at least one valid precursor set was suggested. It was slightly better for *stereo* but above 90% for all the model combinations, which is an important requirement to guarantee the possibility to always offer at least one disconnection

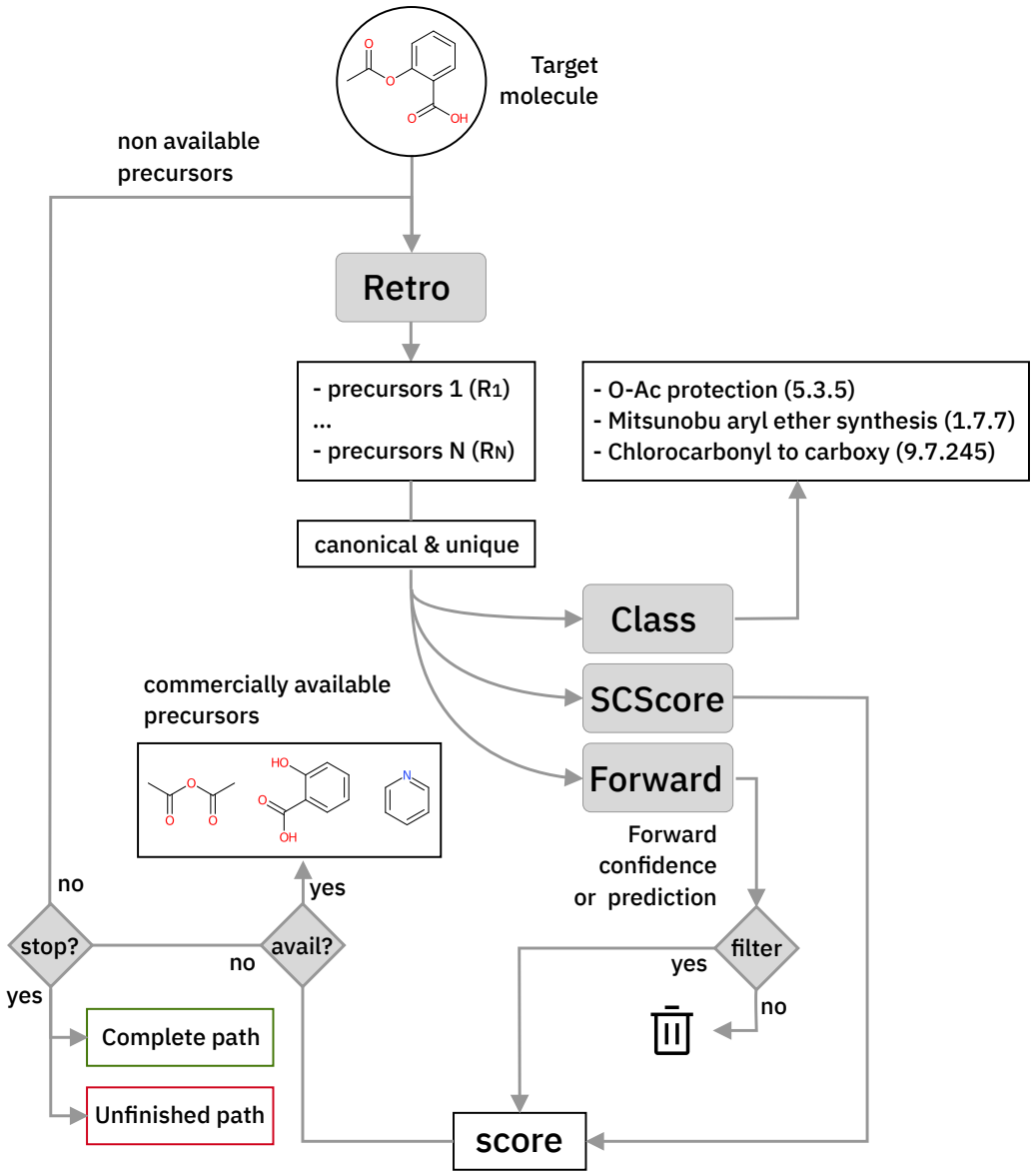

Figure 4.5: **Schematic of the multi-step retrosynthetic workflow.**

strategy. Likewise, the class diversity, which is an average of how many different reaction classes are predicted in a single retrosynthetic step, was comparable for both models with slightly better performance for the *pistachio* model. The round-trip accuracy, which is the percentage of precursor sets leading to the initial target when evaluated with the forward model, was better for *stereo* than for *pistachio*. Despite the *stereo* retrosynthetic model performed better than the *pistachio* model in terms of round-trip accuracy and coverage, the synthesis routes generated with this model were

Table 4.1: **Evaluation of single-step retrosynthetic models.** The test data set consisted of 10K entries. For every reaction we generated 10 predictions. The number of resulting precursor suggestions was 100K. Round-trip accuracy (RT), coverage (Cov.), class diversity (CD), the inverse of the Jensen Shannon divergence of the class likelihood distributions (1/JSD), the percentage of invalid SMILES (ismi) and the human expert evaluation (hu. ev.) are reported in the table. Models with the "_i" suffix were trained on an inchified data set. Models starting with "ste" were trained with the *stereo* data set and the ones with "pist" with the *pistachio* data set.

Model retro	forw.	Test data	RT [%]	Cov. [%]	CD	$\frac{1}{JSD}$	ismi [%]	hu. ev.
ste_i	pist_i	ste	81.2	95.1	1.8	16.5	0.5	-
ste_i	pist_i	pist	79.1	93.8	1.8	20.6	1.1	-
pist_i	pist_i	pist	74.9	95.3	2.1	22.0	0.5	+
pist	pist_i	pist	71.1	92.6	2.1	27.2	0.6	++

of lower quality and often characterised by a sequence of illogical protection/deprotection steps as determined by the human expert assessment (last column in Table 4.1). This apparent paradox became clear when we analysed in detail how humans approach the problem of retrosynthesis.

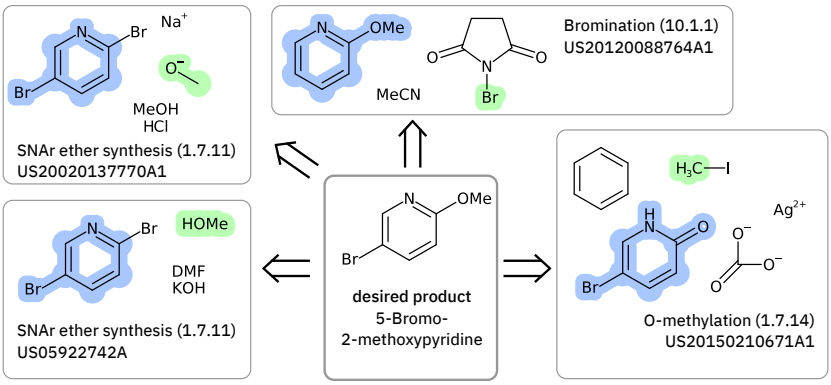

Figure 4.6: **Accuracy metric problem.** Highlighting a few of the precursors and reactions leading to 5-Bromo-2-methoxypyridine that are found in the US Patents data set. The molecules were depicted with CDK [182].

Solving retrosynthetic problems requires a careful analysis of which ones among multiple precursors could lead to the desired product more efficiently, as seen in Figure 4.6 for 5-Bromo-2-methoxypyridine. Humans address this issue by mentally listing and analysing all possible disconnection sites and retaining only the options, for which the corresponding precursors are thought to produce the target molecule most selectively.

For an expert, it is not sufficient to always find at least one disconnection site (coverage) and be sure that the corresponding precursors will selectively lead to the original target (round-trip accuracy). It is necessary to generate a diverse sample of disconnection strategies to cope with competitive functional group reactivity (class diversity). Moreover, most important, every disconnection class needs to have a similar probability distribution to all the other classes (Jensen-Shannon di-

vergence, JSD). Continuing the parallelism with human experts, if one was exposed to the same reaction classes for many years, the use of those familiar schemes in the route planning would appear more frequently, leading to strongly biased retrosynthesis. Therefore, it is essential to reduce any bias in single-step retrosynthetic models to a minimum.

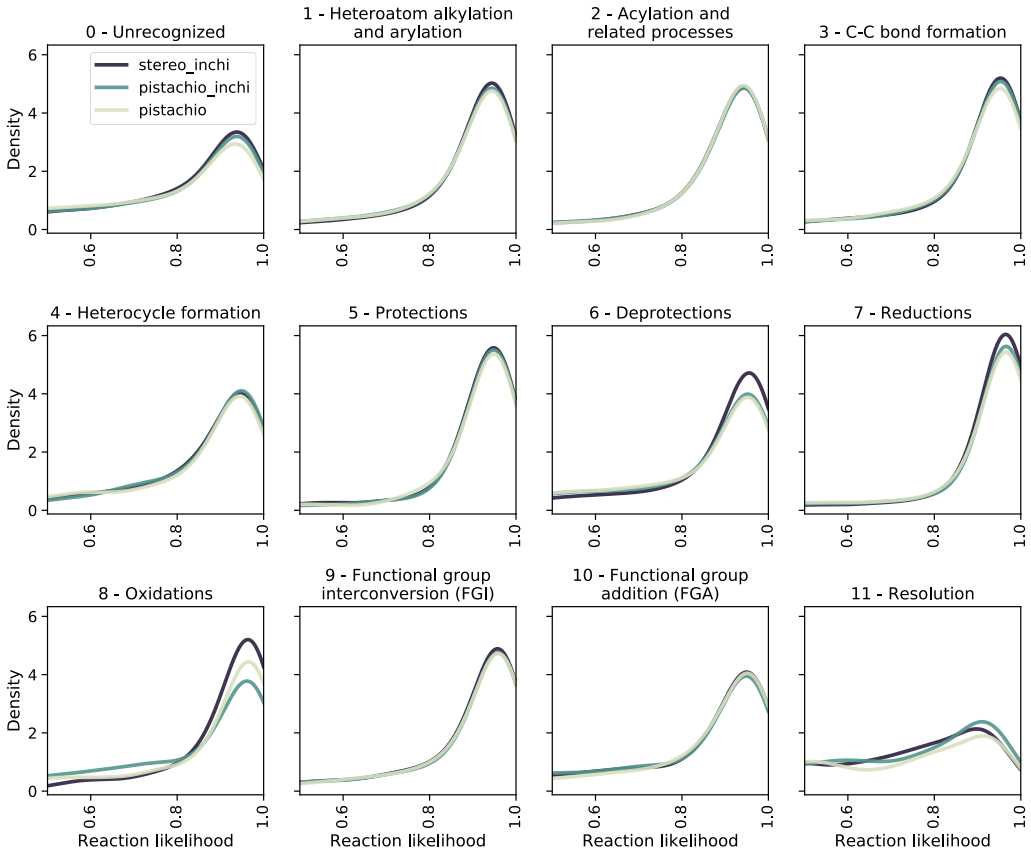

Figure 4.7: **Reaction likelihood distributions.** The likelihood distributions predicted by a forward model (*pistachio_i*) for the reactions suggested by different retro models. We show the likelihood range between 0.5 and 1.0.

To evaluate the bias of single-step models, we use the JSD of the likelihood distributions for the prediction divided in different reaction superclasses, which we report in Table 4.1 as 1/JSD. The larger this number, the more similar the likelihood distributions of the reactions belonging to different classes are and hence, the less dominant (lower bias) individual reaction classes are in the multi-step synthesis. In Figure 4.7, we show the likelihood distributions for the different models in Table 4.1. Except for the resolution class, all of the distributions show a peak close to 1.0, which clearly shows that the model learned how to predict the reaction in those classes. The resolution class is instead relatively flat as a consequence of the poor data quality/quantity for stereochemical reactions both in the *stereo* and *pistachio* data set. Interestingly, one can see that for the *stereo* model the likelihood distributions of the deprotection, reduction and oxidation reactions are dif-

ferent (and generally more peaked) from all other distributions generated with the same model. This statistical imbalance favours those reaction classes and explains the occurrence of illogical loops of protection/deprotection or oxidation/reduction strategies. While peaked distributions are desirable, as this is a consequence of the model learning to predict disconnection strategies in a precise class, the dissimilarity (JSD) between the twelve probability distributions reflects an intrinsic bias, likely due to unbalanced data sets. Among the few models reported, the *pistachio* model was found to have the best similarity (1/JSD) score and is the one analysed in the subsequent part of the manuscript and made available online.

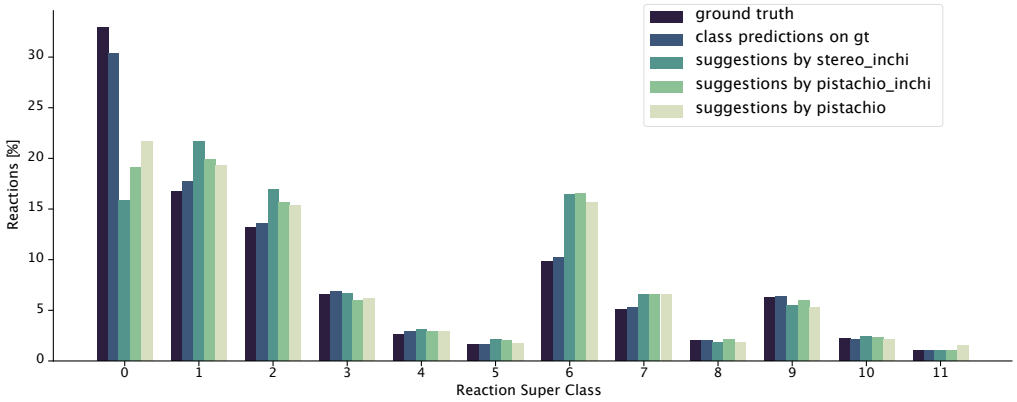

Figure 4.8: **Distribution of reaction superclasses for the ground truth [63].** the predicted superclasses for the ground truth reactions and the predicted superclasses for the reactions suggested by the different retrosynthesis models.

The class diversity and similarity scores require the identification of the reaction class for each prediction. We used a transformer-based reaction classification model, as described in [177]. In Figure 4.8, we report the ground truth classified by the NameRXN [63] tool, the class distribution predicted by our classification model on the ground truth reactions and finally, the class distributions predicted for the reactions suggested by the retrosynthesis models (see Table 4.1). We observe that the classifications made by our class prediction model are in agreement with the ones of NameRXN [63] and match them with an accuracy of 93.8%. The distributions of the single-step retrosynthetic models resemble the original one with the number of unrecognised reactions nearly halved. All of the models learned to predict more recognisable reactions, even for products, for which there was an unrecognised reaction in the ground truth.

4.3.2 A holistic evaluation of the pathway prediction

An evaluation of the model was carried out through performing the retrosynthesis of the compounds reported in Figure 4.9. Some of these are known compounds, for which the synthesis is reported in the literature (1, 2, 5, 7, 8), others are unknown structures (3, 4, 6, 9). For the first group, the evaluation of the model could be made by comparing the proposed retrosynthetic analysis with the known synthetic pathway. For the second group, a critical evaluation of the proposed

retrosynthesis, which takes into account the level of chemo-, regio-, and stereoselectivity for every retrosynthetic step was performed. The parameters used for each retrosynthesis are reported in the supplementary information. In some cases, the default values were changed to increase the hyper-graph exploration and yield better results. As an output, the model generates several retrosynthetic sequences for each compound, each one with a different confidence level. Because the model predicts not only reactants but also reagents, solvents and catalysts, there are several sequences with similar confidence level and identical disconnection strategies and differing only by the suggested reaction solvents in a few steps. Therefore, we report only one of the similar sequences in the supplementary information.

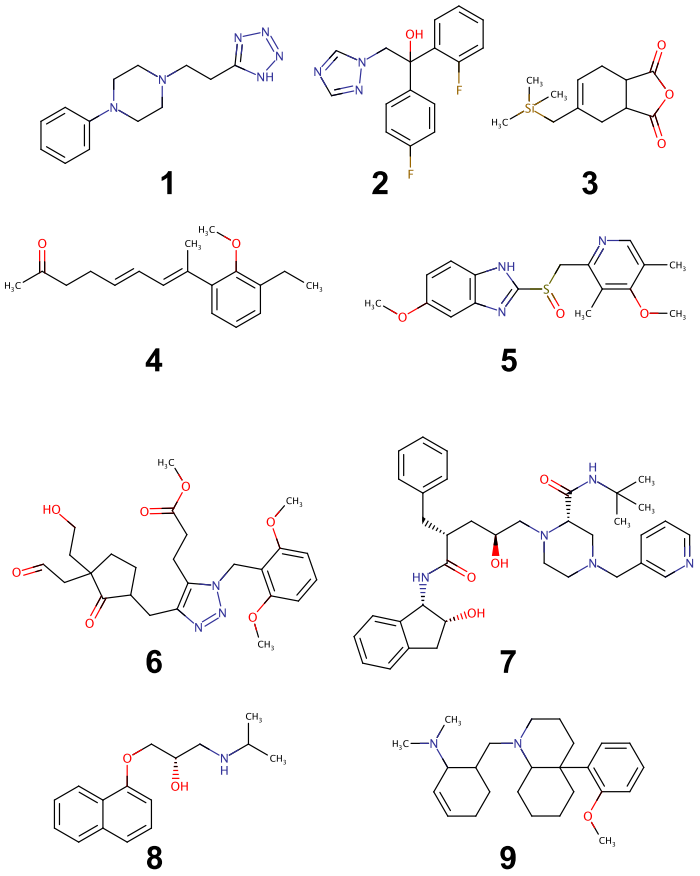

Figure 4.9: **Set of molecules used to assess the quality of retrosynthesis.**

All of the retrosynthetic routes generated for compounds 1, 2 and 3 fulfill the criteria of chemoselectivity. The highest confidence sequence (called "sequence 0") of 1 corresponds to the reported synthesis of the product [183] and starts from the commercially available acrylonitrile. The other two sequences (17 and 22) use synthetic equivalents of acrylonitrile and also show its preparation. For compound 2, the highest confidence retrosynthetic sequence (sequence 0) does not correspond to the synthetic pathway reported in the literature, where the key step is the opening of an epoxide ring. Two other sequences (5 and 23) report this step, and one of them (sequence 5) cor-

responds to the literature synthesis [184]. The retrosynthetic sequence for compound 3 provides a Diels-Alder reaction as the first disconnection strategy and proposes a correct retrosynthetic path for the synthesis of the diene from available precursors. A straightforward retrosynthetic sequence was also found in the case of compound 4, where the diene moiety was disconnected by two olefination reactions and the sequence uses structurally simple compounds as starting material. It may be debatable whether the two olefinations through a Horner-Wadsworth-Emmons reaction, can really be stereoselective towards the E-configurated alkenes or whether the reduction of the conjugate aldehyde by NaBH4 can be completely chemoselective towards the formation of the allylic alcohol. Only experimental work can solve this puzzle and give the correct answer.

The retrosynthesis of racemic omeoprazole 5 returned a sequence consisting of one step only because the model finds in its library of available compounds the sulfide precursor of the final sulfoxide. When repeating the retrosynthesis using benzene as starting molecule in conjunction with a restricted set of available compounds, we obtained a more complete retrosynthetic sequence with some steps in common with the reported one [185]. However, although all of the steps fulfill the chemoselectivity requirement, the sequence is characterised by some avoidable protection-deprotection steps. This sequence nicely reflects the bias present in the likelihood distributions of the different superclasses for the chosen model. Although the single-step retrosynthetic model has the best Jensen-Shannon divergence among all of the trained models, there is still room for improvements that we will explore in the future. A higher similarity across the likelihood distributions will prevent the occurrence of illogical protection-deprotection, estherification/saponification steps.

Besides, the reported sequence for 5 lists a compound not present in the restricted set of available molecules as starting material. A “de novo” retrosynthesis of this compound solved the problem. The retrosynthetic sequence of the structurally complex compound 6 was possible only with wider settings allowing a more extensive hyper-graph exploration. The result was a retrosynthetic route starting from simple precursors: notably, the sequence also showed the synthesis of the triazole ring through a Huisgen cycloaddition. However, we recognised the occurrence of some chemoselectivity problems in step 6, when the enolate of the ketone is generated in the presence of an acetate group, used as protection of the alcohol. This problem could be avoided by using a different protecting group for the alcohol. By contrast, the alkylation of the ketone enolate by means of a benzyl bromide bearing an enolisable ester group in the structure appears less problematic, due to the high reactivity of the bromide. The retrosynthesis of the chiral stereodefined compound indinavir, 7, completed in one step, through finding a very complex precursor in the set of available molecules. Sequences of lower confidence resulted in more retrosynthetic steps, disconnecting the molecule as in the reported synthesis [186] but stopped at the stereodefined epoxide, with no further disconnection paths available. However, when the retrosynthesis was performed on the same racemic molecule, a chemoselective retrosynthetic pathway was found, disconnecting the epoxide and starting from simple precursors. Similarly, for the other optically active compound, propranolol, 8, which was disconnected according to the published synthetic pathway [187] only when the retrosynthesis was performed on the racemic compound. The problem experienced with stereodefined molecules reflects the poor likelihood distribution of the resolution superclass in Figure 4.7. Because all current USPTO derived data sets (*stereo* and *pistachio*) have particularly noisy stereochemical data we decided to retain only few entries in order to avoid jeopardising the overall quality. With a limited number of stereochemical examples available in

the training set, the model was not able to learn reactions belonging to the resolution class, failing to provide disconnection options for stereodefined centres.

The retrosynthesis of the last molecule, 9, succeeded only with intensive hyper-graph exploration settings. However, the retrosynthetic sequence is tediously long, with several avoidable esterification-saponification steps. Similar to 5, the bias in the likelihood distributions is the one reason for this peculiar behavior. In addition, a non-symmetric allyl bromide was chosen as precursor of the corresponding tertiary amine: this choice entails a regioselectivity problem, given that the allyl bromide can undergo nucleophilic displacement not only at the ipso position, giving rise to the correct product, but also at the allylic position, resulting in the formation of the regioisomeric amine. Lastly, the model was unable to find a retrosynthetic path for one complex building block, which was not found in the available molecule set. However, a slight modification of the structure of this intermediate enabled a correct retrosynthetic path to be found, which could also be easily applied to the original problem, starting from 1,3-cycloexanedione instead of cyclohexanone. We also made a comparison of our retrosynthetic architecture with previous work [107, 162], using the same compounds for the assessments (see SI). The model performed well on the majority of these compounds, showing problems in the case of stereodefined compounds as in the previous examples. Retrosynthetic paths were easily obtained only for their racemic structure. The proposed retrosyntheses in some cases are similar to those reported [162] while, for some compounds [107] they are different but still chemoselective. Only in a few cases, the model failed to find a retrosynthesis.

4.4 Discussion

In this work, we presented an extension of our Molecular Transformer architecture combined with a hyper-graph exploration strategy to design retrosynthesis without human intervention. We introduce a single-step retrosynthetic model predicting reactants as well as reagents for the first time. We also introduce four new metrics (coverage, class diversity, round-trip accuracy and Jensen-Shannon divergence) to provide a thorough evaluation of the single-step retrosynthetic model. The optimal synthetic pathway is found through a beam search on the hyper-graph of the possible disconnection strategies and allows to circumvent potential selectivity traps. The hyper-graph is constructed on the fly, and the nodes are filtered, and further expanded based on a Bayesian-like probability score until commercially available building blocks are identified. We assessed the entire framework by reviewing several retrosynthetic problems to highlight strengths and weaknesses. As confirmed by the statistical analysis, the entire framework performs very well for a broad class of disconnections. An intrinsic bias towards a few classes (reduction / oxidation / estherification / saponification) may lead, in some cases, to illogical disconnection strategies that are a peculiar fingerprint of the current learning process. Also, an insufficient ability to handle stereochemical reactions is the result of the poor quality training data set that covers only a few examples in the resolution class. The use of the four new metrics, combined with the critical analysis of the current model, provides a well defined strategy to optimise the retrosynthetic framework by focusing exclusively on the performance of the single-step retrosynthetic model without the need to manually review the quality of entire retrosynthetic routes. A key role in this strategy

will be the construction of statistically relevant training data sets to improve the confidence of the model in different types of reaction classes and disconnections.

5 Mapping the space of chemical reactions using attention-based neural networks

Organic reactions are usually assigned to classes containing reactions with similar reagents and mechanisms. Reaction classes facilitate the communication of complex concepts and efficient navigation through chemical reaction space. However, the classification process is a tedious task. It requires the identification of the corresponding reaction class template via annotation of the number of molecules in the reactions, the reaction center, and the distinction between reactants and reagents. This work shows that transformer-based models can infer reaction classes from non-annotated, simple text-based representations of chemical reactions. Our best model reaches a classification accuracy of 98.2%. We also show that the learned representations can be used as reaction fingerprints that capture fine-grained differences between reaction classes better than traditional reaction fingerprints. The insights into chemical reaction space enabled by our learned fingerprints are illustrated by an interactive reaction atlas providing visual clustering and similarity searching.

This chapter has been published as a scientific article in Nature Machine Intelligence:

P Schwaller, D Probst, AC Vaucher, VH Nair, D Kreutter, T Laino, JL Reymond. Mapping the Space of Chemical Reactions using Attention-Based Neural Networks. *Nat. Mach. Intell.*, 2021, 3, 144–152.

5.1 Introduction

In the last decade, computer-based systems [44, 162, 164] have become an important asset available to chemists. Deep learning methods stand out, not only for reaction prediction tasks [4, 35, 36], but also for synthesis route planning [43, 107, 109] and synthesis procedures to action conversions [188].

Among the few approaches, natural language processing (NLP) methods [2, 3] applied to Simplified molecular-input line-entry system (SMILES) [46, 47] and other text-based representation of molecules and reactions are particularly effective in the chemical domain. Recently, Schwaller et al. [189] demonstrated that neural networks are able to capture the atom rearrangements from precursors to products in chemical reactions without supervision. Figure 5.1 a) shows examples of chemical reactions and the corresponding textual representation in b).

The demand for robust algorithms to categorise chemical reactions is high. The knowledge of the class of a reaction has a great value for expert chemists, for example to assess the quality of the

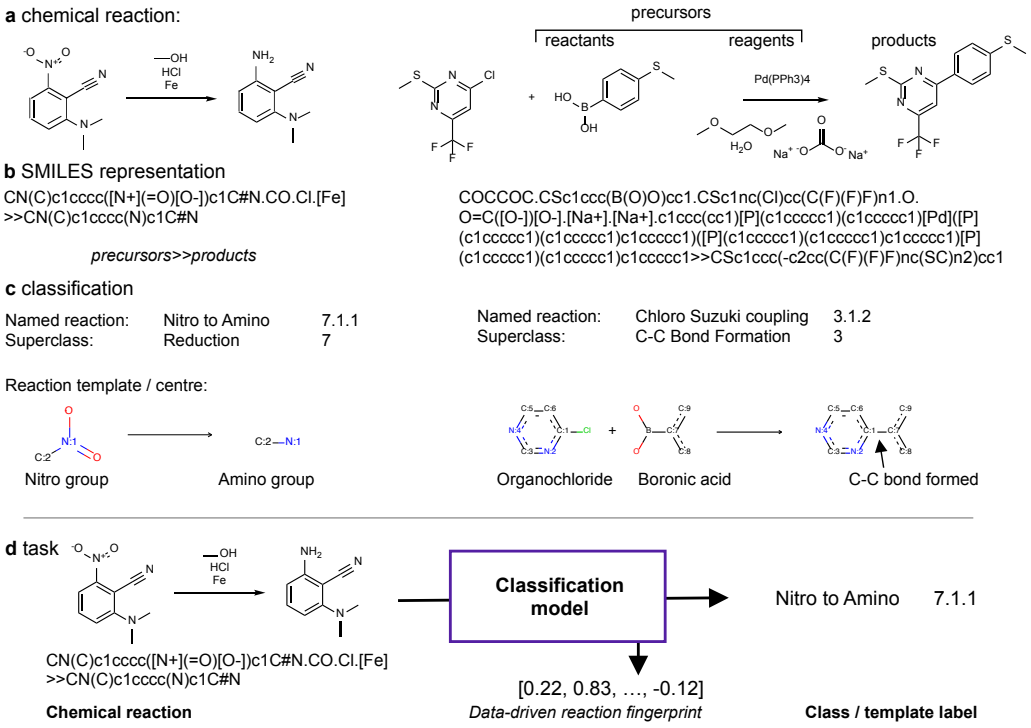

Figure 5.1: **Data representation and task.** Two examples of chemical reactions with associated classification labels and reaction templates describing the transformation. The task is to predict the reaction class or template label from the chemical reaction. The encoded representation of the reaction can be used as data-driven reaction fingerprint.

reaction prediction [190]. Chemists use reaction classes to navigate large databases of reactions and retrieve similar members of the same class to analyse and infer optimal reaction conditions. They also use reaction classes as an efficient way to communicate what a chemical reaction does and how it works in terms of atomic rearrangements. As seen in Figure 5.1 c), reaction classes can be named after the reaction type referring to the changing structural features, such as "Nitro to Amino". Alternatively, they can be named after the persons who discovered the chemical reaction or refined an already known transformation, like the second example in Figure 5.1 c). It is a chloro Suzuki coupling reaction named after Akira Suzuki, who received the Nobel prize in 2010 for his work on palladium-catalysed cross-coupling reactions [191]. The current state-of-the-art in reaction classification is are commercially available tools [63, 192], which classify reactions based on a library of expert-written rules. These tools typically make use of SMIRKS [193], a language for describing transformations in the SMILES format [46, 47]. On the contrary, classifiers based on machine learning have the potential to increase the robustness to noise in the reaction equations and to avoid the need for an the explicit formulation of rules.

Early work in the 90s used self-organising neural networks to map organic reactions and investigate similarities between them [194, 195, 196]. More recently, Schneider et al.[64] developed a reaction classifier based on traditional reaction fingerprints. Molecular and reaction fingerprints

are fixed-size vector encodings of discrete molecular structures and chemical reactions. The currently best performing fingerprint by Schneider et al.[64] combines a products-reactants difference fingerprint with molecular features calculated on the reagents and was tested on a limited set of 50 reaction classes. This difference fingerprint is currently one of the most frequently used hand-crafted ones. It has been successfully applied to reaction conditions predictions [197], where the reagents were not taken into account for the reaction description. Ghiandoni et al. [198] introduced an alternative hierarchical classification scheme and random forest classifier for reaction classification. Their algorithm outputs a confidence score through conformal prediction. The fingerprints developed by Schneider et al. [64] and Ghiandoni et al. [198] both require a reactants-reagents role separation [62], which is often ambiguous and thus limits their applicability.

Traditionally, reaction fingerprints were hand-crafted using the reaction centre or a combination of the reactant, reagent and product fingerprints. ChemAxon [199], for instance, provides eight types of such reaction fingerprints. Based on the differentiable molecular fingerprint by Duvenaud et al. [73], the first example of a learned reaction fingerprint was presented by Wei et al. [31] and used to predict chemical reactions. Unfortunately, their fingerprint was restricted to a fixed reaction scheme consisting of two reactants and one reagent, and hence, only working for reactions conform with that scheme. Similarly, the multiple fingerprint features by Sandfort et al. [200] are made by concatenating multiple fingerprints for a fixed number of molecules.

In the first part of our work, we predict chemical reaction classes using attention-based neural networks from the family of transformers [2, 3]. Our deep learning models do not rely on the formulation of specific rules that require every reaction to be properly atom-mapped. Instead, they learn the atomic motifs that differentiate reactions from differentclasses from raw reaction SMILES without reactant-reagent role annotations (Figure 1d). The transformer-based sequence-2-sequence (seq-2-seq) model [2] matched the ground-truth classification with an accuracy of 95.2% and the Bidirectional Encoder Representations from Transformers (BERT) classifier[3] with 98.2%. We analyse the encoder-decoder attention of the seq-2-seq model and the self-attention of the BERT model. Hereby we observe that atoms involved in the reaction centre, as well as reagents specific to the reaction class, have larger attention weights.

In the second part, we demonstrate that the representations learned by the BERT models, unsupervised and supervised, can be used as reaction fingerprints. The reaction fingerprints we introduce are independent of the number of molecules involved in a reaction. The BERT models trained on chemical reactions can convert any reaction SMILES into a vector without requiring atom-mapping or a reactant-reagent separation. Therefore our reaction fingerprints are universally applicable to any reaction database. Based on those reaction fingerprints and TMAP [201], a method to visualise high-dimensional spaces as tree-like graphs, we were able to map the chemical reaction space and show in our reaction atlases nearly perfect clustering according to the reaction classes. Moreover, our fingerprints enable chemists to efficiently search chemical reaction space and retrieve metadata of similar reactions. The metadata could, for instance, contain typical conditions, synthesis procedures, and reaction yields.

On an imbalanced data set, our fingerprints and classifiers reach an overall classification accuracy of more than 98%, compared to 41 % when using a traditional reaction fingerprint. The ability to accurately classify chemical reactions and represent them as fingerprints, enhances the accessibility of reaction by machines and humans alike. Hence, our work has the potential to

unlock new insights in the field of organic synthesis. In recent studies, our models were used to predict experimentally measured activation energies [202] and reaction yields [203].

5.2 Results

5.2.1 Reaction classification

Classification results

We used a labeled set of chemical reactions as ground truth to train two transformer-based deep learning models as architecture [2, 3]. The first one is an encoder-decoder transformer as introduced by Vaswani et al. [2] for sequence-to-sequence (seq-2-seq) tasks in neural machine translation. The second one is an encoder-only transformer called BERT introduced by Devlin et al. [3]. The latter model with a classification head on top is typically used in NLP for single sentence classification tasks [204, 205]. A visualisation of such a BERT classifier is shown in Figure 5.5.

The ground truth data is composed of chemical transformations represented in text format as SMILES. Their labeling (classification) was taken from the strongly imbalanced Pistachio data set [180], which uses NameRXN for the reaction classification [63]. In an additional experiment, we use reaction template labels derived from open-source data, which we will refer to as USPTO 1k TPL. We analysed the classification performance of our models on the test sets, which contained 132k reactions from 792 different classes in Pistachio, and 45k reactions from 1000 template classes in USPTO 1k TPL. A summary of the results can be found in Table 5.1. On the Pistachio test set, the transformer encoder-decoder model (enc2-dec1) matched the ground truth classification with an accuracy of 95.2%. The reaction BERT classifier predicted the correct name reaction with an accuracy of 98.2%, therefore achieving significantly better results than with the seq-2-seq approach. As a comparison to previous work [64], we computed the transformation fingerprint AP3 (folded) + featureFP on the Pistachio data and used a 5-NearestNeighbour (5-NN) classifier [206] to classify the test set reactions. Even though we separated the reactants and reagents using RDKit [153], the classifier only achieved an overall accuracy of 41.0%. The traditional fingerprint was not able to represent the fine-grained differences between the reaction classes. The "Unrecognised", "Carboxylic acid + amine condensation", "Amide Schotten-Baumann" and "N-Boc deprotection" classes contained the most false positives.

In contrast, our BERT classifier without reactant-reagent separation was the best performing model, when looking at the confusion entropy of a confusion matrix (CEN) [207] and overall Matthews correlation coefficient (MCC) [208, 209].

To show that the inferior performance of the traditional reaction fingerprint did not stem from the choice of the 5-NN classifier, we took the embeddings of the pretrained (*rxnfp (pretrained)*) and finetuned BERT (*rxnfp*) as inputs for the 5-NN classifier. We then classified the test set reactions and computed the scores. As expected, the results for *rxnfp*, which corresponds to the input of the classifier layer in the BERT classifier, perfectly matched the scores of the BERT classifier.

The mismatches in the Pistachio test set are mainly related to "Unrecognised" reactions. When analysing the individual errors, we observed that our models were able to predict the correct reaction class for reactions that had a slight change in the representation between precursors and product (e.g. different tautomers). Such examples were not matched by the brittle rules that gen-

Table 5.1: **Classification results.** The lower the confusion entropy of a confusion matrix (CEN) and the higher the Matthews correlation (MCC) coefficient the better. The traditional fingerprint is an AP3 256 (folded) + agents features developed by Schneider et al. [64].

Pistachio	Accuracy	CEN	MCC
Traditional fp[64] + 5-NN classifier	0.410	0.365	0.305
Transformer enc2-dec1	0.952	0.039	0.946
BERT classifier	**0.982**	**0.014**	**0.980**
rxnfp (pretrained) + 5-NN classifier	0.819	0.121	0.797
rxnfp + 5-NN classifier	**0.989**	**0.010**	**0.988**
USPTO 1k TPL	Accuracy	CEN	MCC
Traditional fp[64] + 5-NN classifier	0.295	0.424	0.292
BERT classifier	**0.989**	**0.006**	**0.989**
rxnfp (pretrained) + 5-NN classifier	0.340	0.392	0.337
rxnfp + 5-NN classifier	**0.989**	**0.006**	**0.989**

erated the ground-truth classes. Hence, they were labeled as "Unrecognised" reactions. Our models show very high robustness against errors in the SMILES representation. In the supplementary information, we report cases where, despite an error in the molecular representation, our model was able to correctly classify the reaction that was originally described by chemists in the patent procedure text.

On the USPTO 1k TPL test set, the traditional and pretrained fingerprint performed worse than on the Pistachio data set. However, the BERT classifier as well as the embeddings of the BERT classifier with the 5-NN classifier matched the performance they had on the Pistachio data set with an accuracy of 98.9%.

An elaborate description of both types of reaction fingerprints is presented in the section on data-driven reaction fingerprints below. A comparison of our data-driven approach to traditional fingerprints on a balanced data set of 50k reactions can be found in the supplementary information. Even when using as little as 10k training reactions from 50 different classes the fine-tuned embeddings are able to outperform traditional fingerprints by increasing precision, recall and F1-score from 0.97 to 0.99.

Visualisation of attention weights

Figure 5.2 shows the layer-wise [CLS] token attention of the BERT classifier (above the reaction) and the encoder-decoder attention of the seq-2-seq model (below the reaction) for two different chemical transformations. We observed that the larger weights were associated with the atoms that are part of the reaction centre or precursors specific to the reaction class. Just like a human expects to see a certain group of atoms based on the classification, for the seq-2-seq model, the decoder learned to focus on the atoms involved in the rearrangement to classify reactions. For the BERT classifier, the initial layers had weak attention on all reaction tokens. The middle layers tended to

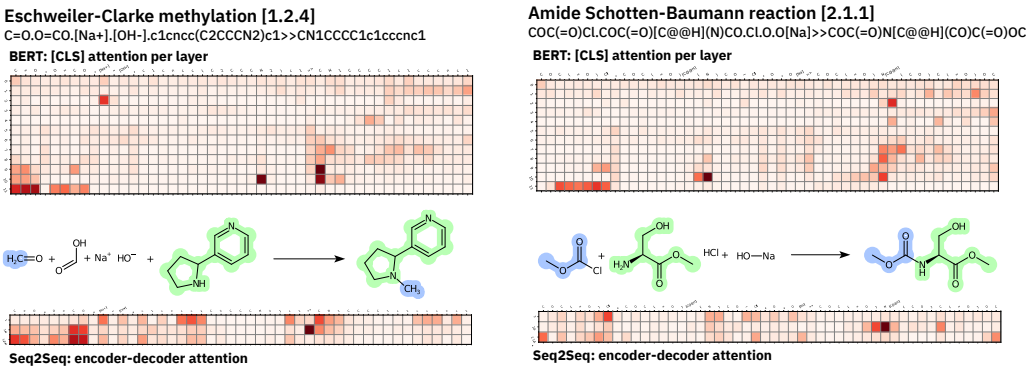

Figure 5.2: **Attention weights interpretation.** Layer-wise [CLS] token attention for the BERT classifier and encoder-decoder attention for the enc2-dec1 transformer model. The horizontal axis contains the SMILES tokens of the input reaction. The darker the token the more attention a specific token had received in that particular layer or output step. The colouring on the reaction depictions created with CDK depict [182] shows the mapping from precursors to product in the ground truth.

attend either the product or the precursors. The last layers focused on the reaction centre and the precursors that are important for the classification.

5.2.2 Mapping chemical reaction space

Data-driven reaction fingerprints

Molecular fingerprints are widely used to screen molecules with similar properties or map chemical space [210]. Our reaction BERT models does not only perform best on the classification task but also allows chemists to generate vectorial representations of chemical reactions. Here we introduce reaction fingerprints based on the embeddings computed by BERT [3] models. They can be applied to any reaction data set, as they do not require a reactant-reagent split or a fixed number of precursors. During the pretraining of the BERT model, individual tokens in the reaction SMILES are masked and then predicted by the model. As the prepended [CLS] token is never masked, the model is always able to attend the representation of this token to recover the masked tokens. The intuition is that the model uses the [CLS] token to embed a global description of the reaction. Before the fine-tuning, the [CLS] token embeddings are learned purely by self-supervision. We refer to this fingerprint as *rxnfp (pretrained)*. For the supervised fine-tuning, the embeddings of the [CLS] token are then taken as input for a one layer classification head and further refined. We refer to the fingerprint fine-tuned on the Pistachio training set as *rxnfp*. In our case, the [CLS] token embedding is a vector of size 256, corresponding to the hidden size of the BERT model. During the supervised classification task, the model has to focus on the reaction centre and certain precursors that are specific to the individual name reactions. For instance, the Eschweiler-Clarke methylation (1.2.4) is a methylation reaction that can be distinguished from other methylation reactions as its precursors contain formaldehyde and formic acid (see Figure 5.2). Another example are Suzuki-type coupling reactions, where the "-type" suffix means that

the metal catalyst is missing but the described reaction would otherwise correspond to a Suzuki coupling reaction.

Reaction atlases

In Figure 5.3, we show an annotated version of a reaction atlas created by using the embeddings of a BERT classifier fine-tuned for three epochs. The colours correspond to the 12 superclasses found in the data set. The individual classes are almost perfectly clustered. It is worth noting that the sub-trees in the TMAP closely group related reaction classes. For instance, in the upper left, one sub-tree contains all "Formylation"-related reactions, Weinreb reactions are clustered in a branch in the lower left and Suzuki-type reactions share the same branch as the corresponding Suzuki reactions. The unannotated reaction atlas was created using the fingerprints computed from a pretrained reaction BERT model without classification fine-tuning. Even after applying a purely unsupervised masked language modeling training, the model was already able to extract features relevant for reaction classification and some clustering can be observed in the figure.

An interactive reaction TMAP [201], visualising the public Schneider 50k [64] data set by using the *rxnfp (10k)* embeddings and highlighting different precursor and product properties, can be found on `https://rxn4chemistry.github.io/rxnfp/tmaps/tmap_ft_10k.html`.

Reaction search

One of the primary use cases for reaction fingerprints is the search for similar reactions in a database. An atom-mapping independent reaction fingerprint is extremely powerful, as it unlocks the possibility of reaction retrieval without the need of knowing the reaction centre. For instance, when a black box model like a forward reaction prediction model [4] or a retrosynthesis model [43] predicts a reaction, the most similar reactions from the training set of those models could be retrieved. Such a retrieval of similar reactions could not only increase the explainability of deep learning models. It would also allow chemists to access the metadata (including yield and reaction conditions) of the closest reactions, if this information is available.

In Figure 5.4 the three approximate nearest neighbours of the BERT classifier fingerprint can be found for four test set reactions from four distinct reaction classes. The nearest neighbours searches on the training set containing 2.4M reactions were performed within milliseconds using unoptimised python code on a MacBook Pro (Processor: 2.7 GHz Intel Core i7, Memory: 16 GB 2133 MHz). They were based on the LSH forest from the TMAP module developed by Probst and Reymond [201] In all searches, the nearest neighbours corresponded to the same class as the query reaction. The similarities between the query reaction and the retrieved nearest neighbours were clearly visible even for non-experts. The reactions share similar, if not the same precursors, and the products show similar features. One of the great advantages of this reaction search method is that it only requires a reaction SMILES as input.

To investigate the robustness of our BERT classifier embeddings we removed three classes from the fine-tuning training set (Number of removed reaction classes: '1.6.4 - Chloro N-alkylation': 24109, '3.9.17 - Weinreb Iodo coupling': 225, '9.7.73 - Hydroxy to azido': 1526) and fine-tuned another BERT classifier. After 5 epochs, we generated the embeddings for the test set reactions from the three removed classes. For the "Chloro N-alkylation" and the "Hydroxy to azido" class the most common prediction was "Unrecognised". All the predictions of the BERT model trained

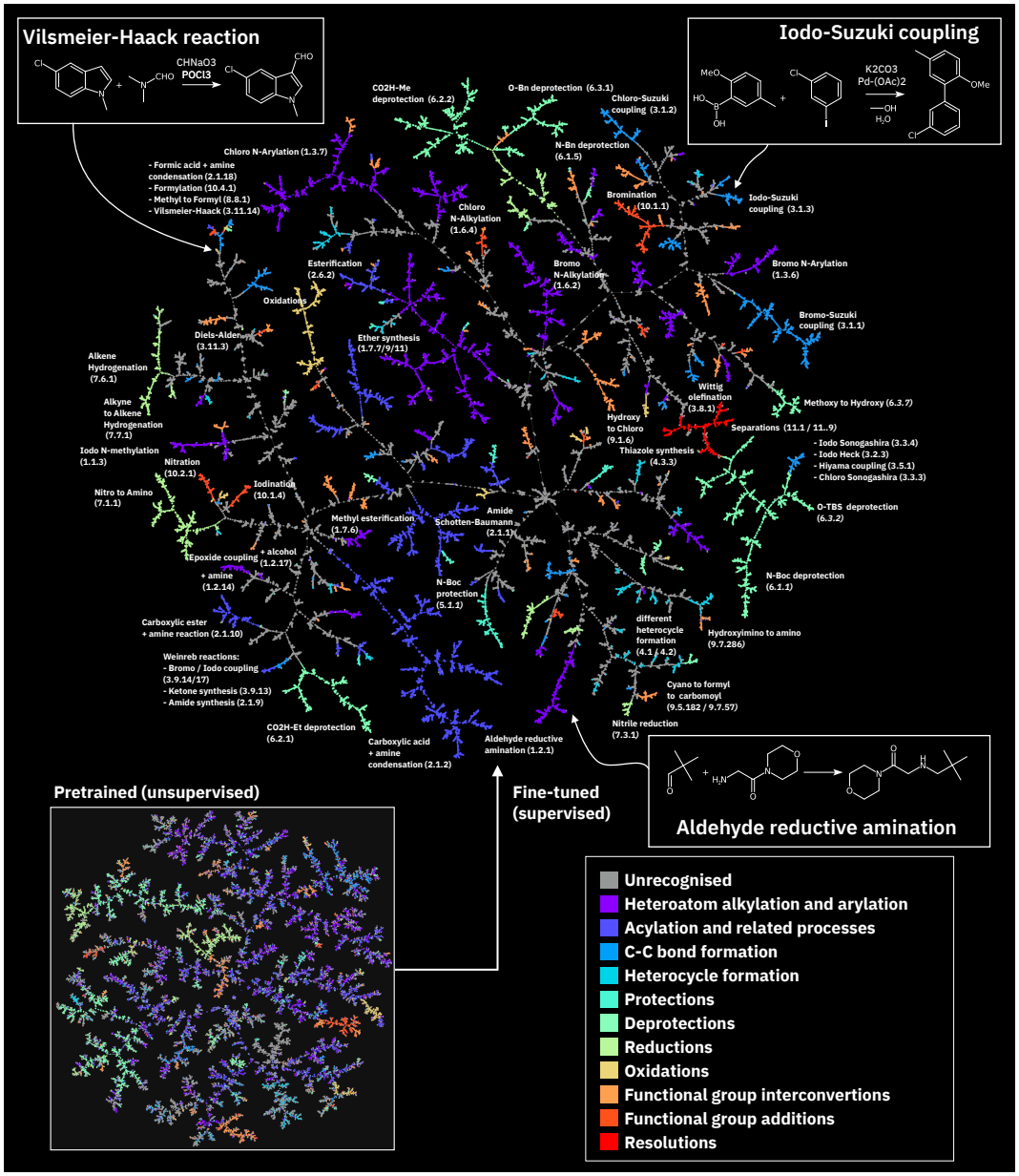

Figure 5.3: **Reaction atlases.** Top: Annotated reaction atlas created from *rxnfp*. Bottom: reaction atlas made from *rxnfp (pretrained)*. The different fingerprints of the test set reactions are visualised using a TMAP algorithm [201] and the Faerun visualiation library [211]. The fingerprints were minhashed using a weighted hashing scheme to make them compatible with the LSH forest.

without the removed classes for the "Weinreb Iodo coupling" were "Weinreb bromo coupling" that differs just by the type of the reacting halogen atom. Another interesting experiment is the retrieval of nearest neighbours from the original training set for the embeddings generated by the BERT model trained without the removed classes. For 1078 out of 1370 "Chloro N-alkylation"

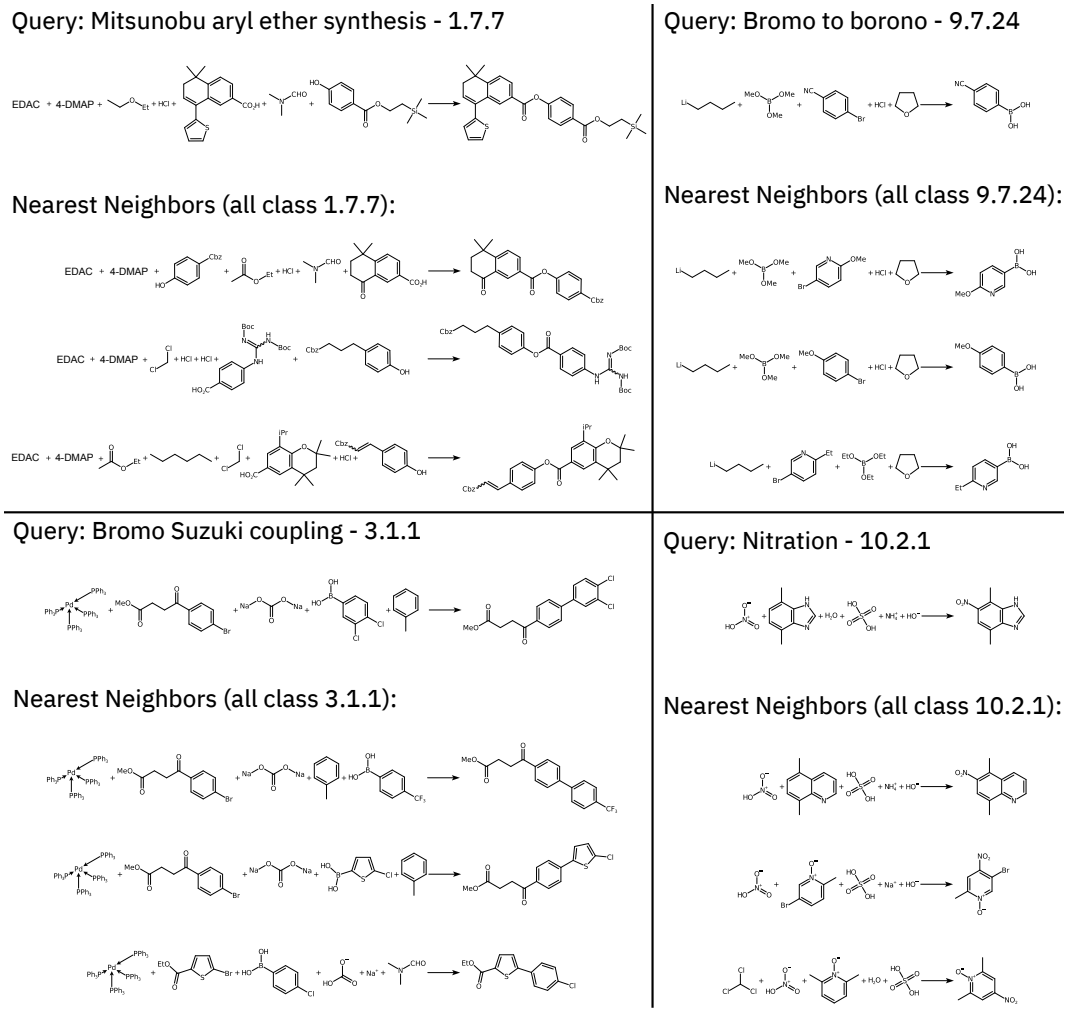

Figure 5.4: **Nearest-neighbour queries.** Four examples of reaction SMILES queries and the three nearest neighbours retrieved from the LSH forest [201] of the training set containing 2.4M reactions. All the retrieved reactions belong to the same reaction class as the query reaction and show similar precursors.

reactions in the test set, the nearest neighbour in the initial training set (including all the reaction classes) was a "Chloro N-alkylation" reaction. For the 10 "Weinreb Iodo coupling" reactions, the nearest neighbours in the original training set were four "Weinreb Bromo coupling" and other four "Bromo Grignard + nitrile ketone synthesis" reactions, which are both closely related reaction types. There was no clearly dominating reaction class in the nearest neighbours with 44 out of 76 reactions being "Unrecognised".

5.3 Discussion

In this work, we focused on the data-driven classification of chemical reactions with natural language processing methods and on the use of their embedded information to design reaction fingerprints. Our transformer-based models were able to learn the classification schemes using a broad set of chemical reactions as ground-truth, labeled by a commercially available reaction classification tool. With the BERT classifier, we match the rule-based classification with an accuracy of 98.2%, compared to 41% for a traditional fingerprint plus 5-nearest neighbours classifier. Our models are able to learn the atomic environment characteristics of each class and provide a rationale that is easily interpretable by chemists. Understanding the reasoning behind each classification by using the attention weights may help the end-user chemists with the adoption process of these technologies. We showed that the representations learned by our BERT models can be used as reaction fingerprints. Those data-driven reaction fingerprints unlock the possibility of mapping the reaction space without knowing the reaction centres or the reactant-reagent split. They also enable efficient nearest neighbour searches on reaction data sets containing millions of reactions. Moreover, our fingerprints were recently used to estimate experimentally measured activation energies [202] and fine-tuned to predict chemical reaction yields [203].

5.4 Methods

5.4.1 Data

The data consisted of 2.6M reactions extracted from the Pistachio database [180] (version 191118), where we removed duplicates and filtered invalid reactions using RDKit [153]. The data set was split into train, validation and test sets (90% / 5% / 5%), with reactions with identical products kept in the same set. The reaction data in Pistachio was classified using NameRXN [63], a rule-based software that classifies roughly 1000 different name reactions. The classification is organised into superclasses [212], reaction categories and name reactions according to the RXNO ontology [213]. For more detail on name reactions and their categories, we refer the reader to the work of Schneider et al.[173]. As common in practice, we represent the chemical reactions with reaction SMILES [46, 47]. We tokenise the reaction SMILES as in Schwaller et al.[4] without enforcing any distinction between reactants and reagents. Therefore, our method is universally applicable, including those reactions where the reactant-reagent distinction is subtle [62]. To compare with previous work and ensure reproducibility, we used the reaction data set published by Schneider et al. [64] with 50k reactions belonging to 50 different reaction classes. We also introduced an open-source reaction classification data set, which we named USPTO 1k TPL, derived from the USPTO data base by Lowe [42]. It consists of 445k reactions divided into 1000 template labels. The data set was randomly split into 90% for training and validation and 10% for testing. The labels were obtained by atom-mapping the USPTO data set with RXNMapper [189]. Subsequently, the template extraction workflow by Thakkar et al. [109, 214] was applied and finally, selecting reactions belonging to the 1000 most frequent template hashes. Those template hashes were used as class labels. Similarly to the Pistachio data set, USPTO 1k TPL is strongly imbalanced.

5.4.2 Models

We trained two different types of deep learning models inspired by recent progress in Natural Language Processing. The first model is an autoregressive encoder-decoder transformer model [2]. We constructed the model with 2 encoder layers and 1 decoder layer. For the prediction target, we split the class prediction into superclass, category and name reaction prediction. This means, for example, that the target string for the name reaction "1.2.3" would be "1 1.2 1.2.3". As the source and target are dissimilar, we did not share encoder and decoder embeddings. We used the same remaining hyperparameter as were used for the training of the Molecular Transformer [4, 156], which is state-of-the-art in chemical reaction prediction.

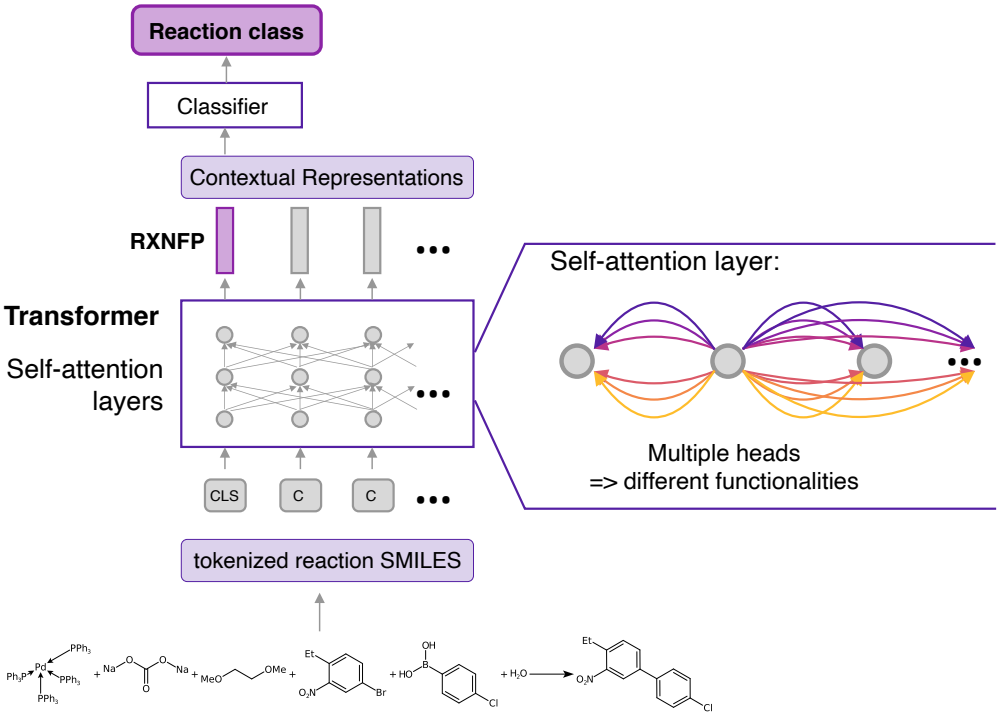

Figure 5.5: **BERT reaction classification model.** The figure illustrates a BERT model with stacks of self-attention layers. All self-attention layers consist of multiple attention heads. Using a classifier head the model was applied to a chemical reaction classification task. The encoding of the [CLS] token can also be used as reaction fingerprint (rxnfp).

One of the major recent advancement in natural language processing is BERT [3], which compared to the seq-2-seq architecture only consists of a transformer encoder with specific heads that can be fine-tuned for different tasks such as multi-class prediction. The model is visualised in Figure 5.5. We pretrained a BERT model using masked language modeling loss on the chemical reactions. The task of the model in masked language modeling consists of predicting individual tokens of the input sequence that have been masked with a probability of 0.15. Same as in the BERT training, a special class token [CLS] was prepended to the tokenised reaction SMILES.

The [CLS] token was never masked during this self-supervised training. In contrast to the original BERT pretraining [3], we did not use the next sentence prediction task. We then fine-tuned the pretrained model with a classifier head on the name reaction classes. The embeddings of the [CLS] token were taken as input for the classifier head. Compared to the hyperparameters of the BERT-Base model in Ref. [215], we decreased the hidden size to 256, the intermediate size to 512, and the number of attention heads to 4. For the pretraining, we set 820k steps with a learning rate of 1e-4 and a maximum sequence length of 512, the rest of the parameters were kept as suggested in Ref. [215]. For the classification fine-tuning, we only changed the learning rate to 2e-5, kept the maximum sequence length of 512 and fine-tuned for 5 epochs. After training, we converted the models to PyTorch [81] models, which matched the Huggingface [216] interface, as it facilitated further analysis.

5.4.3 k-nearest neighbour classifier

The k-nearest neighbour classifier used to assess the quality of the proposed reaction representations is based on the Facebook AI Similarity Search (FAISS) framework developed by Facebook research [206]. As FAISS provides an efficient implementation of brute-force k-nearest neighbour searches that can be applied on relatively large data sets. Possible biases introduced through approximation methods were therefore avoided. The number of nearest neighbours $k = 5$ and the Euclidean metric (L2) are chosen for all tests. The predicted class of the query was assumed to be the one that is represented within most often the result set. Ties were broken using the distance between the query and one or more neighbours.

5.4.4 TMAP

TMAP [201] is a dimensionality reduction algorithm capable of handling millions of data points. The advantage of TMAP compared to other dimensionality reduction algorithms is the 2D tree-like output, which preserves both local and global structures, with a focus of local structure. The algorithm consists of four steps: 1) LSH Forest-based indexing, 2) k-nearest neighbour graph generation, 3) minimum spanning tree calculation using Kurskal's algorithm and 4) creating the tree-like layout. The resulting layout is then displayed using the interactive data visualisation framework Faerun [211].

TMAP [201] and Faerun [211] were originally developed to visualise large molecular data sets, but have been shown to be applicable to a wide range of other data. Here, we extended the framework with a customised version of SmilesDrawer [217] that has been extended to allow for the display of chemical reactions.

5.4.5 Evaluation metrics

To compare the results on the imbalanced classification test set, we used the confusion entropy of the confusion matrix (CEN) [207] calculated as follows,

$$P_{i,j}^{j} = \frac{Matrix(i,j)}{\sum_{k=1}^{|C|}\Big(Matrix(j,k) + Matrix(k,j)\Big)},$$

$$P_{i,j}^{i} = \frac{Matrix(i,j)}{\sum_{k=1}^{|C|}\Big(Matrix(i,k) + Matrix(k,i)\Big)}$$

$$CEN_j = -\sum_{k=1,k\neq j}^{|C|}\left(P_{j,k}^{j} log_{2(|C|-1)}\Big(P_{j,k}^{j}\Big) + P_{k,j}^{j} log_{2(|C|-1)}\Big(P_{k,j}^{j}\Big)\right)$$

$$P_j = \frac{\sum_{k=1}^{|C|}\Big(Matrix(j,k) + Matrix(k,j)\Big)}{2\sum_{k,l=1}^{|C|} Matrix(k,l)}$$

$$CEN = \sum_{j=1}^{|C|} P_j CEN_j$$

where Matrix is the confusion matrix, and the overall Matthews Correlation Coefficient (MCC)[208, 209] is,

$$cov(X,Y) = \sum_{i,j,k=1}^{|C|}\Big(Matrix(i,i)Matrix(k,j) - Matrix(j,i)Matrix(i,k)\Big)$$

$$cov(X,X) = \sum_{i=1}^{|C|}\left[\Big(\sum_{j=1}^{|C|} Matrix(j,i)\Big)\Big(\sum_{k,l=1,k\neq i}^{|C|} Matrix(l,k)\Big)\right]$$

$$cov(Y,Y) = \sum_{i=1}^{|C|}\left[\Big(\sum_{j=1}^{|C|} Matrix(i,j)\Big)\Big(\sum_{k,l=1,k\neq i}^{|C|} Matrix(k,l)\Big)\right]$$

$$MCC = \frac{cov(X,Y)}{\sqrt{cov(X,X)\times cov(Y,Y)}}.$$

Both are recommended metrics for imbalanced multi-class classification problems. We computed the scores using PyCM [218]. For the comparison on the balanced data set, we used the average recall, precision and F1 score, as those metrics were used by Schneider et al. [64]. The recall, precision and F1 score values for the individual classes are shown in the supplementary material.

Data & code availability

The Schneider 50k data set is publicly available [64]. We provide a new reaction data set (USPTO 1k TPL), derived from the work of Lowe [42], containing the 1000 most common reaction templates as classes. It can be accessed through `https://rxn4chemistry.github.io/rxnfp`. The commercial Pistachio (version 191118) data set can be obtained from NextMove Software [180]. Pistachio relies on Leadmine [59] to text-mine patent data. The data set comes with reaction classes assigned using NameRXN (`https://www.nextmovesoftware.com/namerxn.html`). The rxnfp code

and the experiments on the public data sets, as well as an interative TMAP, can be found on https://rxn4chemistry.github.io/rxnfp[219].

6 Prediction of chemical reaction yields using deep learning

Artificial intelligence is driving one of the most important revolutions in organic chemistry. Multiple platforms, including tools for reaction prediction and synthesis planning based on machine learning, successfully became part of the organic chemists' daily laboratory, assisting in domain-specific synthetic problems. Unlike reaction prediction and retrosynthetic models, the prediction of reaction yields has received less attention in spite of the enormous potential of accurately predicting reaction conversion rates. Reaction yields models, describing the percentage of the reactants converted to the desired products, could guide chemists and help them select high-yielding reactions and score synthesis routes, reducing the number of attempts. So far, yield predictions have been predominantly performed for high-throughput experiments using a categorical (one-hot) encoding of reactants, concatenated molecular fingerprints, or computed chemical descriptors. Here, we extend the application of natural language processing architectures to predict reaction properties given a text-based representation of the reaction, using an encoder transformer model combined with a regression layer. We demonstrate outstanding prediction performance on two high-throughput experiment reactions sets. An analysis of the yields reported in the open-source USPTO data set shows that their distribution differs depending on the mass scale, limiting the dataset applicability in reaction yields predictions.

This chapter has been accepted as a scientific article in Machine Learning: Science and Technology:

P Schwaller, AC Vaucher, T Laino, JL Reymond. Prediction of Chemical Reaction Yields using Deep Learning. *Mach. Learn.: Sci. Technol.*, 2021, 2 015016, DOI: 10.1088/2632-2153/abc81d

6.1 Introduction

Chemical reactions in organic chemistry are described by writing the structural formula of reactants and products separated by an arrow, representing the chemical transformation by specifying how the atoms rearrange between one or several reactant molecules and one or several product molecules [189]. Economic, logistic, and energetic considerations drive chemists to prefer chemical transformations capable of converting all reactant molecules into products with the highest yield possible. However, side-reactions, degradation of reactants, reagents or products in the course of the reaction, equilibrium processes with incomplete conversion to a product, or sim-

ply by product isolation and purification undermine the quantitative conversion of reactants into products, rarely reaching optimal performance.

Reaction yields are usually reported as a percentage of the theoretical chemical conversion, i.e., the percentage of the reactant molecules successfully converted to the desired product compared to the theoretical value. It is not uncommon for chemists to synthesise a molecule in a dozen or more reaction steps. Hence, low-yield reactions may have a disastrous effect on the overall route yield because of the individual steps' multiplicative effect. Therefore, it is not surprising that designing new reactions with yields higher than existing ones attracts much effort in organic chemistry research.

In practice, specific chemical reaction classes are characterised by lower or higher yields, with the actual value depending on the reaction conditions (temperature, concentrations, etc.) and on the specific substrates.

Estimating the reaction yield can be a game-changing asset for synthesis planning. It provides chemists with the ability to evaluate the overall yield of complex reaction paths, addressing possible shortcomings well ahead of investing hours and materials in wet-lab experiments. Computational models predicting reaction yields could support synthetic chemists in choosing an appropriate synthesis route among many predicted by data-driven algorithms. Moreover, reaction yields prediction models could also be employed as scoring functions in computer-assisted retrosynthesis route planning tools [43, 107, 125, 126], to complement forward prediction models [4, 43] and in-scope filters [107].

Most of the existing efforts in constructing models for the prediction of reactivity or of reaction yields focused on a particular reaction class: oxidative dehydrogenations of ethylbenzene with tin oxide catalysts [220], reactions of vanadium selenites [221], Buchwald–Hartwig aminations [200, 222, 223], and Suzuki–Miyaura cross-coupling reactions [224, 225, 226]. To the best of our knowledge, there was only one attempt to design a general-purpose prediction model for reactivity and yields, without applicability constraints to a specific reaction class [227]. In this work, the authors design a model predicting whether the reaction yield is above or below a threshold value and conclude that the models and descriptors they consider cannot deliver satisfactory results.

Here, we build on our legacy of treating organic chemistry as a language to introduce a new model that predicts reaction yields starting from reaction SMILES [36]. More specifically, we fine-tune the rxnfp models by Schwaller et al. [177] based on a BERT-encoder [3] by extending it with a regression layer to predict reaction yields. BERT encoders belong to the transformer model family, which has revolutionised natural language processing [2, 3]. These models take sequences of tokens as input to compute contextualised representations of all the input tokens, and can be applied to reactions represented in the SMILES [46] format. In this work, we demonstrate for the first time, that these natural language architectures are very useful not only when working with language tokens, but also to provide descriptors of high quality to predict reaction properties such as reaction yields.

It is possible to train our approach both on data specific to a given reaction class or on data representing different reaction types. Thus, we initially trained the model on two high-throughput experimentation (HTE) data sets. Among the few HTE reaction data sets published in recent years, we selected the data sets for palladium-catalysed Buchwald–Hartwig reactions provided by Ahneman et al. [222] and for Suzuki–Miyaura coupling reactions provided by Perera et al. [228]. Finally, we trained our model on patent data available in the USPTO data set [41, 42].

HTE and Patent data sets are very different in terms of content and quality. HTE data sets typically cover a very narrow region in the chemical reaction space, with chemical reaction data related to one or a few reaction templates applied to large combinations of selected precursors (reactants, solvents, bases, catalysts, etc.). In contrast, patent reactions cover a much wider reaction space. In terms of quality, HTE data sets report reactions represented uniformly and with yields measured using the same analytical equipment, thus providing a consistent and high quality collection of knowledge. In comparison, the yields from patents were measured by different scientists using different equipments. Incomplete information in the original documents, such as unreported reagents or reaction conditions, and the extensive limitation in text mining technologies makes the entire set of patent reactions quite noisy and sparse. An extensive analysis of the USPTO data set revealed that the experimental conditions and reaction parameters, such as scale of the reaction, concentrations, temperature, pressure, or reaction duration, may have a significant effect on the measured reaction yields. The functional dependency of the yields from the reaction conditions poses additional constraints, as the model presented in this work does not consider those values explicitly in the reaction descriptor. The basic assumption is that every reaction yield reported in the data set is optimised for the reaction parameters.

Our best performing model reached an R^2 score of 0.956 on a random split of the Buchwald-Hartwig data set while the highest R^2 score on the smoothed USPTO data was 0.388. These numbers reflect how the intrinsic data set limitations increase the complexity of training a sufficiently good performing model on the patent data, resulting into a more difficult challenge than training a model for the HTE data set.

6.2 Models and experimental pipeline

We base our models directly on the reaction fingerprint (rxnfp) models by Schwaller et al. [177]. We use a fixed size encoder model size, tuning only the hyperparameter for dropout rate and learning rate, thus avoiding often encountered difficulties of neural networks with numerous hyperparameters. During our experiments, we observed good performances for a wide range of dropout rates (from 0.1 to 0.8) and conclude that the initial learning rate is the most important hyperparameter to tune. To facilitate the training, our work uses simpletransformers [229], huggingface transformer [216] and PyTorch framework [81]. The overall pipeline is shown in Figure 6.1.

To provide an input compatible with the rxnfp model we use the same RDKit [153] reaction canonicalisation and SMILES tokenization [4] as in the rxnfp work [177].

6.3 Results

6.3.1 High-throughput experiment yield predictions

Buchwald–Hartwig reactions

Ahneman et al. [222] performed high-throughput experiments on Pd-catalysed Buchwald–Hartwig C-N cross coupling reactions, measuring the yields for each reaction. For the experiments, they used three 1536-well plates spanning a matrix of 15 aryl and heteroaryl halides, 4 Buchwald ligands, 3 bases, and 23 isoxazole additives resulting in 3955 reactions. As inputs for their models,

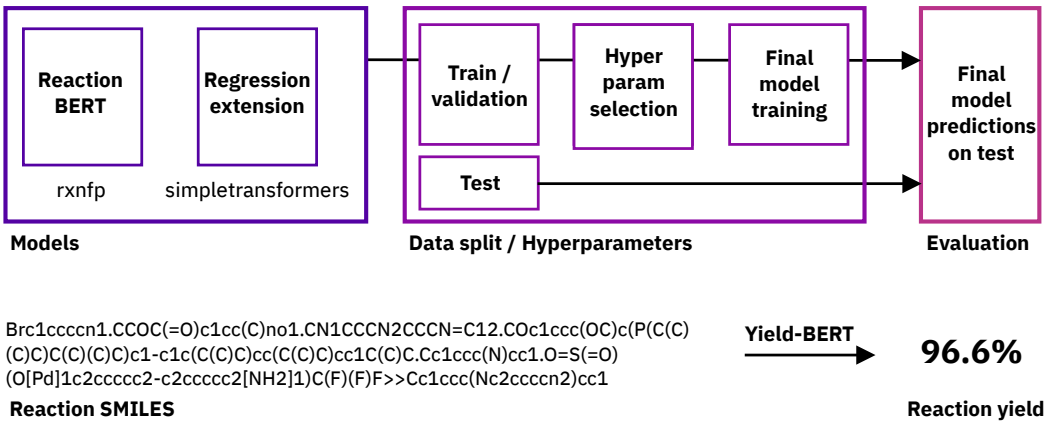

Figure 6.1: **Training/evaluation pipeline and task description.**

Ahneman et al. [222] computed 120 molecular, atomic and vibrational properties with density functional theory using Spartan for every halide, ligand, base and additive combination. The descriptors included highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital LUMO energy, dipole moment, electronegativity, electrostatic charge and NMR shifts for atoms shared by the reagents. Compared to reaction SMILES that can vary in length, the input in the work of Ahneman et al. [222] was a fixed-size vector. They investigated numerous methods, including linear models, k-Nearest-Neighbours, support vector machines, Bayes generalised linear models, artificial neural networks and random forests. Eventually, they selected their random forest model as the best performing. The work of Ahneman et al. [222] was challenged by Chuang and Keiser [223], who pointed out several issues. First, by replacing the computed chemical features with random features of the same length or one-hot encoded vectors Chuang and Keiser got similar performance than the original paper with the chemical features. Therefore, they weakened the original claim that additive features were the most important for the predictions. However, the additive features were on average still estimated to be the most important features by the random forest model when the yields were shuffled [223]. Recently, Sandfort et al. [200] used a concatenation of multiple molecular fingerprints as an alternative reaction representation to demonstrate superior yield prediction performance compared to one-hot encoding.

Unlike previous work, we directly use the reaction SMILES as input to a BERT-based reaction encoder [177] enriched with a regression layer (Yield-BERT). To investigate the suggested method, we used the same splits as Sandfort et al. [200]. In contrast, to their work, we used 1/7 of the training set from the first random split as a validation set to select optimal values for the two hyperparameters, namely, learning rate and dropout probability. Once selected, we kept the hyperparameters identical for all the subsequent experiments.

The results are shown in Table 6.1. Using solely a reaction SMILES representation, our method achieves an average R^2 of 0.951 on the random splits and outperforms not only the MFF by Sandfort et al. [200], but also the chemical descriptors computed with DFT by Ahneman et al. [222]. Moreover, for the out-of-sample tests where the isoxazole additives define the splits our method performs on average better than MFF and one-hot descriptors and comparable to the chemical

Table 6.1: **Comparing methods on the Buchwald-Hartwig data set.** All results shown in this table used the rxnfp pretrained model as base encoder.

R^2	DFT [222]	one-hot [200, 223]	MFF [200]	Yield-BERT
rand 70/30	0.92	0.89	0.927 ± 0.007	$\mathbf{0.951 \pm 0.005}$
rand 50/50	0.9			0.92 ± 0.01
rand 30/70	0.85			0.88 ± 0.01
rand 20/80	0.81			0.86 ± 0.01
rand 10/90	0.77			0.79 ± 0.02
rand 5/95	0.68			0.61 ± 0.04
rand 2.5/97.5	0.59			0.45 ± 0.05
test 1	0.8	0.69	0.85	0.84 ± 0.01
test 2	0.77	0.67	0.71	0.84 ± 0.03
test 3	0.64	0.49	0.64	0.75 ± 0.04
test 4	0.54	0.49	0.18	0.49 ± 0.05
avg. 1-4	0.69	0.59	0.60	0.73

descriptors. As in the work of Sandfort et al. [200], the test 3 split resulted in the worst model performance. For the rest of the out-of-sample, our method performs better than the others. We also reduced the training set to 5% (197 reactions), 10% (395 reactions) and 20% (791 reactions) and observed that the model learned to reasonably predict yields despite the significantly smaller training set.

Suzuki–Miyaura reactions

Perera et al. [228] used HTE technologies to the class of the Suzuki–Miyaura reactions. They considered 15 pairs of electrophiles and nucleophiles, each leading to a different product. For each pair, they varied the ligands (12 in total), bases (8), and solvents (4), resulting in a total of 5760 measured yields. The same data set was also investigated in the work of Granda et al. [224].

Here, we first trained our yield prediction models with the same hyperparameters as for the Buchwald–Hartwig reaction experiment above, achieving an R^2 score of 0.79 ± 0.01. Second, we tuned the dropout probability and learning rate, similarly to the previous experiment, using a split of the training set of the first random split. The resulting hyperparameters were then used for all the splits. The hyperparameter tuning did not lead to better performance compared to the parameters used for the Buchwald–Hartwig reactions. This shows that the models have a stable performance for a wide range of parameters and that they are transferable from one data set to another related data set.

We also compared two different base encoder models that are available from the rxnfp library [177], namely the BERT model pretrained with a masked language modelling task, and the BERT model subsequently fine-tuned on a reaction class prediction task. The results are displayed in Table 6.2. In contrast to the Buchwald–Hartwig data set, where no difference between the two

Table 6.2: **Summary of the average R^2 scores on the Suzuki–Miyaura reactions data set.** Different base encoders for the Yield-BERT were compared. We used 10 different random folds (70/30).

Base encoder rxnfp [177] Hyperparameters	pretrained same as 3.1	pretrained tuned	ft same as 3.1	ft tuned
random 70/30	0.79 ± 0.01	0.79 ± 0.02	$\mathbf{0.81 \pm 0.02}$	$\mathbf{0.81 \pm 0.01}$

base encoders was observed, the ft model achieves an R^2 score of 0.81 ± 0.01, outperforming the pretrained base encoder on the Suzuki–Miyaura reactions.

Discovery of high yielding reactions with reduced training sets

Granda et al. [224] proposed to train on a random (10%) portion of the original data set to evaluate the rest of the reactions with the purpose of selecting the next reactions to test. Similarly, we trained our models on different fractions of the training set and used them to evaluate the yields of the remaining reactions. The aim here is to evaluate how well the models are at selecting high-yielding reactions after having seen a small fraction of randomly chosen reactions.

As can be seen from Figure 6.2, training on only 5% of the reactions already enables a chemist to select some of the highest yielding reactions for the next round of the experiments. With a training set of 10% the yields of the selected reactions are close to the best possible selection marked with "ideal" in the Figure. For the Buchwald–Hartwig reaction, using a model trained on 10% of the data set, the 10 reactions from the remaining unseen data set predicted to have the highest yields, have an average yield of 90 ± 6 %, compared to the ideal selection of 98.7 ± 0.9 %. In contrast, a random selection of 10 reactions would have let to yields of 34 ± 27 %. The selection works similarly for the Suzuki–Miyaura reactions.

We performed a purely greedy selection, as we aimed to find highest yielding reactions after one training round. A wider chemical reaction space exploration with a reaction selection using more elaborate uncertainty estimates and an active learning strategy was investigated by Eyke et al. [226].

6.3.2 Patent yield predictions

In this section, we analyse USPTO data set [41, 42] yields. We started from the same set as in our previous work [116], keeping only reactions for which yields and product mass were reported. In contrast to HTE, where reactions are typically performed in sub-gram scale, the patent data contains reactions spanning a wider range, from grams to sub-grams scales.

Gram versus sub-gram scale

When investigating the yields for different mass scales, we observed that gram and sub-gram scales had statistically different yield distributions, as shown in Figure 6.3. One reason could be that the reaction sub-gram scale reactions are generally less optimised than gram-scale. In sub-gram scale, the primary goal is to show that the desired product is present. To be able to synthesise a specific compound on a larger scale, reactions are optimised and predominantly high yielding

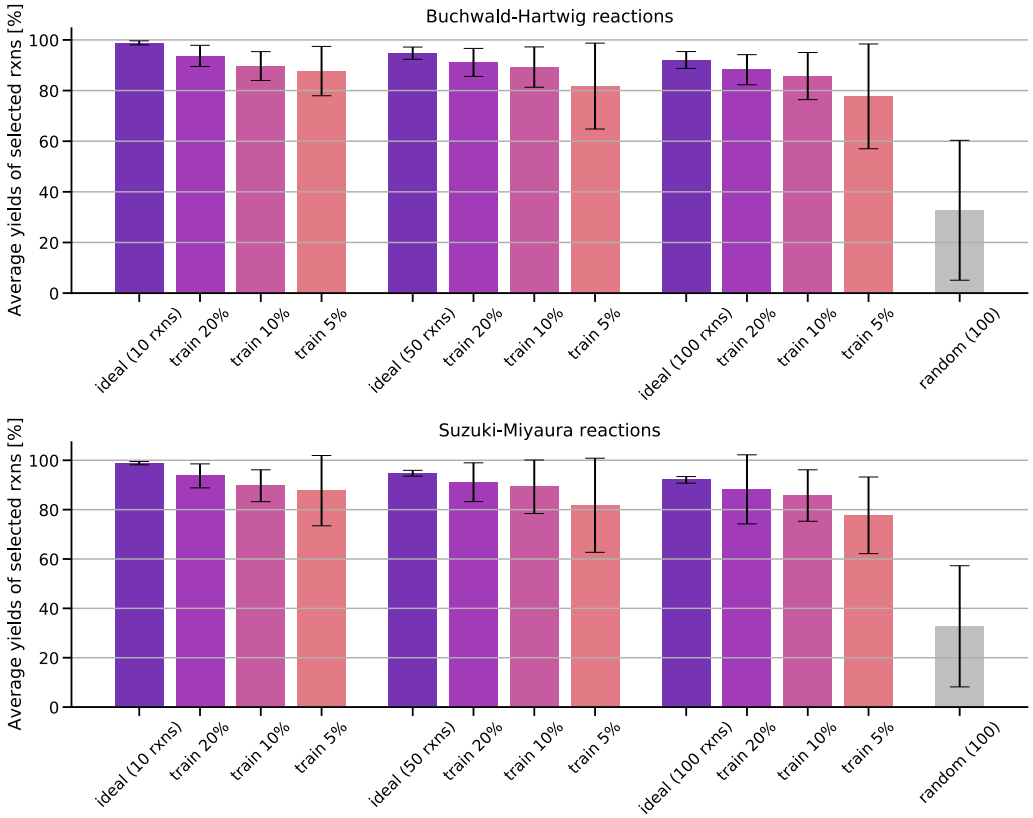

Figure 6.2: **Discovery of high yielding reaction.** Average and standard deviation of the yields for the 10, 50, and 100 reactions predicted to have the highest yields after training on a fraction of the data set (5%, 10%, 20%). The ideal reaction selection and a random selection are plotted for comparison.

reactions are employed. Therefore, we split the USPTO reactions into two data sets according to the product mass. If for the same canonical reaction SMILES multiple yields were reported in the same mass scale, we took the average of those yields.

We performed various experiments summarised in Table 6.3. The R^2 scores for the randomly train-test splits with 0.117 for gram scale and 0.195 low. As expected, the tasks become even more difficult when the time split is used. In our experiment, we took all reactions first published in 2012 and before as training/validation set and the reactions published after 2012 as test set. To show that the model was still able to learn, we performed a sanity check by randomising the yields across the training reactions. The resulting performance on the test set was a R^2 score of 0.

Unfortunately, the yields from the USPTO data set could not be accurately predicted. To better understand why, we further inspected the USPTO reaction yields with a visual analysis using reaction atlases built using TMAP [201], faerun [211] and our reaction fingerprints [177]. Figure 6.4 reveals that globally reaction classes tend to have similar yields. However, if a local neighbourhood is analysed the nearest neighbours often have extremely diverse reaction yields. Those diverse

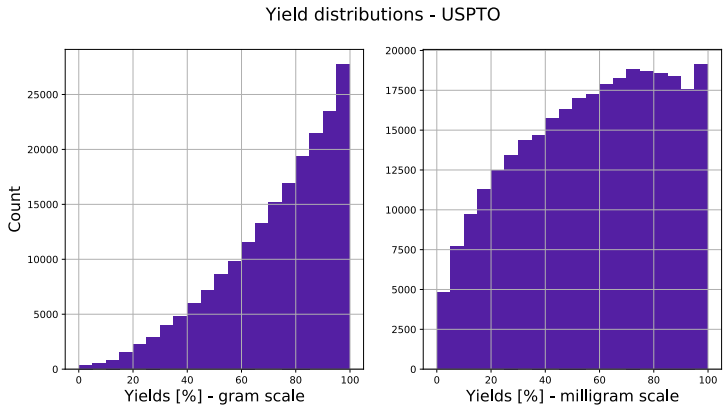

Figure 6.3: **USPTO yields histograms separated in gram and sub-gram scale**

yields make it challenging for the model to learn anything but yield averages for similar reactions and hence, explain the low performance on the patent reactions. This analysis opens up relevant questions on the quality of the reported information (relative to the mass scale) and its extraction accuracy from text, which could severely hamper the development of reaction yield predictive models. The need of cleaned and consistent reaction yields data set is even more important than for other reaction prediction tasks.

Table 6.3: **USPTO yield prediction results.** Summary of the R^2 scores on the different USPTO reaction sets.

scale	gram	sub-gram
random split	0.117	0.195
time split	0.095	0.142
random split (smoothed)	0.277	0.388
randomised yields	0.0	0.0

In Table 6.3, the "random split (smoothed)" row shows an experiment inspired from the observations above. As some of the yields values are probably incorrect in the data set, we smoothed the yields by computing the average of the three nearest neighbour yields plus twice the own yield of the reaction. The nearest neighbours were estimated using the *rxnfp ft* [177] and faiss [230]. On the smoothed data sets, the performance of our models more than triples in the gram scale and doubles on the sub-gram scale, achieving R^2 scores of 0.277 and 0.388, respectively. The removal of noisy reactions [190] or reaction data augmentation techniques [114] could potentially lead to further improvements.

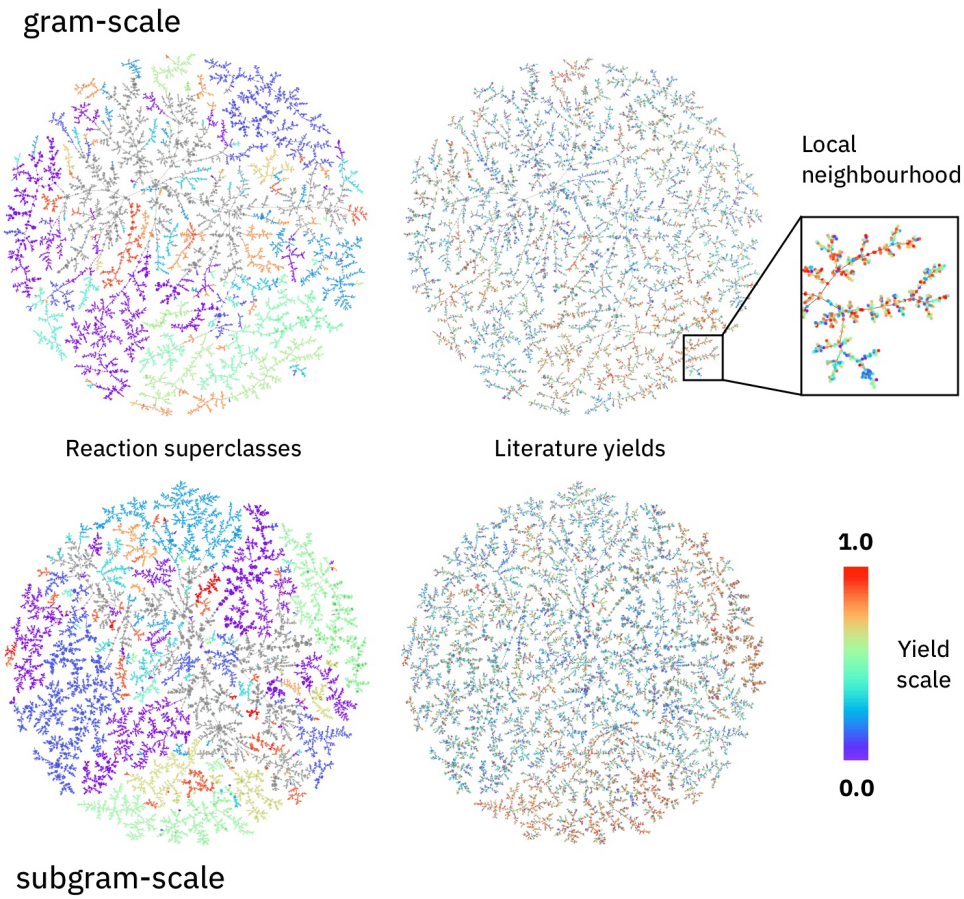

Figure 6.4: **Reaction Yield Atlases.** Top: gram scale. Bottom: sub-gram scale. Left: Reaction superclass distribution, reactions belonging to the same superclass have the same colour. Right: Corresponding reaction yields.

6.4 Discussion

In this work, we combined a reaction SMILES encoder with a reaction regression task to design a reaction yield predictive model. We analysed two HTE reaction data sets, showing excellent results. On the Buchwald–Hartwig reaction data set, our models outperform previous work on random splits and perform similar to models trained on chemical descriptors computed with DFT on test sets where specific additives were held out from the training set. Compared to random forest models, the feature importance can not directly be obtained. Future work could (visually) investigate the attention weights to find out what tokens and molecules contribute the most to the predictions [130, 231].

We analysed the yields in the public patent data and show that the distribution of reported yields strongly differs depending on the reaction scale. Because of the intrinsic lack of consistency and quality in the patent data, our proposed method fails to predict patent reaction yields accurately. While we cannot rule out the existence of any other architecture potentially performing

better than the one presented in this manuscript, we raise the need for a more consistent and better quality public data set for the development of reaction yields prediction models. The suspect that the patent data yields are inconsistently reported is substantiated by the large variability of methods used to purify and report yields by the different reaction mass scales and the different optimisation in each reported reaction. Our reaction atlases [177, 201, 211] reveal globally higher yielding reaction classes. However, nearest neighbours often have significantly scattered yields. We show that better results can be achieved by smoothing the patent data yields using the nearest neighbours.

Our approach to yield predictions can be extended to any reaction regression task, for example, for predicting reaction activation energies [202, 232, 233], and is expected to have a broad impact in the field of organic chemistry.

The code and data are available on `https://rxn4chemistry.github.io/rxn_yields/`.

7 Data augmentation strategies to improve reaction yield predictions and estimate uncertainty

Chemical reactions describe how precursor molecules react together and transform into products. The reaction yield describes the percentage of the precursors successfully transformed into products relative to the theoretical maximum. The prediction of reaction yields can help chemists navigate reaction space and accelerate the design of more effective routes. Here, we investigate the best-studied high-throughput experiment data set and show how data augmentation on chemical reactions can improve yield predictions' accuracy, even when only small data sets are available. Previous work used molecular fingerprints, physics-based or categorical descriptors of the precursors. In this manuscript, we fine-tune natural language processing-inspired reaction transformer models on different augmented data sets to predict yields solely using a text-based representation of chemical reactions. When the random training sets contain 2.5% or more of the data, our models outperform previous models, including those using physics-based descriptors as inputs. Moreover, we demonstrate the use of test-time augmentation to generate uncertainty estimates, which correlate with the prediction errors.

This chapter has been presented as a scientific article at the Machine Learning for Molecules workshop at NeurIPS 2020:

P Schwaller, AC Vaucher, T Laino, JL Reymond. Data augmentation strategies to improve reaction yield predictions and estimate uncertainty. DOI:10.26434/chemrxiv.13286741 (CC BY-NC-ND 4.0).

7.1 Introduction

The synthesis of new chemicals affects numerous aspects of our life, ranging from food and medicine to novel materials for technological applications. The current machine learning revolution in automated synthesis can significantly accelerate novel materials and molecules' development. In the last years, natural language processing methods emerged as robust and effective approaches in the field of organic chemistry, showing promising results in reaction prediction [4, 36, 111, 116], retrosynthesis planning [43, 65, 114, 234], data curation [190] and synthesis action generation [188, 235]. In those studies the encoder-decoder transformer models introduced by Vaswani et al. [2] excel among all other neural network architectures. More recently, the use of encoder-only transformers such as BERT [3, 98] led to advances in reaction classification and fingerprints [177], as well as in unsupervised reaction atom-to-atom mapping [189] and reaction yield predictions [203].

Reaction yields describe the percentage of the reactant molecules converted into the desired product molecule during a chemical reaction. The prediction of reaction yields can guide chemists in selecting the next experiments to perform, and retrosynthetic planning tools in aiming for routes that maximise the overall yield, thus minimising waste. Extensive chemical reaction yield data sets exist for high-throughput experiments (HTE). Examples are the Suzuki–Miyaura coupling reactions by Perera et al. [228] and the palladium-catalysed Buchwald–Hartwig reactions by Ahneman et al. [222], to date the best-studied HTE yield data set. In this work, we study reactions yield prediction using the latter data set [222], containing a total of 3955 Buchwald–Hartwig reactions with measured yields. Figure 6.1 a) provides an overview of the data set.

In a recent manuscript, Schwaller et al. [203] introduced a BERT [3] model with a regression head to predict reactions' yields given as input a reaction SMILES [46, 47], a text-based molecule and reaction representation. We show in Figure 6.1 a) and c) the task description, together with an example of a reaction SMILES. Here, we investigate how different data augmentation techniques (Figure 6.1 b), molecule permutations and SMILES randomisations [114, 236, 237, 238]) improve the performance of the yield prediction models. Moreover, we demonstrate the use of test-time augmentation (Figure 7.1 d)) to provide uncertainty estimates [239] on the reaction yields, that correlate with the predictions' errors.

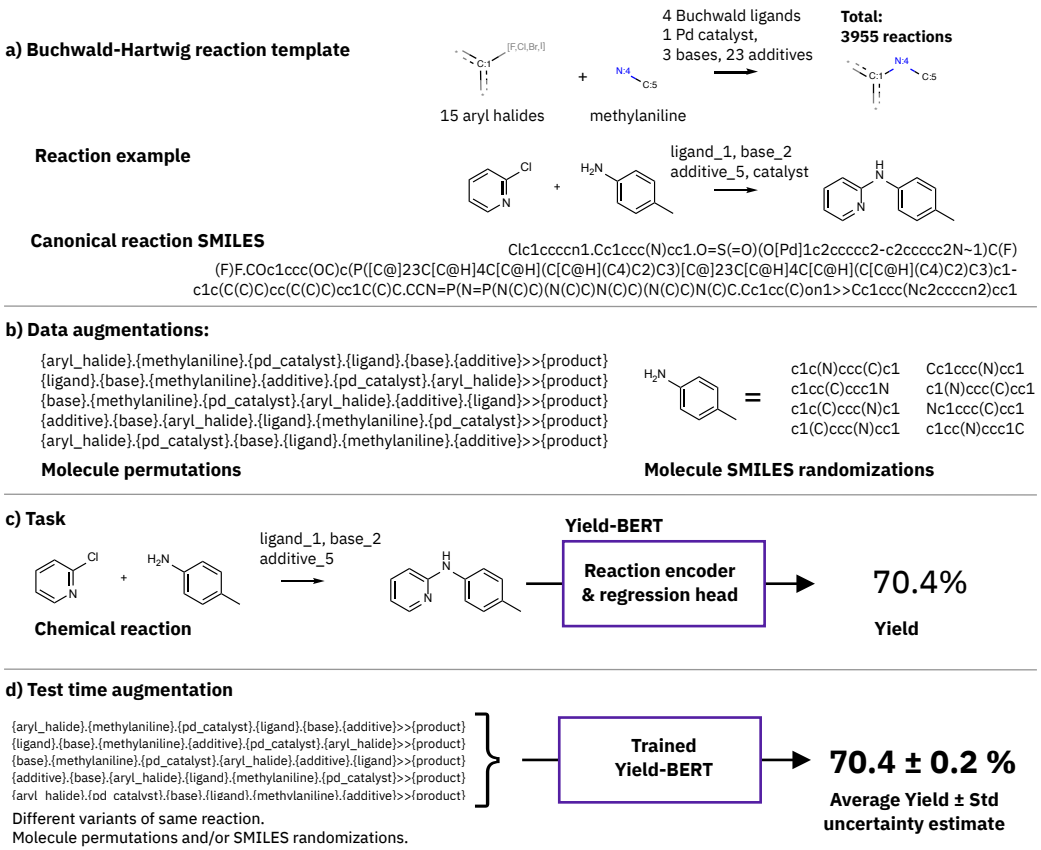

Figure 7.1: **Task overview.** Training/evaluation pipeline and task description.

7.2 Results

Our models were trained using Simpletransformers [229], Huggingface transformers [216], PyTorch [81] and scripts adapted from the RXN yields GitHub repository [203, 240]. Canonicalisations and augmentations were done using RDKit [153]. As described in the work of Schwaller et al. [203], fine-tuning a pretrained reaction BERT model [177] for a specific task provides the advantage of having most of the hyperparameters already optimised and fixed. Schwaller et al. [203] tuned only the dropout probability and the learning rate on the training data of the first random split, further split into a smaller training and validation set. Here, we initialised the dropout and learning rate using the values reported in [203] and we determined the optimal numbers of data augmentations using the same training/validation set. We investigated the two data augmentation techniques: molecule permutations, where we randomly shuffle the order of the precursors, SMILES randomisations, where we generated multiple randomised SMILES for a given molecule [238], and the combination of the two. Examples of augmented reactions and molecules are shown in Figure 6.1 b).

7.2.1 Yield prediction

Most of the results in the literature were published on 70%/30% (training/testing) random splits. In Table 7.1, we compared the results of the canonical order, the permuted precursors, the randomised SMILES and the combination of both permutation plus randomisation to previous studies [200, 203, 222, 223]. While the use of the canonical order SMILES representation plus BERT with a regression head [203] already outperforms one-hot encodings [223], physics-based descriptors [222] and multi-fingerprint features [200] plus a random forest regressor, here we significantly improve the R^2 score using randomisation. The same number of training augmentations, as stated in Table 7.1, was used throughout this work.

Table 7.1: **Random splits 70/30.** The results were averaged over 10 splits.

R^2	# samples/augmentations per rxn	mean	std
canonical	1	0.951	0.005
permuted	5	0.964	0.003
randomised	15	**0.970**	0.003
permuted & randomised (p&r)	15	**0.970**	0.003
MFF + RF [200]		0.927	0.007
DFT + RF [222]		0.92	
one-hot + RF [223]		0.89	

Moreover, we investigated the prediction performance on reduced training sets (Table 7.2), an experiment also performed by Ahneman et al. [222]. We observed that using SMILES randomisation, we outperformed all other approaches, using only 2.5% (or 98 data points). Although deep learning models are typically criticised as being data-hungry, our combination of a pretrained base-encoder [177] and data augmentation leads to accurate predictions in the small data regime.

Table 7.2: **Results in low-data regime.** Reduced training sets, averaged over 10 splits. Compared to the DFT-descriptor plus a RF model by Ahneman et al. [222].

R^2	canonical	permuted	**randomised**	perm & rand	DFT [222]
2.5% train	0.45 ± 0.05	0.47 ± 0.13	$\mathbf{0.61 \pm 0.04}$	0.57 ± 0.08	0.59
5% train	0.61 ± 0.04	0.70 ± 0.06	$\mathbf{0.74 \pm 0.03}$	0.71 ± 0.04	0.68
10% train	0.79 ± 0.02	0.81 ± 0.02	$\mathbf{0.81 \pm 0.02}$	0.81 ± 0.02	0.77
20% train	0.86 ± 0.01	0.87 ± 0.02	$\mathbf{0.89 \pm 0.01}$	0.89 ± 0.01	0.81
30% train	0.88 ± 0.01	0.90 ± 0.01	$\mathbf{0.92 \pm 0.01}$	0.91 ± 0.01	0.85
50% train	0.92 ± 0.01	0.94 ± 0.01	$\mathbf{0.95 \pm 0.01}$	0.95 ± 0.01	0.9

The data set of Ahneman et al. [222] also contains four out-of-sample splits, for which certain additives are only present in the test set. The results in Table 7.3 show that the models trained on canonical reaction SMILES without data augmentation perform best. For Test 4, the additives of the training set are the least representative of the ones in the test data. Therefore, the model trained on randomised SMILES, which better captures the patterns in the training data, unsurprisingly performs worse on that set.

Table 7.3: **Results on out-of-sample test splits.** Our results were averaged over 5 random seeds.

R^2	Test 1	Test 2	Test 3	Test 4	Avg.
canonical	$\mathbf{0.84 \pm 0.01}$	0.84 ± 0.03	$\mathbf{0.75 \pm 0.04}$	0.49 ± 0.05	$\mathbf{0.73 \pm 0.15}$
permuted	0.82 ± 0.01	$\mathbf{0.90 \pm 0.01}$	0.63 ± 0.05	0.43 ± 0.07	0.69 ± 0.19
randomised	0.80 ± 0.01	0.88 ± 0.02	0.56 ± 0.08	0.07 ± 0.04	0.58 ± 0.33
perm&rand	0.79 ± 0.09	$\mathbf{0.90 \pm 0.01}$	0.55 ± 0.05	0.27 ± 0.14	0.63 ± 0.26
MFF [200]	**0.85**	0.71	0.64	0.18	0.60
DFT [222]	0.8	0.77	0.64	**0.54**	0.69
OH [223]	0.69	0.67	0.49	0.49	0.59

7.2.2 Uncertainty estimation

We introduce test-time augmentation to provide an uncertainty estimation on our yield predictions. We input several data augmented versions of the same reaction and output the predicted yield as the average of the predicted yields using their standard deviation as the uncertainty estimate. Doing so does not significantly change the R^2 score. We measure the quality of the uncertainty estimates by computing the spearman's rank correlation coefficient (ρ) between absolute error and standard deviation of predicted yields, similar to the work by Hirschfeld et al. [241] on uncertainty quantification for molecular property predictions. The coefficient ranges between -1 and 1 and measures the monotonic relation between errors and uncertainty estimates. Figure 7.2) shows that ρ increases for all augmentation methods with the number of test-time augmentations and converges to values above 0.4. For the example plots in Figure 7.3 a) and Figure 7.3 b), we used the models trained on randomised SMILES and applied 10 test-time augmentations. In

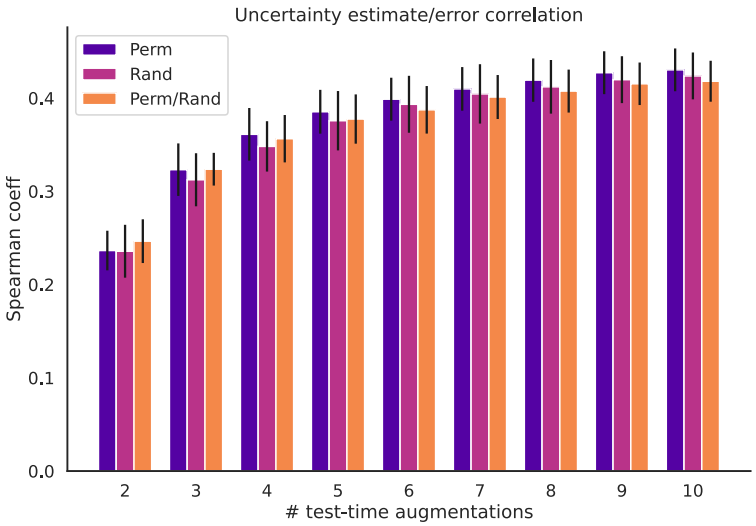

Figure 7.2: **Test-time augmentations for uncertainty estimation.** Spearman's rank correlation coefficient with increasing number of test-time augmentations.

Figure 7.3 a), we show how the predicted values get more certain and precise when increasing the data set from 2.5% to 70%. The out-of-sample test set plots in Figure 7.3 b) show that the uncertainty estimate correlates well with the error. Points with a larger error are generally more uncertain. Moreover, the models consistently predict a high yield for the reaction with the highest experimental yield independently of the split.

7.3 Discussion

In this manuscript, we presented augmentation strategies to increase reaction yield prediction using as input solely a text-based representation of chemical reactions. Even in a small data regime, a reaction BERT with regression head fine-tuned on randomised molecule representations was able to outperform physics-based descriptors plus random forest [222]. Although data augmentations result in worse performance for strongly dissimilar out-of-sample test reactions, we show that test-time data augmentations can provide uncertainty estimates without the need of model retraining. The uncertainty estimates correlate with the error of the predictions and could be used to guide the chemical space exploration [226, 242, 243, 244]. The code and 400 trained models to produce the results described in this work are available for download (`https://github.com/rxn4chemistry/rxn_yields`).

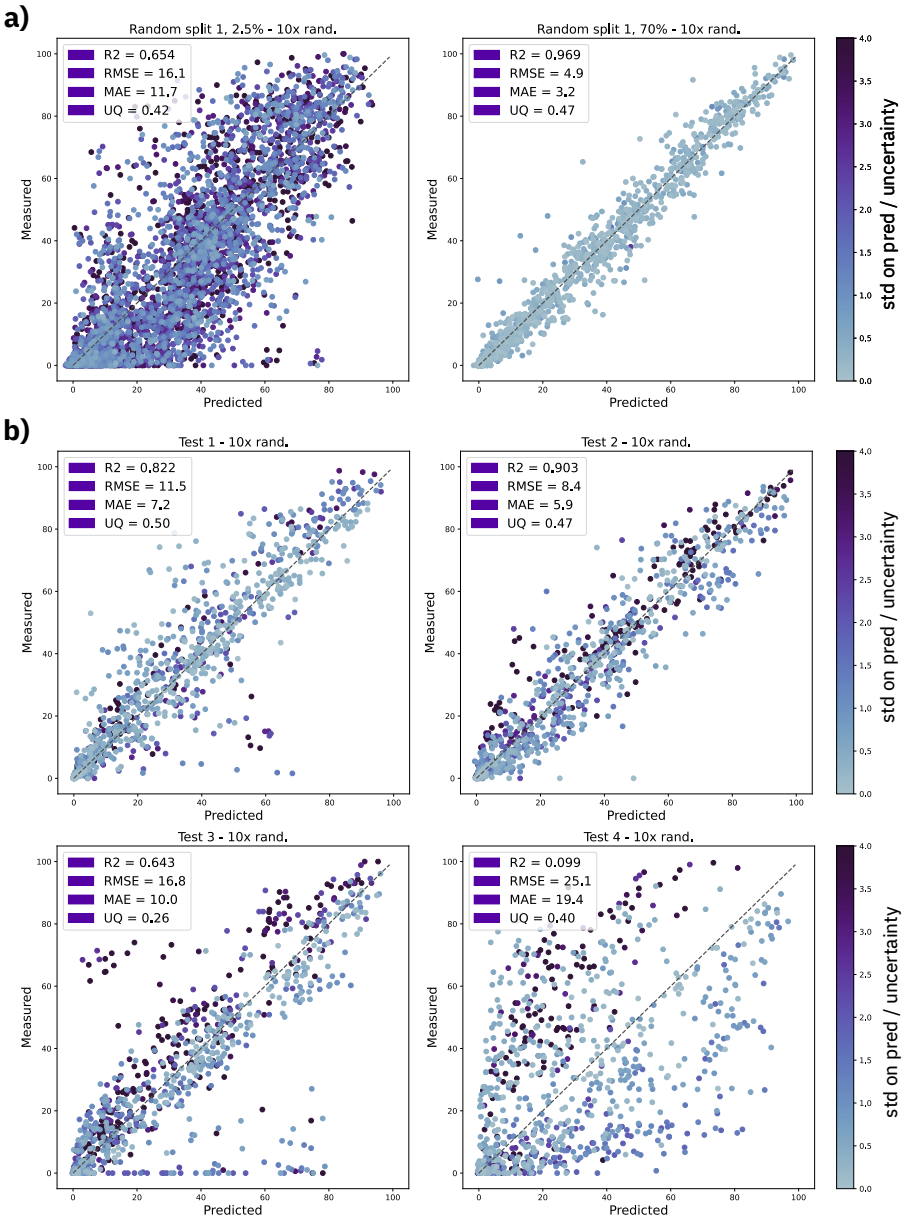

Figure 7.3: **Uncertainty estimation correlation examples.** a) Predictions and uncertainty on random split 01 with 2.5% and 70% training data using a fixed molecule order and 10 SMILES randomisations (randomised). b) Out-of-sample test set predictions using a fixed molecule order and 10 SMILES randomisations (randomised). Uncertainty scale was kept the same for all plots and capped at 4.0. MAE = mean average error, RMSE = root mean squared error, UQ = spearman's coefficient ρ.

8 Extraction of organic chemistry grammar from unsupervised learning of chemical reactions

Humans use different domain languages to represent, explore, and communicate scientific concepts. During the last few hundred years, chemists compiled the language of chemical synthesis inferring a series of "reaction rules" from knowing how atoms rearrange during a chemical transformation, a process called atom-mapping. Atom-mapping is a laborious experimental task and, when tackled with computational methods, requires continuous annotation of chemical reactions and the extension of logically consistent directives. Here, we demonstrate that Transformer Neural Networks learn atom-mapping information between products and reactants without supervision or human labelling. Using the Transformer attention weights, we build a chemically agnostic, attention-guided reaction mapper, and extract coherent chemical grammar from unannotated sets of reactions. Our method shows remarkable performance in terms of accuracy and speed, even for strongly imbalanced and chemically complex reactions with non-trivial atom-mapping. It provides the missing link between data-driven and rule-based approaches for numerous chemical reaction tasks.

This chapter has been presented as a scientific article at the ML Interpretability for Scientific Discovery workshop at ICML 2020 and was accepted in Science Advances:

P Schwaller, B Hoover, JL Reymond, H Strobelt, T Laino. Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. *Sci. Adv.*, 2021, 7, 15, eabe4166 (CC BY-NC). The visualisation tools were developed by Benjamin Hoover and Hendrik Strobelt.

8.1 Introduction

Humans leverage domain-specific languages to communicate and record a variety of concepts. Every language contains structural patterns that can be formalised as a grammar, i.e., a set of rules that describe how words can be combined to form sentences. Through the use of these rules, it is possible to create an infinite number of comprehensible clauses (knowledge) using a set of domain characteristic elements (words) obeying domain-specific rules (grammar and syntax). When applied to scientific and technical domains, a language is often more a method of computation than a method of communication.

Organic chemistry rules, for instance, have been developed over two centuries, in which experimental observations were translated into a specific language where molecular structures are words

and reaction templates the grammar. These grammar rules illustrate the outcome of chemical reactions and are routinely taught using specific diagrammatic representation (Markush representations). More convenient representations like reaction simplified molecular-input line-entry system (SMILES) [46] also exist for information technologies applied to synthesis planning and reaction prediction. In both Markush and SMILES representations, the grammar rules are present as latent knowledge in the historical corpus of raw reaction data.

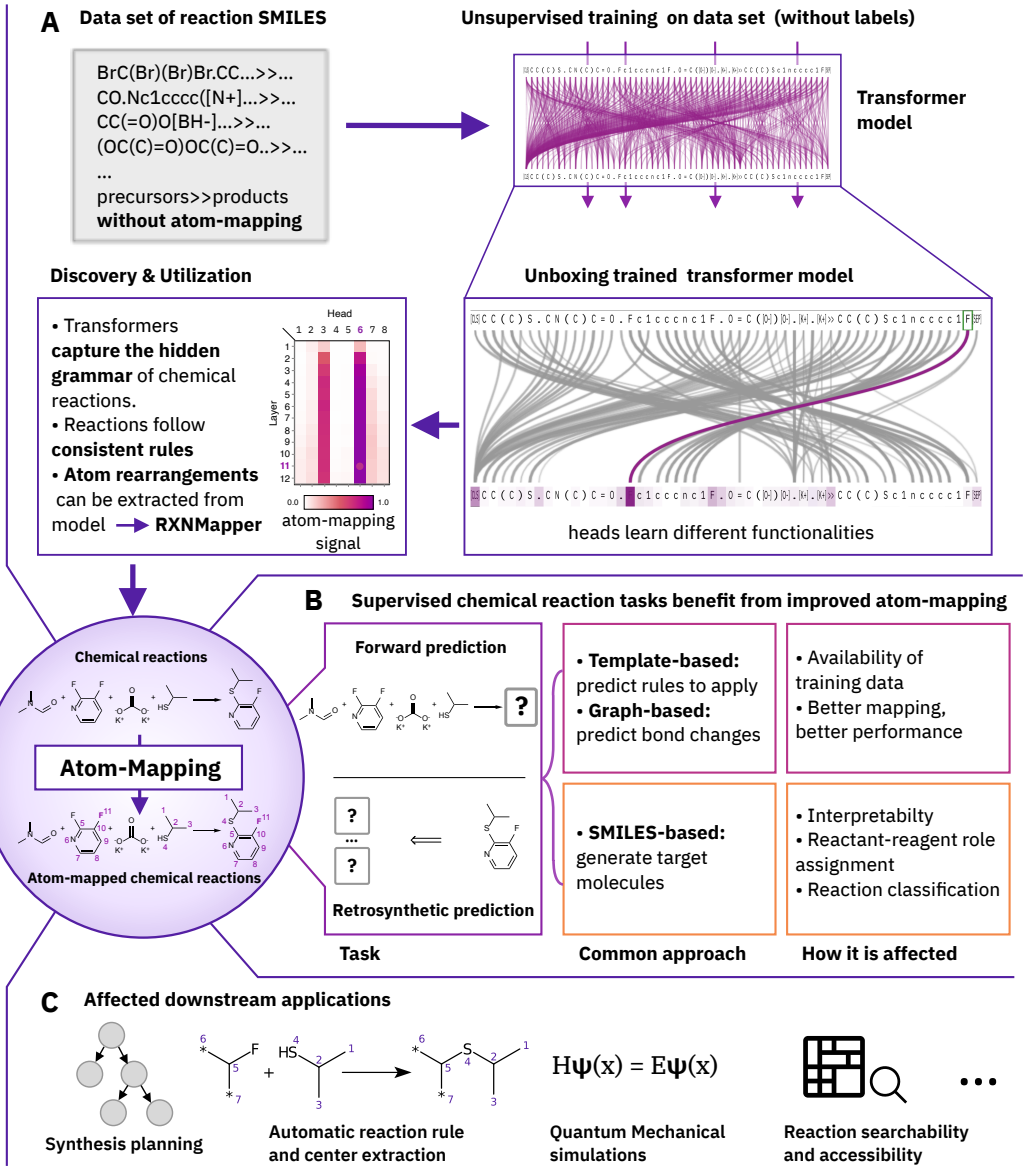

Figure 8.1: **Atom-mapping Overview.** (A) Process that led to the discovery of the atom-mapping signal and ultimately to the development of RXNMapper. (B) Directly affected chemical reaction prediction tasks. (C) Importance of atom-mapping in affected downstream applications.

The digitisation of these rules proved to be a successful approach to design modern computer programs[165] aiding chemists in synthetic laboratory tasks. Compiling reaction rules from domain data is tedious, requiring decades of labour hours and challenging to scale. The availability of an automatic and reliable method for annotating how atoms rearrange in chemical reactions, a process known as atom-mapping, could change profoundly the way organic chemistry is currently digitised. However, the process of atom-mapping is an NP-hard problem, dealt with computational technologies since 1970s [245, 246]. Most atom-mapping solutions are either structure-based [247, 248, 249, 250, 251, 252] or optimisation-based [253, 254, 255, 256, 257]. Most atom-mapping solutions are either structure-based or optimisation-based [245, 246]. The current state-of-the-art is a combination of heuristics, a set of expert-curated rules that precompute candidates for complex reactions, and a graph-theoretical algorithm to generate the final mapping as developed by Jaworski et al. [61]. Nonetheless, brittle preprocessing steps, closed-source code, computationally intensive strategies (more than 100 seconds for some reactions), and the need for expert-curated rules hinder its wider adoption. Most public reaction data comes with rule-based Indigo atom-maps [60], which are taken as ground-truth for subsequent work[5, 6, 35, 37, 113, 258], irrespective of the explicit warnings about atom-maps quality issues[42].

Natural language processing (NLP) models [151] are among the few neural network architectures showing a significant impact on synthetic chemistry [172] and not relying on atom-mapping algorithms. Their ability to encode latent knowledge from a training set of molecules and reactions represented as text (SMILES[46]) avoids the need to codify the chemical reaction grammar. Molecular Transformer models, a recent addition to the NLP family, are the state-of-the-art for forward reaction prediction tasks, achieving an accuracy higher than 90% [2, 4, 114, 116, 138]. Understanding the reasons for this performance requires the analysis of the neural network's hidden weights, which introduces the inherent complexity of interpreting neural networks.

Here, we report for the first time the evidence that Transformer encoder models [3, 98] learn atom-mapping as a key signal when trained on unmapped reactions on the self-supervised task of predicting the randomly masked parts in a reaction sequence, a process depicted in Figure 8.1 A. Transformer architectures can learn the underlying atom-mapping of chemical reactions, without any human labelling or supervision, solely from a large training set of reaction SMILES tokenised by atoms[4, 43]. After establishing an attention-guided atom-mapper and introducing a neighbour attention multiplier, we were able to achieve 99.4% correct full atom-mappings on a test set of 49k strongly unbalanced patent reactions [62] with high-quality atom-maps [63].

The advantage of this approach is its unsupervised nature. In contrast to supervised approaches, here the atom-mapping signal is learned during training as a consistent pattern hidden in the reaction data sets, without ever seeing any example of atom-mapped reactions. As a consequence, the quality of this approach is not limited by the quality of labeled data generated by an existing annotation tool. Moreover, the unsupervised nature allows to scale the extraction of chemical reaction grammar without the need of increasing human resources.

Numerous deep-learning methods developed for organic chemistry, like forward and backward reaction prediction, will benefit from better atom-mapping (8.1 B). From template-based approaches that use atom-mapping to automatically extract the templates from chemical reaction data sets [35, 107, 109, 214], to graph-based approaches, predicting bond changes or graph edits, that require atom-mapped reactions to extract the labels used for training the models [5, 6, 113, 258]. Even the predictions of atom-mapping independent and template-free SMILES-2-

SMILES approaches [4, 43, 116] may benefit from better atom-mapping, thus becoming more transparent and interpretable. In SMILES-2-SMILES approaches, the model generates the product structures sequentially atom-by-atom given the precursors or vice versa, generate the precursors given the product, without any support from atom-mapping information. After adding the atom-mapping in a post-processing step, predictions can be linked back to training reactions with the same reaction template. The atom-maps also enable the use of quantum mechanical simulations to compute reaction energies and the mechanism without human intervention by providing the corresponding atom pairs between precursors and products.

Moreover, our contributions will lead to improvements in the downstream applications that depend on better atom-mapping and chemical reaction rules (Figure 8.1 C): retrosynthesis planning methods [107, 109, 259], chemical reactivity predictions using graph neural network algorithms [6], reactant-reagent role assignments [62], interpretation of predictions [4], and knowledge extraction from reaction databases [260].

The attention-guided reaction mapper (henceforth referred to as RXNMapper) can handle stereochemistry and unbalanced reactions and is in terms of speed and accuracy the state-of-the-art open-source tool for atom-mapping, providing an effective alternative to the time-intensive human extraction of chemical reaction rules. We release RXNMapper together with the atom-mapped public reaction data set of Lowe [42] and a set of retrosynthetic rules[35, 107, 109, 214] extracted from it. The observed atom-mapping performance indicates that a consistent set of atom-mapping grammar rules exists as latent information in large data sets of chemical reactions, providing the link between data-driven/template-free and rule-based systems.

8.2 Results

Self-attention is the major component of algorithms called Transformers that are setting new records on NLP benchmarks, e.g., BERT [3] and ALBERT [98], and even creating breakthroughs in the chemical domain [4, 43, 177]. Transformers use several self-attention modules, called heads, across multiple layers to learn how to represent each token in an input – e.g., each atom and bond in a reaction SMILES – given the tokens around it. Each head learns to attend to the inputs independently. When applied to chemical reactions, Transformers use attention to focus on atoms relevant to understand important molecular structures, describe the chemical transformation, and detect useful latent information. Fortunately, the internal attention mechanisms are intuitive to visualise and interpret using interactive tools [129, 130, 261]. Through visual analysis, we observed that some Transformer heads learn distinct chemical features. Most strikingly, specific heads learned how to connect product atoms to reactant atoms, the process defined above as atom-mapping. We call these Transformer heads atom-mapping heads.

Throughout this work, our Transformer architecture of choice is ALBERT [98]. ALBERT's primary advantage over its predecessor BERT [3] is that it shares network weights across layers during training. This both makes the model smaller and keeps the functionality learned by a head the same across layers and consistent across inputs. Learned functions such as forward and backward scanning of the sequence, focusing on non-atomic tokens (ring openings/closures), and atom-mapping all perform similarly, irrespective of the input.

8.2.1 From raw attention to atom-mapping

To quantify our observations, we developed an attention-guided algorithm that converts the bidirectional attention signal of an atom-mapping head into a products-to-reactants atom-mapping. This specific mapping order ensures that each atom in the products corresponds to an atom in the reactants, which is important given that the most sizable open-source reaction data sets[41, 42] report only major products and show reactions that have fewer product atoms than reactant atoms.

The product atoms are mapped to reactant atoms one at a time, starting with product atoms that have the largest attention to an identical atom in the reactants. At each step, we introduce a neighbour attention multiplier that increases the attention connection from adjacent atoms of the newly mapped product atom to adjacent atoms of the newly mapped reactant atom, boosting the likelihood of an atom having the same adjacent atoms in reactants and products. This process continues until all product atoms are mapped to corresponding reactant atoms. Interestingly, the constraint of mapping only to equivalent atoms led to negligible improvements in terms of atom-mapping correctness, indicating that the model had already learned this rule in its atom-mapping function.

We selected the best performing model/layer/head combination after evaluation on a curated set of 1k patent reactions by Schneider et al. [62] originally mapped with the rule-based NameRXN tool [63]. We used the remaining 49k reactions as a test set. We consider the atom maps in NameRXN [63] to be of high quality because they are a side product of successfully matched reaction rules humanly designed. We used our best ALBERT model (total 12 layers, 8 heads) configuration (at layer 11, head 6, and multiplier 90) for RXNMapper.

8.2.2 Atom-mapping evaluation

The predominant use case for atom-mapping algorithms is to map heavily imbalanced reactions, such as those in patent reaction data sets[41, 42], or those predicted by data-driven reaction prediction models[4]. After training RXNMapper on unmapped reactions [42], we investigated the chemical knowledge our model had extracted by comparing our predicted atom maps to a set of 49k test reactions [62]. The majority (96.8%) of the atom-mappings matched the reference, including methylene transfers, epoxidations, and Diels-Alder reactions (Figure 8.2). We manually annotated the remaining discrepancies to discover edge cases where RXNMapper seemingly failed. A more careful analysis showed that out of the 1551 non-matching reactions, only 284 predictions were incorrect. In 415 reactions, RXNmapper gave atom-maps equivalent to the original (e.g, tautomers), and in 436, the atom-maps were better than the reference. In 369 cases, the original reaction was questionable and likely wrongly extracted from patents. For 47 reactions, the key reagents to determine the reaction mechanisms were missing. After removing questionable reactions from the statistics and counting the equivalent mappings as correct, the overall correctness increased to 99.4%.

Among the most frequent failures of RXNMapper, we find examples of wrong atom ordering in rings and azide compounds (Figure 8.2, (d)). In others, the model assigns wrong mappings to a single oxygen atom, like in reductions (Figure 8.2, (e)), or in Mitsunobu reactions (Figure 8.2, (f)), where the phenolic oxygen should become part of the product, but the model maps the primary

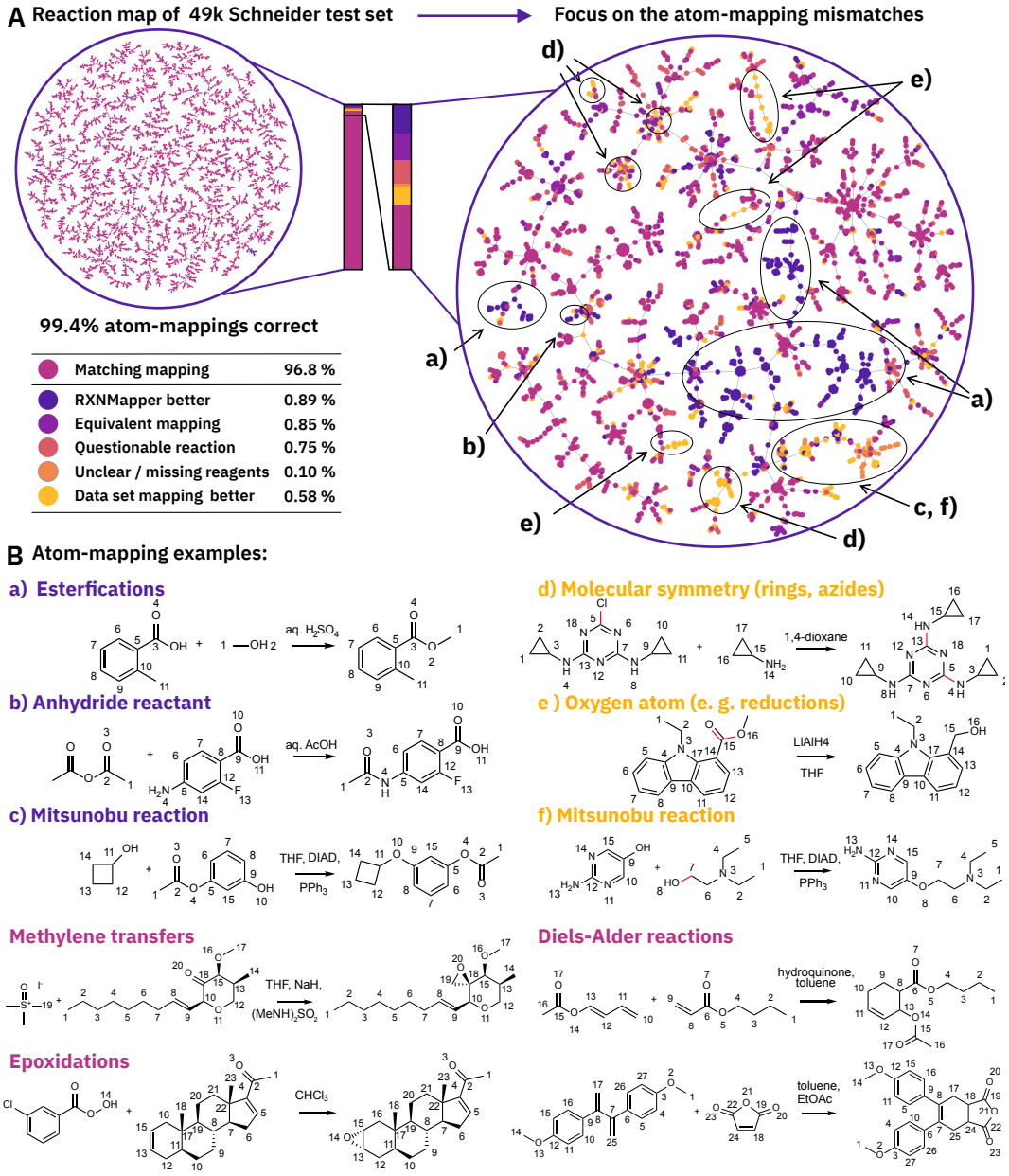

Figure 8.2: **Atom-mapping predictions.** (A) Visualising the results on the whole 49k Schneider test set with a focus on the mismatched atom-mappings (together with 1.5k matches for context) using reaction TMAPs [177, 201]. (B) Examples of atom-mappings generated by RXNMapper. Reactants and reagents were not separated in the inputs.

or secondary alcohol instead. We also observed counterexamples of Mitsunobu reactions (Figure 8.2, (c)) for which our model correctly mapped the reacting oxygen while the rule-based reference contained the wrong mapping as a result of the reaction not matching the Mitsunobu reaction rule. Although the overall quality of the reference atom-maps in the 49k test set [64]

is high, we were able to identify few important advantages of using RXNMapper instead of the rule-based mapped data set. RXNMapper correctly assigns the oxygen of the primary alcohols to be part of the major product for esterification reactions (Figure 8.2, (a)) like Fischer-Speier and Steglich esterifications as opposed to the annotated ground truth. It also correctly recognises anhydrides (Figure 8.2, (b)) and peroxides as reactants in acylation and oxidation reactions where the ground truth favored formic acid and water.

RXNMapper not only excels on patent reactions but performs remarkably well on reactions involving rearrangements of the carbon skeleton where humans require an understanding of the reaction mechanism to correctly atom-map. Striking examples include an intramolecular Claisen rearrangement used to construct fused 7-8 membered ring in the synthesis of the natural product micrandilactone A (Figure 8.3 a)[262, 263]), and the tandem Palladium-catalyzed semipinacol rearrangement / direct arylation used for a stereoselective synthesis of benzodiquinanes from cyclobutanols (Figure 8.3 b)[264]). In both cases, RXNMapper completes the correct atom mapping despite the entirely rearranged carbon skeletons resulting in different ring sizes and connections. ReactionMap, Marvin, ChemDraw and Indigo, failed at this atom-mapping task. RXNMapper also succeeds in atom mapping the ring rearrangement metathesis of a norbornene to form a bicyclic enone under catalysis by Grubbs-(I) catalyst (Figure 8.3 c)[265]). In this case, ChemDraw successfully completes the mapping, while the other tools failed. Furthermore, RXNMapper performs well with multicomponent reactions such as the Ugi 4-component condensation of isonitriles, aldehydes, amines and carboxylic acids to form acylated aminoacid amides (Figure 8.3 d), [266]). Here, RXNmapper maps all atoms correctly except for the carbonyl oxygen atom of the isonitrile derived carboxamide. RXNMapper assigns this oxygen atom to the oxygen atom of the carbonyl group of the aldehyde reagent, though this atom actually comes from the hydroxyl group of the carboxylic acid reagent. All other tools failed this atom-mapping task except for Mappet.

Similar to Jaworski et al. [61], we analyzed the atom-mapping in USPTO patent reactions according to the number of bond changes. RXNMapper performs better than Mappet [61] on all reactions except for those involving only one bond change. With an average time to solution of 7.7 ms/reaction on GPU accelerators and 36.4 ms/reaction on CPU, RXNMapper's speed is similar to the Indigo toolkit [60] on balanced reactions and far exceeds Indigo on unbalanced ones. As a comparison, Mappet [61] takes more than 10 seconds per reaction for 3.2% of their balanced test set reactions and for few of the reactions even more than 100 seconds per reaction. Additionally, RXNMapper outputs a confidence score for the generated atom-maps. An analysis of the confidence scores and more detailed comparisons are available in the supplementary materials.

The advantages of RXNMapper compared to the open-source Indigo [60] and the closed-source Mappet [61] are summarised in Table 8.1. RXNMapper is noticeably faster than other tools, handles strongly unbalanced reactions, performs well even on complex reactions and is open-source. It can also be used for compiling retrosynthetic rules, which are of crucial importance for several reaction and retrosynthesis prediction schemes like Chematica [165], in which a multitude of Ph.D. students and Postdocs across 15 years of continuous worked to extract reactions from literature and convert them into retrosynthetic rules. With unsupervised schemes such as RXNMapper, the extraction of retrosynthetic rules can be completed in a matter of weeks, with little human intervention. We demonstrate that by atom-mapping the entire USPTO data sets and by extracting the retrosynthetic rules using the approach described by Thakkar et al. [109].

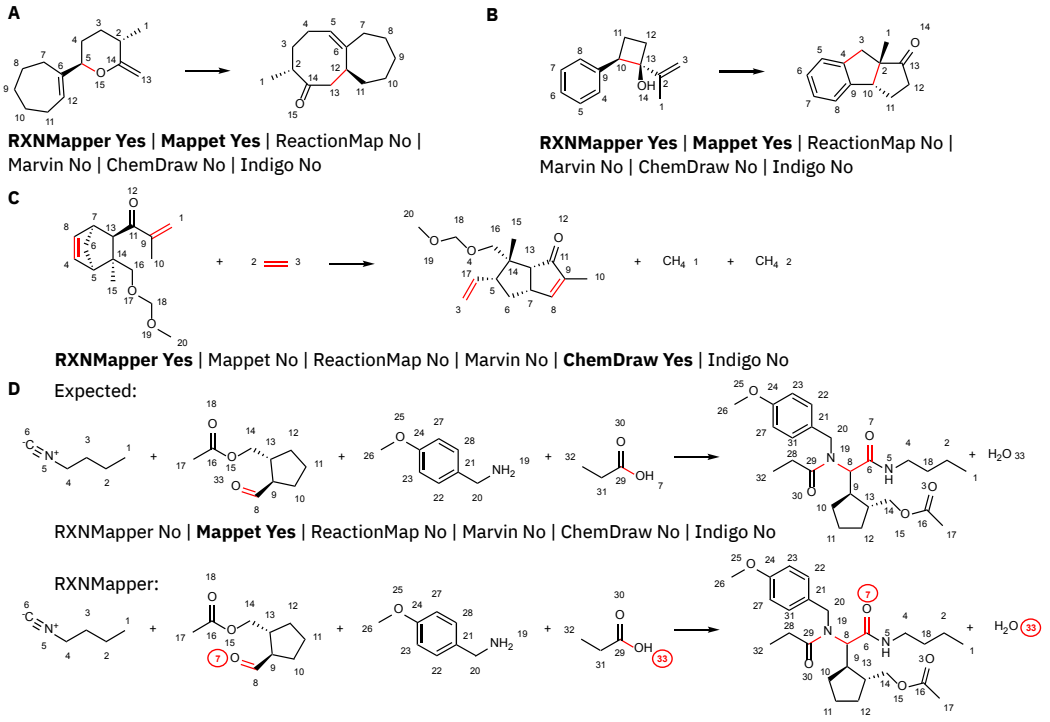

Figure 8.3: **Atom-mapping examples.** Examples and results for commercially available tools from the complex reactions data set by Jaworski et al. [61]. (A) Bu3Al-promoted Claisen rearrangement [262, 263] (B) Palladium-Catalyzed Semipinacol Rearrangement and Direct Arylation [264]. (C) Grubbs-catalyzed ring rearrangement metathesis reaction [265] (D) Ugi reaction [266]

We make available the corresponding atom-mappings of the USPTO data set and the 21k most frequently extracted retrosynthetic rules along with the most commonly used reagents, the corresponding patent numbers, and the first year of appearance. The application of unsupervised schemes demonstrates the feasibility of running a completely unassisted construction of retrosynthetic rules in just a few days – three orders of magnitude faster than previous human curation protocols. The use of unsupervised schemes will facilitate the compilation of new retrosynthetic rules in existing rule-bases systems.

8.3 Discussion

We have shown that the application of unsupervised, attention-based language models to a corpus of organic chemistry reactions provides a way to extract the organic chemistry grammar without human intervention. We unboxed the neural network architecture to extract the rules governing atom rearrangements between products and reactants/reagents. Using this information, we developed an attention-guided reaction mapper that exhibits remarkable performance in both speed and accuracy across many different reaction classes. We showed how to create a state-of-the-art atom-mapping tool within two days of training without the need for tedious and potentially

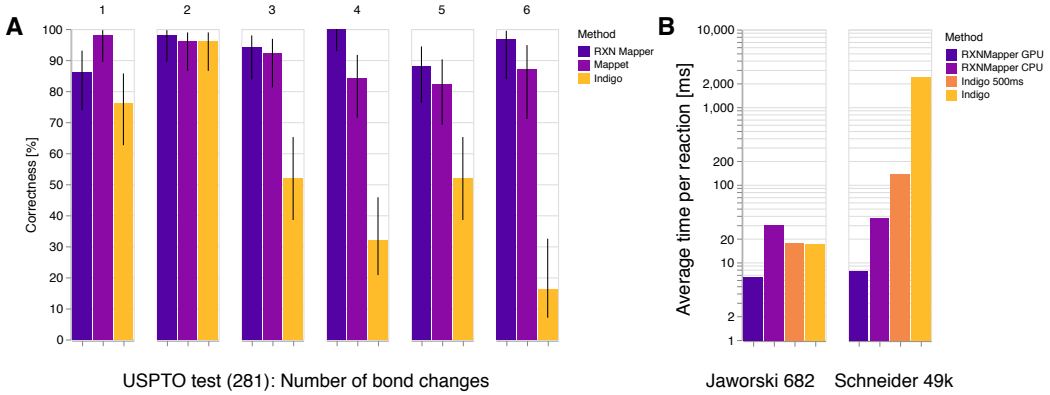

Figure 8.4: **Comparison with other tools.** (A) Comparison of RXNMapper, Mappet [61], and the original Indigo mapping from the USPTO data set (281 reactions). The error bars show the Wilson confidence interval [267]. (B) Mapping speed comparison between RXNMapper and Indigo [60], which is orders of magnitude faster than Mappet [61]. For Indigo 500ms, we set a timeout of 500 ms, after which the tool would return an incomplete mapping. We averaged the timing on the imbalanced reactions for Indigo without timeout on 20k reactions.

	RXNMapper	Indigo [60]	Mappet [61]
Avg time (short)	6.4 ms	17.0 ms	Slower than Indigo
Avg time (strongly unbalanced)	7.7 ms	2400 ms	Not handled
Quality on complex reactions	High	Low	High
Quality on strongly unbalanced reactions	High	Low	–
Open Source code?	Yes	Yes	No

Table 8.1: **Comparison of different atom-mapping tools.**

biased human encoding or curation. Because the entire approach is completely unsupervised, the use of specific reaction datasets can improve the atom-mapping performance on corner cases. The resulting atom-mapping tool is significantly faster and more effective than existing tools, especially for strongly imbalanced reactions. Finally, our work provides the first evidence that unannotated collections of chemical reactions contain all the relevant information necessary to construct a coherent set of atom-mapping rules. Numerous applications built on atom-mapping will immediately benefit from our findings [6, 107, 109, 113], and others will become more interpretable exploiting the potential of unsupervised atom-mappings [4, 43].

The use of symbolic representations and the means to learn autonomously from rich chemical data led to the design of valuable assistants in chemical synthesis[172]. A strengthened trust between human and interpretable data-driven assistants will spark the next revolutions in chem-

istry, where domain patterns and knowledge can be easily extracted and explained from the inner architectures of trained models.

8.4 Methods

Transformers

Transformers are a class of deep neural network architectures that relies on multiple and sequential applications of *self-attention* layers [2]. These layers are composed of one or more *heads*, each of which learns a square attention matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$ of weights that connect each token's embedding Y_i in an input sequence Y of length N to every other token's embedding Y_j. Thus, each element $\mathbf{A}_{ij}$ is the attention weight connecting Y_i to Y_j. This formulation makes the attention weights in the Transformer architecture amenable to visualizations as the curves connecting an input sequence to itself, where a thicker, darker line indicates a higher attention value.

The calculation of the attention matrix of each head can be easily interpreted as a probabilistic hashmap or lookup table over all other elements Y_j. Each head in a self-attention layer will first convert the vector representation of every token Y_i into a key, query, and value vector using the following operations:

$$K_i = \mathbf{W}_k Y_i \qquad Q_i = \mathbf{W}_q Y_i \qquad V_i = \mathbf{W}_v Y_i \tag{8.1}$$

where $W_k \in \mathbb{R}^{d_k \times d_e}$, $W_q \in \mathbb{R}^{d_k \times d_e}$, and $W_v \in \mathbb{R}^{d_v \times d_e}$ are learnable parameters. $\mathbf{A}_i$, or the vector of attention out of token Y_i, is then a discrete probability distribution over the other input tokens, and it is calculated by taking a dot product over that token's query vector and every other token's key vector followed by a softmax to convert the information into probabilities:

$$\mathbf{A}_i = \text{softmax}\Big(\frac{Q_i(\mathbf{W}_k Y^\top)}{\sqrt{d_k}}\Big). \tag{8.2}$$

Note that one can define input sequence Y as an $N \times d_e$ matrix and matrix $\mathbf{W}_k$ as a $d_k \times d_e$ matrix, where d_e is the embedding dimension of each token and d_k is the embedding dimension shared by the query and the key.

Each head must learn a unique function to accomplish the masked language modelling task, and some of these functions are inherently interpretable to the domain of the data. For example, in Natural Language Processing (NLP), it has been shown that certain heads learn dependency and part of speech relationships between words [231, 268]. Using visual tools can make exploring these learned functions easier [130].

Model details

For our experiments, we used PyTorch (v1.3.1) [81] and huggingface transformers (v2.5.0) [216]. The ALBERT model was trained for 48 hours on a single Nvidia P100 GPU with the hyperparameters stated in the supplementary information. Schwaller et al. [4] developed the tokenisation regex used to tokenise the SMILES. We expect further performance improvements when using more extensive data sets (e.g., commercially available ones). The RXNMapper model uses 12 layers, 8 heads, a hidden size of 256, an embedding size of 128, and an intermediate size of 512. In

contrast to ALBERT base [98] with 12M parameters, our model is small and contains only 770k trainable parameters.

Data

The work by Lowe [42] provides the data sets used for training, composed of chemical reactions extracted from both grants and patent applications. We removed the original atom-mapping from this dataset, canonicalised the reactions with RDKit [153], and removed any duplicate reactions. The data set includes reactions with fragment information twice, once with and once without fragment bonds, as defined in the work of Schwaller et al. [43]. The final training set for the masked language modelling task contained a total of 2.8M reactions. For the evaluation and the model selection, we sampled 996 random reactions from the Schneider et al. [62] data set.

To test our models, we first used the remaining 49k reactions from the Schneider50k patents data set [62]. We do not distinguish between reactants and reagents in the inputs of our models. We also used the human-curated test sets that were introduced by Jaworski et al. [61] to compare our approach to previous methods. Table 8.2 shows an overview of the test sets. Note that patent reactions differ from the reactions in Jaworski et al. [61] because the latter removes most reactants and reagents in an attempt to balance the reactions.

Test set	Number of reactions	Avg. number of reactant atoms	Avg. number of product atoms
Simple reactions [61]	100	27.1	27.1
Typical reactions [61]	100	19.9	19.6
Complex reactions [61]	201	25.7	24.8
USPTO bond changes [61]	281	26.0	23.7
Schneider50k test [62]	49000	43.3	26.1

Table 8.2: Data sets used for testing

Attention-guided atom-mapping algorithm

The attention-guided algorithm relies on the construction of the attention matrix for a selected layer and head, where we sum the product-to-reactant and the corresponding reactant-to-product atom attentions. Algorithm 1 provides the exact atom-mapping algorithm. By default, after matching a product-reactant pair, the attentions to those atoms are zeroed. Optionally, atoms in product and reactants can have multiple corresponding atoms. We always mask out attention to atoms of different types.

Atom-mapping curation

Chemically equivalent atoms exist in many chemical reactions. Most of the chemically equivalent atoms could be matched after canonicalising the atom-mapped reaction using RDKit [153, 269].

Algorithm 1: Attention-guided atom-mapping algorithm

Data: Reaction SMILES S, multiplier W, model M
Result: Product $\rightarrow$ reactant atom-mapping P
begin
  $A \longleftarrow$ M(S) `// compute attention matrix`
  for $i \in range(len(P))$ `// iterate through product atoms`
  do
    Mask invalid atoms (not same type; optionally, already mapped)
    Select i, j pair with highest attention A_{ij}
    if $A_{ij} \neq 0$ **then**
      $P_i \longleftarrow$ j `// Map product atom` i `to reactant atom` j
      multiply attention of adjacent atoms of i to adjacent atoms of j by W
        `// Increase neighbour attentions`
    else
      $P_i \longleftarrow -1$ `// No corresponding reactant atom`
      break

Exceptions were atoms of the same type connected to another atom with different bond types, which would form a resonance structure with delocalised electrons. We manually curated these exceptions and added them as alternative maps in the USPTO bond changes test set [61].

Data and code availability

All our generated atom-mappings, including those for the largest open-source patent data set [42], the unmapped training, validation, and test set reactions, can be found in the following repository https://github.com/rxn4chemistry/rxnmapper. The code is available at https://github.com/rxn4chemistry/rxnmapper and a demo at http://rxnmapper.ai.

9 Conclusion and outlook

This chapter first summarises the main contributions of the thesis and then provides an outlook of the remaining challenges and opportunities.

9.1 Summary of the contributions

In this thesis, I made use of the similarities between written human language and organic chemistry to build linguistics-inspired tools that help chemists to accelerate chemical synthesis. More specifically, I developed transformer-based models for different chemical reaction tasks. With encoder-decoder transformers [2, 4] I approached forward reaction prediction and multi-step synthesis planning and with encoder-only transformers [3, 98] reaction classification, fingerprints, yield prediction and atom-mapping.

In chapter 3, I focused on the limitations of previous reaction prediction models and I investigated stereo- and regioselective reaction in a small carbohydrate reaction data set. I showed that using transfer learning, I could overcome some of the weaknesses of previous models. By specialising a Molecular Transformer model on carbohydrate reactions using either multi-task and sequential transfer learning, the model could learn to predict transformations challenging for chemists and models alike. Not only did the transfer learning approach improve on the in distribution test set from 43.3% with the baseline model to 71.2%, also on smaller out-of-distribution similar improvements were observed. One of those test sets consisted of an unpublished synthesis of a lipid-linked oligosaccharide. My results show that the open-source reaction data is enough to be leveraged to train models that perform well on more specific and challenging reaction subspaces, where less data is available.

In chapter 4, I discussed the approach behind the multi-step retrosynthesis tool in the IBM RXN for Chemistry platform [44], which uses two Molecular Transformers. The first is trained on forward reaction prediction and used to score suggestions by the second model, which suggest different reactant and reagent combination that might lead to the target product or a non-commercially available molecule required for the synthesis. I reported the ineffectiveness of using a top-N accuracy to optimise single-step retrosynthesis models in the multi-step setting, which led to the introduction of four newly designed metrics: coverage, class diversity, round-trip accuracy, and Jensen-Shannon divergence metrics. The newly defined metrics improved the comparison with the observations of my experimental collaborators and made it possible an effective optimisation of single-step retrosynthesis models for a multi-step setting. The Jensen-Shannon divergence metric was recently revised making it cumulative and non-parametric [190]. Unlike other multi-step retrosynthesis tools, my approach not only predicts the largest fragments, synthons or reactants, but also reagents simultaneously. To date, it is still the only data-driven atom-mapping independent retrosynthesis approach. My models are freely accessible through the IBM RXN

for Chemistry platform in a fully automated mode and in a interactive mode. In the interactive mode, human chemists get suggestions by the AI models and then, select the most suitable given their expertise. Hence, the design of synthesis routes becomes like a human-AI interaction game. Such collaborative procedures could facilitate the adoption of machine learning tools by synthetic chemists [270].

After introducing approaches for forward reaction prediction in the low-data regime and a multi-step synthesis planning using black-box models, in chapter 5, I made the predictions of those models more explainable by developing a reaction classification model. Synthetic chemists commonly use reaction classes to communicate reaction characteristics efficiently. My transformer-based models were trained on predicting reaction classes from reaction SMILES without reactant-reagent separation and can directly be applied to the outputs of the models in chapter 3 and 4. On a test set from the Pistachio [180] data set originally classified with rule-based NameRXN tool [63], the best classifier achieved an accuracy of 98.9%. My models reached the same performance on the USPTO 1k TPL classification data set, which I derived from the open-source USPTO data [41, 42, 59]. The classification models are used in the IBM RXN for Chemistry [44] to group similar reactions in the individual steps of predicted synthesis routes. Based on encoder-only transformer classifiers, I introduced atom-mapping independent reaction fingerprints. My reaction fingerprints are available in open-source and enable efficient similarity searches and reaction clustering.

Economic, logistic and energetic considerations motivate chemists to optimise reactions to convert most of the reactants into the desired product. In chapters 6 and 7, I investigated reaction yield prediction models. I used the yields of two high-throughput experiment data sets[222, 228] and the yields data extracted from USPTO [41, 42, 59] to develop models that predict yields using canonical reaction SMILES as input in chapter 6. While my approach worked well on reaction data originating from the same source, as in the high-throughput experiment, the USPTO yield data turned out to be too noisy for accurate predictions. In chapter 7, I focused on the best-studied high-throughput experiment data set containing Buchwald–Hartwig reactions [222]. Using data augmentation techniques, I showed that the linguistics-inspired models could consistently outperform methods using physics-based descriptors, even in the low-data regime. Moreover, I proposed a novel way of estimating epistemic uncertainty through test-time augmentation.

Finally, in chapter 8, I investigated what encoder-only transformer models learn while being trained on chemical reaction data with a self-supervised mask language modelling task. Based on a visual inspection of the attention weights and an analysis of the functionalities that different heads had learned, I discovered an atom-mapping pattern consistently present in at least one head in all trained models. My models managed to capture the hidden grammar of chemical reactions without explicitly being told to do so. Based on the attention weights of the atom-mapping head, I developed an atom-mapping tool called RXNMapper. RXNMapper efficiently produces high-quality atom-maps even on strongly imbalanced reaction equations and chemically complex reactions. Using the atom-mapping generated with RXNMapper, I extracted consistent reaction rules from unlabelled chemical reaction data sets. The open-source RXNMapper was recently selected as the best atom-mapping tool in a benchmarking study conducted by an independent group [271] - even better than commercially available tools. This result is remarkable as the models learned underlying atom-mapping signal without supervision or human labelling. Moreover, RXNMapper was used to improve the reactant-reagent split in the open-source USPTO reaction data [272] show-casing its immediate impact on the community.

The challenges I addressed, including prediction of carbohydrate reactions, multi-step synthesis planning, and atom-mapping, went beyond simple regression tasks. I showed that natural language processing models could learn chemical knowledge from text-based representations of molecules and chemical reactions. Hence, predicting chemical reactivity, which long was reckoned to be an art only human experts could predict, has become within reach of data-driven learning systems. Starting from my pretrained models and recipes, presented in chapter 3, the fine-tuning of a specific Molecular Transformer can be done in a few hours on any reaction subspace of interest. Similarly, new single-step retrosynthesis models from 4 could be trained using transfer learning and integrated multi-step synthesis planning tools to extend their applicability domain. The individual predictions can not only be classified into reaction classes with the models introduced in 5, they can easily be linked back to the most similar reactions in the training data using my reaction fingerprint. This procedure can give chemist direct access to additional metadata like the patent numbers, reaction procedures, and conditions of similar reactions and explain the predictions. Further information, such as reaction centres, reaction rules and molecules' roles (reactant/reagent), can be obtained by analysing the atom-maps generated by RXNMapper developed in 8. Hence, I overcame some of the major limitations of entirely data-driven reaction prediction and retrosynthesis approaches. Through my efforts in explainability, the IBM RXN platform and the RXNMapper (`http://rxnmapper.ai`), chemical reaction language models became increasingly approachable for chemists.

9.2 Outlook

The demonstrated advances in machine learning for organic synthesis were made possible through powerful hardware, new machine learning algorithms and frameworks, and open data availability. The last part of the conclusion is dedicated to open challenges and opportunities in machine learning for chemical reactions. I will briefly discuss data quality, chemical representations, the broader adoption of machine learning models in chemistry and synthesis automation.

9.2.1 Data quality

Although our transformer-based models have shown to cope relatively well with noisy chemical data, improvements can be expected once better quality data becomes available. As pointed out in chapter 2, there are numerous steps from the bench chemists performing the reaction and reporting it in an ELN to the reaction ending up in a reaction database. An inaccurate description of the procedure, typos in an IUPAC name or an erroneous text-mining and conversion can introduce noise or completely invalidate a reaction. Even partly human-curated databases like Reaxys [39] have quality issues. Due to their reaction format, where only reactants and products are described in SMILES, extensive preprocessing is required to make their data valuable. If at all, reaction conditions and yields are recorded in non-standardised formats.

An example of the impact of the knowledge represented in the training data comes from the work of Kovacs et al. [273], who observed that Friedel-Crafts reactions, reported in the open-source USPTO data, predominantly result in a substitution of the hydrogen in the para position. As a consequence, data-driven models trained on that data will favour this substitution over the substitution in the meta position, independent of the functional groups, and hence, can easily be

fooled with underrepresented meta-directing groups. Another weakness of data-driven models originates from incomplete reactions. The palladium catalysts are missing in some of the Suzuki-coupling reactions in the data set, but present in others. Therefore, the models will learn that the catalysts are not important for the reaction to occur and ignore them.

Recently, myself and coworkers [190] suggested an approach how to automatically clean reaction data without human intervention using forgetting events. In another work, I took an orthogonal approach [274], where a model automatically completes information missing from corrupted reaction equations. While those approaches can improve the quality of existing data set to a certain extend, there is the urgent need for a better chemical data publication pipeline. Many of the error-prone steps could be circumvented if the data would be recorded in ELNs from the start and submitted in a standardised format as part of a scientific publication. However, the definition and wide adoption of standards are challenging. Currently, the Open Reaction Database [275] is being developed, which could be a key step forward towards better reaction data and standardisation. Moreover, the publication of low-yielding and failed reactions would enable the development of models that would learn from negative examples. To date, reaction data sets are heavily biased towards frequently used and successful reactions. An effort, similar to the one in the computational material science community with projects like Materials Cloud [276], NOMAD [277] and the Materials Genome Initiative [278], is required to develop distributed solutions, to agree on chemical reaction representation standards, and make data accessible and reusable [279]. Open-access publications in organic chemistry accompanied by machine-readable experimental data, including information on failed experiments, would enable the development of better performing data-driven models.

9.2.2 Better molecular and reaction representations

The approaches presented in this thesis were based on a text-based molecule, and reaction representations called SMILES [46, 47]. However, sterics and the 3-dimensional shape of molecules, which may play a crucial role in chemical reactivity are poorly captured by the SMILES representation. Similarly, molecular graph representations, which have a more substantial inductive bias on covalent bonds, fail to correctly represent the 3D information, as long as bond angles and lengths, are not determined. Future studies should focus on better molecular representations. An interesting approach could be the usage 3D-roto equivariant neural networks [280], which, for example, lead to ground-breaking results in protein structure prediction [281]. Besides more accurate predictions, 3D-roto equivariant neural networks could potentially lead to improved reaction fingerprints when applied to the individual molecules in a reaction. However, Cartesian coordinates introduce other challenges, for example, the one of correctly determining the right conformers [282]. Another representation challenge is that current cheminformatics tools, like RDKit [153], have only limited support for non-covalent bonds in molecules. Those bonds are particularly important for organometallic complexes, which are often used as catalysts in organic synthesis. Already a standardised canonical representation of organometallic compounds could help. At the moment, the same metal catalysts often exist in multiple variants in the same databases making it more difficult for data-driven models to learn their effects.

For chemical reactions, I made a first step towards enabling efficient quantum simulations for holistic transition state and energy barrier calculations with RXNMapper [189]. Knowing the

atom-mapping and therefore, the bond changes occurring during a reaction could significantly reduce the number of possible transition states that have to be considered. Frameworks like ReactionPredictor [30], Reaction Mechanism Generator [283, 284], global route reaction mapping [285, 286], autodE [287] could be coupled with machine learning-based synthesis planning tools, automatically calculate reaction profiles and guide the selection process towards more favourable routes [202, 288]. In this regard, it would be practical to predict balanced reaction equations with full product information and not only the major product.

9.2.3 Education and collaborative approaches

Eventually, machine learning tools have to be applied in practice. Performing in silico experiments on overly simplified and unrealistic data sets, as it is often done to compare machine learning approaches is only of limited interest. Methods should either be validated experimentally or made easily accessible to the researchers that can best benefit from them. To date, only a few research groups have both a strong synthesis and machine learning expertise. For synthetic chemists, even when open-sourced, downloading code from GitHub and spending time to make the programs work on their chemical subspace of interest is seldom an option. Platforms such as IBM RXN for Chemistry [44], and ASKCOS [125] are good examples of how a wider adoption can be facilitated. On those platforms, chemists can use familiar tools to draw molecules as inputs to fully trained machine learning models and get back predictions. They can then assess the predictions to get a feeling of how useful the implemented machine learning models could be for them. Examples shown in scientific papers can be exciting but are frequently cherry-picked and not statistically relevant.

Machine learning researchers should communicate with synthetic chemists and learn how to make their approaches more relevant for solving real-world problems. Too often, machine learning researchers are only interested in the beauty of the algorithm and simplify the task at hand to make it work with their algorithm, ignoring the practical end-use completely. Researchers understanding the challenges in chemistry and machine learning will make the most significant contributions to the field. Along the same lines, it would be great if the next generation of synthetic chemists and material scientists were educated in programming and machine learning. Hence, they could better understand the advantages and limitations of different machine learning approaches, integrate them into their workflow, and tap the machine learning tools' full potential. As Derek Lowe [289] once said *"It is not that machines are going to replace chemists. It's that the chemists who use machines will replace those that don't."* Ground-breaking work could originate from close collaborations between synthetic chemists and machine learning researchers. Generic machine learning models that can easily be adapted to specific chemists needs, for example, through few-shot learning [96] could be particularly useful to achieve this goal.

9.2.4 Automated synthesis

The rise in automation and robotic platforms for synthesis that we are currently witnessing is expected to impact the quality of the produced data profoundly [224, 290, 291, 292]. Automation comes with the advantages of being more efficient, less error-prone than human labour, and more reproducible. Moreover, all reaction conditions and parameters, such as temperature or pressure, can be recorded and ideally made available in a machine-readable format.

Most automation studies, such as, for example, the work from which I obtained the data for the yield prediction models [222, 228], were restricted to a well-defined chemical subspace. Ahneman et al. [222], and Perera et al. [228] restricted it to a reaction type described with one reaction template and a fixed number of precursors. The less constrained the search space and more flexible the automation platform, the more extensive the possibility for discoveries. Recently, more modular discovery-oriented platforms were studied. For instance, Coley et al. [125] developed reconfigurable flow apparatus. Although it was connected to a retrosynthesis prediction algorithm, it still required human input to determine reagents and amounts. Steiner et al. [293] designed a language to describe modular synthesis operations and automate bench chemistry. Another modular linear approach was investigated by Bédard et al. [294]. Chatterjee et al.[295], instead, favoured a more versatile radial synthesis approach, where the modules were placed around a central switching station.

Going towards autonomous systems, myself and coworkers [235] recently introduced models that could infer all actions required for a robot (or human alike) to run a reaction only from its chemical equation. Paired with the retrosynthesis approach presented in chapter 4, the models lead to successful automatic syntheses on the IBM RoboRXN platform [296]. This development was made possible by previous work, where synthesis procedures were converted into structured synthesis actions [188]. Similarly, Mehr et al. [297], presented an approach for converting reaction procedures into synthesis actions. Although they validated some converted procedures by executing the reactions on a robotic platform, the predicted procedures still required human-made modifications before the execution. Unlike previous approaches, Burger et al. [298] developed a platform-independent mobile robotic system. Focusing on a narrow chemical space, the so-called robotic chemist autonomously performed 688 experiments over eight days searching for photocatalyst mixtures. When such systems become more affordable, scalable and reliable, the productivity and discovery in synthetic chemistry could tremendously increase [299].

With improving automation and data collection, one challenge we will face will be the combination of data from different sources with varying noise levels to best guide exploration of chemical space. Nevertheless, the potential of a feedback loop between automation platforms or mobile robotic chemists and data-driven models is enormous and will likely revolutionise the way chemistry is done today.

A Appendix: Recent Data-Driven Learning Systems for Chemical Reactions

Different neural network-based chemical reaction prediction approaches up to 2019 are shown in Table A.1.

Table A.1: **Different chemical reaction prediction approaches.** Comparison of the input, output, data and model architecture of the data-driven reaction prediction approaches analysed in this work.

	Input	Output	Data	Model
Kayala et al., Fooshee et al. [29, 30, 34]	Atomic and molecular features, reaction conditions	Electron sources/sinks, mechanistic steps	Generated using rules	Feed-forward neural network
Wei et al. [31]	Neural fingerprint of 2 reactants + 1 reagent	Template ranking (16 rules)	Generated using rules	Feed-forward neural network
Segler and Waller [32]	Extended-connectivity fingerprints	Links in knowledge graph	Binary reactions from Reaxys	Graph reasoning
Segler and Waller [33]	Extended-connectivity fingerprints	Template ranking (8820 rules)	Extracted from Reaxys	Feed-forward neural network
Coley et al. [35]	Edit-based, applying templates	Product ranking	USPTO-15k (15k random reactions)	Feed-forward neural network
Jin et al. [5]	Molecular graph	Bond changes	USPTO_MIT data set (480k reactions)	Graph Convolutional Neural Network
Schwaller et al. [36]	SMILES, separated reagents	Product molecule generation	USPTO_MIT, USPTO_STEREO (1M reactions)	Seq2Seq model with attention
Bradshaw et al. [37]	Molecular graph, separated reagents	Bond changes	USPTO_LEF (350k reactions)	Gated Graph Neural Networks
Do et al. [38]	Molecular graph, separated reagents	Bond changes	USPTO-15k, USPTO_MIT	Graph Transformation Policy Network
Coley et al. [6]	Molecular graph	Bond changes	USPTO_MIT	Graph Convolutional Neural Network
Schwaller et al. [4]	SMILES	Product molecule generation	USPTO_MIT, USPTO_LEF, USPTO_STEREO	Transformer network

B Appendix: Transfer learning enables the molecular transformer to predict regio- and stereoselective reactions on carbohydrates

B.1 Data

Recent advancement in machine learning for reaction prediction were made possible thanks to the vast availability of chemical reaction data. The largest open-source reaction data set was constructed by Lowe [42] and subsequently filtered and cleaned by different groups [5, 36, 37]. A general overview of the different reaction data sets can be found in [138]. To have a large set covering a broad range of chemical reaction classes, we started from the raw data of Lowe and constructed the reaction smiles from the extracted components. We filtered out all reactions for which we cannot match all the components to a SMILES structure. For instance, if the metal catalyst in a Suzuki coupling reaction could not be mapped to a structure because of a wrong IUPAC name in the patent, the reaction was tagged as incomplete and removed. After canonicalising the reactions using RDKit [153] and removing duplicates the generic data set (USPTO) yielded 1.2M reaction smiles. We split the data into training, validation and test sets (1.09M / 0.6M /0.6M), making sure that same products remained in the same set. Trained models and our reaction data can be found on `https://github.com/rxn4chemistry/OpenNMT-py/tree/carbohydrate_transformer`.

The second data set we use in this work is specific to carbohydrates chemistry. We manually extracted reactions from papers of 26 authors in the field of carbohydrate chemistry using Reaxys [39]. We considered full reactions with preparation and filtered out multi-step and enzymatic reactions. Reagents, solvents and catalysts, for which in the Reaxys database only the chemical names are available, were converted to SMILES structures and added to the precursors in the reaction SMILES. We only kept reactions, for which we could convert all relevant names to SMILES. We removed reactions with multiple products, the reactions without stereocentre in the product and those with just a single precursor. After the removal of duplicate reactions, the carbohydrate reactions data set (CARBO) yielded 25k reaction smiles. Similar as for the USPTO data set, we split the data into training, validation and test sets (19.7k / 2.4k / 2.5k). We also make sure that all reactions resulting in the same product molecule are in the same set. On average the products contained 6.4 stereo centres.

The following is the list of the 50 most commonly appearing authors in our Carbo data set (ordered alphabetically, the ones used for the query are highlighted with a star): Ando, Hiromune*;

Bandiera, Tiziano; Beau, Jean-Marie*; Bernet, Bruno; Bertozzi, Carolyn R.*; Bertozzi, Fabio; Bols, Mikael*; Boons, Geert-Jan*; Cheng, Ting-Jen R.; Crich, David*; Davis, Benjamin G.*; Demchenko, Alexei V.*; Ernst, Beat*; Fang, Jim-Min; Fujimoto, Yukari; Fukase, Koichi*; Hasegawa, Akira; Hotha, Srinivas*; Hui, Yongzheng; Hung, Shang-Cheng; Imamura, Akihiro; Ishida, Hideharu*; Jung, Karl-Heinz; Kajihara, Yasuhiro*; Kajimoto, Tetsuya; Kiso, Makoto; Kulkarni*, Suvarn S.*; Kusumoto, Shoichi; Li, Qin; Lin, Chun-Cheng; Oscarson, Stefan*; Pedersen, Christian Marcus; Pornsuriyasak, Papapida; Schmidt, Richard R*.; Schwardt, Oliver; Seeberger, Peter H.*; Shie, Jiun-Jie; Stuetz, Arnold E.; Suda, Yasuo; Sun, Jiansong; Urban, Dominique; Vasella, Andrea; Vincent, Stephane P.*; Withers, Stephen G.*; Wong, Chi-Huey*; Xiong, De-Cai; Yang, Jin-Song*; Ye, Xin-Shan*; Yu, Biao*; Zhang, Li-He.

B.2 Hyperparameters and training details

The training and evaluation was performed using OpenNMT-py[156, 157].

Anaconda Environment

To reproduce our results and run our models create the following conda environment:

```
conda create -n carbo python=3.6 -y
conda activate carbo
conda install -c rdkit rdkit=2019.03.2 -y
conda install -c pytorch pytorch=1.2.0 -y
pip install OpenNMT-py==1.0.0.rc2
```

B.2.1 Preprocessing of reactions

Prepare the OpenNMT input files running:

```
onmt_preprocess -train_src $DATADIR/src-train.txt \
    -train_tgt $DATADIR/tgt-train.txt \
    -valid_src $DATADIR/src-valid.txt \
    -valid_tgt $DATADIR/tgt-valid.txt \
    -save_data $DATADIR/preprocessed_onmt36 -share_vocab \
    -src_seq_length 3000 -tgt_seq_length 3000 \
    -src_vocab_size 3000 -tgt_vocab_size 3000
```

The tokenisation function, which is used to split the reaction Smiles into tokenised reactions, is available from [4, 158].

B.2.2 Training

We used OpenNMT-py and trained the *multi-task* and single data set models with the following hyperparameters.

```
onmt_train -data $DATADIR/preprocessed_onmt36  \
    -save_model  uspto_MT384 \
    -seed $SEED -gpu_ranks 0  \
    -train_steps 250000 -param_init 0 \
    -param_init_glorot -max_generator_batches 32 \
    -batch_size 6144 -batch_type tokens \
     -normalization tokens -max_grad_norm 0  -accum_count 4 \
    -optim adam -adam_beta1 0.9 -adam_beta2 0.998 -decay_method noam  \
    -warmup_steps 8000 -learning_rate 2 -label_smoothing 0.0 \
    -layers 4 -rnn_size  384 -word_vec_size 384 \
    -encoder_type transformer -decoder_type transformer \
    -dropout 0.1 -position_encoding -share_embeddings  \
    -global_attention general -global_attention_function softmax \
    -self_attn_type scaled-dot -heads 8 -transformer_ff 2048
```

The weights for the data sets can be set using the arguments,

```
-data_ids uspto carbo --data_weights $w1 $w2
```

the weights in what proportion examples from the two data sets are shown within a batch.

For the fine-tuning phase we started from the last checkpoint of the training on the USPTO data set and trained for 6k steps on the CARBO dataset:

```
-train_from /path/to/checkpoint
```

Predicting reaction outcomes

We test our models and predict reactions with a beam size of 5 and a max_length of 300 tokens using the *onmt_translate* script from OpenNMT-py [156].

```
onmt_translate -model uspto_model_pretrained.pt \
    -src $DATADIR/src-test.txt -output predictions.txt \
    -n_best 1 -beam_size 5 -max_length 300 \
    -batch_size 64
```

B.3 Supplementary Tables

	USPTO	CARBO (before 2016)	CARBO (2016 and after)
mean	0.36	6.43	6.24
std	1.01	5.06	5.00
min	0	1	1
25%	0	4	4
50%	0	5	5
75%	0	8	6
max	4	50	39

Table B.1: **Product stereo centres per reaction statistics in USPTO and CARBO data sets.** Statistics on the number of stereo centres in the products of the different reaction data sets. While the USPTO data set has 0.36 stereo centres in the product on average, there are over 6 in the CARBO data set.

C Appendix: Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy

C.1 Hyper-graph exploration

Algorithm 2 provides an overview of the hyper-graph expansion strategy, where given a starting node (N), the graph is expanded by predicting the reactions and precursors (R_i) leading to the molecule N. The single-step retrosynthetic model uses a beam-search to explore the possible disconnections and we retain the top-15 predicted sets of precursors (thus, $i = \{1, 2, ..., 15\}$). The SMILES corresponding to these predictions are canonicalised and duplicate entries removed. Any SMILES that fails in the canonicalisation step or contains the target molecule is also removed. The remaining sets of precursors are further filtered by using the forward model to assess reaction viability and selectivity. Regarding viability, we retain only those precursors (R_i) whose top-1 forward model predictions match the molecule N. This guarantees that, in the presence of multiple functional groups, the recommended disconnection leads to the desired targets. While this is a necessary condition, it is not a sufficient one as competitive reactions (top-2 and following) may lead to a mixture of molecules different from the desired target. In order to enforce chemoselectivity, we use the likelihood of the top-1 forward prediction model and select only top-1 predictions with a likelihood larger than the subsequent top-2 by at least 0.2. As the sum of likelihoods for the predictions of different sets of precursors (R_i) leading to a target N is one, any prediction likelihood higher than 0.6 automatically satisfies the requirements above and passes our filter. This filtering protocol increases the occurrence of chemo-selective reactions along the retrosynthetic path, penalising disconnections that are highly competitive.

Moreover, precursor sets are clustered together to identify similar disconnection strategies and reduce tree complexity. Within the same cluster, the precursors related to the highest forward prediction likelihood are used as starting nodes for further tree expansion. Every precursor molecule, unless already present in the graph, will generate a new node, and every reaction will connect each of the reactants to the target molecule by means of a new hyper-arc.

Every hyper-arc in the tree is scored with a so-called optimisation score, which is used to define the "best" retrosynthetic route. The total score of a retrosynthetic pathway is calculated by multiplying the scores of all the arcs contained in the path. The definition of the score for a single arc is:

$$S(\mathrm{C} \Rightarrow \mathrm{A} + \mathrm{B}) = P(\mathrm{A} + \mathrm{B} \rightarrow \mathrm{C}) \frac{s(\mathrm{A}) * s(\mathrm{B})}{s(\mathrm{C})} \tag{C.1}$$

Algorithm 2: Hyper-graph expansion algorithm

Data: Existing Node N, Beam Size B, retrosynthesis model, forward model
Result: New Nodes connected to N
begin
  $R = \{R_i | i = 1..B\} \longleftarrow$ Predict possible retrosynthesis steps (top-B) `// ` R_i `are represented as SMILES`

  for $R_i \in R$ `// select precursor sets for expansion` **do**
    $R_i \longleftarrow$ Try to canonicalise R_i, discard if not canonicalizable
    Discard R_i, if N is a precursor in R_i
    $L_{R_i \to N} \longleftarrow$ Compute likelihood of reaction $R_i \to N$
    if $L_{R_i \to N} > 0.6$ **then**
      Attach R_i to N with a hyper-arc
    else
      $F_{top-1}, F_{top-2} \longleftarrow$ Predict top-2 forward reactions from R_i
      if *Product of* F_{top-1} *is* N *and* $Likelihood(F_{top-1}) > 0.2 + Likelihood(F_{top-2})$ **then**
        Attach R_i to N with a hyper-arc
      else
        discard R_i

where $S(\mathrm{C} \Rightarrow \mathrm{A} + \mathrm{B})$ denotes the score for a single retrosynthetic step: the higher the score the higher the preference towards that step. $P(\mathrm{A} + \mathrm{B} \rightarrow \mathrm{C})$ is the likelihood of the forward chemical reaction computed by the forward prediction model. $s(\mathrm{X}) | X \in \{\mathrm{A}, \mathrm{B}, \mathrm{C}\}$ is the simplicity score of molecule X:

$$s(\mathrm{X}) = 1 - \frac{SC(\mathrm{X}) - 1}{4} \tag{C.2}$$

where $SC(\mathrm{X})$ is the SCScore [176] of molecule X. The SCScore of a molecule increases from 1 to 5 with an increasing complexity of the synthetic route. In this framework, the SCScore constitutes the driving force that pulls a retrosynthetic pathway towards simpler molecules.

Equation C.1 closely resembles the definition of the Bayesian probability. In fact, assuming access to the set of all possible reactions, the likelihood of a retrosynthetic step would be defined as the conditional probability of observing the product when given the reactants, weighted by the ratio between the occurrence of the reagents and the occurrence of the product.

Even with a multi-million entry database, the evaluation of the individual components would still be quite inaccurate. In fact, any molecule unreported in this database will contribute a value of zero to the evaluation of the Bayesian probability, with important drawbacks for the hyper-tree exploration. Therefore, the definition of the score for a single retrosynthetic step was only inspired by the Bayesian probability. We replaced the conditional probability with the likelihood of the forward prediction model and the probability of observing either reactants or products with a simplicity score. Similar to the Bayesian probability, the use of this heuristic favours those reaction that give more simple products (compared to reactants) under the same forward prediction likelihood.

The search for the optimal retrosynthetic route starts with the definition of a target molecule and uses a beam-search approach. The beam-search method is a greedy version of the best-first search: while best-first explores the entire graph and sorts all the possible paths according to some heuristic score, the beam search limits the exploration to a defined number of paths, thus limiting the computational cost without offering any guarantee of identifying the globally optimal path. The beam-search, as implemented in our software, relies on the following steps:

1. Expand the graph at every node contained in one of the possible pathways discovered up to this point and not yet expanded.
2. Create a new pathway for each of the arcs created by the last expansion.
3. Repeat steps 1 and 2 for a given number of times.
4. Assign a score to every pathway and discard the ones with the lowest score until the total number of "un-terminated" pathways correspond to the number of beams imposed by the user.
5. Restart from point 1 until all of the pathways meet one of the terminating conditions.

Each pathway of the beam-search may end because all the molecules needed to start the synthesis are found in a database of commercially available chemicals; or because the number of synthetic steps (which corresponds to the number of "expansion phases") exceeds the number of maximum steps defined by the user; or finally because there is no possibility to further expand the needed

nodes. The last condition may result from none of the set of precursors (R_i) surviving the filtering or from all the hyper-arcs generated by the expansion forming a cycle in the tree. From a chemical point of view, this means that one of the precursors of the product requires the product to synthesise itself.

Every time a pathway enters a cycle, the pathway itself is considered terminated. The tree exploration returns all the possible paths leading to a successful retrosynthesis, sorted by the optimisation score.

C.2 Molecule representation

Similar to our previous works we use SMILES to represent molecules, taking more advantage of the auxiliary fragment information in which the grouped fragment indices are written after the label 'f:'. The different groups are separated by a ',' and the connected fragments within a group are separated by '.'. An example would be '|f:1.2,4.5|'. , where the fragments 1 and 2 as well as 4 and 5 belong together. There is nothing that enforces closeness of fragments in the SMILES string, hence different fragments belonging to the same compound could end up at opposite ends of the string. Typical examples are metallorganic compounds. Here, we relate the fragments within a group with a '∼' character instead of a '.'. Consequently, the fragmented molecules are kept together in the reaction string.

Atom-mapping as well as reactant-reagent roles, are a rich source of information generated by highly complicated tasks [62], the assignment often being subjectively made by humans. Schwaller et al. [4] recently proposed to ignore reactant and reagent roles for the reaction prediction task. In contrast to previous works [166, 167, 169, 170], the single-step retrosynthetic model presented here predicts reactants and reagents. In an effort to simplify the prediction task, the most common precursors with a length of more than 50 tokens were replaced by molecule tokens. Those molecules were turned back into the usual tokenisation before calculating the likelihood with the forward model. Moreover, to ensure a basic tautomer standardisation we inchified our molecules, as described in [300], to improve the quality of the forward prediction model. In contrast to previous work [65], we never use a reaction class token as input for the retrosynthesis model.

The data sets used to train the different models in this work are derived from the open source USPTO reaction database by Lowe [41, 42] and the Pistachio database by NextMove Software [180]. We preprocessed both data sets to filter out incomplete reactions and keep 1M and 2M entries, respectively. As done previously in [4, 111], we added 800k textbook reactions to the training of specific forward and retrosynthetic models.

C.3 Models

C.3.1 Forward reaction prediction model

The forward prediction model was trained with the same hyperparameters as the original Molecular Transformer [4], apart from the number of the attention layers, which was increased from 256 to 384. Thanks to the increase in capacity, a higher validation accuracy could be reached. For the final model we used a data set derived from Pistachio3.0 [180] where all the molecules were

inchified. As described in the work of Schwaller et al. [4] we augmented the training data with the addition of random SMILES and textbook reactions to the training set.

The forward prediction model can be used in two modes. First, when given a precursor set, the most likely products can be predicted. Second, when given a precursor set and a target product, the likelihood of this specific reaction can be estimated. In this work, we set the beam size of the forward model to 3.

As described previously, we use the forward chemical prediction model as a digital domain expert for evaluating the correctness of the predictions generated by the retrosynthetic model. As recently published [4], the accuracy of this model is higher than 90% when compared with a public data set. In order to calibrate the forward prediction model within the entire retrosynthetic framework, 50 random forward reaction predictions were analyzed by human experts. The assessment gave an accuracy of 78% which should be compared to an accuracy of 80% given by the trained model. Although the data set is too limited to claim any statistical relevance, this assessment offers strong evidence in favour of using the forward prediction model as a digital twin of human chemists.

C.3.2 Reaction classification model

To classify reactions, we used a data-driven reaction classification model [177] that was trained similarly to the Molecular Transformer forward and retrosynthetic model. It is characterised by four encoder layers and one decoder layer and trained using the same hyperparameters. The main difference is that the inputs were made up of the complete reaction string (precursors→products) and the outputs of the split reaction class identifier from NameRXN, consisting of three numbers corresponding to superclass, classes/categories and named reaction. More details on reaction classes can be found in [173]. The classification model used in this work matches the same class as the NameRXN tool [63] for 93.8% of the reactions.

C.3.3 Experiments on single-step retrosynthesis models

In Table C.1 we show how different metrics develop during the training of the *stereo* retro model. After 100k time steps the round-trip accuracy and the coverage plateau and only a slight improvement of the invalid SMILES percentage can be observed, when training for longer.

Table C.2 shows a comparison of models trained on different data sets and evaluated with the beam sizes 5, 10 and 20. The beam size defines how many precursor set suggestions output. The more data is used in the training set the less invalid SMILES the models tend to generate. As expected the coverage increases with larger beam sizes, while the round-trip accuracy and the percentage of invalid SMILES worsen only slightly. *stereo only* means that the model was trained purely on the 1M reactions derived from the open USPTO dataset [41, 42]. The *stereo* model was trained on the USPTO dataset and 800K textbook reactions from Nam & Kim [111]. For the *augmented* model we performed a SMILES data augmentation for the source molecules by using non canonical SMILES [236]. The target always consisted of canonical SMILES. In contrast to reaction prediction [4], the augmentation seemed not to be beneficial in our retrosynthesis model training experiments.

Table C.1: **Model performance during training.**Development of the round-trip accuracy, coverage and percentage of invalid SMILES during training of the retrosynthesis model, evaluated with a forward model trained on *stereo only*.

Model	Beam	Total rxns	Round-trip accuracy	Coverage	Invalid SMILES
stereo only 10k	10	100k	56.9%	87.4%	4.03 %
stereo only 20k	10	100k	73.8%	93.8%	1.72 %
stereo only 50k	10	100k	78.7%	95.0%	0.81 %
stereo only 100k	10	100k	81.6%	95.8%	0.65 %
stereo only 150k	10	100k	81.3%	95.8%	0.62 %
stereo only 200k	10	100k	81.0%	95.8%	0.59 %
stereo only 250k	10	100k	81.5%	95.9%	0.58 %

Table C.2: **Model performance varying the beam size.** Evaluation of retrosynthesis models with different training data, evaluated on the same validation set with different beam sizes.

Model	Beam	Total	Round-trip accuracy	Coverage	Invalid SMILES
stereo only	5	50k	82.4%	93.5%	0.57 %
stereo	5	50k	83.6%	94.2%	0.52 %
augmented	5	50k	81.8%	94.0%	0.43 %
stereo only	10	100k	81.5%	95.9%	0.59 %
stereo	10	100k	82.4%	96.4%	0.49 %
augmented	10	100k	80.7%	96.2%	0.42 %
stereo only	20	200k	79.8%	97.1%	0.65 %
stereo	20	200k	80.8%	97.5%	0.87 %
augmented	20	200k	78.9%	97.5%	0.49 %

C.4 Synthesis routes

On the subsequent pages, the synthesis routes discussed in the main text are presented. The routes were predicted by the model, which is openly available on the IBM RXN for Chemistry platform [44]. Figure C.1 shows a screenshot of the results page for an example retrosynthesis route prediction.

C.4.1 Index of generated retrosynthetic routes

The targets from Coley et al. [125] are extracted from: http://ibm.biz/Coley-Test, where corresponding retrosynthesis are also made available.
Both Segler Test-1 and Test-2 are instead from the supporting information [107]: `http://ibm.biz/Segler-Test1-2` , with fully reported synthesis.

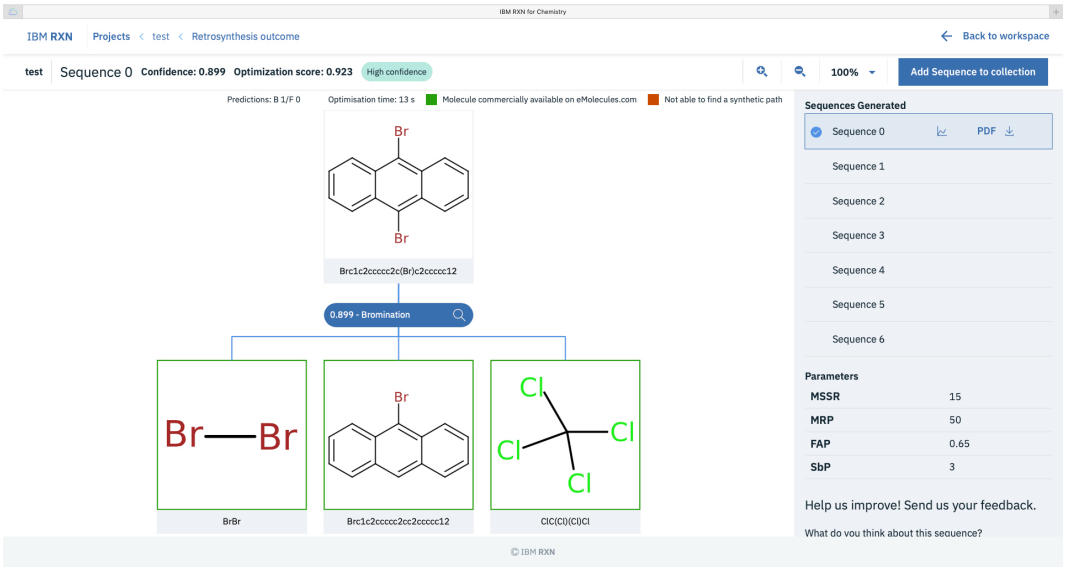

Figure C.1: **IBM RXN for Chemistry platform.** Retrosynthesis route prediction results view.

The generated by IBM RXN for Chemistry can be found in the supplementary information of [43].

D Appendix: Mapping the space of chemical reactions using attention-based neural networks

D.1 Reaction properties atlases

Supplementary Figure D.1 shows the chemical reaction found in the 50k set by Schneider et al. [64] visualised with TMAP [201] using the *rxnfp (10k)*. The BERT model, which generated this reaction fingerprint was trained on the 10k training reactions. The reaction maps are made of the 10k training reactions plus 40k unseen reactions. The reactions corresponding to same reaction classes are well clustered together. We highlight reactions that contain specific elements in the precursors and observe that they found in the same branches of the map. Moreover, we visualise product properties and also observe defined clustering.

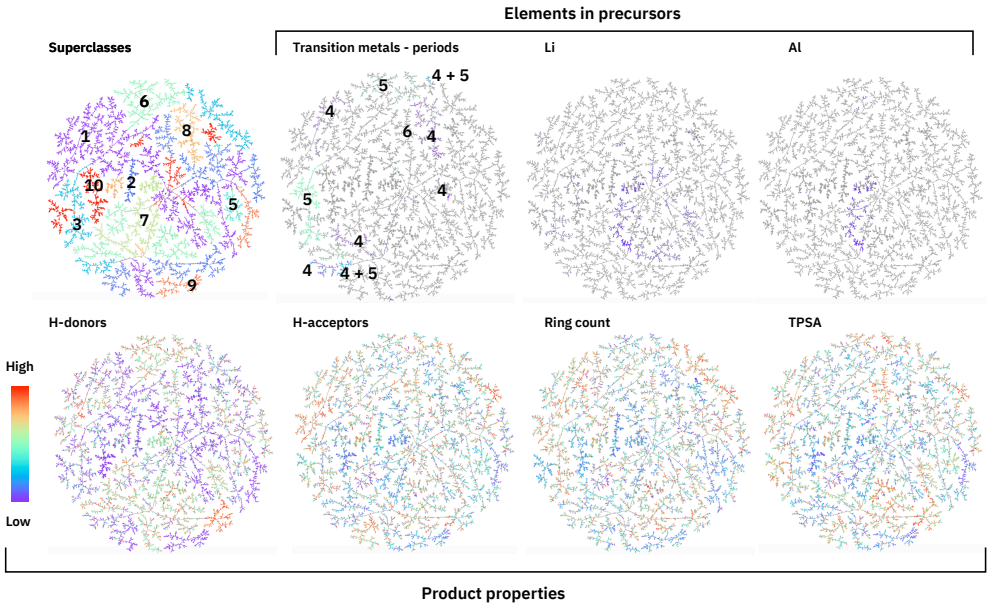

Figure D.1: **Reaction properties** TMAP [201] of the Schneider 50k set using the *rxnfp (10k)* embeddings. The superclasses, as well as specific metallic elements in the precursors and product properties are highlighted in the different maps. An interactive version of this map is also available as a separate file.

D.2 Analysis of Pistachio predictions

We analysed the BERT classifier in more detail and compared it to the seq-2-seq transformer model. First, we identified different types of incorrect predictions by the transformer BERT classifier model, which are summarised in Table D.1. Most errors are related to the "Unrecognised" class of the RXNO ontology. The most frequent error type is the prediction of a reaction class for a reaction classified as "Unrecognised" (47.9% of all incorrect predictions), and the second most frequent error type is predicting "Unrecognised" when a class should be predicted (22.8%). The third most frequent error is predicting the incorrect name reaction (third number of the class string, 17.5%). The remaining errors are predicting an incorrect superclass (first number of the class string, 8.3%) and predicting an incorrect category (second number of the class string, 3.5%).

Table D.1: **Incorrect predictions.** Types of incorrect predictions of the BERT model on the test set consisting of a total of 132213 reactions.

	Count	Percentage
Correctly predicted	129892	98.24%
Model predicts name reaction instead of "Unrecognised"	1111	0.84%
Model predicts "Unrecognised" instead of name reaction	529	0.40%
Incorrect name rxn	407	0.31%
Incorrect superclass	193	0.15%
Incorrect category	81	0.06%

In Table D.2, we show the reaction classes for which our model makes incorrect predictions most frequently. Due to statistical sampling, we restricted this analysis to reactions with at least 20 occurrences in the test set. For 12 out of 15 of these reaction classes, the most common error source is the failure to assign a reaction class, thus predicting "Unrecognised". Among the other most common failures, there is the "Bouveault-Blanc reduction", where an ester is reduced to a primary alcohol. Hence, it is very similar to the Ester to alcohol reduction class, with which it is most mistaken. The difference lies in the specific precursors used in the "Bouveault-Blanc reduction", such as sodium and ethanol or methanol. The "1,3-Dioxane synthesis" reaction class has an overall accuracy of 88.9%. However, there are some reactions mistaken for "Dioxolane synthesis", for which the newly formed heterocycle in the product has an additional carbon atom.

Although the large number of "Unrecognised" reactions in Pistachio makes an extensive analysis difficult, the inspection of a few dozen cases provides interesting insights. Part of the "Unrecognised" reactions should actually belong to a name reaction. The data-driven approach can be more robust than rule-based models and assign the correct reaction class. For example, in contrast to rule-based models, data-driven ones are often able to capture the reaction class despite changes in the tautomeric state between precursors and product. Another part of those "Unrecognised" reactions belongs to the category for which multiple transformations occur simultaneously. In this case, the reaction cannot be classified into a single name reaction, and our model predicts one of the corresponding reactions. Such examples can be found in deprotection reactions where more than one distinct functional group is removed. Another interesting aspect comes from molecules that are incorrectly parsed in Pistachio. If the SMILES string of a molecule involved in the re-

Table D.2: **Detailed failure analysis.** Worst-predicted reaction classes with more than 20 occurrences in the test set for the BERT classifier.

Reaction class		Accuracy [%]	Most frequent incorrectly predicted class	
1.1.2	Menshutkin reaction	62.1	0.0	Unrecognised
3.9.41	Decarboxylative coupling	72.1	0.0	Unrecognised
9.7.140	Defluorination	75.6	0.0	Unrecognised
7.4.2	Bouveault-Blanc reduction	76.4	7.4.1	Ester to alcohol reduction
11.1	Chiral separation	83.6	0.0	Unrecognised
8.8.11	Hydroxylation	83.7	0.0	Unrecognised
4.3.11	Thiazoline synthesis	85.7	0.0	Unrecognised
3.9.12	Olefin metathesis	85.8	0.0	Unrecognised
2.5.5	Nitrile + amine reaction	86.0	0.0	Unrecognised
9.7.42	Chloro to fluoro	86.4	0.0	Unrecognised
10.4.2	Methylation	88.9	0.0	Unrecognised
4.2.39	1,3-Dioxane synthesis	88.9	4.2.20	Dioxolane synthesis
4.1.53	1,2,4-Triazole synthesis	90.0	0.0	Unrecognised
1.1.6	Chloro Menshutkin reaction	90.6	0.0	Unrecognised
5.1.2	N-Cbz protection	90.9	2.1.1	Amide Schotten-Baumann

action was incorrectly derived from the name, rule-based approaches fail to recognise the atomic rearrangements and thus to classify the reaction. For minor parsing errors, our model shows its potential, recognising the correct transformation in several instances.

The accuracy of the enc2-dec1 seq-2-seq model was 3% worse than the one of the BERT classifier. When comparing the predictions of the two models, we observe that most of the differences are related to the "Unrecognised" class. 3511 out of 5108 reactions that were correctly predicted by the BERT classifier but not the seq-2-seq model belong to the "Unrecognised" class. Moreover, the three classes containing the most examples of reaction classes predicted correctly by the BERT classifier but not by the seq-2-seq model were "Carboxylic acid + amine condensation" (2.1.2), "Methylation" (10.4.2) and "Williamson ether synthesis" (1.7.9) reactions with 90, 61 and 37 examples respectively. In contrast, the seq-2-seq model was able to classify 474 reactions as "Unrecognised", which were classified as recognised name reactions by the BERT model. Besides the "Unrecognised" reactions, the three reaction types with the most examples that were correctly predicted by the seq-2-seq model but not by the BERT classifier were "Bouveault-Blanc reduction" (7.4.2), "Ester to alcohol reduction" (7.4.1) reactions with 33 and 15 examples respectively. The seq-2-seq seems to capture the subtle difference between the two distinct "Ester to alcohol" (7.4) classes better.

D.3 Analysis of 50k set predictions

Schneider et al. [64] evaluated their reaction fingerprints by analysing how well it could classify chemical reactions using a logistic regression classifier [301]. For a given reaction input, they

trained their classifier to predict 1 out of 50 named reaction classes using 200 training/validation and 800 testing examples per class. To be able to directly compare to the results of Ref. [64], we investigated our learned fingerprints on their data sets, pretrained and fine-tuned on the same 10k training reactions resulting in *rxnfp (10k)*. A summary where we report recall, precision and F-score averaged over the 50 classes can be found in Supplementary Table D.3. While the *rxnfp (pretrained)* does not suffice to match the performance of the handcrafted fingerprint on this balanced data set, *rxnfp (10k)*, generated after fine-tuning the model on as little as the 10k reactions, is able to reach scores of 0.99 compared to 0.97 for the hand-crafted fingerprint.

Table D.3: Comparing fingerprints on the 50k reactions classification benchmark by Schneider et al. [64] (50 classes, 1000 reactions per class, 200 for training/validation and 800 for testing)

Fingerprint	recall	precision	F-score	
AP3 256 (folded) [64] + Agent features	0.97	0.97	0.97	handcrafted, reactants-reagents separation
rxnfp (pretrained)	0.90	0.90	0.90	after pretraining
rxnfp (10k)	0.99	0.99	0.99	fine-tuning on 10k reactions training set[64]

Supplementary Table D.4 and Supplementary Figure D.2 show the detailed results for *rxnfp (10k)*. Supplementary Table D.5 and Supplementary Figure D.3 show the results of for *rxnfp (pretrained)* computed by the model never fine-tuned on reaction classification.

For both data-driven fingerprints the methylation class seems to be the hardest to predict correctly. Using the pretrained fingerprint it is hard to distinguish between reaction classes that differ only by one atom, like “CO2H-Et deprotection” and “CO2H-Me deprotection”. “Carboxylic acid + amine condensation” are confused with “Amide Schotten-Baumann” reactions and “Mitsunobu aryl ether synthesis” with “Williamson ether synthesis” reactions. It is likely that in future unsupervised reaction fingerprints will be developed that capture this fine-grained information better.

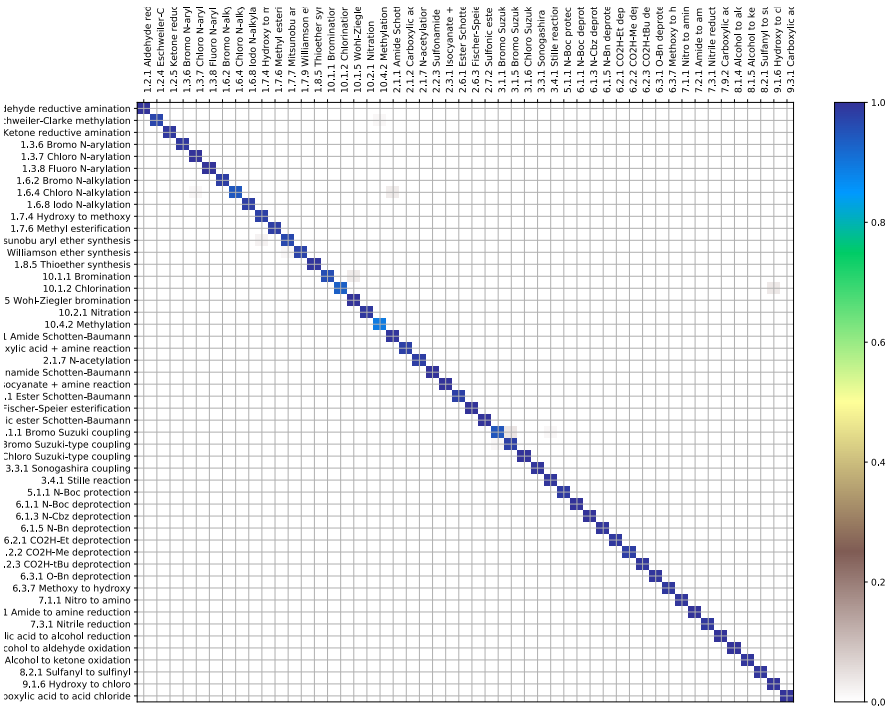

Figure D.2: Confusion matrix for *rxnfp (10k)* train

Table D.4: *rxnfp (10k)* train: 50k reactions classification benchmark by Schneider et al. [64]

	recall	prec	F-score	reaction class	
0	0.9988	0.9901	0.9944	Aldehyde reductive amination	1.2.1
1	0.9712	0.9848	0.9780	Eschweiler-Clarke methylation	1.2.4
2	0.9888	0.9950	0.9918	Ketone reductive amination	1.2.5
3	0.9912	0.9863	0.9888	Bromo N-arylation	1.3.6
4	0.9962	0.9827	0.9894	Chloro N-arylation	1.3.7
5	0.9975	0.9876	0.9925	Fluoro N-arylation	1.3.8
6	0.9825	0.9788	0.9807	Bromo N-alkylation	1.6.2
7	0.9437	0.9921	0.9673	Chloro N-alkylation	1.6.4
8	0.9838	0.9825	0.9831	Iodo N-alkylation	1.6.8
9	0.9775	0.9678	0.9726	Hydroxy to methoxy	1.7.4
10	0.9838	0.9838	0.9838	Methyl esterification	1.7.6
11	0.9675	0.9639	0.9657	Mitsunobu aryl ether synthesis	1.7.7
12	0.9750	0.9665	0.9708	Williamson ether synthesis	1.7.9
13	0.9938	0.9938	0.9938	Thioether synthesis	1.8.5
14	0.9575	0.9935	0.9752	Bromination	10.1.1
15	0.9313	0.9868	0.9582	Chlorination	10.1.2
16	0.9988	0.9685	0.9834	Wohl-Ziegler bromination	10.1.5
17	0.9888	0.9987	0.9937	Nitration	10.2.1
18	0.8938	0.9483	0.9202	Methylation	10.4.2
19	0.9950	0.9522	0.9731	Amide Schotten-Baumann	2.1.1
20	0.9788	0.9899	0.9843	Carboxylic acid + amine reaction	2.1.2
21	0.9838	0.9975	0.9906	N-acetylation	2.1.7
22	0.9975	0.9975	0.9975	Sulfonamide Schotten-Baumann	2.2.3
23	1.0000	0.9950	0.9975	Isocyanate + amine reaction	2.3.1
24	0.9775	0.9726	0.9751	Ester Schotten-Baumann	2.6.1
25	0.9962	0.9815	0.9888	Fischer-Speier esterification	2.6.3
26	1.0000	1.0000	1.0000	Sulfonic ester Schotten-Baumann	2.7.2
27	0.9463	0.9818	0.9637	Bromo Suzuki coupling	3.1.1
28	0.9800	0.9596	0.9697	Bromo Suzuki-type coupling	3.1.5
29	1.0000	0.9950	0.9975	Chloro Suzuki-type coupling	3.1.6
30	0.9925	0.9937	0.9931	Sonogashira coupling	3.3.1
31	0.9925	0.9778	0.9851	Stille reaction	3.4.1
32	0.9850	0.9975	0.9912	N-Boc protection	5.1.1
33	1.0000	0.9780	0.9889	N-Boc deprotection	6.1.1
34	0.9975	1.0000	0.9987	N-Cbz deprotection	6.1.3
35	0.9950	0.9925	0.9938	N-Bn deprotection	6.1.5
36	0.9888	0.9875	0.9881	CO2H-Et deprotection	6.2.1
37	0.9825	0.9800	0.9813	CO2H-Me deprotection	6.2.2
38	0.9950	0.9925	0.9938	CO2H-tBu deprotection	6.2.3
39	0.9950	0.9925	0.9938	O-Bn deprotection	6.3.1
40	0.9888	0.9900	0.9894	Methoxy to hydroxy	6.3.7
41	0.9938	0.9925	0.9931	Nitro to amino	7.1.1
42	0.9975	0.9803	0.9888	Amide to amine reduction	7.2.1
43	0.9912	0.9925	0.9919	Nitrile reduction	7.3.1
44	0.9988	0.9938	0.9963	Carboxylic acid to alcohol reduction	7.9.2
45	1.0000	0.9963	0.9981	Alcohol to aldehyde oxidation	8.1.4
46	0.9950	0.9987	0.9969	Alcohol to ketone oxidation	8.1.5
47	0.9950	0.9962	0.9956	Sulfanyl to sulfinyl	8.2.1
48	0.9962	0.9614	0.9785	Hydroxy to chloro	9.1.6
49	0.9975	0.9888	0.9932	Carboxylic acid to acid chloride	9.3.1
	0.99	0.99	0.99	Average	

Table D.5: *rxnfp (pretrained)*: 50k reactions classification benchmark by Schneider et al. [64]

	recall	prec	F-score	reaction class	
0	0.9012	0.8990	0.9001	Aldehyde reductive amination	1.2.1
1	0.8063	0.8323	0.8190	Eschweiler-Clarke methylation	1.2.4
2	0.9213	0.9213	0.9213	Ketone reductive amination	1.2.5
3	0.8600	0.8632	0.8616	Bromo N-arylation	1.3.6
4	0.8712	0.7938	0.8308	Chloro N-arylation	1.3.7
5	0.9225	0.9498	0.9360	Fluoro N-arylation	1.3.8
6	0.8113	0.8353	0.8231	Bromo N-alkylation	1.6.2
7	0.7600	0.7696	0.7648	Chloro N-alkylation	1.6.4
8	0.8125	0.7908	0.8015	Iodo N-alkylation	1.6.8
9	0.8500	0.8662	0.8580	Hydroxy to methoxy	1.7.4
10	0.9200	0.9258	0.9229	Methyl esterification	1.7.6
11	0.8413	0.8519	0.8465	Mitsunobu aryl ether synthesis	1.7.7
12	0.8000	0.7960	0.7980	Williamson ether synthesis	1.7.9
13	0.9225	0.8902	0.9061	Thioether synthesis	1.8.5
14	0.9437	0.9461	0.9449	Bromination	10.1.1
15	0.9463	0.9232	0.9346	Chlorination	10.1.2
16	0.9838	0.9633	0.9734	Wohl-Ziegler bromination	10.1.5
17	0.9738	0.9725	0.9731	Nitration	10.2.1
18	0.6625	0.7172	0.6888	Methylation	10.4.2
19	0.8175	0.7861	0.8015	Amide Schotten-Baumann	2.1.1
20	0.8013	0.8250	0.8129	Carboxylic acid + amine reaction	2.1.2
21	0.9600	0.9588	0.9594	N-acetylation	2.1.7
22	0.9450	0.9345	0.9397	Sulfonamide Schotten-Baumann	2.2.3
23	0.9725	0.9569	0.9647	Isocyanate + amine reaction	2.3.1
24	0.8625	0.8582	0.8603	Ester Schotten-Baumann	2.6.1
25	0.9525	0.9658	0.9591	Fischer-Speier esterification	2.6.3
26	0.9700	0.9395	0.9545	Sulfonic ester Schotten-Baumann	2.7.2
27	0.9437	0.9333	0.9385	Bromo Suzuki coupling	3.1.1
28	0.9113	0.9045	0.9078	Bromo Suzuki-type coupling	3.1.5
29	0.9550	0.9340	0.9444	Chloro Suzuki-type coupling	3.1.6
30	0.9625	0.9686	0.9655	Sonogashira coupling	3.3.1
31	0.9150	0.9150	0.9150	Stille reaction	3.4.1
32	0.9613	0.9661	0.9637	N-Boc protection	5.1.1
33	0.9100	0.9089	0.9094	N-Boc deprotection	6.1.1
34	0.8600	0.9005	0.8798	N-Cbz deprotection	6.1.3
35	0.9700	0.9293	0.9492	N-Bn deprotection	6.1.5
36	0.7688	0.7437	0.7560	CO2H-Et deprotection	6.2.1
37	0.7150	0.7259	0.7204	CO2H-Me deprotection	6.2.2
38	0.9450	0.9486	0.9468	CO2H-tBu deprotection	6.2.3
39	0.8962	0.9459	0.9204	O-Bn deprotection	6.3.1
40	0.9313	0.9418	0.9365	Methoxy to hydroxy	6.3.7
41	0.9663	0.9898	0.9779	Nitro to amino	7.1.1
42	0.9613	0.9470	0.9541	Amide to amine reduction	7.2.1
43	0.9900	0.9888	0.9894	Nitrile reduction	7.3.1
44	0.9838	0.9887	0.9862	Carboxylic acid to alcohol reduction	7.9.2
45	0.9750	0.9750	0.9750	Alcohol to aldehyde oxidation	8.1.4
46	0.9600	0.9540	0.9570	Alcohol to ketone oxidation	8.1.5
47	0.9700	0.9898	0.9798	Sulfanyl to sulfinyl	8.2.1
48	0.9663	0.9748	0.9705	Hydroxy to chloro	9.1.6
49	0.9875	0.9925	0.9900	Carboxylic acid to acid chloride	9.3.1
	0.90	0.90	0.90	Average	

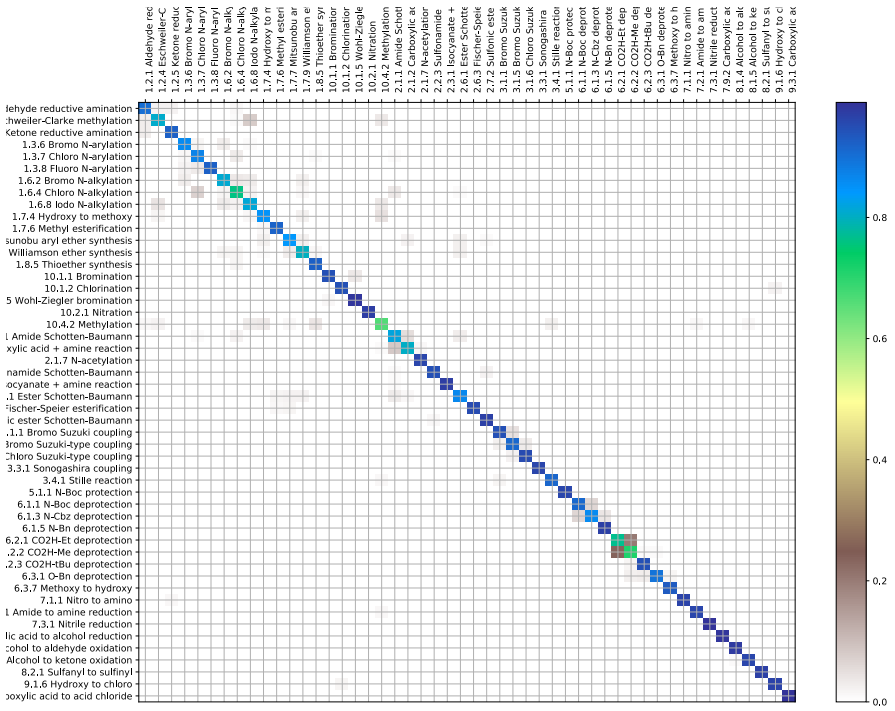

Figure D.3: Confusion matrix for *rxnfp (pretrained)*

E Appendix: Prediction of chemical reaction yields using deep learning

E.1 Detailed results on Buchwald–Hartwig reactions

Figure E.1-E.14 show the correlation between the measured yields and the predicted yields for the different splits published by Sandfort et al. [200]. Moreover, the root mean squared error (RMSE) and the mean average error (MAE) are shown in the figures.

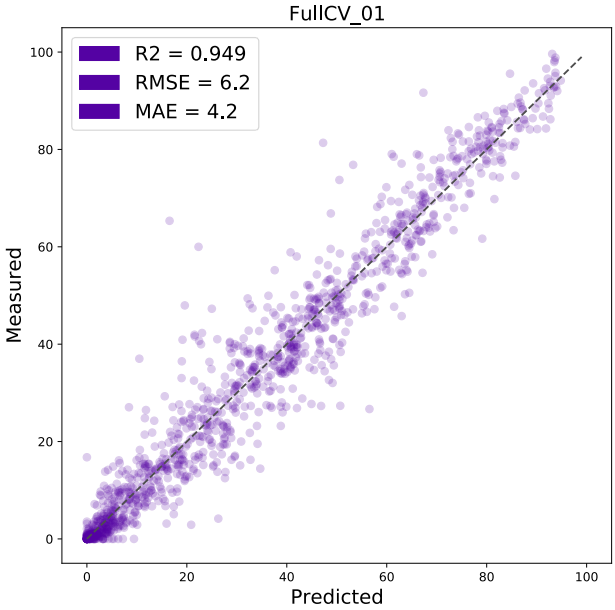

Figure E.1: Measured vs predicted yields [%] - FullCV_01

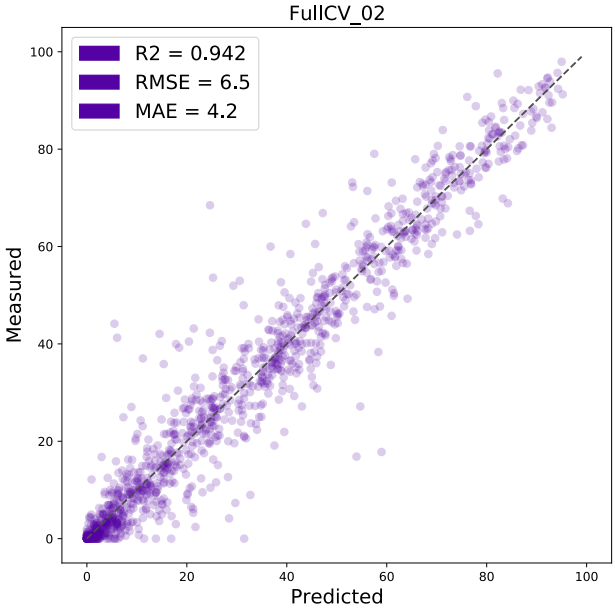

Figure E.2: Measured vs predicted yields [%] - FullCV_02

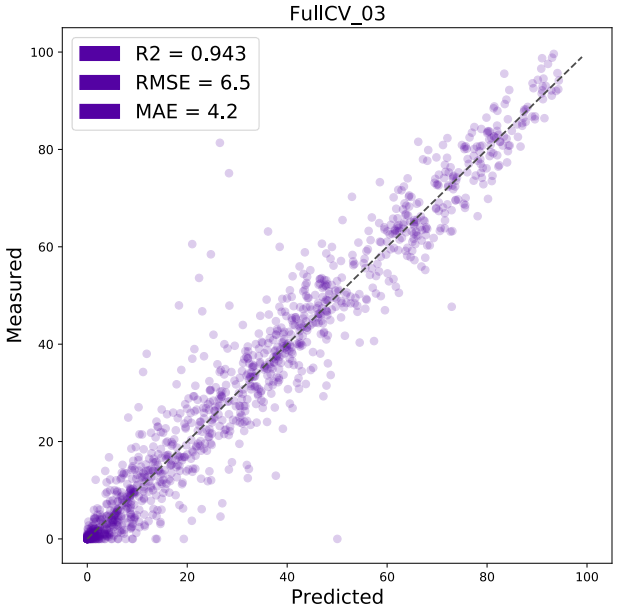

Figure E.3: Measured vs predicted yields [%] - FullCV_03

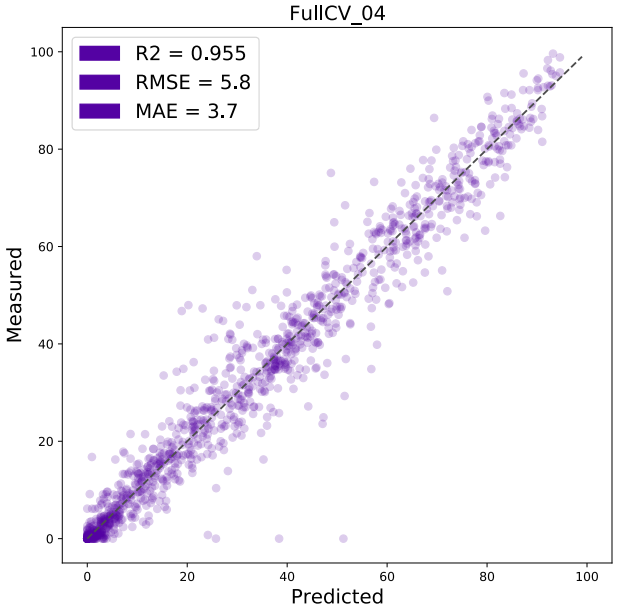

Figure E.4: Measured vs predicted yields [%] - FullCV_04

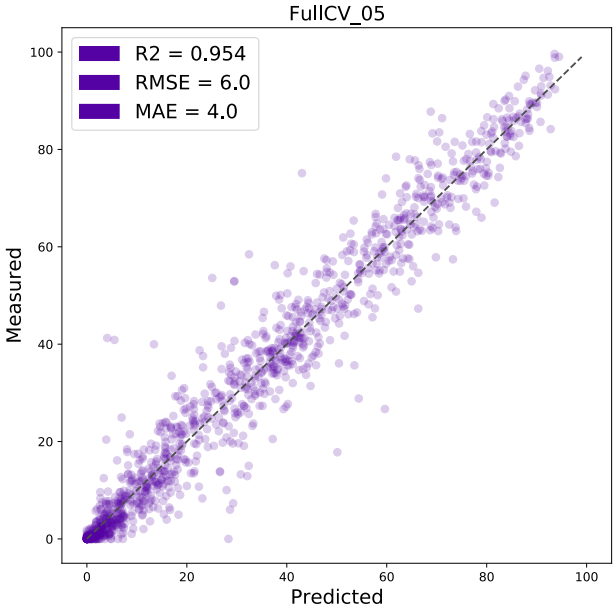

Figure E.5: Measured vs predicted yields [%] - FullCV_05

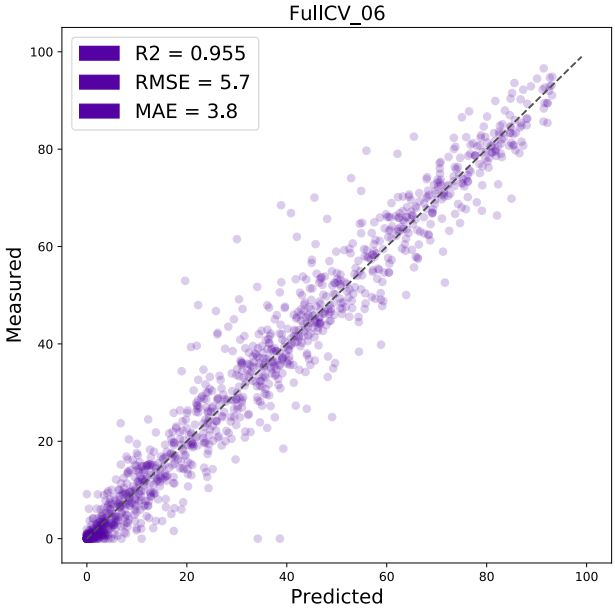

Figure E.6: Measured vs predicted yields [%] - FullCV_06

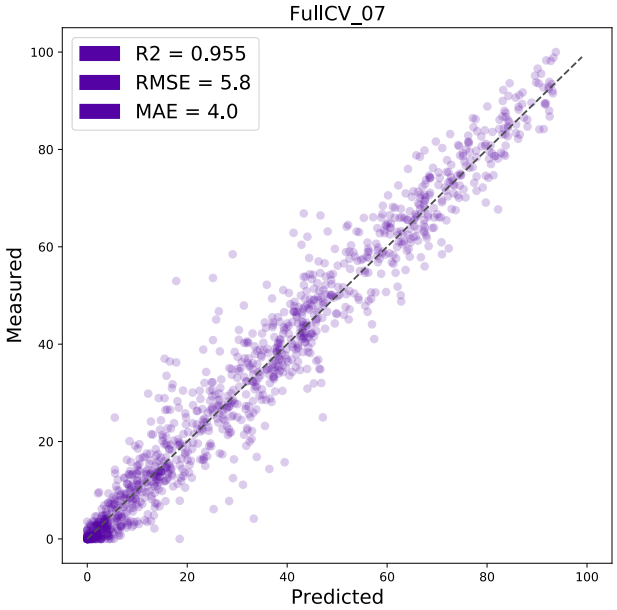

Figure E.7: Measured vs predicted yields [%] - FullCV_07

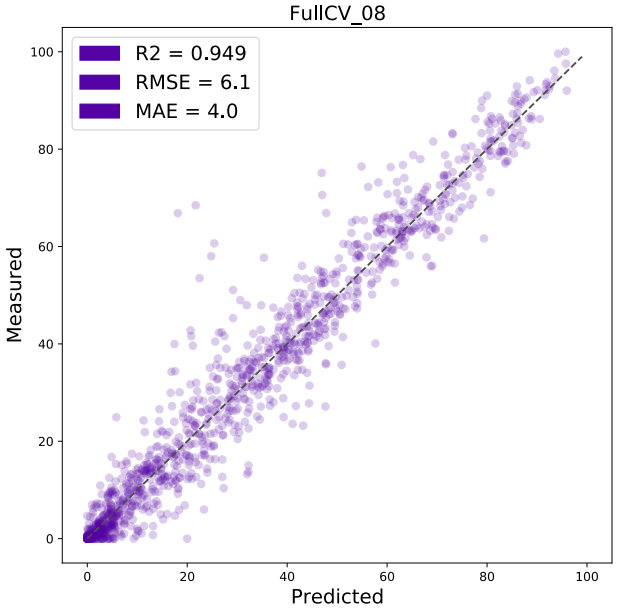

Figure E.8: Measured vs predicted yields [%] - FullCV_08

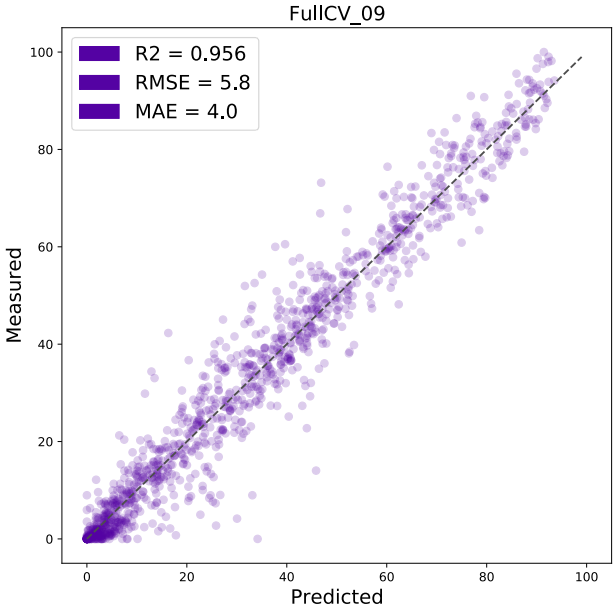

Figure E.9: Measured vs predicted yields [%] - FullCV_09

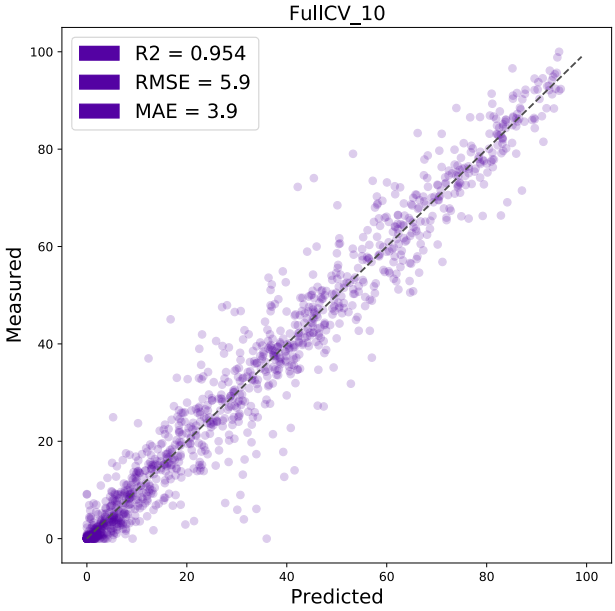

Figure E.10: Measured vs predicted yields [%] - FullCV_10

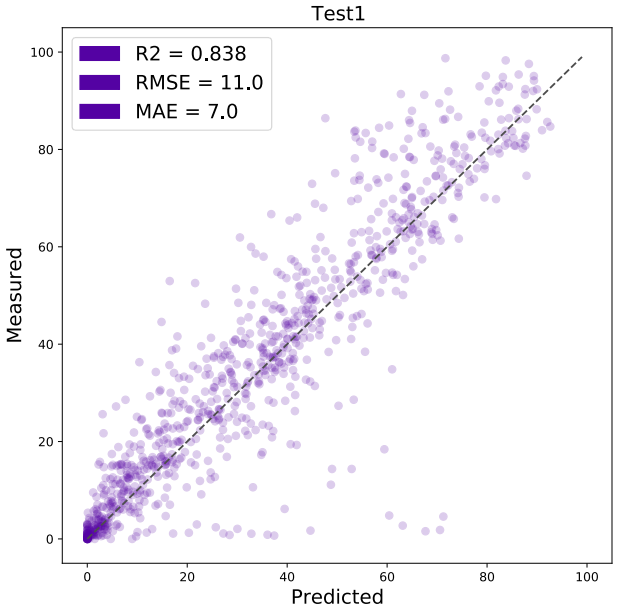

Figure E.11: Measured vs predicted yields [%] - Test1

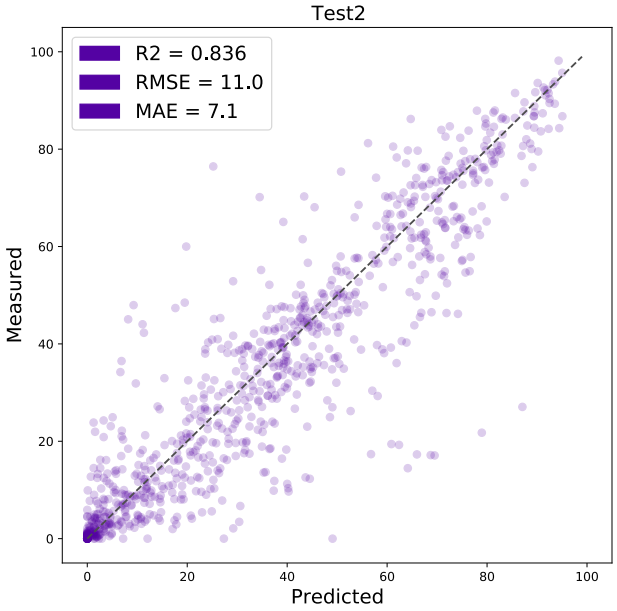

Figure E.12: Measured vs predicted yields [%] - Test2

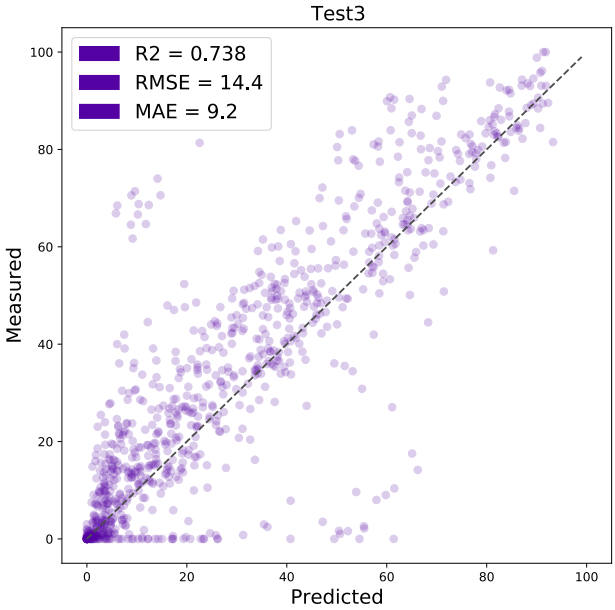

Figure E.13: Measured vs predicted yields [%] - Test3

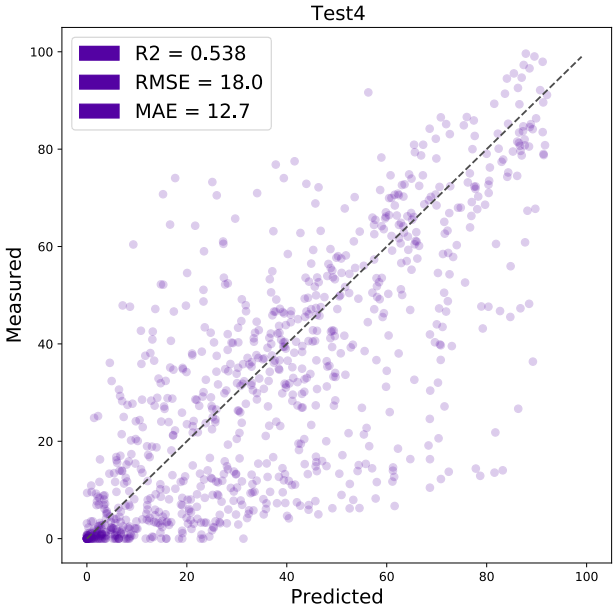

Figure E.14: Measured vs predicted yields [%] - Test4

E.2 Detailed results on Suzuki–Miyaura reactions

Figure E.15-E.24 show the correlation between the measured yields and the predicted yields for model with the *rxnfp ft* base encoder on the 10 random splits.

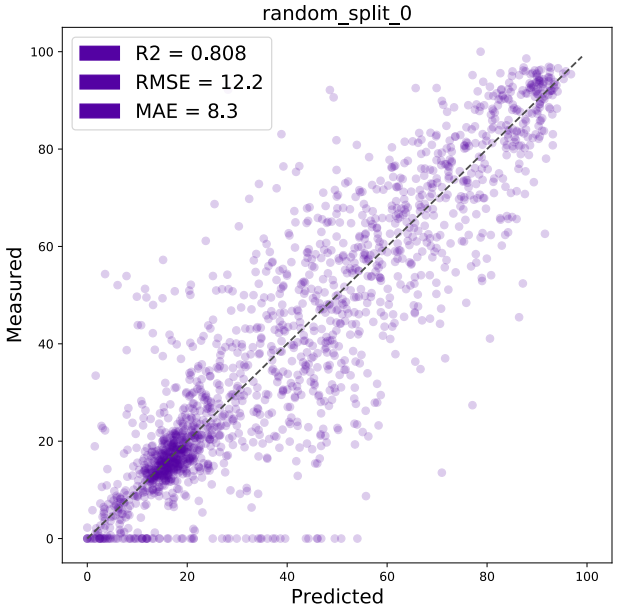

Figure E.15: Measured vs predicted yields [%] - random_split_0

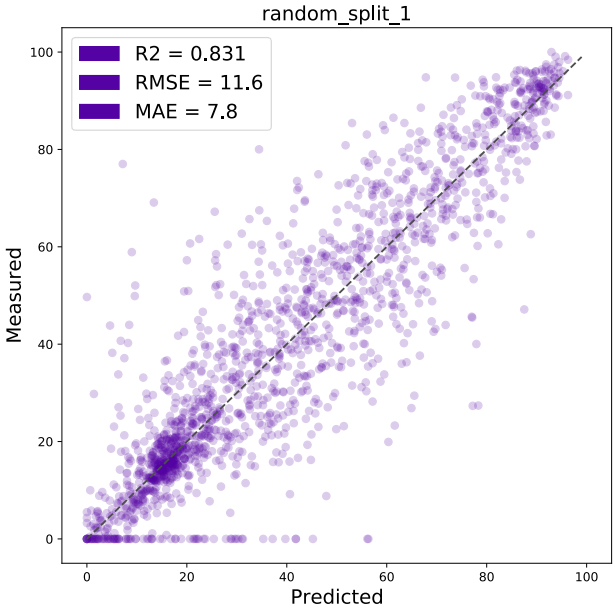

Figure E.16: Measured vs predicted yields [%] - random_split_1

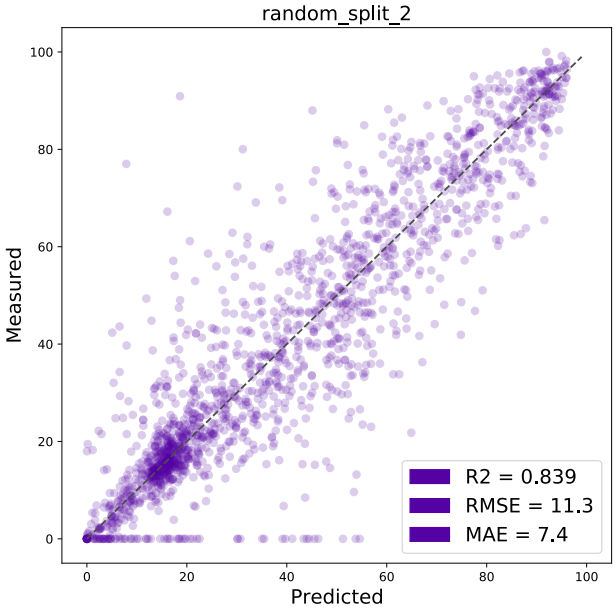

Figure E.17: Measured vs predicted yields [%] - random_split_2

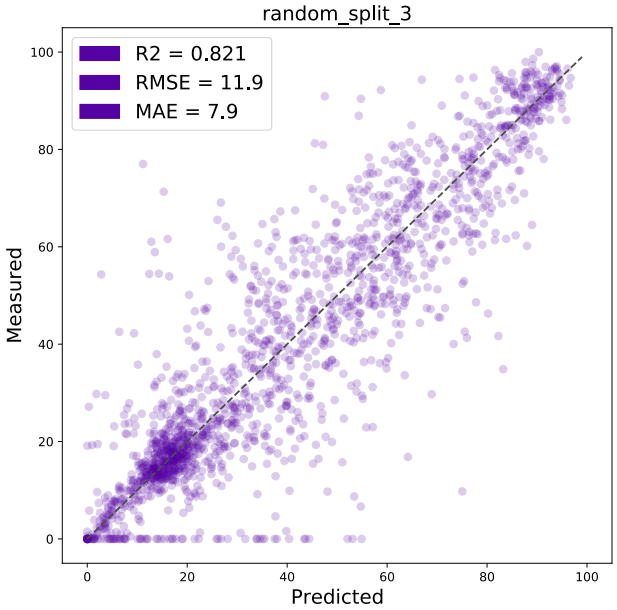

Figure E.18: Measured vs predicted yields [%] - random_split_3

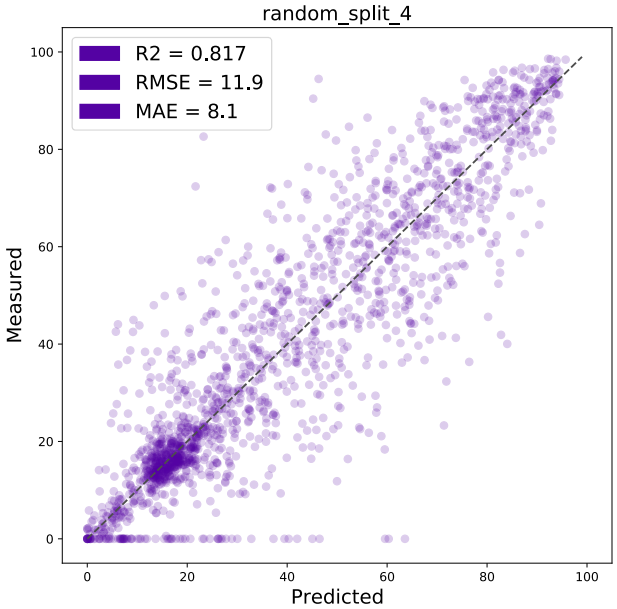

Figure E.19: Measured vs predicted yields [%] - random_split_4

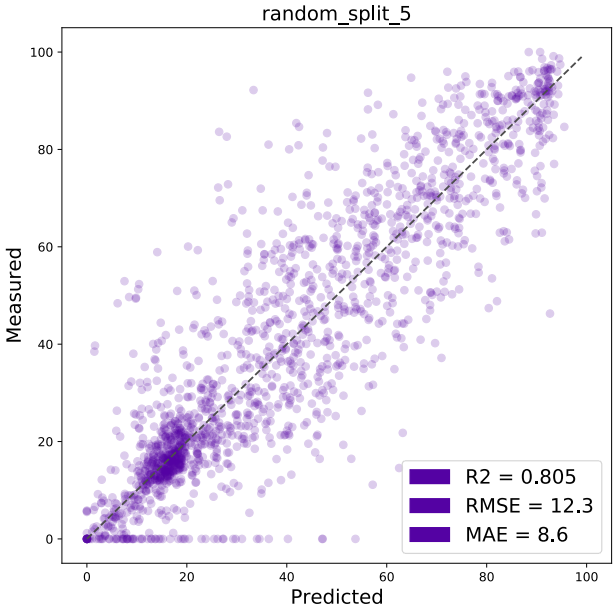

Figure E.20: Measured vs predicted yields [%] - random_split_5

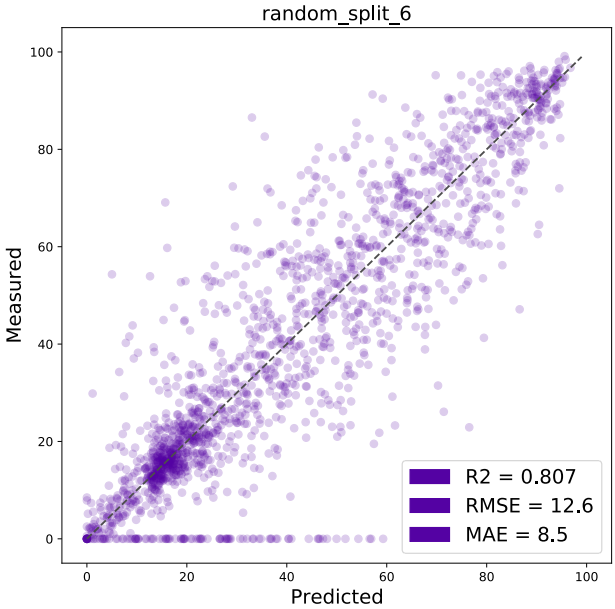

Figure E.21: Measured vs predicted yields [%] - random_split_6

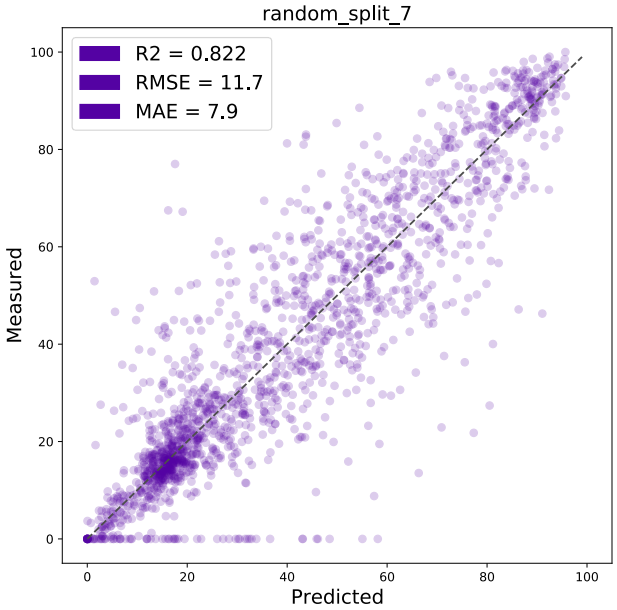

Figure E.22: Measured vs predicted yields [%] - random_split_7

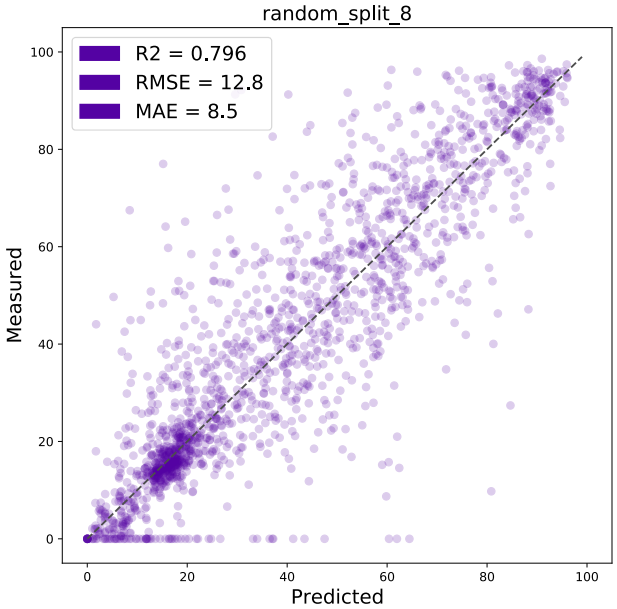

Figure E.23: Measured vs predicted yields [%] - random_split_8

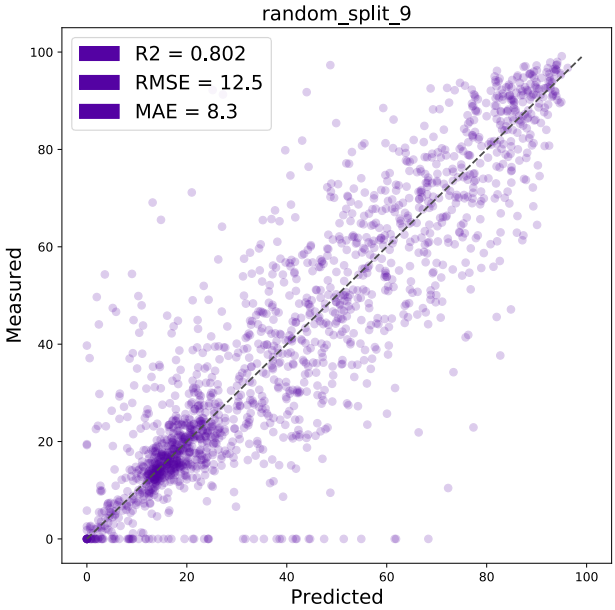

Figure E.24: Measured vs predicted yields [%] - random_split_9

E.3 Detailed analysis of USPTO yields data

Table E.1 show global statistics on the gram scale and sub-gram scale USPTO yields data sets.

Table E.1: USPTO yield statistics

	gram scale	subgram scale
count	197619	302040
mean	73.2	56.8
std	20.9	26.6
min	0.0	0.0
25%	60.2	35.5
50%	78.0	58.9
75%	90.3	79.5
max	100.0	100.0

Tables E.2 and E.3 show the yields average in the random split test set for the different reaction superclasses.

Figure E.25 shows the distributions of the smoothed yields. To smooth the yields of the USPTO data set [41, 42] we calculated the average of the 3 nearest-neighbours of the reaction, computed using the *rxnfp ft* [177] and faiss [230], and twice the own reaction yield.

Table E.2: **Test set sub-gram scale.** Average and standard deviation per class.

Class	Name	Mean [%]	Std	Count
0	Unrecognised	52.1	26.8	12359
1	Heteroatom alkylation and arylation	53.3	25.8	12995
2	Acylation and related processes	54.8	25.6	10583
3	C-C bond formation	53.2	25.6	5111
4	Heterocycle formation	48.0	25.1	2043
5	Protections	69.8	22.3	527
6	Deprotections	68.7	25.2	8542
7	Reductions	67.5	26.1	3528
8	Oxidations	63.4	25.3	1078
9	Functional group interconversion (FGI)	62.3	25.2	2779
10	Functional group addition (FGA)	56.2	25.1	863

Table E.3: **Test set gram scale.** Average and standard deviation per class.

Class	Name	Mean [%]	Std	Count
0	Unrecognised	69.4	22.0	10327
1	Heteroatom alkylation and arylation	71.9	20.9	7912
2	Acylation and related processes	74.5	19.7	4745
3	C-C bond formation	70.7	20.0	2547
4	Heterocycle formation	67.1	22.9	1417
5	Protections	79.9	18.5	1154
6	Deprotections	82.2	16.9	3332
7	Reductions	81.2	18.2	3105
8	Oxidations	76.0	18.8	742
9	Functional group interconversion (FGI)	74.9	20.1	2751
10	Functional group addition (FGA)	71.7	21.7	1491

E.4 Hyperparameter tuning

The two hyperparameters we tuned were dropout rate (between 0.05 and 0.8) and learning rate (between 1e-6 and 1e-4). For the *rxnfp pretrained* model on the Buchwald-Hartwig reactions a learning rate of 9.659e-05 and dropout probability of 0.7987 led to the highest validation R^2 score. We observe high R^2 scores for a wide range of dropout probabilities. The hyperparameter tuning was performed on a single *Nvidia RTX 2070 super* GPU and the optimal hyperparameters were found in less than 12 hours. A typical training run (10 epochs) on the same hardware takes 4 minutes and 30 seconds. We trained the final models for 15 epochs.

On the Suzuki-Miyaura reactions, we selected a learning rate of 5.812e-05 and dropout probability of 0.5848 for the *rxnfp pretrained* base encoder and a learning rate of 9.116e-05 and dropout probability 0.7542 for the *rxnfp ft* base encoder model.

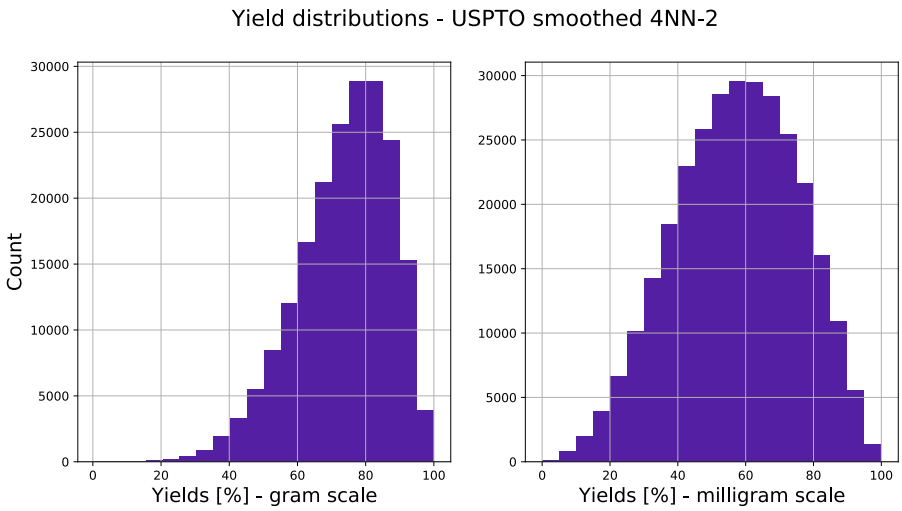

Figure E.25: **Smoothed USPTO yields.** Distribution separated in gram and sub-gram scale

On the USPTO data we performed a hyperparameter search using a reduced training set of 50k reactions and only 3 epochs. We selected a learning rate of 1.562e-05 and dropout probability of 0.5237 for the gram scale and 2.958e-05 and 0.5826 respectively, for the sub-gram scale. The final models were trained for 2 epochs on the complete training data, as an evaluation showed signs of over-fitting from the third epochs on.

Figure E.26 – E.30 show the hyperparameters with the corresponding R^2 values on the validation set. The validation was made on subsplit of the training set of the first random split for all three data sets. Overall, the learning rate seemed to be more important to tune than the dropout probability.

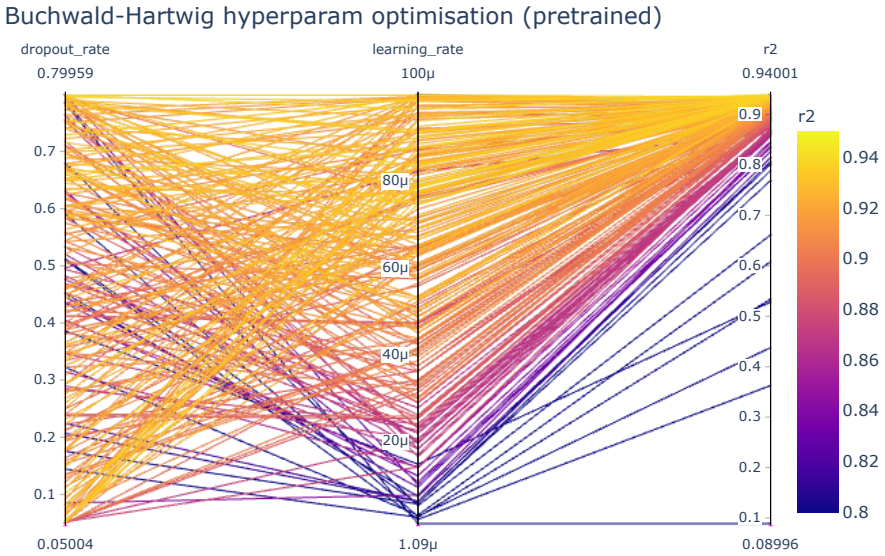

Figure E.26: Hyperparameter optimisation on Buchwald-Hartwig data set (pretrained base encoder)

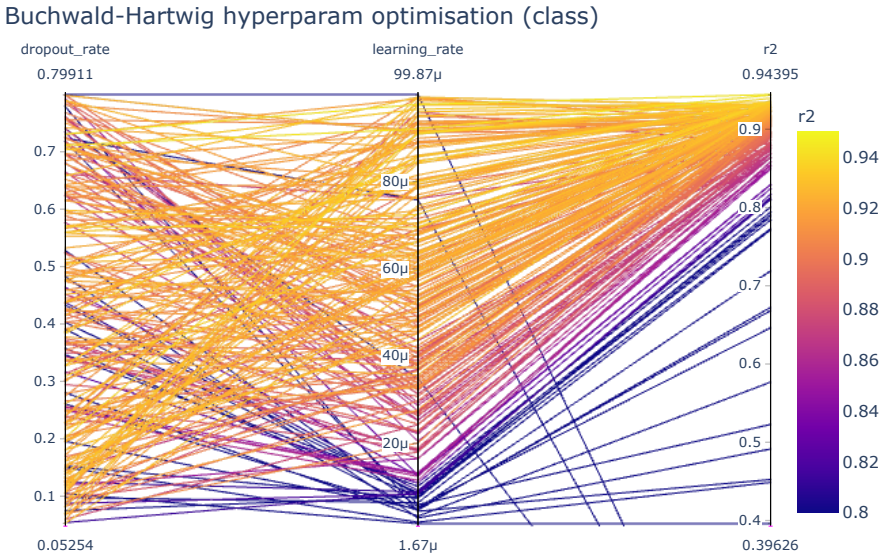

Figure E.27: Hyperparameter optimisation on Buchwald-Hartwig data set (class base encoder)

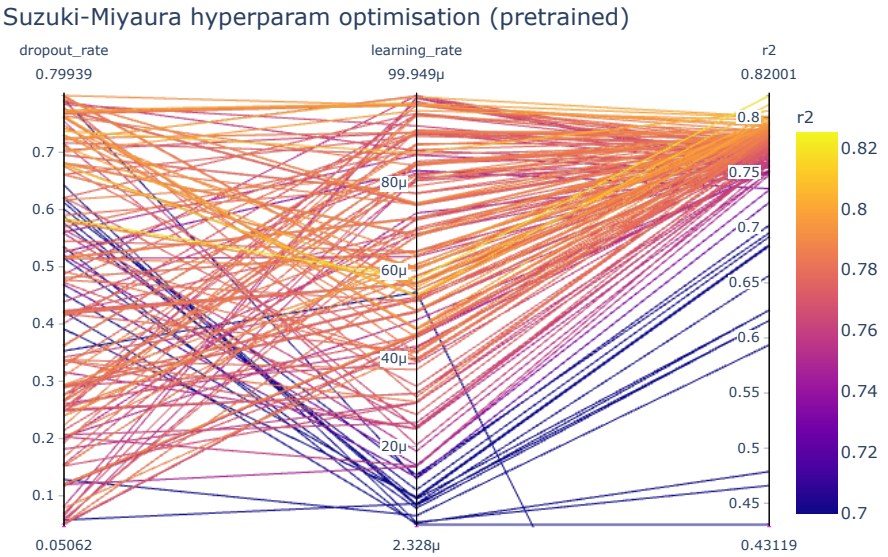

Figure E.28: Hyperparameter optimisation on Suzuki-Miyaura data set (pretrained base encoder)

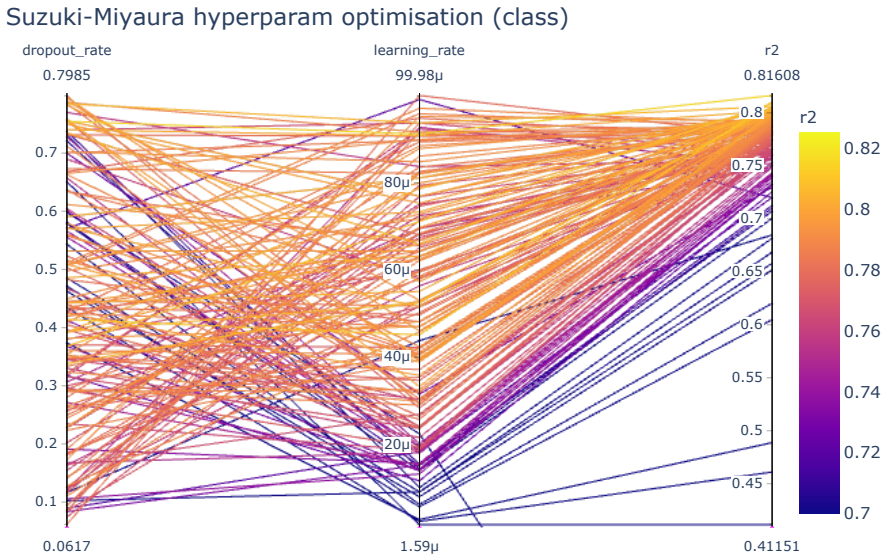

Figure E.29: Hyperparameter optimisation on Suzuki-Miyaura data set (class base encoder)

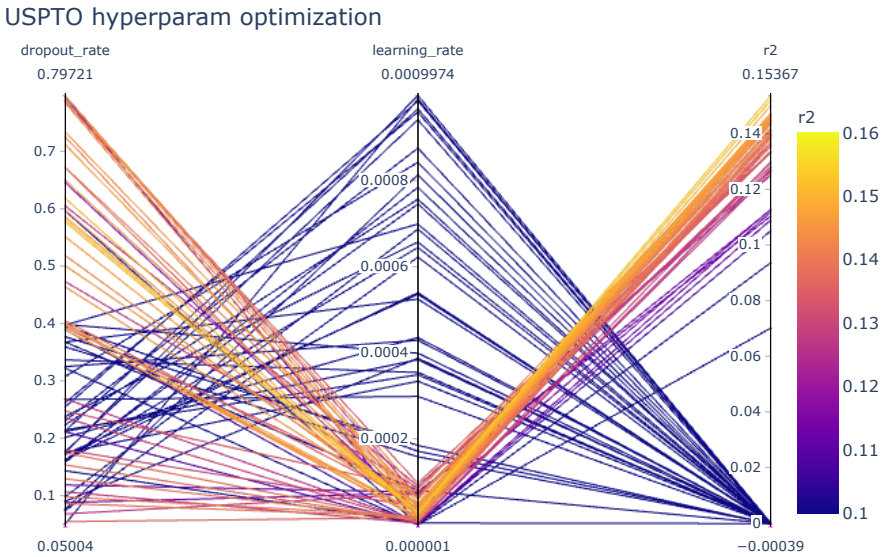

Figure E.30: Hyperparameter optimisation on USPTO subgram data set (pretrained base encoder)

F Appendix: Extraction of organic chemistry grammar from unsupervised learning of chemical reactions

F.1 Detailed evaluation

49k Schneider test set

Reaction class	Total (curated)	Matching [%]	Correct [%]
Heteroatom alkylation and arylation	14836 (14698)	96.8	99.2
Acylation and related processes	11670 (11593)	95.7	99.8
C-C bond formation	5550 (5502)	98.0	99.4
Heterocycle formation	889 (881)	90.6	94.7
Protections	655 (652)	97.4	98.6
Deprotections	8055 (7983)	98.1	99.9
Reductions	4499 (4466)	97.6	99.1
Oxidations	809 (805)	98.0	99.9
Functional group interconversion (FGI)	1809 (1775)	96.2	99.8
Functional group addition (FGA)	228 (228)	89.0	99.1
All	49000 (48583)	96.8	99.4

Table F.1: Results on the 49k patent test set

Table F.1 provides results on the 49k patent test set. Overall, the generated atom-maps exactly match the original atom-maps in 96.8% of the cases. After removing questionable reactions from the statistics and counting the equivalent mappings as correct, the overall correctness increased to 99.4%. Table F.1 shows the atom-mapping correctness divided into the different superclasses, where heterocycle formations were the most challenging superclass with 94.7% correctness.

While analyzing the discrepancies in the atom-mapping generated on the 49k patents test set, we labelled 369 as questionable and 47 as unclear. Questionable reactions typically contain multiple products similar to reactants, as in Figure F.1 a). The reason could be a wrong extraction from patents. Unclear reactions, on the other hand, have correct reactants but miss reagents, which are crucial to determine the reaction mechanism. The example shown in Figure F.1 b) looks like

a Mitsunobu reaction but the DEAD or DIAD reagents are not present. Despite the missing reagents, RXNMapper would have correctly mapped the phenolic alcohol.

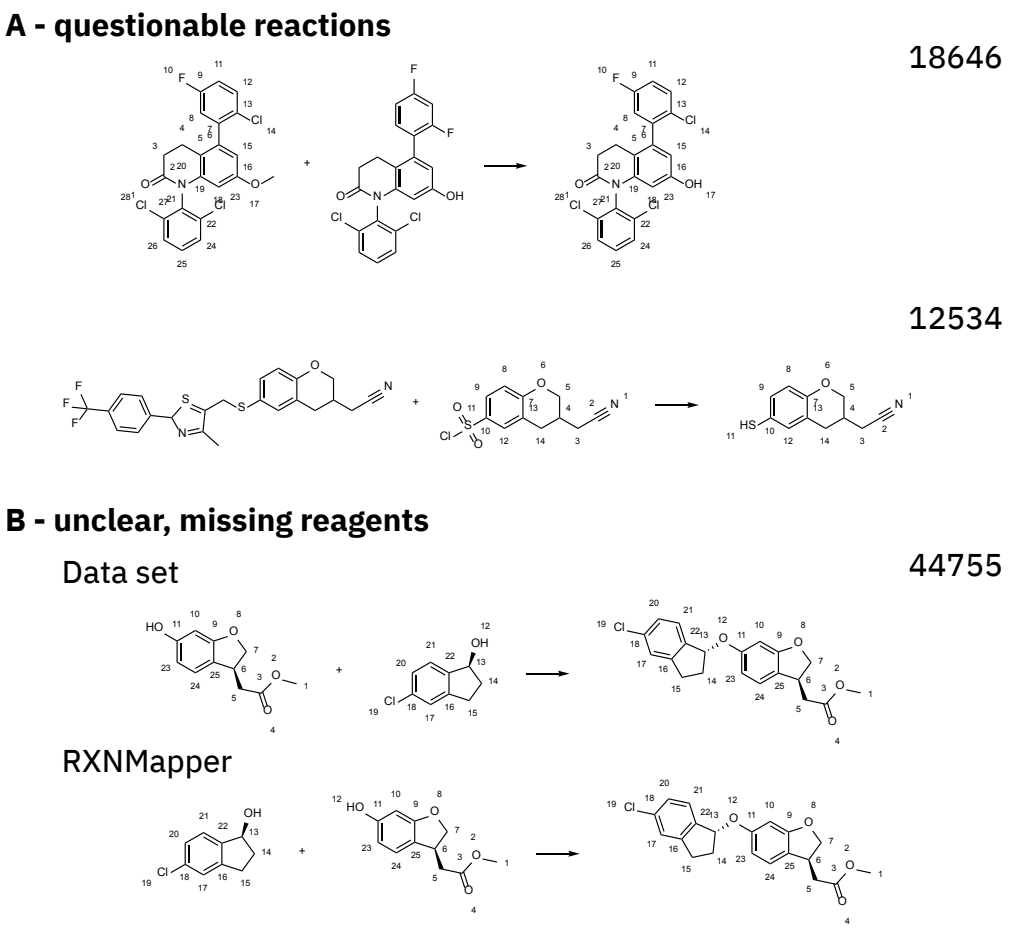

Figure F.1: Examples of (A) reactions that were classified as questionable. (B) a reaction for which the correct atom-mapping is unclear as critical reagents are missing

Figure F.2 shows reactions that were counted as correct even though the atom-mapping was not identical with the one in the data set. Such reactions typically have two equivalent atoms or symmetry operations that make the atom maps equivalent. If there was twice the same molecule on the product side, the atom-mappings in the original data set pointed for both molecules to the same atoms in the reactants. In contrast, our algorithm in the default configuration mapped different atoms in the reactants.

Comparison with atom-mapping tools

Recently, Jaworski et al. [61] developed an atom-mapper based on graph-theoretical approach augmented with human-expert written rules. They compared their tool called Mappet[61] to other

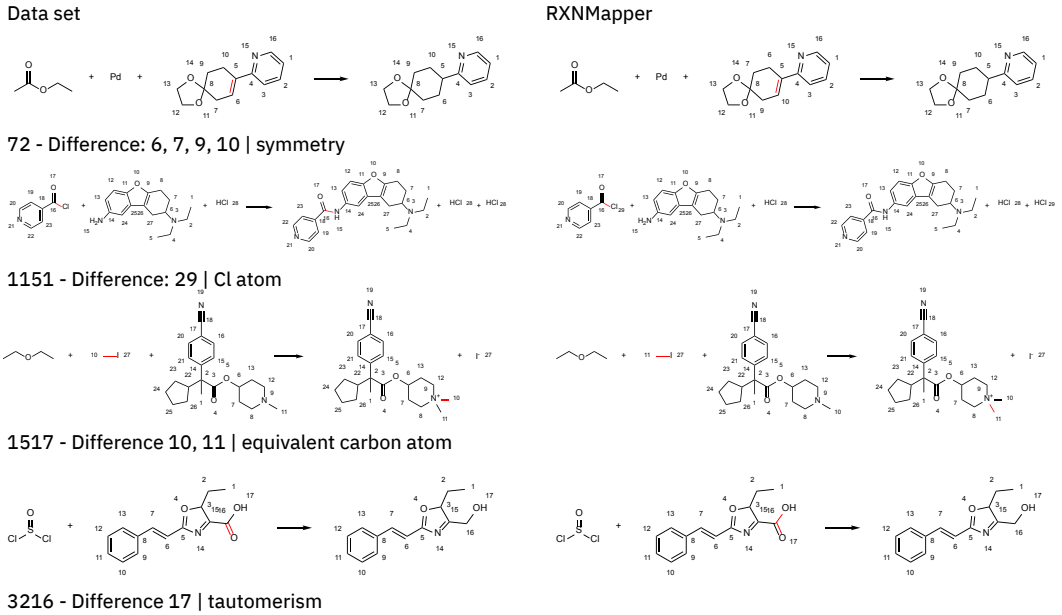

Figure F.2: Examples of atom-mappings that differed from the data set but were counted as equally correct.

methods. We performed the same tests using our RXNMapper. Figure F.3 shows the correctness on three different test sets of our attention-based RXNMapper, Mappet [61], Marvin JS (version 16.4.18) [302], ReactionMap [34], ChemDraw Prime (version 16.0.0.82), and Indigo (version 1.3.0 beta)[60]. The simple reactions set consists of 100 reactions from total syntheses reported in Org. Lett., J. Am. Chem. Soc., and J. Org. Chem., whereas the typical reactions set consists of 100 almost, but not fully, balanced patent reactions. RXNMapper achieves correctness scores similar to Mappet on both these sets. On the complex reaction set, which consists of 201 mechanistically complex reactions from recent literature, we perform slightly worse than Mappet but better than other reported methods. Still, the results are impressive as RXNMapper was not tuned specifically for any of these test sets. An overview of the test sets can be found in Table F.2.

Test set	Number of reactions	Avg. number of reactant atoms	Avg. number of product atoms
Simple reactions [61]	100	27.1	27.1
Typical reactions [61]	100	19.9	19.6
Complex reactions [61]	201	25.7	24.8

Table F.2: Data sets for the comparison with other tools.

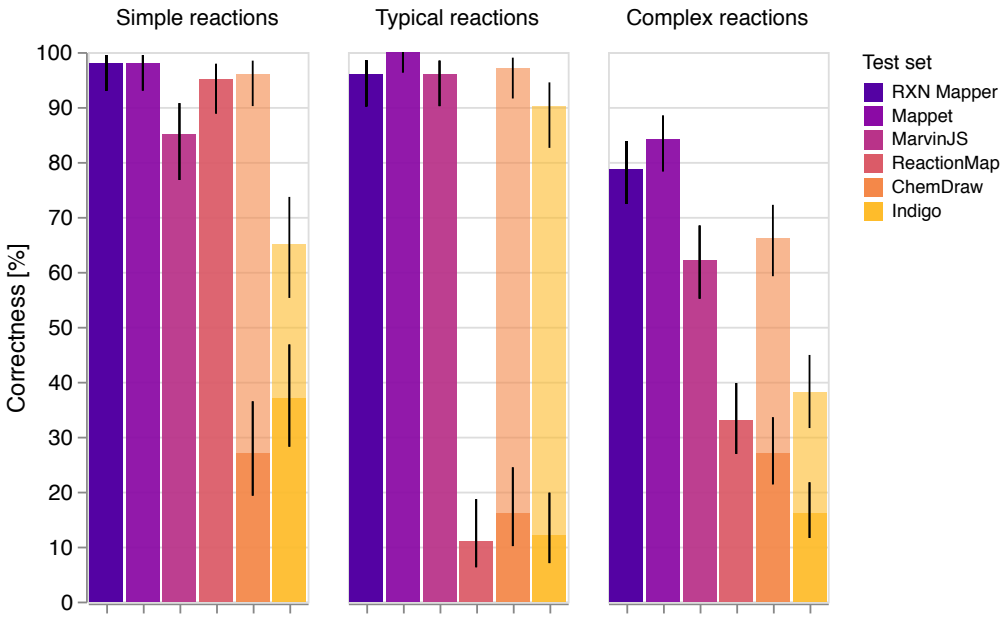

Figure F.3: Tool comparison, test originally published by Jaworski et al. [61]. The error bars show the Wilson confidence interval [267].

Computational performance

In contrast to previous methods, RXNMapper does not require balanced or almost balanced reactions. It can compute atom-mapping for both patent reactions and reactions predicted by template-free reaction prediction models. RXNMapper maps the 682 balanced reactions from the work of Jaworski et al. [61] at 33.3 reactions per second (30 ms/reaction) on a MacBook Pro: 2.7 GHz Intel Core i7, 16 GB 2133 MHz LPDD and reaches 156.2 reactions per second (6.4 ms/reaction), when the attention model inference is accelerated using a GPU (Nvidia RTX 2070 super). The computational performance is nearly the same when mapping reactions from the 49k patent reaction data set, which are mapped at a speed of 27.5 reactions per second (36.4 ms/reaction) on CPU only and 130 reactions per second (7.7 ms/reaction) using a GPU. In terms of speed RXNMapper performs similar to Indigo toolkit [60] on the balanced reactions, RXNMapper significantly outperforms Indigo on the patent reactions that contain many more reactants. The computational performance makes it feasible to apply RXNMapper to large reaction data sets in a reasonable time. We remapped the largest open-source reaction data set [42] at an average speed of 7.37 ms/reaction and made it available at `https://github.com/rxn4chemistry/rxnmapper`.

F.2 Confidence score

The confidence score for atom-mapping is computed by multiplying the selected attention scores for all the mapped product atoms. As seen in Figure F.4, correctly generated atom-mappings have,

on average, a higher confidence score than those that contain mistakes. Questionable reactions (e.g., where the reaction was wrongly extracted from patents) contain the lowest confidence scores.

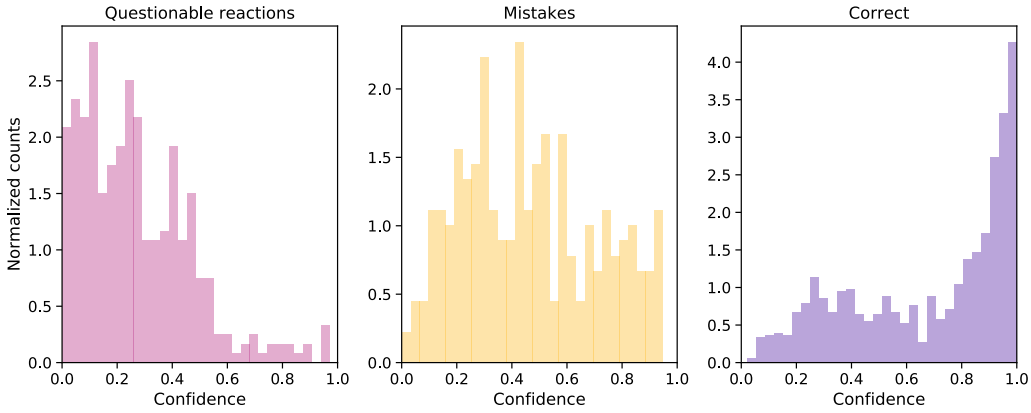

Figure F.4: Normalized histograms of confidence scores on three categories of atom-mappings: atom-mappings on questionable reactions, wrongly generated atom-mappings and correct atom mappings.

F.3 Hyperparameters and model selection

Hyperparameters

We trained the models for 48 hours on a single Nvidia P100 GPU with a masked language masking probability of 0.15. We used the training scripts from huggingface [216] adapted to work with a SmilesTokenizer, which we made available. For the ALBERT models, we fixed the number of layers to 12, the activation function to GELU, the dropout probability for 0.1, the embedding size to 128, the intermediate size to 512. We varied both the hidden size and the number of heads. The model with 8 heads uses a hidden size of 256, the model with 10 heads uses a hidden size of 320, and the model with 12 heads uses a hidden size of 384. We experimented with larger models, but the differences in atom mapping correctness were marginal. Our final model has only 770k trainable parameters, which is small compared to BERT base [3] with 108M and ALBERT base [98] with 12M parameters.

Model selection

The improvement of the atom-mapping correctness may increase up to 30% when changing the neighbour attention multiplier from 1 (basic algorithm) to a value of 20. Figure F.5 shows the atom-mapping correctness on the validation reactions for all the heads and layers of different models. For the ALBERT pre-trained model, at least one head learned atom-mapping, and the position and role of the heads remained constant across all layers. The atom-mapping correctness increased in the first layers and is more or less constant from layer 7 to 11. In contrast, for the BERT model

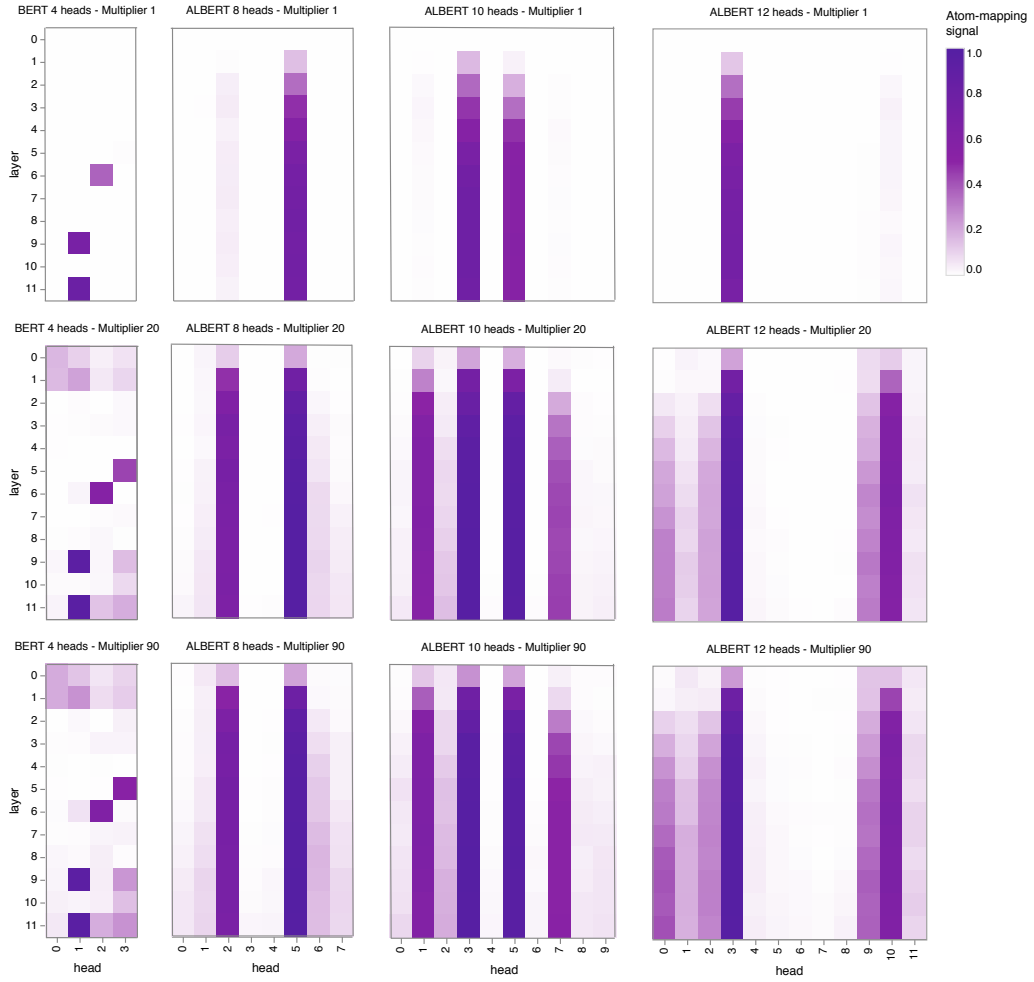

Figure F.5: Atom-mapping performance of all layers and heads of one BERT and 3 ALBERT models on the patent validation set with multipliers of 1, 20 and 90.

does not share weights across layers and only particular heads in particular layers had learned an atom-mapping signal.

As shown in Figure F.6, the atom-mapping correctness steeply increases in the first 100k training steps then continues to increase more slowly. We observed this behaviour for all models we trained. Moreover, models with more heads seemed to learn the atom-mapping signal faster, but the models with fewer heads quickly beat the performance of the larger models.

The top-20 model combinations are shown in Table F.3. We selected checkpoint 1310k (layer 10, head 5) as the best performing model on the 1k patent validation set. We used this model to perform all experiments in the main paper.

As shown in Figure F.7, increasing the nearest neighbour multiplier increases the atom-wise and full reaction atom-mapping correctness.

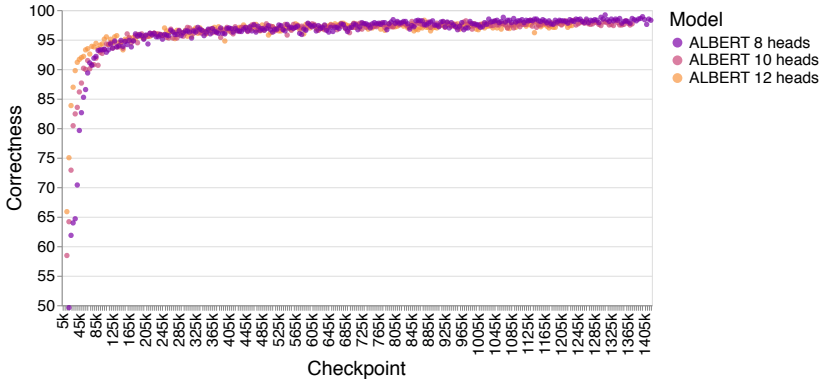

Figure F.6: Evaluation atom-mapping correctness for checkpoints every 5k training steps on the validation set for ALBERT models with 8, 10 and 12 heads. The layer was fixed to 10, the multiplier to 90 and the head with the largest atom-mapping signal was selected.

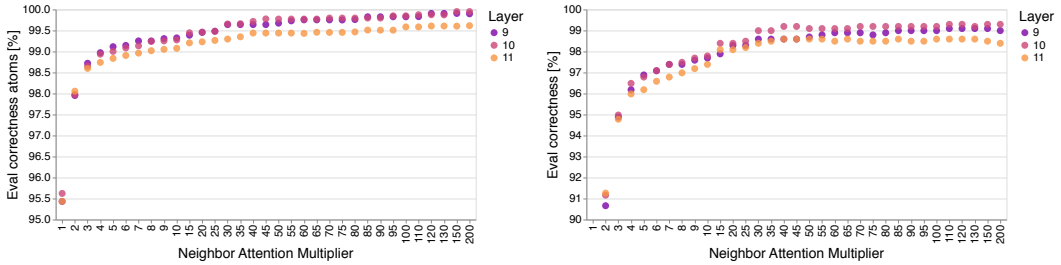

Figure F.7: Evaluation atom-mapping correctness per atom (left) and per reaction (right) for different multiplier.

F.4 Visualisation of self-attention

Visual inspection of the attention weights enabled the initial discovery that molecular Transformer models learned atom-mapping as a key signal. We release a tool called RXNMapper-Vis that allows others to explore the attentions of the ALBERT model behind RXNMapper interactively and make new hypotheses. RXNMapper-Vis maps the attentions from the tokenised SMILES onto a 2D skeletal structure to ease interpretation. The tool has been made available at https://rxnmapper.ai.

RXNMapper-Vis was inspired by previous work to visualise the attentions of Transformer models in the natural language processing (NLP) [129, 130, 261]. These tools can reveal learned but hidden behaviours of Transformers such as hidden language dependencies and parts of speech (e.g., attentions linking root Verbs to their Direct Objects), coreference (e.g., "she" attending to "mother"), entities (e.g., "Elon Musk" or "Iran"), and gender biases associated with particular roles (e.g., models predicting "he" as the necessary pronoun for "doctor"). Some of these learned patterns correlate to properties within the chemical domain. For example, coreference correlates to the learned atom-mapping behaviour discussed in this paper. We hope that others will be able to

	name	checkpoint	layer	head	Atom acc. [%]	Correctness [%]
11740	ALBERT 8 heads	1310k	10	5	99.8	99.2
12009	ALBERT 8 heads	1400k	9	5	99.9	99.1
11739	ALBERT 8 heads	1310k	9	5	99.8	99.0
12010	ALBERT 8 heads	1400k	10	5	99.7	98.9
11709	ALBERT 8 heads	1300k	9	5	99.7	98.9
11710	ALBERT 8 heads	1300k	10	5	99.7	98.8
11005	ALBERT 8 heads	1065k	10	5	99.6	98.8
11291	ALBERT 8 heads	1160k	11	5	99.8	98.7
11604	ALBERT 8 heads	1265k	9	5	99.8	98.6
11845	ALBERT 8 heads	1345k	10	5	99.8	98.6
11995	ALBERT 8 heads	1395k	10	5	99.7	98.6
11996	ALBERT 8 heads	1395k	11	5	99.7	98.6
11006	ALBERT 8 heads	1065k	11	5	99.7	98.6
11935	ALBERT 8 heads	1375k	10	5	99.6	98.6
11679	ALBERT 8 heads	1290k	9	5	99.6	98.6
11381	ALBERT 8 heads	1190k	11	5	99.5	98.6
11080	ALBERT 8 heads	1090k	10	5	99.4	98.6
11289	ALBERT 8 heads	1160k	9	5	99.8	98.5
11560	ALBERT 8 heads	1250k	10	5	99.7	98.5
11725	ALBERT 8 heads	1305k	10	5	99.7	98.5

Table F.3: Top-20 model/layer/head combinations by correctness on the validation set for a multiplier of 90.

use RXNMapper-Vis to find meaningful patterns in the layers and heads of the molecular Transformer model and that these discoveries can enrich our knowledge and improve our tooling for the chemical domain.

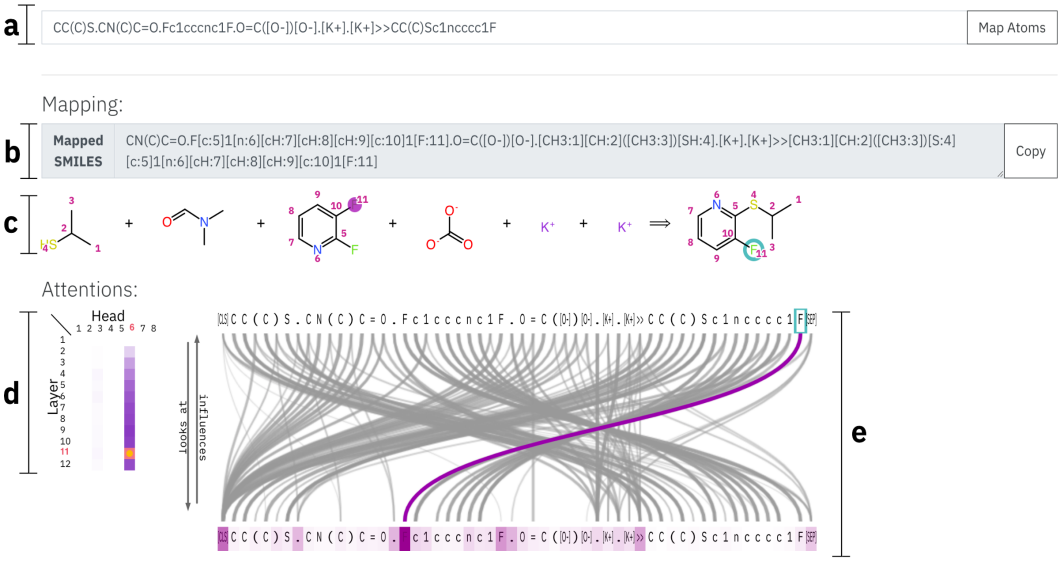

Figure F.8: An overview of RXNMapper-Vis. Users can insert their reaction SMILES in (a), and the tool will display the atom-mapped string in (b). A 2D skeletal structure depiction of the SMILES is shown in (c). Hovering over any atom will show the attention weights out of that atom and onto all the other atoms. Clicking on an atom will freeze that particular attention view. The attentions of different heads and layers can be inspected in (d), where darker backgrounds of each cell indicate a higher performance at atom-mapping. Note that atom labels in (c) only show for the atom-mapping head. Changing the selected layer/head combination will update the attentions in (c) and (e). The attention graph in (e) shows the self-attention of the input as a connected graph, where darker and thicker curves indicate a higher attention weight out of tokens in the top row into each token in the bottom row. Hovering over any token highlights the connected attentions in the graph and the corresponding atoms in (c). Here, the Fluorine in the product is selected, and both the attention graph and the skeletal structure show the greatest attention to the correct reactant atom. The complete discrete probability distribution of the attentions is shown as a purple background over the input sequence.

Abbreviations

AI	Artificial intelligence
ALBERT	A Lite BERT
API	Application programming interface
BERT	Bidirectional Encoder Representations from Transformers
CAMEO	Computer-assisted mechanistic evaluation of organic reactions
CARBO	Carbohydrate reactions data set
CEN	Confusion entropy of confusion matrix
CML	Chemical Markup Language
CPU	Central processing unit
DFT	Density functional theory
DL	Deep learning
ECFP	Extended-Connectivity Fingerprint
ELN	Electronic lab notebooks
EROS	Elaboration of Reactions for Organic Synthesis
ESI	Electron spray ionisation
FAISS	Facebook AI Similarity Search
GCNN	Graph convolutional neural network
GLUE	General Language Understanding Evaluation
GPT	Generative Pre-trained Transformer
GPU	Graphics processing unit
GRU	Gated recurrent unit
HOMO	Highest occupied molecular orbital
HRMS	High resolution mass spectra
HTE	High-throughput experimentation
IBM	International Business Machines Corporation
InChI	International Chemical Identifier
IUPAC	International Union of Pure and Applied Chemistry
JSD	Jensen-Shannon divergence
LHASA	Logic and Heuristics Applied to Synthetic Analysis
LLO	Lipid linked oligosacccharide
LSH	Locality-sensitive hashing
LSTM	Long-Short Term Memory
MCC	Matthews correlation coefficient
MCTS	Monte Carlo tree search
MFF	Multiple fingerprint features
ML	Machine learning
MLM	Masked Language Modelling

MT	Molecular Transformer
NLP	Natural language processing
NMR	Nuclear magnetic resonance
OST	Oligosaccharil transferase
R&D	Research and development
ReLU	Rectified linear unit
RF	Random forest
RInChI	Reaction International Chemical Identifier
RXN	Reaction
RXNFP	Reaction fingerprint
SCscore	Synthetic complexity score
SECS	Simulation and Evaluation of Chemical Synthesis
SELFIES	SELF-referencing Embedded Strings
seq-2-seq	Sequence-to-sequence
SMARTS	SMILES arbitrary target specification
SMILES	Simplified molecular-input line-entry system
SOPHIA	System for organic reaction prediction by heuristic approach
TLC	Thin layer chromatography
TMAP	Tree map
TPL	Templates
TTA	Test-time augmentation
USPTO	United States Patent and Trademark Office
WLDN	Weisfeiler-Lehman difference network
WLN	Weisfeiler-Lehman network
WODKA	Workbench for the Organization of Data for Chemical Applications
XML	Extensible Markup Language

Bibliography

1. D. C. Blakemore, L. Castro, I. Churcher, D. C. Rees, A. W. Thomas, D. M. Wilson, and A. Wood. "Organic synthesis provides opportunities to transform drug discovery." *Nat. Chem.* 10:4, 2018, pp. 383–394.
2. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. "Attention is all you need". In: *Advances in Neural Information Processing Systems*. 2017, pp. 5998–6008.
3. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2019, pp. 4171–4186.
4. P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas, and A. A. Lee. "Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction". *ACS Cent. Sci.* 5:9, 2019, pp. 1572–1583.
5. W. Jin, C. Coley, R. Barzilay, and T. Jaakkola. "Predicting organic reaction outcomes with weisfeiler-lehman network". In: *Advances in Neural Information Processing Systems*. 2017, pp. 2607–2616.
6. C. W. Coley, W. Jin, L. Rogers, T. F. Jamison, T. S. Jaakkola, W. H. Green, R. Barzilay, and K. F. Jensen. "A graph-convolutional neural network model for the prediction of chemical reactivity". *Chem. Sci.* 10:2, 2019, pp. 370–377.
7. M. Campbell, A. J. Hoane, and F.-h. Hsu. "Deep Blue". *Artif. Intell.* 134:1, 2002, pp. 57–83. ISSN: 0004-3702.
8. D. Ferrucci, E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. A. Kalyanpur, A. Lally, J. W. Murdock, E. Nyberg, J. Prager, N. Schlaefer, and C. Welty. "Building Watson: An Overview of the DeepQA Project". en. *AI Mag.* 31:3, 2010, pp. 59–79. ISSN: 2371-9621.
9. D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis. "Mastering the game of Go with deep neural networks and tree search". en. *Nature* 529:7587, 2016, pp. 484–489.
10. D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, and D. Hassabis. "Mastering the game of Go without human knowledge". en. *Nature* 550:7676, 2017, pp. 354–359.

11. E. J. Corey. "General methods for the construction of complex molecules". *Pure Appl. Chem.* 14:1, 1967, pp. 19–38.
12. E. J. Corey, A. K. Long, and S. D. Rubenstein. "Computer-assisted analysis in organic synthesis". en. *Science* 228:4698, 1985, pp. 408–418.
13. E. Corey, W. T. Wipke, R. D. Cramer III, and W. J. Howe. "Computer-assisted synthetic analysis. Facile man-machine communication of chemical structure by interactive computer graphics". *J. Am. Chem. Soc.* 94:2, 1972, pp. 421–430.
14. W. T. Wipke and T. M. Dyott. "Simulation and evaluation of chemical synthesis. Computer representation and manipulation of stereochemistry". *J. Am. Chem. Soc.* 96:15, 1974, pp. 4825–4834.
15. H. L. Gelernter, A. F. Sanders, D. L. Larsen, K. K. Agarwal, R. H. Boivie, G. A. Spritzer, and J. E. Searleman. "Empirical Explorations of SYNCHEM". en. *Science* 197:4308, 1977, pp. 1041–1049.
16. H. Gelernter, J. R. Rose, and C. Chen. "Building and refining a knowledge base for synthetic organic chemistry via the methodology of inductive and deductive machine learning". *J. Chem. Inf. Comput. Sci.* 30:4, 1990, pp. 492–504.
17. J. Gasteiger and C. Jochum. "EROS A computer program for generating sequences of reactions". en. *Org. Compd.*, 1978, pp. 93–126.
18. J. Gasteiger, W. D. Ihlenfeldt, and P. Röse. "A collection of computer methods for synthesis design and reaction prediction". en. *Recl. Trav. Chim. Pays-Bas* 111:6, 1992, pp. 270–290.
19. J. Dugundji and I. Ugi. "An algebraic model of constitutional chemistry as a basis for chemical computer programs". en. In: *Computers in Chemistry*. Fortschritte der Chemischen Forschung. Springer Berlin Heidelberg, 1973, pp. 19–64. ISBN: 978-3-540-38510-3.
20. W. L. Jorgensen, E. R. Laird, A. J. Gushurst, J. M. Fleischer, S. A. Gothe, H. E. Helson, G. D. Paderes, and S. Sinclair. "CAMEO: a program for the logical prediction of the products of organic reactions". *Pure Appl. Chem.* 62:10, 1990, pp. 1921–1932.
21. H. Satoh and K. Funatsu. "SOPHIA, a knowledge base-guided reaction prediction system-utilization of a knowledge base derived from a reaction database". *J. Chem. Inf. Comput. Sci.* 35:1, 1995, pp. 34–44.
22. B. A. Grzybowski, S. Szymkuć, E. P. Gajewska, K. Molga, P. Dittwald, A. Wołos, and T. Klucznik. "Chematica: A Story of Computer Code That Started to Think like a Chemist". *Chem* 4:3, 2018, pp. 390–398.
23. S. Szymkuć, E. P. Gajewska, T. Klucznik, K. Molga, P. Dittwald, M. Startek, M. Bajczyk, and B. A. Grzybowski. "Computer-Assisted Synthetic Planning: The End of the Beginning". *Angew. Chem. Int. Ed.* 55:20, 2016, pp. 5904–5937.
24. B. Mikulak-Klucznik, P. Gołębiowska, A. A. Bayly, O. Popik, T. Klucznik, S. Szymkuć, E. P. Gajewska, P. Dittwald, O. Staszewska-Krajewska, W. Beker, et al. "Computational planning of the synthesis of complex natural products". *Nature*, 2020, pp. 1–6.

25. O. Engkvist, P.-O. Norrby, N. Selmi, Y.-h. Lam, Z. Peng, E. C. Sherer, W. Amberg, T. Erhard, and L. A. Smyth. "Computational prediction of chemical reactions: current status and outlook". *Drug Discovery Today* 23:6, 2018, pp. 1203–1218.
26. W.-D. Ihlenfeldt and J. Gasteiger. "Computer-assisted planning of organic syntheses: the second generation of programs". *Angew. Chem. Int. Ed.* 34:23-24, 1996, pp. 2613–2633.
27. M. H. Todd. "Computer-aided organic synthesis". *Chem. Soc. Rev.* 34:3, 2005, pp. 247–266.
28. A. Cook, A. P. Johnson, J. Law, M. Mirzazadeh, O. Ravitz, and A. Simon. "Computer-aided synthesis design: 40 years on". *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* 2:1, 2012, pp. 79–107.
29. M. A. Kayala, C.-A. Azencott, J. H. Chen, and P. Baldi. "Learning to predict chemical reactions." *J. Chem. Inf. Model.* 51:9, 2011, pp. 2209–2222.
30. M. A. Kayala and P. Baldi. "ReactionPredictor: prediction of complex chemical reactions at the mechanistic level using machine learning." *J. Chem. Inf. Model.* 52:10, 2012, pp. 2526–2540.
31. J. N. Wei, D. Duvenaud, and A. Aspuru-Guzik. "Neural networks for the prediction of organic chemistry reactions". *ACS Cent. Sci.* 2:10, 2016, pp. 725–732.
32. M. H. S. Segler and M. P. Waller. "Modelling Chemical Reasoning to Predict and Invent Reactions." *Chem. Eur. J.* 23:25, 2017, pp. 6118–6128.
33. M. H. Segler and M. P. Waller. "Neural-symbolic machine learning for retrosynthesis and reaction prediction". *Chem. Eur. J.* 23:25, 2017, pp. 5966–5971.
34. D. Fooshee, A. Andronico, and P. Baldi. "ReactionMap: An efficient atom-mapping algorithm for chemical reactions". *J. Chem. Inf. Model.* 53:11, 2013, pp. 2812–2819.
35. C. W. Coley, R. Barzilay, T. S. Jaakkola, W. H. Green, and K. F. Jensen. "Prediction of organic reaction outcomes using machine learning". *ACS Cent. Sci.* 3:5, 2017, pp. 434–443.
36. P. Schwaller, T. Gaudin, D. Lanyi, C. Bekas, and T. Laino. ""Found in Translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models". *Chem. Sci.* 9:28, 2018, pp. 6091–6098.
37. J. Bradshaw, M. Kusner, B. Paige, M. Segler, and J. Hernández-Lobato. "A generative model for electron paths". In: *7th International Conference on Learning Representations, ICLR 2019*. 2019.
38. K. Do, T. Tran, and S. Venkatesh. "Graph transformation policy network for chemical reaction prediction". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2019, pp. 750–760.
39. *Reaxys database*. (Accessed Oct 29, 2019). URL: https://www.reaxys.com.
40. *Scifinder database*. (Accessed Oct 29, 2019). URL: hhttps://www.cas.org/products/scifinder.
41. D. M. Lowe. "Extraction of chemical structures and reactions from the literature". PhD thesis. University of Cambridge, 2012.

42. D. Lowe. "Chemical reactions from US patents (1976-Sep2016)", 2017. DOI: 10.6084/m9.figshare.5104873.v1.

43. P. Schwaller, R. Petraglia, V. Zullo, V. H. Nair, R. A. Haeuselmann, R. Pisoni, C. Bekas, A. Iuliano, and T. Laino. "Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy". *Chem. Sci.* 11, 2020, pp. 3316–3325.

44. *IBM RXN for Chemistry*. (Accessed Sep 13, 2019). URL: https://rxn.res.ibm.com.

45. A. Dalby, J. G. Nourse, W. D. Hounshell, A. K. Gushurst, D. L. Grier, B. A. Leland, and J. Laufer. "Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited". *J. Chem. Inf. Comput. Sci.* 32:3, 1992, pp. 244–255.

46. D. Weininger. "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules". *J. Chem. Inf. Comput. Sci.* 28:1, 1988, pp. 31–36.

47. D. Weininger, A. Weininger, and J. L. Weininger. "SMILES. 2. Algorithm for generation of unique SMILES notation". *J. Chem. Inf. Comput. Sci.* 29:2, 1989, pp. 97–101.

48. P. Murray-Rust and H. S. Rzepa. "Chemical markup, XML, and the Worldwide Web. 1. Basic principles". *J. Chem. Inf. Comput. Sci.* 39:6, 1999, pp. 928–942.

49. G. L. Holliday, P. Murray-Rust, and H. S. Rzepa. "Chemical markup, XML, and the world wide web. 6. CMLReact, an XML vocabulary for chemical reactions". *J. Chem. Inf. Model.* 46:1, 2006, pp. 145–157.

50. G. Grethe, J. Goodman, and C. Allen. "International chemical identifier for chemical reactions". *J. Cheminf.* 5:1, 2013, pp. 1–1.

51. G. Grethe, G. Blanke, H. Kraut, and J. M. Goodman. "International chemical identifier for reactions (RInChI)". *J. Cheminf.* 10:1, 2018, pp. 1–9.

52. S. R. Heller, A. McNaught, I. Pletnev, S. Stein, and D. Tchekhovskoi. "InChI, the IUPAC international chemical identifier". *J. Cheminf.* 7:1, 2015, p. 23.

53. P.-M. Jacob, T. Lan, J. M. Goodman, and A. A. Lapkin. "A possible extension to the RInChI as a means of providing machine readable process data". *J. Cheminf.* 9:1, 2017, p. 23.

54. D. Rogers and M. Hahn. "Extended-connectivity fingerprints". *J. Chem. Inf. Model.* 50:5, 2010, pp. 742–754.

55. N. O'Boyle and A. Dalke. "DeepSMILES: An Adaptation of SMILES for Use in Machine-Learning of Chemical Structures", 2018. DOI: 10.26434/chemrxiv.7097960.v1.

56. M. Krenn, F. Hase, A. Nigam, P. Friederich, and A. Aspuru-Guzik. "Self-Referencing Embedded Strings (SELFIES): A 100% robust molecular string representation". *Mach. Learn.: Sci. Technol.*, 2020.

57. B. Sanchez-Lengeling and A. Aspuru-Guzik. "Inverse molecular design using machine learning: Generative models for matter engineering". *Science* 361:6400, 2018, pp. 360–365.

58. L. David, A. Thakkar, R. Mercado, and O. Engkvist. "Molecular representations in AI-driven drug discovery: a review and practical guide". *J. Cheminf.* 12:1, 2020, pp. 1–22.

59. D. M. Lowe and R. A. Sayle. "LeadMine: a grammar and dictionary driven approach to entity recognition". *J. Cheminf.* 7:1, 2015, pp. 1–9.

60. *Indigo Toolkit*. (Accessed Apr 02, 2020). URL: https://lifescience.opensource.epam.com/indigo/.

61. W. Jaworski, S. Szymkuć, B. Mikulak-Klucznik, K. Piecuch, T. Klucznik, M. Kaźmierowski, J. Rydzewski, A. Gambin, and B. A. Grzybowski. "Automatic mapping of atoms across both simple and complex chemical reactions". *Nat. Commun.* 10:1, 2019, pp. 1–11.

62. N. Schneider, N. Stiefl, and G. A. Landrum. "What's what: The (nearly) definitive guide to reaction role assignment". *J. Chem. Inf. Model.* 56:12, 2016, pp. 2336–2346.

63. *Nextmove Software NameRXN*. (Accessed Jul 29, 2019). URL: http://www.nextmovesoftware.com/namerxn.html.

64. N. Schneider, D. M. Lowe, R. A. Sayle, and G. A. Landrum. "Development of a novel fingerprint for chemical reactions and its application to large-scale reaction classification and similarity". *J. Chem. Inf. Model.* 55:1, 2015, pp. 39–53.

65. B. Liu, B. Ramsundar, P. Kawthekar, J. Shi, J. Gomes, Q. Luu Nguyen, S. Ho, J. Sloane, P. Wender, and V. Pande. "Retrosynthetic reaction prediction using neural sequence-to-sequence models". *ACS Cent. Sci.* 3:10, 2017, pp. 1103–1113.

66. W. S. McCulloch and W. Pitts. "A logical calculus of the ideas immanent in nervous activity". *Bull. Math. Biophys.* 5:4, 1943, pp. 115–133.

67. F. Rosenblatt. "The perceptron: a probabilistic model for information storage and organization in the brain." *Psychol. Rev.* 65:6, 1958, p. 386.

68. V. Nair and G. E. Hinton. "Rectified linear units improve restricted boltzmann machines". In: *Proceedings of the 27th International Conference on International Conference on Machine Learning*. 2010, pp. 807–814.

69. P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner, et al. "Relational inductive biases, deep learning, and graph networks". *arXiv: 1806.01261*, 2018.

70. Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel. "Handwritten digit recognition with a back-propagation network". *Advances in Neural Information Processing Systems* 2, 1989, pp. 396–404.

71. N. Kalchbrenner, E. Grefenstette, and P. Blunsom. "A Convolutional Neural Network for Modelling Sentences". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2014, pp. 655–665.

72. S. Ji, W. Xu, M. Yang, and K. Yu. "3D convolutional neural networks for human action recognition". *IEEE Trans. Pattern Anal. Mach. Intell.* 35:1, 2012, pp. 221–231.

73. D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams. "Convolutional networks on graphs for learning molecular fingerprints". In: *Advances in Neural Information Processing Systems*. 2015, pp. 2224–2232.

74. K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. "Learning phrase representations using RNN encoder-decoder for statistical machine translation". *arXiv: 1406.1078*, 2014.

75. S. Hochreiter and J. Schmidhuber. "Long short-term memory". *Neural Comput.* 9:8, 1997, pp. 1735–1780.

76. D. E. Rumelhart, G. E. Hinton, and R. J. Williams. "Learning representations by back-propagating errors". *Nature* 323:6088, 1986, pp. 533–536.

77. J. Kiefer, J. Wolfowitz, et al. "Stochastic estimation of the maximum of a regression function". *Ann. Math. Stat.* 23:3, 1952, pp. 462–466.

78. D. P. Kingma and J. Ba. "Adam: A method for stochastic optimization". *arXiv: 1412.6980*, 2014.

79. I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

80. M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. "Tensorflow: A system for large-scale machine learning". In: *12th USENIX symposium on operating systems design and implementation (OSDI 16)*. 2016, pp. 265–283.

81. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. "Pytorch: An imperative style, high-performance deep learning library". In: *Advances in Neural Information Processing Systems*. 2019, pp. 8026–8037.

82. W. P. Walters. "Code Sharing in the Open Science Era". *J. Chem. Inf. Model.* 60:10, 2020, pp. 4417–4420.

83. M.-T. Luong and C. D. Manning. "Achieving Open Vocabulary Neural Machine Translation with Hybrid Word-Character Models". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2016, pp. 1054–1063.

84. I. Sutskever, O. Vinyals, and Q. V. Le. "Sequence to sequence learning with neural networks". In: *Advances in Neural Information Processing Systems*. 2014, pp. 3104–3112.

85. T. Mikolov, K. Chen, G. Corrado, and J. Dean. "Efficient estimation of word representations in vector space". *arXiv: 1301.3781*, 2013.

86. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. "Distributed representations of words and phrases and their compositionality". *Advances in Neural Information Processing Systems* 26, 2013, pp. 3111–3119.

87. D. Bahdanau, K. Cho, and Y. Bengio. "Neural machine translation by jointly learning to align and translate". *arXiv: 1409.0473*, 2014.

88. M.-T. Luong, H. Pham, and C. D. Manning. "Effective approaches to attention-based neural machine translation". *arXiv: 1508.04025*, 2015.

89. A. M. Dai and Q. V. Le. "Semi-supervised sequence learning". In: *Advances in Neural Information Processing Systems*. 2015, pp. 3079–3087.

90. M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. "Deep contextualized word representations". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2018, pp. 2227–2237.
91. J. Howard and S. Ruder. "Universal Language Model Fine-tuning for Text Classification". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2018, pp. 328–339.
92. S. Ruder, M. E. Peters, S. Swayamdipta, and T. Wolf. "Transfer learning in natural language processing". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*. 2019, pp. 15–18.
93. S. Ruder. "Neural transfer learning for natural language processing". PhD thesis. NUI Galway, 2019.
94. A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. "Language models are unsupervised multitask learners". *OpenAI blog* 1:8, 2019, p. 9.
95. A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding". In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. 2018, pp. 353–355.
96. T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. "Language models are few-shot learners". *arXiv: 2005.14165*, 2020.
97. E. Strubell, A. Ganesh, and A. McCallum. "Energy and Policy Considerations for Deep Learning in NLP". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019, pp. 3645–3650.
98. Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations". In: *International Conference on Learning Representations*. 2020. URL: https://openreview.net/forum?id=H1eA7AEtvS.
99. V. Sanh, L. Debut, J. Chaumond, and T. Wolf. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter". *arXiv: 1910.01108*, 2019.
100. G. Hinton, O. Vinyals, and J. Dean. "Distilling the knowledge in a neural network". *arXiv: 1503.02531*, 2015.
101. N. Kitaev, Ł. Kaiser, and A. Levskaya. "Reformer: The efficient transformer". *arXiv: 2001.04451*, 2020.
102. I. Beltagy, M. E. Peters, and A. Cohan. "Longformer: The long-document transformer". *arXiv: 2004.05150*, 2020.
103. S. Wang, B. Li, M. Khabsa, H. Fang, and H. Ma. "Linformer: Self-Attention with Linear Complexity". *arXiv: 2006.04768*, 2020.

104. A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret. "Transformers are rnns: Fast autoregressive transformers with linear attention". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 5156–5165.

105. K. Choromanski, V. Likhosherstov, D. Dohan, X. Song, A. Gane, T. Sarlos, P. Hawkins, J. Davis, A. Mohiuddin, L. Kaiser, et al. "Rethinking attention with performers". *arXiv: 2009.14794*, 2020.

106. J. H. Chen and P. Baldi. "No electron left behind: a rule-based expert system to predict chemical reactions and reaction mechanisms". *J. Chem. Inf. Model.* 49:9, 2009, pp. 2034–2043.

107. M. H. Segler, M. Preuss, and M. P. Waller. "Planning chemical syntheses with deep neural networks and symbolic AI". *Nature* 555:7698, 2018, pp. 604–610.

108. C. W. Coley, W. H. Green, and K. F. Jensen. "Machine Learning in Computer-Aided Synthesis Planning." *Acc. Chem. Res.* 51:5, 2018, pp. 1281–1289.

109. A. Thakkar, T. Kogej, J.-L. Reymond, O. Engkvist, and E. J. Bjerrum. "Datasets and their influence on the development of computer assisted synthesis planning tools in the pharmaceutical domain". *Chem. Sci.* 11:1, 2020, pp. 154–168.

110. M. H. Segler, T. Kogej, C. Tyrchan, and M. P. Waller. "Generating focused molecule libraries for drug discovery with recurrent neural networks". *ACS Cent. Sci.* 4:1, 2018, pp. 120–131.

111. J. Nam and J. Kim. "Linking the neural machine translation and the prediction of organic chemistry reactions". *arXiv: 1612.09529*, 2016. (Accessed Aug 29, 2019).

112. R.-R. Griffiths, P. Schwaller, and A. Lee. "Dataset Bias in the Natural Sciences: A Case Study in Chemical Reaction Prediction and Synthesis Design", 2018. DOI: 10.26434/chemrxiv.7366973.v1.

113. W. W. Qian, N. T. Russell, C. L. Simons, Y. Luo, M. D. Burke, and J. Peng. "Integrating Deep Neural Networks and Symbolic Inference for Organic Reactivity Prediction", 2020. DOI: 10.26434/chemrxiv.11659563.v1.

114. I. V. Tetko, P. Karpov, R. Van Deursen, and G. Godin. "State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis". *Nat. Commun.* 11:1, 2020, pp. 1–11.

115. M. Sacha, M. Błaż, P. Byrski, P. Włodarczyk-Pruszyński, and S. Jastrzębski. "Molecule Edit Graph Attention Network: Modeling Chemical Reactions as Sequences of Graph Edits". *arXiv: 2006.15426*, 2020.

116. G. Pesciullesi, P. Schwaller, T. Laino, and J.-L. Reymond. "Transfer learning enables the molecular transformer to predict regio-and stereoselective reactions on carbohydrates". *Nat. Commun.* 11:1, 2020, pp. 1–8.

117. J. S. Schreck, C. W. Coley, and K. J. Bishop. "Learning Retrosynthetic Planning through Simulated Experience". *ACS Cent. Sci.*, 2019.

118. I. A. Watson, J. Wang, and C. A. Nicolaou. "A retrosynthetic analysis algorithm implementation". *J. Cheminf.* 11:1, 2019, p. 1.

119. R. Fagerberg, C. Flamm, R. Kianian, D. Merkle, and P. F. Stadler. "Finding the K best synthesis plans." *J. Cheminf.* 10:1, 2018, p. 19.

120. D. Lowe. "AI designs organic syntheses." *Nature* 555:7698, 2018, pp. 592–593.

121. F. Feng, L. Lai, and J. Pei. "Computational Chemical Synthesis Analysis and Pathway Design." *Front. Chem.* 6, 2018, p. 199.

122. J. Savage, A. Kishimoto, B. Buesser, E. Diaz-Aviles, and C. Alzate. "Chemical reactant recommendation using a network of organic chemistry". In: *Proceedings of the Eleventh ACM Conference on Recommender Systems*. 2017, pp. 210–214.

123. A. Masoumi, M. Soutchanski, and A. Marrella. "Organic Synthesis as Artificial Intelligence Planning". In: *International Workshop on Semantic Web Applications and Tools for Life Sciences SWATLS*. 2013.

124. J. Law, Z. Zsoldos, A. Simon, D. Reid, Y. Liu, S. Y. Khew, A. P. Johnson, S. Major, R. A. Wade, and H. Y. Ando. "Route Designer: a retrosynthetic analysis tool utilizing automated retrosynthetic rule generation." *J. Chem. Inf. Model.* 49:3, 2009, pp. 593–602.

125. C. W. Coley, D. A. Thomas, J. A. Lummiss, J. N. Jaworski, C. P. Breen, V. Schultz, T. Hart, J. S. Fishman, L. Rogers, H. Gao, et al. "A robotic platform for flow synthesis of organic compounds informed by AI planning". *Science* 365:6453, 2019, eaax1566.

126. S. Genheden, A. Thakkar, V. Chadimová, J.-L. Reymond, O. Engkvist, and E. Bjerrum. "AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning". *J. Cheminf.* 12:1, 2020, pp. 1–9.

127. A. Cadeddu, E. K. Wylie, J. Jurczak, M. Wampler-Doty, and B. A. Grzybowski. "Organic chemistry as a language and the implications of chemical linguistics for structural and retrosynthetic analyses." *Angew. Chem. Int. Ed.)* 53:31, 2014, pp. 8108–8112.

128. L. Pattanaik, O. E. Ganea, I. Coley, K. F. Jensen, W. H. Green, and C. W. Coley. "Message Passing Networks for Molecules with Tetrahedral Chirality". *arXiv: 2012.00094*, 2020.

129. J. Vig. "A Multiscale Visualization of Attention in the Transformer Model". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 2019, pp. 37–42.

130. B. Hoover, H. Strobelt, and S. Gehrmann. "exbert: A visual analysis tool to explore learned representations in transformers models". *arXiv: 1910.05276*, 2019.

131. E. J. Corey. "The logic of chemical synthesis: multistep synthesis of complex carbogenic molecules (Nobel Lecture)". *Angew. Chem. Int. Ed.* 30:5, 1991, pp. 455–465.

132. D. C. Blakemore, L. Castro, I. Churcher, D. C. Rees, A. W. Thomas, D. M. Wilson, and A. Wood. "Organic synthesis provides opportunities to transform drug discovery". *Nat. Chem.* 10:4, 2018, pp. 383–394.

133. Y.-C. Lo, S. E. Rensi, W. Torng, and R. B. Altman. "Machine learning in chemoinformatics and drug discovery". *Drug Discovery Today* 23:8, 2018, pp. 1538–1546.

134. J. L. Melville, E. K. Burke, and J. D. Hirst. "Machine learning in virtual screening". *Comb. Chem. High Throughput Screening* 12:4, 2009, pp. 332–343.

135. A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi. "Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons". *Phys. Rev. Lett.* 104:13, 2010, p. 136403.

136. P. O. Dral. "Quantum Chemistry in the Age of Machine Learning". *J. Phys. Chem. Lett.* 11:6, 2020, pp. 2336–2347.

137. V. H. Nair, P. Schwaller, and T. Laino. "Data-driven Chemical Reaction Prediction and Retrosynthesis". *CHIMIA Int. J. Chem.* 73:12, 2019, pp. 997–1000.

138. P. Schwaller and T. Laino. "Data-Driven Learning Systems for Chemical Reaction Prediction: An Analysis of Recent Approaches". In: *Machine Learning in Chemistry: Data-Driven Algorithms, Learning Systems, and Predictions*. ACS Publications, 2019, pp. 61–79.

139. B. Ernst, G. W. Hart, and P. Sinaÿ. *Carbohydrates in chemistry and biology*. Wiley Blackwell, 2008.

140. P. Stallforth, B. Lepenies, A. Adibekian, and P. H. Seeberger. "Carbohydrates: a frontier in medicinal chemistry". *J. Med. Chem.* 52:18, 2009, pp. 5561–5577.

141. J. M. Boilevin and J.-L. Reymond. "Synthesis of lipid-linked oligosaccharides (LLOs) and their phosphonate analogues as probes to study protein glycosylation enzymes". *Synthesis* 50:14, 2018, pp. 2631–2654.

142. R. Mettu, C.-Y. Chen, and C.-Y. Wu. "Synthetic carbohydrate-based vaccines: challenges and opportunities". *J. Biomed. Sci. Eng.* 27:1, 2020, pp. 1–22.

143. F. Broecker and P. H. Seeberger. "Identification and Design of Synthetic B Cell Epitopes for Carbohydrate-Based Vaccines". In: *Methods Enzymol.* Vol. 597. Elsevier, 2017, pp. 311–334.

144. L.-A. Barel and L. A. Mulard. "Classical and novel strategies to develop a Shigella glycoconjugate vaccine: from concept to efficacy in human". *Hum. Vaccines Immunother.* 15:6, 2019, pp. 1338–1356.

145. M. N. Kamat and A. V. Demchenko. "Revisiting the Armed- Disarmed Concept Rationale: S-Benzoxazolyl Glycosides in Chemoselective Oligosaccharide Synthesis". *Org. Lett.* 7:15, 2005, pp. 3215–3218.

146. B. Dhakal and D. Crich. "Synthesis and stereocontrolled equatorially selective glycosylation reactions of a pseudaminic acid donor: importance of the side-chain conformation and regioselective reduction of azide protecting groups". *J. Am. Chem. Soc.* 140:44, 2018, pp. 15008–15015.

147. B. Zoph, D. Yuret, J. May, and K. Knight. "Transfer Learning for Low-Resource Neural Machine Translation". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 2016, pp. 1568–1575.

148. P. Ramachandran, P. Liu, and Q. Le. "Unsupervised Pretraining for Sequence to Sequence Learning". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 383–391. DOI: 10.18653/v1/D17-1039. URL: https://www.aclweb.org/anthology/D17-1039.

149. K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu. "MASS: Masked Sequence to Sequence Pre-training for Language Generation". In: *International Conference on Machine Learning*. 2019, pp. 5926–5936.

150. J. S. Smith, B. T. Nebgen, R. Zubatyuk, N. Lubbers, C. Devereux, K. Barros, S. Tretiak, O. Isayev, and A. E. Roitberg. "Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning". *Nat. Commun.* 10:1, 2019, pp. 1–8.

151. H. Öztürk, A. Özgür, P. Schwaller, T. Laino, and E. Ozkirimli. "Exploring chemical space using natural language processing methodologies for drug discovery". *Drug Discovery Today*, 2020.

152. A. Behera, D. Rai, and S. S. Kulkarni. "Total Syntheses of Conjugation-ready Trisaccharide Repeating Units of Pseudomonas aeruginosa O11 and Staphylococcus aureus Type 5 Capsular Polysaccharide for Vaccine Development". *J. Am. Chem. Soc.* 142:1, 2019, pp. 456–467.

153. G. Landrum, P. Tosco, B. Kelley, sriniker, gedeck, NadineSchneider, R. Vianello, A. Dalke, Ric, B. Cole, AlexanderSavelyev, S. Turk, M. Swain, A. Vaucher, D. N, M. Wójcikowski, A. Pahl, JP, F. Berenger, strets123, JLVarjo, N. O'Boyle, D. Cosgrove, P. Fuller, J. H. Jensen, G. Sforna, DoliathGavid, K. Leswing, S. Leung, and J. van Santen. *rdkit/rdkit: 2019_03_4 (Q1 2019) Release*. 2019. DOI: 10.5281/zenodo.3366468. URL: https://doi.org/10.5281/zenodo.3366468.

154. A. S. Ramıérez, J. Boilevin, R. Biswas, B. H. Gan, D. Janser, M. Aebi, T. Darbre, J.-L. Reymond, and K. P. Locher. "Characterization of the single-subunit oligosaccharyltransferase STT3A from Trypanosoma brucei using synthetic peptides and lipid-linked oligosaccharide analogs". *Glycobiol.* 27:6, 2017, pp. 525–535.

155. J. S. Bloch, G. Pesciullesi, J. Boilevin, K. Nosol, R. N. Irobalieva, T. Darbre, M. Aebi, A. A. Kossiakoff, J.-L. Reymond, and K. P. Locher. "Structure and mechanism of the ER-based glucosyltransferase ALG6". *Nature*, 2020, pp. 1–5.

156. G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush. "OpenNMT: Open-Source Toolkit for Neural Machine Translation". In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 2017. DOI: 10.18653/v1/P17-4012.

157. *OpenNMT-py*. (Accessed Oct 29, 2019). URL: https://github.com/OpenNMT/OpenNMT-py.

158. *Molecular Transformer*. (Accessed Aug 29, 2019). URL: https://github.com/pschwllr/MolecularTransformer.

159. A. Suzuki. "Recent advances in the cross-coupling reactions of organoboron derivatives with organic electrophiles, 1995–1998". *J. Organomet. Chem.* 576:1, 1999, pp. 147–168.

160. Y. Ai, N. Ye, Q. Wang, K. Yahata, and Y. Kishi. "Zirconium/Nickel-Mediated One-Pot Ketone Synthesis". *Angew. Chem. Int. Ed.* 129:36, 2017, pp. 10931–10935.

161. X. Liu, X. Li, Y. Chen, Y. Hu, and Y. Kishi. "On Ni Catalysts for Catalytic, Asymmetric Ni/Cr-Mediated Coupling Reactions". *J. Am. Chem. Soc.* 134:14, 2012, pp. 6136–6139.

162. C. W. Coley, L. Rogers, W. H. Green, and K. F. Jensen. "Computer-assisted retrosynthesis based on molecular similarity". *ACS Cent. Sci.* 3:12, 2017, pp. 1237–1245.

163. H. Gao, T. J. Struble, C. W. Coley, Y. Wang, W. H. Green, and K. F. Jensen. "Using Machine Learning To Predict Suitable Conditions for Organic Reactions." *ACS Cent. Sci.* 4:11, 2018, pp. 1465–1476.

164. B. A. Grzybowski, K. J. M. Bishop, B. Kowalczyk, and C. E. Wilmer. "The 'wired' universe of organic chemistry". *Nat. Chem.* 1:1, 2009, pp. 31–36.

165. T. Klucznik, B. Mikulak-Klucznik, M. P. McCormack, H. Lima, S. Szymkuć, M. Bhowmick, K. Molga, Y. Zhou, L. Rickershauser, E. P. Gajewska, A. Toutchkine, P. Dittwald, M. P. Startek, G. J. Kirkovits, R. Roszak, A. Adamski, B. Sieredzińska, M. Mrksich, S. L. Trice, and B. A. Grzybowski. "Efficient Syntheses of Diverse, Medicinally Relevant Targets Planned by Computer and Executed in the Laboratory". *Chem* 4:3, 2018, pp. 522–532. ISSN: 2451-9294.

166. S. Zheng, J. Rao, Z. Zhang, J. Xu, and Y. Yang. "Predicting Retrosynthetic Reaction using Self-Corrected Transformer Neural Networks". *arXiv: 1907.01356*, 2019.

167. P. Karpov, G. Godin, and I. V. Tetko. "A transformer model for retrosynthesis". In: *International Conference on Artificial Neural Networks*. Springer. 2019, pp. 817–830.

168. X. Liu, P. Li, and S. Song. "Decomposing Retrosynthesis into Reactive Center Prediction and Molecule Generation". *bioRxiv*, 2019, p. 677849.

169. K. Lin, Y. Xu, J. Pei, and L. Lai. "Automatic Retrosynthetic Pathway Planning Using Template-free Models". *arXiv: 1906.02308*, 2019.

170. A. A. Lee, Q. Yang, V. Sresht, P. Bolgar, X. Hou, J. L. Klug-McLeod, and C. R. Butler. "Molecular Transformer unifies reaction prediction and retrosynthesis across pharma chemical space". *Chem. Commun.*, 2019, pp. 12152–12155.

171. H. Duan, L. Wang, C. Zhang, and J. Li. "Retrosynthesis with Attention-Based NMT Model and Chemical Analysis of the" Wrong" Predictions". *arXiv: 1908.00727*, 2019.

172. A. F. de Almeida, R. Moreira, and T. Rodrigues. "Synthetic organic chemistry driven by artificial intelligence". *Nat. Rev. Chem.* 1, 2019, pp. 1–16.

173. N. Schneider, D. M. Lowe, R. A. Sayle, M. A. Tarselli, and G. A. Landrum. "Big data from pharmaceutical patents: a computational analysis of medicinal chemists' bread and butter". *J. Med. Chem.* 59:9, 2016, pp. 4385–4402.

174. K. Mao, P. Zhao, T. Xu, Y. Rong, X. Xiao, and J. Huang. "Molecular Graph Enhanced Transformer for Retrosynthesis Prediction". In: Cold Spring Harbor Laboratory, 2020.

175. B. Chen, T. Shen, T. S. Jaakkola, and R. Barzilay. "Learning to Make Generalizable and Diverse Predictions for Retrosynthesis". *arXiv: 1910.09688*, 2019.

176. C. W. Coley, L. Rogers, W. H. Green, and K. F. Jensen. "SCScore: Synthetic complexity learned from a reaction corpus". *J. Chem. Inf. Model.* 58:2, 2018, pp. 252–261.

177. P. Schwaller, D. Probst, A. C. Vaucher, V. H. Nair, D. Kreutter, T. Laino, and J.-L. Reymond. "Mapping the Space of Chemical Reactions using Attention-Based Neural Networks". *Nat. Mach. Intell.* 3, 2021, pp. 144–152. DOI: 10.1038/s42256-020-00284-w.

178. T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. "Improved techniques for training gans". In: *Advances in Neural Information Processing Systems*. 2016, pp. 2234–2242.

179. J. Nieminen and M. Peltola. "Hypertrees". *Appl. Math. Lett.* 12:2, 1999, pp. 35–38.

180. *Nextmove Software Pistachio*. (Accessed Jul 29, 2019). URL: http://www.nextmovesoftware.com/pistachio.html.

181. J. Lin. "Divergence measures based on the Shannon entropy". *IEEE Trans. Inf. Theory* 37:1, 1991, pp. 145–151.

182. E. L. Willighagen, J. W. Mayfield, J. Alvarsson, A. Berg, L. Carlsson, N. Jeliazkova, S. Kuhn, T. Pluskal, M. Rojas-Chertó, O. Spjuth, et al. "The Chemistry Development Kit (CDK) v2. 0: atom typing, depiction, molecular formulas, and substructure searching". *J. Cheminf.* 9:1, 2017, p. 33.

183. D. Lednicer and L. A. Mitscher. *The organic chemistry of drug synthesis. 2*. A Wiley-Interscience publication. OCLC: 310877189. Wiley, New York, 1980. ISBN: 978-0-471-04392-8.

184. P. A. Worthington. "Synthesis and Fungicidal Activity of Triazole Tertiary Alcohols". In: *Synthesis and Chemistry of Agrochemicals*. Chap. 27, pp. 302–317.

185. H. Cotton, T. Elebring, M. Larsson, L. Li, H. Sörensen, and S. von Unge. "Asymmetric synthesis of esomeprazole". *Tetrahedron: Asymmetry* 11:18, 2000, pp. 3819–3825.

186. J. F. Larrow, E. Roberts, T. R. Verhoeven, K. M. Ryan, C. H. Senanayake, P. J. Reider, E. N. Jacobsen, S. A. Lodise, and A. B. Smith. "(1S, 2R)-1-aminoindan-2-ol [1H-Inden-2-ol, 1-amino-2, 3-dihydro-(1S-cis)-]". *Org. Synth.* 76, 1999.

187. A. F. Crowther and L. H. Smith. ".beta.-Adrenergic blocking agents. II. Propranolol and related 3-amino-1-naphthoxy-2-propanols". *J. Med. Chem.* 11:5, 1968. PMID: 5697060, pp. 1009–1013.

188. A. C. Vaucher, F. Zipoli, J. Geluykens, V. H. Nair, P. Schwaller, and T. Laino. "Automated extraction of chemical synthesis actions from experimental procedures". *Nat. Commun.* 11:1, 2020, p. 3601.

189. P. Schwaller, B. Hoover, J.-L. Reymond, H. Strobelt, and T. Laino. "Unsupervised Attention-Guided Atom-Mapping". *ChemRxiv preprint*. DOI: 10.26434/chemrxiv.12298559.v1.

190. A. Toniato, P. Schwaller, A. Cardinale, J. Geluykens, and T. Laino. "Unassisted Noise-Reduction of Chemical Reactions Data Sets". *ChemRxiv preprint*, 2020. DOI: 10.26434/chemrxiv.12395120.v1.

191. N. Miyaura and A. Suzuki. "Palladium-catalyzed cross-coupling reactions of organoboron compounds". *Chem. Rev.* 95:7, 1995, pp. 2457–2483.

192. H. Kraut, J. Eiblmaier, G. Grethe, P. Löw, H. Matuszczyk, and H. Saller. "Algorithm for reaction classification". *J. Chem. Inf. Model.* 53:11, 2013, pp. 2884–2895.

193. *Daylight Theory Manual, Chapter 5*. (accessed May 25, 2014). URL: http://www.daylight.com/dayhtml/doc/theory/theory.smirks.html.

194. L. Chen and J. Gasteiger. "Organic Reactions Classified by Neural Networks: Michael Additions, Friedel–Crafts Alkylations by Alkenes, and Related Reactions". *Angew. Chem. Int. Ed.* 35:7, 1996, pp. 763–765.

195. L. Chen and J. Gasteiger. "Knowledge discovery in reaction databases: Landscaping organic reactions by a self-organizing neural network". *J. Am. Chem. Soc.* 119:17, 1997, pp. 4033–4042.

196. H. Satoh, O. Sacher, T. Nakata, L. Chen, J. Gasteiger, and K. Funatsu. "Classification of organic reactions: similarity of reactions based on changes in the electronic features of oxygen atoms at the reaction sites". *J. Chem. Inf. Comput. Sci.* 38:2, 1998, pp. 210–219.

197. H. Gao, T. J. Struble, C. W. Coley, Y. Wang, W. H. Green, and K. F. Jensen. "Using Machine Learning To Predict Suitable Conditions for Organic Reactions." *ACS Cent. Sci.* 4:11, 2018, pp. 1465–1476.

198. G. M. Ghiandoni, M. J. Bodkin, B. Chen, D. Hristozov, J. E. Wallace, J. Webster, and V. J. Gillet. "Development and Application of a Data-Driven Reaction Classification Model: Comparison of an Electronic Lab Notebook and Medicinal Chemistry Literature". *J. Chem. Inf. Model.* 59:10, 2019, pp. 4167–4187.

199. *ChemAxon*. (Accessed Dec 21, 2019). URL: https://docs.chemaxon.com/display/ltsargon/Reaction+fingerprint+RF.

200. F. Sandfort, F. Strieth-Kalthoff, M. Kühnemund, C. Beecks, and F. Glorius. "A structure-based platform for predicting chemical reactivity". *Chem*, 2020.

201. D. Probst and J.-L. Reymond. "Visualization of very large high-dimensional data sets as minimum spanning trees". *J. Cheminf.* 12:1, 2020, pp. 1–13.

202. K. Jorner, T. Brinck, P.-O. Norrby, and D. Buttar. "Machine learning meets mechanistic modelling for accurate prediction of experimental activation energies". *Chem. Sci.*, 2020.

203. P. Schwaller, A. C. Vaucher, T. Laino, and J.-L. Reymond. "Prediction of Chemical Reaction Yields using Deep Learning". *ChemRxiv preprint*, 2020. DOI: 10.26434/chemrxiv.12758474.

204. R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts. "Recursive deep models for semantic compositionality over a sentiment treebank". In: *Proceedings of the 2013 conference on empirical methods in natural language processing*. 2013, pp. 1631–1642.

205. A. Warstadt, A. Singh, and S. R. Bowman. "Neural network acceptability judgments". *Trans. Assoc. Comp. Ling.* 7, 2019, pp. 625–641.

206. J. Johnson, M. Douze, and H. Jégou. "Billion-scale similarity search with GPUs". *arXiv: 1702.08734*, 2017.

207. J.-M. Wei, X.-J. Yuan, Q.-H. Hu, and S.-Q. Wang. "A novel measure for evaluating classifiers". *Expert Syst. Appl.* 37:5, 2010, pp. 3799–3809.

208. B. W. Matthews. "Comparison of the predicted and observed secondary structure of T4 phage lysozyme". *Biochim. Biophys. Acta Prot. Struc.* 405:2, 1975, pp. 442–451.

209. J. Gorodkin. "Comparing two K-category assignments by a K-category correlation coefficient". *Comput. Biol. Chem.* 28:5-6, 2004, pp. 367–374.

210. A. Capecchi, D. Probst, and J.-L. Reymond. "One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome". *J. Cheminf.* 12:1, 2020, pp. 1–15.

211. D. Probst and J.-L. Reymond. "FUn: a framework for interactive visualizations of large, high-dimensional datasets on the web". *Bioinf.* 34:8, 2017, pp. 1433–1435.

212. J. S. Carey, D. Laffan, C. Thomson, and M. T. Williams. "Analysis of the reactions used for the preparation of drug candidate molecules". *Org. Biomol. Chem.* 4:12, 2006, pp. 2337–2347.

213. *RSC's RXNO Ontology*. (Accessed Sep 13, 2019). URL: http://www.rsc.org/ontologies/RXNO/index.asp.

214. C. W. Coley, W. H. Green, and K. F. Jensen. "RDChiral: An RDKit wrapper for handling stereochemistry in retrosynthetic template extraction and application". *J. Chem. Inf. Model.* 59:6, 2019, pp. 2529–2537.

215. *BERT code*. (Accessed Oct 15, 2019). URL: https://github.com/google-research/bert%5C#sentence-and-sentence-pair-classification-tasks.

216. T. Wolf, J. Chaumond, L. Debut, V. Sanh, C. Delangue, A. Moi, P. Cistac, M. Funtowicz, J. Davison, S. Shleifer, et al. "Transformers: State-of-the-art natural language processing". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 2020, pp. 38–45.

217. D. Probst and J.-L. Reymond. "Smilesdrawer: parsing and drawing SMILES-encoded molecular structures using client-side javascript". *J. Chem. Inf. Model.* 58:1, 2018, pp. 1–7.

218. S. Haghighi, M. Jasemi, S. Hessabi, and A. Zolanvari. "PyCM: Multiclass confusion matrix library in Python". *J. Open Source Software* 3:25, 2018, p. 729.

219. *RXNFP repository (v0.0.7)*. (Accessed Nov 17, 2020). URL: https://dx.doi.org/10.5281/zenodo.4277570.

220. S. Kite, T. Hattori, and Y. Murakami. "Estimation of catalytic performance by neural network — product distribution in oxidative dehydrogenation of ethylbenzene". *Appl. Catal., A* 114:2, 1994, pp. L173–L178.

221. P. Raccuglia, K. C. Elbert, P. D. F. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier, and A. J. Norquist. "Machine-learning-assisted materials discovery using failed experiments". *Nature* 533:7601, 2016, pp. 73–76.

222. D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher, and A. G. Doyle. "Predicting reaction performance in C–N cross-coupling using machine learning". *Science* 360:6385, 2018, pp. 186–190.

223. K. V. Chuang and M. J. Keiser. "Comment on "Predicting reaction performance in C–N cross-coupling using machine learning"". *Science* 362:6416, 2018.

224. J. M. Granda, L. Donina, V. Dragone, D.-L. Long, and L. Cronin. "Controlling an organic synthesis robot with machine learning to search for new reactivity". *Nature* 559:7714, 2018, pp. 377–381.

225. Z. Fu, X. Li, Z. Wang, Z. Li, X. Liu, X. Wu, J. Zhao, X. Ding, X. Wan, F. Zhong, D. Wang, X. Luo, K. Chen, H. Liu, J. Wang, H. Jiang, and M. Zheng. "Optimizing chemical reaction conditions using deep learning: a case study for the Suzuki–Miyaura cross-coupling reaction". *Org. Chem. Front.*, 2020.

226. N. S. Eyke, W. H. Green, and K. F. Jensen. "Iterative experimental design based on active machine learning reduces the experimental burden associated with reaction screening". *React. Chem. Eng.*, 2020.

227. G. Skoraczyński, P. Dittwald, B. Miasojedow, S. Szymkuć, E. P. Gajewska, B. A. Grzybowski, and A. Gambin. "Predicting the outcomes of organic reactions via machine learning: are current descriptors sufficient?" *Sci. Rep.* 7:1, 2017, p. 3582.

228. D. Perera, J. W. Tucker, S. Brahmbhatt, C. J. Helal, A. Chong, W. Farrell, P. Richardson, and N. W. Sach. "A platform for automated nanomole-scale reaction screening and micromole-scale synthesis in flow". *Science* 359:6374, 2018, pp. 429–434.

229. *Simpletransformers*. (Accessed Jul 02, 2020). URL: https://simpletransformers.ai.

230. J. Johnson, M. Douze, and H. Jégou. "Billion-scale similarity search with GPUs". *arXiv: 1702.08734*, 2017.

231. J. Vig and Y. Belinkov. "Analyzing the Structure of Attention in a Transformer Language Model". In: *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. 2019, pp. 63–76.

232. C. A. Grambow, L. Pattanaik, and W. H. Green. "Reactants, products, and transition states of elementary chemical reactions based on quantum chemistry". *Sci. Data* 7:1, 2020, pp. 1–8.

233. G. F. von Rudorff, S. Heinen, M. Bragato, and A. von Lilienfeld. "Thousands of reactants and transition states for competing E2 and SN2 reactions". *Mach. Learn.: Sci. Technol.*, 2020.

234. K. Lin, Y. Xu, J. Pei, and L. Lai. "Automatic retrosynthetic route planning using template-free models". *Chem. Sci.* 11:12, 2020, pp. 3355–3364.

235. A. C. Vaucher, P. Schwaller, J. Geluykens, V. H. Nair, A. Iuliano, and T. Laino. "Inferring Experimental Procedures from Text-Based Representations of Chemical Reactions". *ChemRxiv preprint*, 2020. DOI: 10.26434/chemrxiv.13118423.v1.

236. E. J. Bjerrum. "Smiles enumeration as data augmentation for neural network modeling of molecules". *arXiv: 1703.07076*, 2017.

237. J. Arús-Pous, S. V. Johansson, O. Prykhodko, E. J. Bjerrum, C. Tyrchan, J.-L. Reymond, H. Chen, and O. Engkvist. "Randomized SMILES strings improve the quality of molecular generative models". *J. Cheminf.* 11:1, 2019, pp. 1–13.

238. G. Lambard and E. Gracheva. "SMILES-X: autonomous molecular compounds characterization for small datasets without descriptors". *Mach. Learn.: Sci. Technol.* 1:2, 2020, p. 025004.

239. G. Wang, W. Li, M. Aertsen, J. Deprest, S. Ourselin, and T. Vercauteren. "Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks". *Neurocomputing* 338, 2019, pp. 34–45.

240. *RXN Yields repo*. (Accessed Oct 02, 2020). 2020. URL: https://rxn4chemistry.github.io/rxn_yields/.

241. L. Hirschfeld, K. Swanson, K. Yang, R. Barzilay, and C. W. Coley. "Uncertainty quantification using neural networks for molecular property prediction". *J. Chem. Inf. Model.* 60:8, 2020, pp. 3770–3780.

242. A. R. Thawani, R.-R. Griffiths, A. Jamasb, A. Bourached, P. Jones, W. McCorkindale, A. A. Aldrick, and A. A. Lee. "The Photoswitch Dataset: A Molecular Machine Learning Benchmark for the Advancement of Synthetic Chemistry". *arXiv: 2008.03226*, 2020.

243. K. Felton, J. Rittig, and A. Lapkin. "Summit: Benchmarking Machine Learning Methods for Reaction Optimisation". *ChemRxiv preprint*, 2020. DOI: 10.26434/chemrxiv.12939806.v1.

244. F. Häse, M. Aldeghi, R. J. Hickman, L. M. Roch, M. Christensen, E. Liles, J. E. Hein, and A. Aspuru-Guzik. "Olympus: a benchmarking framework for noisy optimization and experiment planning". *arXiv: 2010.04153*, 2020.

245. W. L. Chen, D. Z. Chen, and K. T. Taylor. "Automatic reaction mapping and reaction center detection". *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* 3:6, 2013, pp. 560–593.

246. G. A. P. Gonzalez, L. R. El Assal, A. Noronha, I. Thiele, H. S. Haraldsdóttir, and R. M. Fleming. "Comparative evaluation of atom mapping algorithms for balanced metabolic reactions: application to Recon 3D". *J. Cheminf.* 9:1, 2017, p. 39.

247. M. F. Lynch and P. Willett. "The automatic detection of chemical reaction sites". *J. Chem. Inf. Comput. Sci.* 18:3, 1978, pp. 154–159.

248. J. J. McGregor and P. Willett. "Use of a maximum common subgraph algorithm in the automatic identification of ostensible bond changes occurring in chemical reactions". *J. Chem. Inf. Comput. Sci.* 21:3, 1981, pp. 137–140.

249. T. E. Moock, J. G. Nourse, D. Grier, and W. D. Hounshell. "The implementation of atom-atom mapping and related features in the reaction access system (REACCS)". In: *Chemical structures*. Springer, 1988, pp. 303–313.

250. K. Funatsu, T. Endo, N. Kotera, and S.-I. Sasaki. "Automatic recognition of reaction site in organic chemical reactions". *Tetrahedron Comput. Method.* 1:1, 1988, pp. 53–69.

251. R. Körner and J. Apostolakis. "Automatic determination of reaction mappings and reaction center information. 1. The imaginary transition state energy approach". *J. Chem. Inf. Model.* 48:6, 2008, pp. 1181–1189.

252. J. Apostolakis, O. Sacher, R. Körner, and J. Gasteiger. "Automatic determination of reaction mappings and reaction center information. 2. Validation on a biochemical reaction database". *J. Chem. Inf. Model.* 48:6, 2008, pp. 1190–1198.

253. C. Jochum, J. Gasteiger, and I. Ugi. "The principle of minimum chemical distance (PMCD)". *Angew. Chem. Int. Ed.* 19:7, 1980, pp. 495–505.

254. T. Akutsu. "Efficient extraction of mapping rules of atoms from enzymatic reaction data". *J. Comput. Biol.* 11:2-3, 2004, pp. 449–462.

255. J. D. Crabtree and D. P. Mehta. "Automated reaction mapping". *ACM J. Exp. Algorithmics* 13, 2009, pp. 1–15.

256. E. L. First, C. E. Gounaris, and C. A. Floudas. "Stereochemically consistent reaction mapping and identification of multiple reaction mechanisms through integer linear optimization". *J. Chem. Inf. Model.* 52:1, 2012, pp. 84–92.

257. M. Latendresse, J. P. Malerich, M. Travers, and P. D. Karp. "Accurate atom-mapping computation for biochemical reactions". *J. Chem. Inf. Model.* 52:11, 2012, pp. 2970–2982.

258. V. R. Somnath, C. Bunne, C. W. Coley, A. Krause, and R. Barzilay. "Learning Graph Models for Template-Free Retrosynthesis". *arXiv: 2006.07038*, 2020.

259. M. Fortunato, C. W. Coley, B. C. Barnes, and K. F. Jensen. "Data Augmentation and Pretraining for Template-Based Retrosynthetic Prediction in Computer-Aided Synthesis Planning". *J. Chem. Inf. Model.*, 2020.

260. L. Chen, J. G. Nourse, B. D. Christie, B. A. Leland, and D. L. Grier. "Over 20 years of reaction access systems from MDL: a novel reaction substructure search algorithm". *J. Chem. Inf. Comput. Sci.* 42:6, 2002, pp. 1296–1310.

261. S. Wiegreffe and Y. Pinter. "Attention is not not Explanation". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019, pp. 11–20.

262. Y.-D. Zhang, W.-W. Ren, Y. Lan, Q. Xiao, K. Wang, J. Xu, J.-H. Chen, and Z. Yang. "Stereoselective Construction of an Unprecedented 7- 8 Fused Ring System in Micrandilactone A by [3, 3]-Sigmatropic Rearrangement". *Org. Lett.* 10:4, 2008, pp. 665–668.

263. T.-W. Sun, W.-W. Ren, Q. Xiao, Y.-F. Tang, Y.-D. Zhang, Y. Li, F.-K. Meng, Y.-F. Liu, M.-Z. Zhao, L.-M. Xu, et al. "Diastereoselective Total Synthesis of (±)-Schindilactone A, Part 1: Construction of the ABC and FGH Ring Systems and Initial Attempts to Construct the CDEF Ring System". *Chem. Asian J.* 7:10, 2012, pp. 2321–2333.

264. A. Schweinitz, A. Chtchemelinine, and A. Orellana. "Synthesis of benzodiquinanes via tandem palladium-catalyzed semipinacol rearrangement and direct arylation". *Org. Lett.* 13:2, 2011, pp. 232–235.

265. R. K. Acharyya, R. K. Rej, and S. Nanda. "Exploration of Ring Rearrangement Metathesis Reaction: A General and Flexible Approach for the Rapid Construction [5, n]-Fused Bicyclic Systems en Route to Linear Triquinanes". *J. Org. Chem.* 83:4, 2018, pp. 2087–2103.

266. L. Moni, L. Banfi, A. Basso, L. Carcone, M. Rasparini, and R. Riva. "Ugi and Passerini reactions of biocatalytically derived chiral aldehydes: application to the synthesis of bicyclic pyrrolidines and of antiviral agent telaprevir". *J. Org. Chem.* 80:7, 2015, pp. 3411–3428.

267. S. Wallis. "Binomial confidence intervals and contingency tests: mathematical fundamentals and the evaluation of alternative methods". *J. Quant. Ling.* 20:3, 2013, pp. 178–208.

268. K. Clark, U. Khandelwal, O. Levy, and C. D. Manning. "What Does BERT Look At? An Analysis of BERT's Attention". *CoRR* abs/1906.04341, 2019. arXiv: 1906.04341. URL: http://arxiv.org/abs/1906.04341.

269. N. Schneider, R. A. Sayle, and G. A. Landrum. "Get Your Atoms in Order - An Open-Source Implementation of a Novel and Robust Molecular Canonicalization Algorithm". *J. Chem. Inf. Model.* 55:10, 2015, pp. 2111–2120.

270. E. J. Griffen, A. G. Dossetter, and A. G. Leach. "Chemists: AI is here, unite to get the benefits". *J. Med. Chem.*, 2020.

271. T. Madzhidov, A. I. Lin, R. Nugmanov, N. Dyubankova, T. Gimadiev, J. K. Wegner, A. Rakhimbekova, T. Akhmetshin, Z. Ibragimova, A. Varnek, et al. "Atom-to-Atom Mapping: A Benchmarking Study of Popular Mapping Algorithms and Consensus Strategies", 2020. DOI: 10.26434/chemrxiv.13012679.v1.

272. Y. Mo, P. Verma, J. Guo, M. E. Fortunato, Z. Lu, C. W. Coley, K. F. Jensen, et al. "Evaluating and clustering retrosynthesis pathways with learned strategy". *Chem. Sci.*, 2020.

273. D. P. Kovacs, W. McCorkindale, et al. "Quantitative Interpretation Explains Machine Learning Models for Chemical Reaction Prediction and Uncovers Bias", 2020. DOI: 10.26434/chemrxiv.13061402.v1.

274. A. C. Vaucher, P. Schwaller, and T. Laino. "Completion of partial reaction equations", 2020. DOI: 10.26434/chemrxiv.13273310.v1.

275. *Open reaction database*. (Accessed Dec 21, 2020). 2020. URL: https://github.com/Open-Reaction-Database.

276. L. Talirz, S. Kumbhar, E. Passaro, A. V. Yakutovich, V. Granata, F. Gargiulo, M. Borelli, M. Uhrin, S. P. Huber, S. Zoupanos, et al. "Materials Cloud, a platform for open computational science". *arXiv: 2003.12510*, 2020.

277. C. Draxl and M. Scheffler. "NOMAD: The FAIR concept for big data-driven materials science". *MRS Bulletin* 43:9, 2018, pp. 676–682.

278. J. J. de Pablo, N. E. Jackson, M. A. Webb, L.-Q. Chen, J. E. Moore, D. Morgan, R. Jacobs, T. Pollock, D. G. Schlom, E. S. Toberer, et al. "New frontiers for the materials genome initiative". *npj Comput. Mater.* 5:1, 2019, p. 41.

279. M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al. "The FAIR Guiding Principles for scientific data management and stewardship". *Sci. Data* 3:1, 2016, pp. 1–9.

280. F. Fuchs, D. Worrall, V. Fischer, and M. Welling. "SE (3)-Transformers: 3D Roto-Translation Equivariant Attention Networks". In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020.

281. *CASP14 - Alpha Fold 2*. (Accessed Dec 21, 2020). 2020. URL: https://predictioncenter.org/casp14/.

282. S. Wang, J. Witek, G. A. Landrum, and S. Riniker. "Improving Conformer Generation for Small Rings and Macrocycles Based on Distance Geometry and Experimental Torsional-Angle Preferences". *J. Chem. Inf. Model.* 60:4, 2020, pp. 2044–2058.

283. C. W. Gao, J. W. Allen, W. H. Green, and R. H. West. "Reaction Mechanism Generator: Automatic construction of chemical kinetic mechanisms". *Comput. Phys. Commun.* 203, 2016, pp. 212–225.

284. M. Liu, A. G. Dana, M. Johnson, M. Goldman, A. Jocher, A. M. Payne, C. Grambow, K. Han, N. W.-W. Yee, E. Mazeau, K. Blondal, R. West, F. Goldsmith, and W. H. Green. "Reaction Mechanism Generator v3.0: Advances in Automatic Mechanism Generation". *ChemRxiv*, 2020. DOI: 10.26434/chemrxiv.13489656.v1.

285. W. Sameera, S. Maeda, and K. Morokuma. "Computational catalysis using the artificial force induced reaction method". *Acc. Chem. Res.* 49:4, 2016, pp. 763–773.

286. S. Maeda, Y. Harabuchi, M. Takagi, K. Saita, K. Suzuki, T. Ichino, Y. Sumiya, K. Sugiyama, and Y. Ono. "Implementation and performance of the artificial force induced reaction method in the GRRM17 program". *J. Comput. Chem.* 39:4, 2018, pp. 233–251.

287. T. A. Young, J. J. Silcock, A. J. Sterling, and F. Duarte. "autodE: Automated Calculation of Reaction Energy Profiles–Application to Organic and Organometallic Reactions". *Angew. Chem. Int. Ed.*, 2020.

288. D. Rappoport and A. Aspuru-Guzik. "Predicting feasible organic reaction pathways using heuristically aided quantum chemistry". *J. Chem. Theory Comput.* 15:7, 2019, pp. 4099–4112.

289. *Making New Drugs With a Dose of Artificial Intelligence*. (Accessed Dec 21, 2020). 2020. URL: https://www.nytimes.com/2019/02/05/technology/artificial-intelligence-drug-research-deepmind.html.

290. A. G. Godfrey, T. Masquelin, and H. Hemmerle. "A remote-controlled adaptive medchem lab: an innovative approach to enable drug discovery in the 21st Century". *Drug Discovery Today* 18:17-18, 2013, pp. 795–802.

291. J. Li, S. G. Ballmer, E. P. Gillis, S. Fujii, M. J. Schmidt, A. M. Palazzolo, J. W. Lehmann, G. F. Morehouse, and M. D. Burke. "Synthesis of many different types of organic small molecules using one automated process". *Science* 347:6227, 2015, pp. 1221–1226.

292. A. Adamo, R. L. Beingessner, M. Behnam, J. Chen, T. F. Jamison, K. F. Jensen, J.-C. M. Monbaliu, A. S. Myerson, E. M. Revalor, D. R. Snead, et al. "On-demand continuous-flow production of pharmaceuticals in a compact, reconfigurable system". *Science* 352:6281, 2016, pp. 61–67.

293. S. Steiner, J. Wolf, S. Glatzel, A. Andreou, J. M. Granda, G. Keenan, T. Hinkley, G. Aragon-Camarasa, P. J. Kitson, D. Angelone, et al. "Organic synthesis in a modular robotic system driven by a chemical programming language". *Science* 363:6423, 2019.

294. A.-C. Bédard, A. Adamo, K. C. Aroh, M. G. Russell, A. A. Bedermann, J. Torosian, B. Yue, K. F. Jensen, and T. F. Jamison. "Reconfigurable system for automated optimization of diverse chemical reactions". *Science* 361:6408, 2018, pp. 1220–1225.

295. S. Chatterjee, M. Guidi, P. H. Seeberger, and K. Gilmore. "Automated radial synthesis of organic molecules". *Nature* 579:7799, 2020, pp. 379–384.

296. *IBM RoboRXN*. (Accessed Dec 21, 2020). 2020. URL: https://rxn.res.ibm.com/rxn/robo-rxn/welcome.

297. S. H. M. Mehr, M. Craven, A. I. Leonov, G. Keenan, and L. Cronin. "A universal system for digitization and automatic execution of the chemical synthesis literature". *Science* 370:6512, 2020, pp. 101–108.

298. B. Burger, P. M. Maffettone, V. V. Gusev, C. M. Aitchison, Y. Bai, X. Wang, X. Li, B. M. Alston, B. Li, R. Clowes, et al. "A mobile robotic chemist". *Nature* 583:7815, 2020, pp. 237–241.

299. A. Thakkar, S. Johansson, K. Jorner, D. Buttar, J.-L. Reymond, and O. Engkvist. "Artificial intelligence and automation in computer aided synthesis planning". *React. Chem. Eng.*, 2020.

300. N. M. O'Boyle. "Towards a Universal SMILES representation-A standard method to generate canonical SMILES based on the InChI". *J. Cheminf.* 4:1, 2012, p. 22.

301. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. "Scikit-learn: Machine Learning in Python". *J. Mach. Learn. Res.* 12, 2011, pp. 2825–2830.

302. *Marvin JS, ChemAxon*. (Accessed Apr 02, 2020). URL: https://chemaxon.com.

Declaration of consent

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

Name/First Name: Schwaller Philippe

Registration Number: 11-802-337

Study program: PhD in Chemistry and Molecular Sciences

Bachelor ☐ Master ☐ Dissertation ☑

Title of the thesis: Learning the Language of Chemical Reactions – Atom by Atom

Supervisor: Prof Jean-Louis Reymond, University of Bern
Dr Teodoro Laino, IBM Research - Europe

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.

02.02.2021, Bern

Place/Date

Signature