Towards Sustainable Synthesis: Integrating Biocatalysis and Language Models in Computer-Aided Synthesis Planning

Inaugural dissertation
of the Faculty of Science,
University of Bern

presented by

David Kreutter

from France

Supervisor of the doctoral thesis:
Prof. Dr. Jean-Louis Reymond
Department of Chemistry, Biochemistry, and Pharmaceutical Science

Towards Sustainable Synthesis: Integrating Biocatalysis and Language Models in Computer-Aided Synthesis Planning

Inaugural dissertation
of the Faculty of Science,
University of Bern

presented by

David Kreutter

from France

Supervisor of the doctoral thesis:
Prof. Dr. Jean-Louis Reymond
Department of Chemistry, Biochemistry, and Pharmaceutical Science

Accepted by the Faculty of Science.

Bern, 30th November 2023

The Dean,
Prof. Dr. Marco Herwegh

ABSTRACT

This thesis is motivated by the objective of condensing experimental biocatalysis knowledge into machine-learning models. It also seeks to bridge the gap between Computer-Aided Synthesis Planning (CASP) tools and biocatalysis. The aim is to develop and implement a multistep synthesis planning software that can incorporate and explore both chemo- and biocatalysis reactions. The resulting tool should provide chemists with mixed catalytic synthetic route options, unlocking biocatalytic opportunities in chemical synthesis.

First, I explore the capabilities of a natural language processing model to learn biocatalysis reactions and perform forward reaction predictions. Similar to how chemists would learn biocatalysis, I provide the starting materials combined with a textual description of the enzyme and train the model to predict the product of the enzymatic transformation. I investigate the influence of transfer learning methods and demonstrate the model's performance with insightful examples. Additionally, I present practical use cases and investigate the limits of the Enzymatic Transformer.

Secondly, I report the creation of the first multistep Transformer-based computer-aided synthesis planning software, leveraging disconnection-aware models for a broader exploration of the chemical space. I design tagging strategies that automatically generate disconnection prompts, balancing diversity and computing time. I employ a triple transformer loop, predicting starting materials (T1), reagents, catalysts, and solvents (T2), followed by a forward validation model (T3) to limit unrealistic predictions. The resulting single-step framework explores a significantly more diverse chemical space while maintaining a critical assessment of the chemical feasibility of the predicted reactions. I detail the implementation of a multistep search using a best-first tree search algorithm guided by a new route penalty score, prioritizing short and efficient routes while exploring diverse retrosynthetic options. I showcase the performance of the CASP tool with insightful retrosynthesis examples of drug molecules. The models, along with the code, are available on GitHub as a Python package.

Finally, I integrate an independent triple transformer loop for biocatalysis into the previously designed CASP software. The reported implementation explores both chemo- and biocatalysis in parallel and builds mixed synthetic routes. This allows the suggestion of the most efficient synthetic routes to chemists, incorporating biocatalytic steps whenever possible, opening the door to more sustainable synthesis route design.

Contents

Contents

Contents

List of Figures

1   Introduction

1.1   Thesis Objectives

Chemistry, as a foundational scientific discipline, permeates every aspect of our daily lives, from the food we consume to the clothes we wear. At its core, molecule synthesis serves as a linchpin across various domains, including pharmaceuticals, food chemistry, agrochemistry, biochemistry, materials science, polymer chemistry, textiles, cosmetics, and perfumery.

With the goal of synthesizing novel materials, chemists engage in retrosynthesis — a pivotal process that involves deconstructing a target molecule into smaller fragments.[1] These fragments are then reassembled from simpler starting materials (SM). Retrosynthesis stands as a crucial step in the design of new molecules and materials, demanding a profound level of expertise and experience. Traditionally, chemists undertook retrosynthesis manually, relying on their knowledge and intuition. However, the escalating complexity of molecules coupled with the multitude of potential synthetic routes, makes this process increasingly challenging. To address this challenge, computer-aided synthesis planning (CASP) tools emerged as promising assistants for chemists.[2, 3] Those tools are designed to provide a starting point for synthesis design, leveraging computational capabilities by navigating the chemical space.

In the realm of advancing sustainable chemistry, biocatalysis emerges as a promising alternative to traditional metal- and organo-catalysis. The increasing accessibility of technologies and automation in enzyme engineering, particularly through directed evolution,[4] makes biocatalysis a more viable option for chemical synthesis. This is demonstrated by various examples of incorporating biocatalysis steps within chemocatalysis synthesis in the past years.[5, 6] However, a critical challenge persists — chemists lack awareness of enzyme capabilities and training in integrating biocatalysis into their synthesis planning. The absence of retrosynthesis planning tools that incorporate biocatalysis further worsens this limitation, resulting in missed opportunities for more sustainable practices. There is a pressing need to integrate biocatalysis into synthesis planning, encouraging the consideration of enzymatic transformation steps at earlier stages of synthesis design.

This thesis aims to explore the concept of encapsulating the expertise of biocatalysis within deep-learning models. The objective is to develop tools that assist chemists in integrating biocatalysis transformations into their synthesis designs. The approach involves leveraging experimen-

tally reported reactions from the literature and employing advanced deep-learning techniques, incorporating attention-based mechanisms along with template-based aspects. This innovative combination aims to enhance the prediction diversity and provide practical tools for chemists.

The thesis also dives into the development of state-of-the-art computer-aided synthesis planning software. This software is designed not only to advance the field but also to integrate biocatalysis. The envisioned result is a hybrid catalytic approach that facilitates a shift towards more sustainable synthesis practices. The goal is to provide chemists with a starting point and envision the potential evolution of enzymes via directed evolution. Ultimately, this comprehensive approach aims to contribute to the broader objective of achieving more sustainable and efficient chemical synthesis practices.

1.2 THESIS OUTLINE

The thesis is outlined as follows:

- Chapter 2 introduces foundational concepts of the thesis, including retrosynthesis, computer-aided synthesis planning (CASP) tools, and biocatalysis. The chapter begins with a brief overview of the historical evolution of CASP tools, providing insights into their development and limitations. Subsequently, it defines biocatalysis and reviews recent applications of CASP tools in the domain of biocatalysis, highlighting associated challenges and limitations of current implementations.

- Chapter 3 will describe the first attempt to instruct a Transformer model on enzymatic transformation reactions. It focuses on small molecule synthesis, utilizing experimentally validated biocatalysis reactions extracted from the Reaxys database. The chapter includes an analysis of the dataset, insights into the model's performance, its limitations, and potential use cases.

- Chapter 4 centers on the development of a multistep retrosynthesis framework. Combining various elements, it seeks to identify concise and efficient retrosyntheses. The tool is intended for route optimization through a more exhaustive exploration of the chemical space, leveraged by a Transformer model trained with tagged reactive centers, allowing a prediction augmentation by automatic tagging procedures. Noteworthy features include the Triple Transformer Loop (TTL) for single-step retrosynthesis prediction, incorporating reagent prediction and forward validation models. The chapter details the use of a heuristic best-first tree search, guided by a penalty-oriented scoring function. It concludes with a demonstration of the framework's capabilities on drug-like molecules and a discussion on future perspectives.

- Chapter 5 concludes the thesis by integrating biocatalysis into the retrosynthesis framework introduced in Chapter 4. It highlights the advantages of using experimental data from Reaxys over metabolic databases. The chapter introduces a Triple Transformer Loop (TTL) dedicated to enzymatic transformations, enabling a parallelizable multistep retrosynthesis planning with the tree search capable of selecting between chemocatalysis and biocatalysis steps. The chapter concludes with a showcase of the hybrid retrosynthesis tool on unseen molecules.

- Chapter 6 summarizes the work undertaken in the thesis, concludes and provides a discussion on future perspectives.

1.3 PUBLICATIONS

This thesis consists of first-author publications presented in separate chapters:

- Chapter 3: D. Kreutter, P. Schwaller, J.-L. Reymond. Predicting Enzymatic Reactions with a Molecular Transformer. *Chem. Sci.* 2021, **12** (25), 8648–8659. DOI: 10.1039/D1SC02362D.

- Chapter 4: D. Kreutter, J.-L. Reymond. Multistep Retrosynthesis Combining a Disconnection Aware Triple Transformer Loop with a Route Penalty Score Guided Tree Search. *Chem. Sci.* 2023, **14** (36), 9959–9969. DOI: 10.1039/D3SC01604H.

The following publication was co-authored but not incorporated into the thesis:

- P. Schwaller, D. Probst, A. C. Vaucher, V. H. Nair, D. Kreutter, T. Laino, J.-L. Reymond. Mapping the Space of Chemical Reactions Using Attention-Based Neural Networks. *Nat Mach Intell* 2021, **3** (2, 2), 144–152. DOI: 10.1038/s42256-020-00284-w.

1.4 FUNDING ACKNOWLEDGEMENTS

2 COMPUTER-AIDED SYNTHESIS PLANNING AND APPLICATIONS IN BIOCATALYSIS

Chemical retrosynthesis is the concept of working in reverse to design a synthetic route for a target molecule. It is a fundamental problem in organic chemistry and a key step in the development of new drugs, materials, and other chemical products. The process of retrosynthesis planning is a complex task that requires the consideration of multiple factors, such as the availability of starting materials, the cost of reagents, and the feasibility of the reactions. The development of computer-aided synthesis planning (CASP) tools has been a long-standing goal in the field of synthetic chemistry. However, despite the significant progress made in recent years, the problem remains unsolved. In this chapter, I introduce key concepts essential to the understanding of the thesis. I start with retrosynthesis, then the field of CASP tools, and discuss the challenges associated with it. I highlight current methods and tools and discuss their limitations. I also discuss the importance of biocatalysis in organic synthesis and the challenges associated with the integration of biocatalysis in CASP tools.

2.1 INTRODUCTION

Chemistry is the scientific study of matter, its properties, and its transformations. It has a rich history dating back to ancient civilizations, but it truly began to emerge as a distinct field in the 17th and 18th centuries. Through the work of pioneering scientists such as Robert Boyle and Antoine Lavoisier, chemistry developed into a discipline that seeks to understand the composition, structure, and behavior of substances. In the mid-19th century, groundbreaking contributions to the understanding of organic molecules emerged from the works of pioneers such as August Kekulé and Archibald Scott Couper. Kekulé proposed the cyclic structure for benzene,[7] while Couper played a crucial role in formalizing the structural formulas of organic compounds.[8] Their independent yet foundational insights laid the groundwork for representing molecules as interconnected scaffolds of carbon atoms, a concept that remains integral to modern chemical understanding.

In the modern era, chemistry assumes a pivotal role, essential in comprehending the natural world and propelling advancements in novel materials, pharmaceuticals, and technologies. Synthesis is a fundamental aspect of chemistry and plays a central role in organic chemistry by the creation of complex organic molecules from simpler precursors.[1] However, navigating the chemical space is a complex task that involves considerations of various factors, including the availability of starting materials, the catalysts and reagents selection, the economic feasibility, and the practical viability of the reactions involved.

Addressing these challenges, the development of computer-aided synthesis planning (CASP) tools emerges as a longstanding goal.[2, 3] The aim is to provide novel insights and assistance to chemists on how they approach and select synthetic routes. While significant progress has been made in recent years, the intrinsic complexities of organic synthesis pose persistent challenges, remaining ongoing challenges.

Another facet of chemistry is the substantial environmental footprint left by the chemical industry, attributed to the utilization of non-renewable resources and the generation of chemical waste.[9] To mitigate this impact, principles of green chemistry must be implemented, aiming to minimize the environmental impact of industrial chemical processes.[10, 11] Within these principles, biocatalysis stands as a recognized and desirable alternative due to the multitude of advantages it offers, including high stereoselectivity, mild reaction conditions, and the use of renewable resources.[12] Recent technological advancements, combined with automation and enzyme-directed evolution, have facilitated a more efficient and rapid integration of enzymatic steps into synthesis routes.[13]

In industry, the integration of biocatalysis into a synthesis plan remains a manual and expert-dependent task. To encourage the greater use of biocatalysis in process chemistry, reactions catalyzed by enzymes also need to be incorporated into CASP tools. This integration aims to provide alternative biocatalysis opportunities for chemists who may not be experts in biocatalysis.

In this chapter, we will review the current state of the art in CASP tools and discuss the general challenges beginning with data and molecular representations. We will also discuss the history and importance of biocatalysis in organic synthesis and examples of successful implementations. Finally, we will discuss the recent attempts and advances in integrating biocatalysis in CASP tools, highlighting methods and limitations of current approaches.

2.2 Chemical Representations

If humans can visualize molecules as drawings, it is a different story for computers for which we need to encode molecules. While this thesis will exclusively focus on the SMILES representation, it is important to note that there are other representations of molecules. For example feature-based

representations,[14, 15, 16] computer-learned,[17, 18, 19] other linear notations,[20, 21] or chemical table representations,[22] mostly discussed in the following reviews.[23, 24]

The Simplified Molecular Input Line Entry System (SMILES) is a linear notation system invented by David Weininger in 1988.[25, 26] Designed for faster computer processing and using more natural grammar, SMILES is a string of characters that encodes the structure of a molecule, offering a human-readable and compact representation. It represents atoms by their corresponding letters, along with bonds and connectivity, constructing the sequence by traversing the molecular graph. The choice of the starting atom for sequence construction may vary, resulting in different SMILES representations for the same molecule, a principle that proves useful for data augmentation.[27, 28] Canonicalization was introduced to represent a molecule with identical SMILES.[29] However, depending on the algorithm of various implementations, such as RDKit,[30] variations might still occur.

Overall, the SMILES notation aided in cheminformatics applications, chemical database integration, and computational tools, despite potential variations and limitations in capturing three-dimensional information.

2.3 Chemical Retrosynthesis and Related Challenges

Chemical synthesis can be envisioned as a process that bridges the conceptualization of a molecule on paper to the realization of that substance in a laboratory flask. To achieve this transformation, a chemist must adopt a strategic approach, stepping back to plan the route leading to the target molecule. The synthesis of complex molecules poses a formidable challenge, demanding a profound understanding of molecular intricacies and the skill to navigate through a labyrinth of chemical transformations.

While the initial use of the term "retrosynthesis" did not originate in the realm of chemistry,[31] it signifies the process of breaking down a problem into simpler components, offering an overall perspective and enhancing comprehension of the problem. Importantly, a retrosynthesis analysis also provides a view of the construction process, a crucial aspect for chemists in understanding and designing synthetic routes.

The term "retrosynthesis" is now exclusively used in the context of chemistry. As described by Corey,[1] it is the process of working backward through strategic operations on the target molecule to reach simpler and accessible starting materials (SM). The retrosynthetic analysis involves manipulating bonds and identifying functional groups for molecular simplification. The selection of transformations is guided by factors such as reaction feasibility, efficiency, and compatibility with other steps and functional groups. Various reaction operations can be applied to the target molecule, including: (i) modification of the carbon skeleton through carbon-carbon bond formation or

cleavage; (ii) formation of rings; (iii) rearrangements; (iv) introduction, interconversion, or removal of functional groups; (v) stereochemical manipulations; (vi) use of protecting groups. These operations are strategically chosen to maximize yield and minimize the number of steps required for the synthesis.

Overall, chemical retrosynthesis involves a careful analysis of the target molecule, thoughtful selection of transformations, and strategic planning to achieve the desired synthesis of complex organic molecules. With the reaction space allowing more than 300 reactions,[32] it requires a level of intellectual agility, comparable to strategic thinking in chess or finding one's way through a multidimensional maze, but with more complex rules. Retrosyntheses of complex molecules, such as paclitaxel and vitamin B_{12}, exemplify the sophistication achievable through strategic retrosynthetic planning.[33, 34, 35]

The relevance of retrosynthetic analysis extends to promoting sustainability in chemical synthesis with green chemistry concepts.[10, 11] By minimizing waste and optimizing resource utilization, retrosynthesis contributes to environmentally conscious synthetic strategies. In essence, retrosynthesis is more than a theoretical exercise, it is an art guiding toward inventive and efficient synthesis strategies.

2.4 Computer-Aided Synthesis Planning Tools

In 2016, AlphaGo beat the world champion in the game of Go.[36] Given the complexity of the game with an immense number of potential moves, finding the optimal strategy is challenging, even for computers, as it is practically impossible to simulate all potential states for numerous moves. Nevertheless, AlphaGo successfully identifies the best strategies through self-play and experiential learning, employing a combination of a Monte Carlo tree search (MCTS) algorithm and a deep neural policy network. The MCTS algorithm, a heuristic search method, navigates the search space by sampling the most promising moves while the deep neural network was trained using a substantial database of expert moves to predict the most likely moves.

Analogous to the way computers mastered the complex game of Go, computational tools have the potential to navigate the vast chemical space through strategic retrosynthetic tree searches. Computer-aided synthesis planning tools aim to assist chemists by proposing synthetic routes for a specified target molecule. The tool is designed to recommend a set of starting molecules, along with a set of reactions that could be employed to synthesize the target molecule from the identified precursors. Much like finding a path through a maze, the tool may propose multiple solutions, but ideally, it should propose a synthetic route that is short, efficient, cost-effective, and environmentally friendly.

CASP tools typically consist of several components whose composition may vary based on the selected strategy. These commonly include (i) a single-step retrosynthesis core, which may comprise a set of encoded rules — either manually encoded or data-driven — or be entirely rule-free and data-driven; (ii) a tree search algorithm guiding the iterative node expansion; (iii) a scoring system to assess potential solutions; and (iv) a database of commercially available starting materials.

2.4.1 HISTORY OF SYNTHESIS PLANNING PROGRAMS

Following the description of general retrosynthetic methods by Corey in 1967,[1] a subsequent challenge arose: Could computers emulate the art of retrosynthesis? In response, in 1969, Corey introduced the Organic Chemical Simulation of Synthesis (OCSS), a computer assistant designed to function like a chemist,[2] later updated and renamed as Logic and Heuristics Applied to Synthetic Analysis (LHASA) in 1977.[3] The tool featured a tablet interface enabling chemists to interact by drawing the target molecule as a structural diagram, subsequently processed by the tool as a list of atom connections.

To conduct the retrosynthesis analysis, LHASA employs "transforms" encoded using the developed languages CHMTRN (CHeMistry TRaNslator) and PATRAN (PAttern TRANslator). These "transforms" constitute a set of rules describing reactions, such as "hydrolysis of esters", "reduction of ketones", etc. Chemists encoded these rules as a knowledge base. The tool proposes retrosyntheses by applying these "transforms" when a match is found. The generated precursors undergo further assessment by additional rules based on the functional groups of the resulting molecules and incompatibility rules that judge the likelihood of a reaction occurring. The options are then presented to the chemist, who selects which one to apply interactively. Corey *et al.* continued to develop the program along with providing general guidelines for retrosynthetic strategies.[37, 38, 39]

Expanding on this pioneering work, additional tools emerged. Todd Wipke, a contributor to the development of LHASA alongside Corey, introduced the SECS program (Simulation and Evaluation of Chemical Synthesis). SECS also utilized a reaction knowledge dataset encoded with the ALCHEM language.[40] However, in this case, SECS went further by incorporating stereochemistry information through connection tables. While SYNNLMA shared a similar approach with its predecessors, it distinguished itself as the first to provide a detailed description of the components inherent in CASP tools. Moreover, it drew attention to the combinatorial explosion problem in retrosynthesis, prompting a critical examination of strategies for more effectively selecting and expanding leaf nodes.[41]

In parallel, SYNCHEM was established, incorporating a fully autonomous multistep scoring-guided search algorithm that prioritized the expansion of the most promising nodes first. Only in the event of failure of the top-scoring routes did it expand the lower-scoring ones.[42] Recognizing

the necessity to account for stereochemistry, a second iteration, SYNCHEM2, was subsequently developed.[43]

Notably, the SYNGEN program was introduced to tackle the challenges of convergent synthesis, a strategy that was minimally supported by precedent CASP tools but is acknowledged for its higher efficiency compared to linear approaches.[44] The SOS (Simulated Organic Synthesis) program was designed to address routes for heterocyclic compounds. Employing an interactive strategy, users can choose from a selection of suggested commercially available precursors. Then the program performs forward reaction simulations attempting to connect the set of starting materials to synthesize the target molecule.[45] CONAN (CONnectivity ANalysis) does not aim to suggest a complete retrosynthetic pathway, instead, its focus is on providing scaffold disconnections. This is particularly valuable for complex structures where visualizing such disconnections can be challenging for chemists.[46]

Subsequently, the "second generation" of CASP tools emerged, aimed at empowering computers to assist chemists in designing synthesis routes rather than relying on chemists to guide the computer to a solution.[47] Mathematical models, exemplified by programs like CICLOPS,[48] later succeeded by EROS (Elaboration of Reactions for Organic Synthesis),[49] and WODCA (Workbench for the Organization of Data for Chemical Applications),[50] utilized *BE*-matrices (bond and electron) for molecular representation, initially neglecting stereochemistry. These tools could perform bond breaking of the target molecule, followed by reaction evaluation based on thermodynamic estimations, resembling a self-evaluation in a round-trip manner. Continuing this trend, the CAMEO program was developed to predict the reaction outcome given the starting material and reaction conditions.[51] Similarly to SOPHIA,[52] it uses a knowledge base of different classes of reactions and a ranking system.

2.4.2 History of reaction databases

Simultaneously, as CASP tools faced challenges in integrating into chemists' routines,[53] alternative approaches emerged, focusing on the utilization of reaction databases. These databases have proven immensely valuable, offering searching algorithms and similarity searches that greatly assist chemists in their daily workflows. Some notable examples include the REACCS database and search system,[54] the ORAC database,[55] SYNLIB,[56] the CASREACT database,[57, 58] the Beilstein database,[59] and the ChemInform database.[60, 61]

In the present day, a handful of commercial solutions leveraging literature-based reaction datasets have become prominent in the field. The SciFinder database, founded on CASREACT and curated by the Chemical Abstracts Service (CAS), stands out as a substantial repository of chemical information. It encompasses a vast collection, housing over 100 million organic and inorganic substances and approximately 50 million reactions.[62, 63]

In parallel, Reaxys, constructed from the Gmelin and Beilstein Handbooks, represents a web-based search and retrieval system designed by Elsevier. This platform encompasses a comprehensive array of chemical compounds, bibliographic data, and chemical reactions. With over 270 million organic, inorganic, and organometallic compounds, as well as 64 million chemical reactions, Reaxys stands as a formidable resource in the chemical research landscape.[64]

Beyond literature-reported databases, Electronic Laboratory Notebook (ELN) systems have emerged to simplify the storage and retrieval of experimental data, widely utilized in daily industrial operations. However, these databases are typically neither publicly accessible nor utilized beyond internal applications,[65] except in the case of projects like MELLODDY, which seeks to leverage the accumulated experimental knowledge through federated learning, enabling the development of more advanced models.[66]

Freely accessible data is also abundant, with resources like the USPTO database, housing all patents granted by the United States Patent and Trademark Office since 1976. However, until the mining efforts by Lowe, accessing reaction data from patents was not straightforward.[67] Lowe's work made patent reactions more readily available to the public, significantly contributing to the catalysis of advancements in reaction predictions, retrosynthesis, and CASP tools.[68, 69] Commercial databases with higher levels of data curation have also emerged, with the Pistachio dataset extending the mining efforts to other patent offices and accumulating over 13 million reactions.[70]

2.4.3 RECENT RULE-BASED COMPUTER-AIDED SYNTHESIS PLANNING

In this overview of selected CASP approaches relevant to the thesis, I will categorize the various methods based on their single-step retrosynthesis core. This classification will begin with rules- or template-based approaches and then transition to machine learning-based methods.

Despite extensive efforts to encode chemistry into rules, none of the aforementioned tools has demonstrated a performance level suitable for daily use by chemists.[47] Only in recent years have commercial solutions, such as Chematica/Synthia™, achieved an effectiveness level suitable for industrial applications.[71, 72] Although the software may seem recent, it required years of research to attain this level of performance. In 2001, the Grzybowski group initiated the encoding of chemical rules for the analysis and exploration of reaction networks.[73, 74] With over 100 000 reaction rules, it was subsequently used for retrosynthesis within the Chematica software,[71] and applied for the retrosynthetic planning of medicinal compounds with synthesis experimental validation.[75]

Including Chematica/Synthia™, most previously discussed rule-based methods relied on chemists manually encoding the rules of chemistry. This task is very complex, as chemistry encompasses multiple factors and exceptions that are difficult to generalize.[76] Recent advances

in databases, coupled with progress in machine learning and computer power, have enabled the development of CASP tools capable of extracting knowledge from data.

This is achieved, for example, through the automatic extraction of rules or templates.[77, 78, 79, 80] Route Designer is one such example of a rule-based synthesis planning tool that utilizes the MOS reaction database from Accelrys and the Beilstein Crossfire reaction database from Elsevier. It addresses the combinatorial explosion problem by employing heuristics while balancing a somewhat exhaustive search.[81]

Building on Route Designer and on the ICSYNTH program reported by Bogevig *et al.*,[82] Segler *et al.* proposed the combination of template-based retrosynthesis with a policy neural networks.[83, 84] This neural network was trained to predict the most suitable templates to apply given the target molecule, among a large selection of applicable templates. This work significantly enhanced the performance of template-based retrosynthesis planning.

The ASKCOS retrosynthesis tool, as reported by Coley *et al.*, drew inspiration from previous concepts and implemented an open-source template-based approach using a graph neural network and a classifier for reaction conditions.[85, 86, 87]

AiZynthFinder reported by Genheden *et al.* is also an open-source synthesis planning tool that employs reaction templates.[88] It is guided by a Monte Carlo tree search and leverages an artificial neural network as a policy to select the most appropriate template to apply.

More recently, Dai reported the use of a conditional graphical model trained to learn when a given template should be applied. The model also considers if the template represents a good multistep opportunity by evaluating the retrosynthetic state and the feasibility of the resulting reaction, demonstrating improved performance.[89]

2.4.4 Recent machine learning approaches

Machine learning approaches first emerged in the field of forward reaction prediction, where multiple template-free methods have been reported,[90, 91, 92, 93, 94] followed by models for the retrosynthesis prediction tasks.[95, 96]

A recent breakthrough by Vaswani *et al.* in natural language processing (NLP) led to significantly faster training times through the parallelization of the attention mechanism and the avoidance of recurrent connections, while also improving performance and accuracy.[97] This model, known as the Transformer, was adapted to the field of chemistry for reaction prediction,[98, 99] and subsequently for retrosynthesis predictions.[100]

The transition from single-step retrosynthesis models to multistep was exemplified by Schwaller *et al.* who reported a Transformer retrosynthesis model combined with graph exploration, introducing the concepts of forward validation and round-trip accuracy.[101, 102] The graph exploration is guided by a simplicity score, derived from the Synthetic Complexity Score (SCScore) by Coley *et*

al..[103] Although the tool is freely usable online, the underlying graph exploration algorithm has not been made publicly available.

Another example, reported by Lin *et al.*, utilized a Transformer model for retrosynthesis in conjunction with a Monte Carlo tree search (MCTS) algorithm.[104] The scoring function is a heuristic that combines the decoding log probability from the model, the change in SMILES length, and the change in the number of rings to rank and score different pathways in retrosynthetic analysis.

Finally, the Retro* synthesis planning tool reported by Chen *et al.*, which is freely available and open source,[105] utilizes an AND-OR tree search and a neural network trained on historical planning paths. However, the tool lacks forward validation of the proposed retrosynthetic pathways, and the scoring function does not appear to consider the molecular structures.

2.5 Biocatalysis: a Solution for Greener Chemistry

2.5.1 Introduction to biocatalysis

Enzymatic reactions have been employed for thousands of years in the production of fermented beverages and food products. The oldest evidence of biocatalysis is the production of mead, a fermented beverage made from honey and water, dating back to approximately 7000 BC in China.[106] However, the term "enzyme" was only coined by Kühne in 1877 upon the discovery of trypsin, a digestive protein.[107] A detailed history of discoveries and uses of enzymes until the 20th century can be found in the review of Wisniak.[108]

Enzymes are biological molecules, typically proteins, that catalyze chemical reactions in living organisms. Thanks to evolution, enzymes are highly substrate-specific, each enzyme catalyzes a particular reaction or a group of closely related reactions.[109, 110, 111] The reaction occurs after binding to specific substrates, by stabilizing the transition state, lowering the activation energy or lowering the energy transition state required for the reaction to occur.[112, 113, 114] Enzymes play a crucial role in various cellular processes, including metabolism, signal transduction, and the synthesis of biomolecules

In modern research, enzymatic reactions are studied in various scientific fields, encompassing the investigation of cellular function and energy regulation, kinetics and mechanisms of reactions, structural biology, biosynthesis, drug discovery, immunology, metabolic pathway analysis, agriculture, and the food industry. This section will specifically highlight the application of enzymatic transformation reactions in organic synthesis, with a focus on the synthesis of drug-like small molecular-weight compounds.

2.5.2 Biocatalysis for green chemistry

Driven by the economic challenges associated with chemical waste disposal, the environmental concerns surrounding chemical synthesis, and the growing demand for sustainable practices, the principles of greener chemistry have gained attention in recent decades.[115] Notably, the fine chemicals and pharmaceutical industries continue to exhibit the highest E factor rate among chemical sectors, emphasizing the urgency for more sustainable approaches in these domains.[116] In addition to the broader scope of Green Chemistry,[10] enzymatic transformations, or biocatalysis, offer numerous advantages for chemical synthesis in the industrial landscape. Enzymes demonstrate versatility, functioning effectively in diverse conditions, including aqueous environments, at ambient temperature and pressure, and in the presence of air. Moreover, enzymes are biodegradable and can be sourced from renewable materials. These inherent properties position enzymes as optimal candidates for developing sustainable chemical processes, addressing various industrial considerations such as safety and cost.

From a chemical standpoint, enzyme catalysis presents several advantages compared to metal- and organocatalysis in chemical synthesis. Enzymes exhibit high selectivity, and the reaction conditions are mild, enabling the use of functional groups that may be incompatible with traditional chemical catalysts. Enzymes are well-suited for cascade reactions, where the product of one enzymatic reaction serves as the substrate for the next, facilitating the synthesis of complex molecules in a single step. Additionally, enzymes excel at catalyzing reactions that pose challenges for traditional chemical catalysts, including achieving regioselectivity, stereoselectivity, or functional group selectivity.[117, 118, 119]

2.5.3 Directed evolution for enzyme engineering

The utilization of enzyme catalysis in chemical synthesis is not consistently applied due to several factors. First, the available enzymes are limited in number, and their catalytic scope is often even more restricted. Secondly, not all available enzymes are adaptable for a given substrate, and the development of new enzymes is time-consuming and a tedious process.

Frances Arnold was honored with the Nobel Prize for her groundbreaking work on directed evolution, a method that facilitates the rapid development of enzymes with desired catalytic activity.[4, 120, 121] This method draws inspiration from Darwinian evolution, wherein enzymes undergo random mutations, and the mutants exhibiting improved desired activity are selectively chosen. Subsequently, these selected mutants undergo further rounds of mutation and selection until the desired properties are reached. Directed evolution has proven successful in enhancing the activity, stability, and substrate specificity of various enzymes, including lipases, oxidoreductases, transferases, hydrolases, and lyases.[122, 123, 124]

2.5.4 Biocatalysis in industry

The industrial-scale use of enzymes for synthesizing small molecules is not a novel concept, and numerous examples are detailed in various reviews.[13, 125] Typically, biocatalysis is incorporated at an advanced stage, once a product is established and commercialized, with the aim of ensuring long-term cost efficiency. Until recently, the integration of biocatalysis into synthesis planning tools was not a priority, and research in this domain was limited. However, recent progress in biocatalysis, robotics, and DNA sequencing has significantly simplified enzyme engineering, facilitating the inclusion of biocatalytic reactions in earlier stages of synthesis planning.[126, 127]

These days, biocatalysis finds application in diverse reactions for synthesizing small molecules,[128, 129, 130, 131] with an increasing number of instances documented in the literature and enzymatic reaction datasets.[132, 133] However, the substrate specificity, catalytic applicability, and the continual emergence of mutant enzymes pose challenges in keeping track of all potential enzymatic reactions, making it difficult to identify a suitable enzyme for a desired reaction.[134] Solutions are needed to enable chemists to efficiently navigate this enzymatic reaction landscape, to easily find the appropriate enzyme for a specific reaction, and to integrate it into a synthesis plan.[135]

2.6 Computer-Aided Synthesis Planning for Biocatalysis

The utilization of biocatalysis in synthesis planning tools is a relatively recent area of research that has gained interest with the lastest advances in biocatalysis. These improvements have allowed the integration of biotransformation processes at an earlier stage of product development, thanks to the faster development of custom enzymes. This development underscores the necessity of creating and integrating synthesis planning tools. Notably, enzyme-catalyzed reactions can accomplish tasks that are challenging or costly to perform through chemocatalysis processes[136]. The potential benefits of integrating a hybrid synthesis are particularly evident in enantioselective reactions. In many cases, nearly half of the product is discarded in the purification step due to the poor stereoselectivity of metal catalytic processes. Therefore, a hybrid synthesis approach could offer a more efficient and sustainable solution.[137, 138]

Enzymatic reactions have been explored for diverse purposes beyond small molecule synthesis. Consequently, the historical record of reactions, especially those associated with metabolic pathway studies representing a significant aspect of the field. Numerous repositories exist, offering comprehensive listings of enzymes and their corresponding reactions. Founded in 1987, BRENDA (BRaunschweig ENzyme DAtabase) stands out as a noteworthy database. It encompasses a collection of over 90 000 enzymes and 8424 EC numbers, compiling information from more than 157 000 references as of January 2023.[139, 140] Another important resource is the Kyoto Encyclopedia of Genes and Genomes (KEGG), initiated in 1995. This database includes data on genomes,

biological pathways, diseases, drugs, and chemical substances, providing details on over 20 000 reactions.[141, 142] MetaCyc is also a valuable repository containing metabolic pathways and over 18 000 reactions.[143, 144] Rhea provides valuable resources of expert-curated biochemical and transport reactions, annotated with EC numbers, reaction stoichiometry, and cofactors.[145, 146] Unlike others, PathBank provides a visualization interface for exploring pathways using supported source databases such as KEGG.[147, 148] MetaNetX provides cross-references between metabolites and biochemical reactions.[149, 150, 151] EzCatDB focuses specifically on enzyme active-sites and mechanism studies, including amino acid sequences of reported enzymes.[152, 153] Also, the UniProt Knowledgebase (UniProtKB) is a comprehensive resource for protein sequence and functional information, including enzyme classification and annotation.[154]

While these datasets have proven extremely valuable in the field of metabolic pathway prediction or exploration,[155, 156, 157, 158, 159, 160, 161, 162, 163] they have also found applications in biocatalysis. For instance, Probst *et al.* combined a selection of four datasets — BRENDA, Rhea, PathBank, and MetaNetX.[164] After data curation, the resulting dataset of 62 222 reactions named ECREACT was employed to train a Transformer model using multitask transfer learning. Both retrosynthesis and forward models were trained, allowing a round-trip validation as previously reported.[101, 102] The retrosynthesis model was designed to predict the precursors and the enzyme commission (EC) number simultaneously. The resulting single-step model was integrated into the multistep framework of Schwaller *et al.* enabling multistep enzymatic-exclusive route predictions. While Transformer models were not the initial attempt at multistep enzymatic-exclusive pathway prediction, the work of Finnigan *et al.* named RetroBioCat describes the use of a set of rules created by expert chemists.[165] They provide a "human-led" approach, where the user can dictate the direction of the retrosynthesis search, known as network explorer. Also, they report a pathway exploration mode that allows automatic exploration based on a scoring system. However, both strategies proposed by Probst *et al.* and Finnigan *et al.* were not implemented in combination with chemocatalysis, limiting their ability to predict reasonable multistep synthesis routes as they exclusively predict enzymatic transformations.

A notable positive aspect of RetroBioCat is its public availability, offering both the CASP source code and the expert-created reaction rules. This accessibility makes it a valuable resource to leverage and build upon. For instance, Sankaranarayanan *et al.* [166] integrated the RetroBioCat reaction templates [165] with the ASKCOS retrosynthesis planning tool.[87] They post-process chemocatalytic routes predicted by ASKCOS, identifying single-steps or sequences of steps that can be substituted by enzymatic reactions. While the approach is innovative, there remains a need for the chemocatalytic planning tool to propose an initial route suitable for step substitution, which is not always the case. Addressing this concern, Levin *et al.* reported a significant advancement — the first CASP tool based on reaction templates offering hybrid retrosynthesis planning incorporating

both biocatalysis and chemocatalysis.[167] They utilized the BKMS database,[168] extracting reaction templates and training a template prioritization neural network to determine when to apply enzymatic templates.

With the exception of a few discussed instances,[165, 166] the majority of retrosynthesis strategies for biocatalysis reported in existing literature relies on metabolic reaction datasets and exhibit notable limitations. These datasets predominantly feature natural products and heavy moieties, rendering them less applicable to the typical small molecules targeted for biocatalytic applications. Furthermore, these datasets often lack representation of engineered enzymes, which are valuable instances showcasing the extension of enzyme capabilities.

Additionally, while template-based approaches have demonstrated impressive performance, their generalizability to new reactions remains challenging, particularly for enzymes with intricate substrate specificity.[169] Conversely, recent deep-learning approaches with attention mechanisms appear to be more aware of the substrate context in which a given enzyme is employed. This higher awareness holds promise for extrapolating enzyme reaction predictions to unseen molecules while maintaining substrate specificity considerations. However, there is no deep-learning approach that combines both chemocatalysis and biocatalysis. More critically, none of these approaches is founded on real experimental data that includes insights from engineered enzymes.

3 PREDICTING ENZYMATIC REACTIONS WITH A MOLECULAR TRANSFORMER

The use of enzymes for organic synthesis allows for simplified, more economical and selective synthetic routes not accessible to conventional reagents. However, predicting whether a particular molecule might undergo a specific enzyme transformation is very difficult. Here we used multi-task transfer learning to train the molecular transformer, a sequence-to-sequence machine learning model, with one million reactions from the US Patent Office (USPTO) database combined with 32 181 enzymatic transformations annotated with a text description of the enzyme. The resulting enzymatic transformer model predicts the structure and stereochemistry of enzyme-catalyzed reaction products with remarkable accuracy. One of the key novelties is that we combined the reaction SMILES language of only 405 atomic tokens with thousands of human language tokens describing the enzymes, such that our enzymatic transformer not only learned to interpret SMILES, but also the natural language as used by human experts to describe enzymes and their mutations.

3.1 INTRODUCTION

The use of enzymes for organic synthesis, commonly referred to as the field of biocatalysis, greatly contributes to organic synthesis methodology by providing the possibility to carry out highly chemo-, regio-, stereo- and enantio-selective transformations under mild and environmentally friendly conditions, often allowing the redesign and simplification of synthetic routes by enabling reactions that are not possible with conventional chemical reagents.[13, 170] The advent of directed enzyme evolution as a tool to increase enzyme performance has also greatly contributed to improve the range and efficiency of enzyme catalyzed reactions for organic synthesis.[121] However, the

implementation of biocatalytic steps in synthetic processes remains challenging because it is very difficult to predict whether a particular substrate might actually be converted by an enzyme to the desired product.

Computer-assisted synthetic planning (CASP) comprises a range of artificial intelligence approaches to predict reaction products from reactant or reagents, or vice versa, and to plan retrosynthesis.[84, 85, 94, 96, 171, 172, 173, 174, 175] Here we asked the question whether CASP might be exploited to predict the outcome of enzymatic reactions for organic synthesis. Recent efforts in predicting enzymatic reactions focused on metabolic reactions from the KEGG enzymatic reaction database and predictions of drug metabolism,[176, 177, 178] as well as retrosynthetic planning with enzymatic reactions using a template based approach.[165] Here we considered the molecular transformer,[97, 98, 99]which is a sequence-to-sequence prediction model operating on text representations of reactions as reaction SMILES (Simplified Molecular Input Line Entry System)[25] including stereochemistry. We set out to use multi-task transfer learning combining the USPTO dataset[179]as a source of general chemistry knowledge with a few thousand enzymatic reactions collected from the scientific literature as a source of specialized knowledge (Figure 3.1).

We used transfer learning previously to enable the molecular transformer to predict complex regio- and stereo-selective reactions at the example of carbohydrates.[180] In this former study transfer learning was performed on a dataset of reactions described as SMILES, which are based on a vocabulary of only a few hundred atomic tokens identical to the vocabulary describing the general USPTO dataset used for primary training. One of the novelties of the present work on enzyme reactions is that we combine SMILES language for the substrates with human language for the enzyme descriptions. Those more diverse inputs result in an increase from 405 atomic tokens for SMILES only to a few thousand atomic and language tokens when describing enzyme reactions, implying that our transformer model had to learn to interpret not only the SMILES language but also natural language, as used by human experts to describe enzymes and their mutations.

3.2 Result and Discussion

3.2.1 Reaction datasets

As a general chemistry dataset, we used the previously reported "USPTO stereo augmented" dataset derived from the patent mining work of Lowe, which contains, for each of the one million reactions in the USPTO dataset, the original reaction SMILES and a randomized SMILES version, both conserving stereochemical information.[68, 69] To compose a specialized dataset of enzymatic reactions, we extracted 70 096 reactions labeled as "enzymatic reactions" from the Reaxys database.[64] We collected the data columns corresponding to reactant SMILES, product SMILES, and enzyme description ("reaction", "reagent" and "catalyst"). Canonicalizing all SMILES and removing reac-
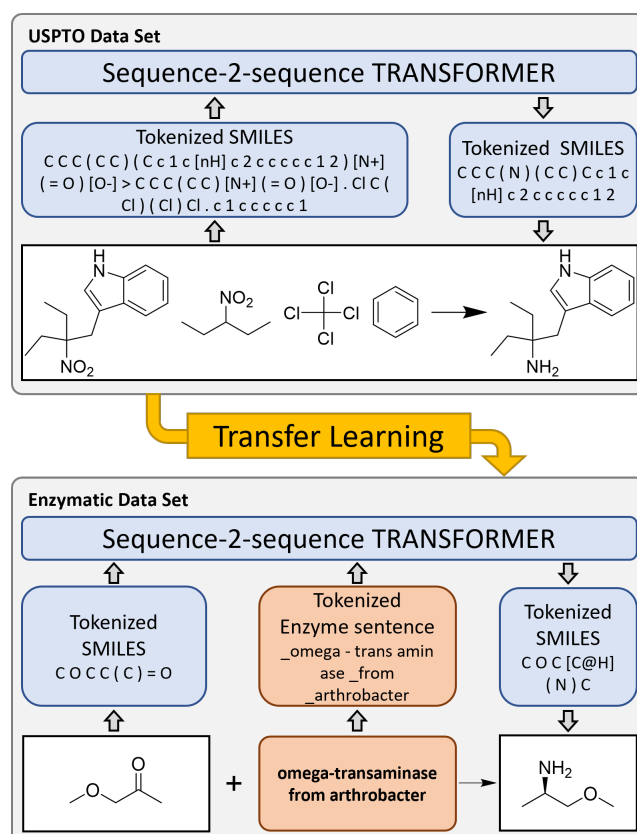
Figure 3.1: **General concept of the enzymatic transformer training.** The USPTO data set contains reactions SMILES describing reactants, reagents and products. The ENZR data set contains reaction SMILES as well as an additional text component.

tions lacking either reactants or products as well as duplicate entries (identical reactants, products and enzyme description) left 32 181 unique enzymatic reactions, each annotated with an enzyme description, referred to here as the ENZR dataset.

Although Reaxys does not cover the full spectrum of scientific literature about enzymes, the ENZR dataset contains a broad range of enzymes covering diverse reaction types, including not only highly specific enzymes such as glucose oxidases and dehydrogenases used in glucose monitoring devices,[181] but also enzymes with a documented broad substrate scope for organic synthesis including mechanistically promiscuous enzymes,[182] such as lipases used to promote aldol and Michael addition reactions,[183] or ene-reductases capable of reducing oximes,[184] thus providing a broad basis for training our model about the scope and specificity of different enzymes. We did not consider the enzyme databases KEGG[185] or BRENDA[140] because their data format is not homogeneous and many of the listed reactions are template-based and not assigned to documented examples.

To better understand our ENZR dataset, we analyzed enzyme reactions in terms of the frequency of occurrence of words with the suffix "-ase", which are the enzyme names, in the enzyme description. Across all enzyme reactions, 81.9% (26 348) contained a single "-ase" word, and 98.4% (31 663) contained one, two, or three "-ase" words (Figure 3.2a). The largest group of single "-ase" word reactions involved a lipase (17%), a type of enzyme which is almost exclusively used alone. By contrast, dehydrogenases and reductases were most frequent in reactions involving two or more "-ase" words, reflecting that such enzymes are often used in processes involving enzyme-coupled cofactor regeneration systems. The ten most frequent "-ase" words corresponded to well-known enzyme families and together covered 50.3% of all enzyme reactions (the 15 most frequent "-ase" words covered 57.0% of all reactions, Figure 3.2b). A finer analysis of enzyme families considering the complete enzyme description, which typically includes the enzyme source and the substrate type, showed that each enzyme family comprised a number of different enzymes (Figure A.1).

To visualize our ENZR dataset, we used our recently reported TMAP (tree-map) algorithm, a powerful tool to represent very large high-dimensional datasets containing up to millions of datapoints as connected trees in two dimensions.[186] In a first TMAP, we connected enzymatic reactions, each represented as a point, according to their similarity measured by the reaction fingerprint RXNFP, a recently reported reaction fingerprint derived from a neural network trained to classify patent chemical reactions.[19] This analysis considered the transformation of substrates into product molecules but not the enzyme description in each ENZR entry. Color-coding the TMAP by the 10 most frequent "-ase" words mentioned above, corresponding to the most abundant enzyme families in the ENZR dataset, showed that these enzyme families formed relatively well separated clusters of reactions, illustrating that, similarly to organic reagents, enzymes carry out well-defined functional group transformations (Figure 3.2c).
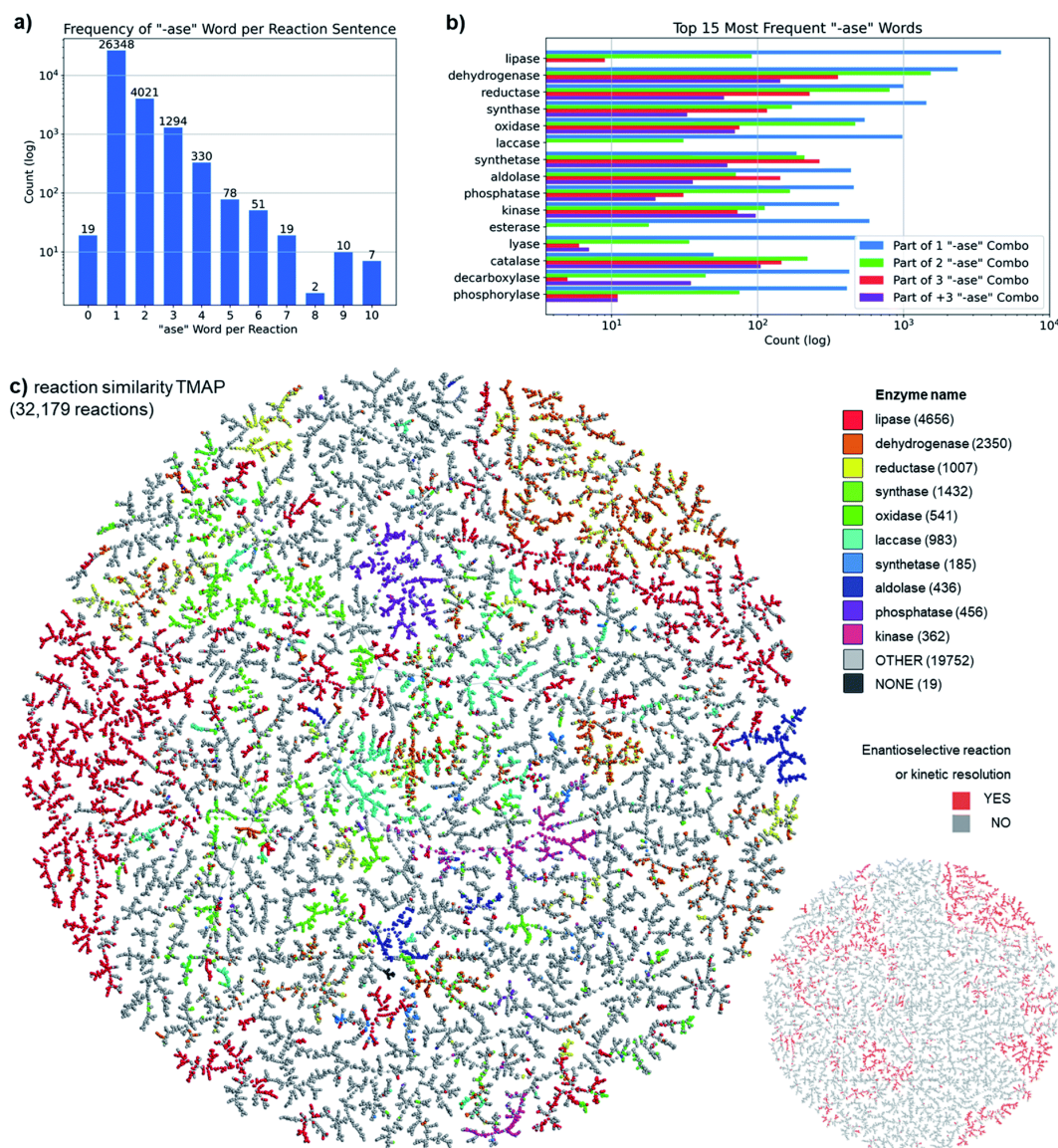
Figure 3.2: **Analysis of the ENZR dataset. (a)** Number of reactions depending on how many "-ase" words are present in the sentence. **(b)** Frequency of the top 15 "-ase" words depending on the count of enzyme name per reaction. **(c)** TMAP of reactions similarity color-coded by the 10 most frequent "-ase" words as listed in **(d)** combinations. The "other" category groups reactions with "-ase" words other than the top 10 "-ase" words as well as reactions containing more than one "-ase" word. Inset lower right: TMAP highlighting enantioselective and kinetic resolution reactions.

In a second color-coded version of the TMAP we labeled all enantioselective and kinetic resolution reactions, identified as reactions SMILES with no "@" characters in the reactants, indicating either the absence of chiral centers or an undefined stereochemistry at chiral centers, but the presence of at least one "@" character in the products SMILES, indicating a specific absolute configuration for chiral centers.[187] This color-code showed that enantioselective and kinetic resolution reactions also formed defined clusters corresponding to biotransformations with mostly dehydrogenases, lipases and reductases (Figure 3.2c, inset lower right).

The different enzymes also formed identifiable clusters in a different TMAP grouping reactions by substructure similarity of the reacting substrates using the extended connectivity fingerprint MHFP6 (Figure A.2).[15] This illustrated that enzymatic reactions in the ENZR dataset followed the well-known trend that enzymes only react with certain types of substrates, in contrast to chemical reagents which are usually only specific for functional groups. The range of substrates utilized by the enzymes covered a broad range of sizes from very small molecules such as pyruvate up to relatively large peptides (Figure A.2, inset).

Taken together, the analysis above indicated that the ENZR dataset contained a diverse set of enzymatic reactions, with the expected biases towards the most frequently used enzymes in the field of biocatalysis such as lipases and dehydrogenases.

3.2.2 Training and evaluation of transformer models for enzymatic reactions

Training a transformer model first requires tokenizing the input and output character strings to allow the model to learn which series of input tokens produces which series of output tokens. For the reaction SMILES in both USPTO and ENZR datasets, we used the approach reported previously for the general molecular transformer, which considers each character of the reaction SMILES as a separate token except Cl, Br, and character strings in square brackets, which denote special elements.[99] The set of tokens necessary for describing reaction SMILES in the USPTO amounted to 405 so-called atomic tokens, and did not increase for describing the reaction SMILES portion of our ENZR dataset, which we first canonicalized using RDKit.[188] To incorporate the enzyme information into our model, we tokenized the sentences describing the enzymes in the ENZR dataset using the Hugging Face Tokenizers library,[189] which after preprocessing resulted in a vocabulary of 3004 atomic and language tokens to describe the ENZR dataset.

In view of evaluating transformer models, we split the USPTO stereo augmented dataset randomly into a training set (900 000 reactions, 90%, 1.8 million reactions after adding for each canonical training reaction a duplicate using non-canonical precursor SMILES), a validation and a test set (each 50 000 reactions, 5%).[69] For the ENZR dataset, we first grouped reactions having the same product in different groups, and then split these groups into a training set (25 700 re-

actions, 80%), a validation and a test set (each 3200 reactions, 10%). Distributing these reaction groups rather than individual reactions into the different sets ensured that products which must be predicted in the validation or test sets have not been seen by the transformer during training or validation sets, respectively.

We then trained various models using OpenNMT[190] and PyTorch,[191] and evaluated them by presenting them with substrate SMILES, optionally together with the partial or full description of the enzyme, for each of the 3200 reactions in the test set. In each case, the model was challenged to write out the SMILES of the reaction product, including the correct stereochemistry, none of which had been seen by the model in the training or validation set. We analyzed whether the correct product was written out within the first one or first two solutions proposed by the model, as well as the percentage of invalid product SMILES, detected using RDKit, appearing among the first one or two solutions (top 1 and top 2 accuracy, blue and cyan bars, top 1 and top 2 invalid SMILES, red and orange bars, Figure 3.3A).

We first evaluated if transformer models could be trained to predict reaction products from only the substrate by omitting any enzyme information during training. The UPSTO only model showed approximately 10% accuracy but a very low percentage of invalid SMILES, indicating that this model understood chemistry but lacked expertise in biotransformations (Figure 3.3A, entry (a)). The ENZR only model also performed poorly (~20% accuracy) and produced ~10% invalid SMILES, reflecting that general chemistry training was insufficient with this relatively small dataset (Figure 3.3A, entry (b)). Nevertheless, training with both models using sequential transfer learning (STL) or multi-task transfer learning (MTL) reached ~50% accuracy, indicating that substrate structure was partially predictive of the outcome of enzymatic reactions even in the absence of any enzyme information (Figure 3.3A, entries (c) and (d)). This partial prediction based on only the substrate reflects the fact that certain types of substrate molecules are only documented to react with specific enzymes in the ENZR dataset. For example, many alcohols are only documented to react with alcohol dehydrogenases to produce the corresponding ketone, such that a transformer model trained with the reaction SMILES learns to predict the ketone as the most likely product even without enzyme information, a prediction which is most of the time the correct one.

Adding enzyme information in form of "-ase" words alone did not significantly increase prediction performance when using only ENZR, however combining the data with the USPTO by transfer learning increased in terms of top 1 accuracy to 51.7% with STL and 54.0% with MTL (Figure 3.3A, entries (e)–(g)). Top 1 prediction accuracy increased further up to 59.5% with STL and 62.2% with MTL when using the complete enzyme information as full sentence (Figure 3.3A, entry (j)). Note that the model trained with ENZR alone only reached 34.3% top 1 accuracy with full enzyme names and produced ~10% invalid SMILES, showing that the general chemistry training learned from USPTO was essential even with full enzyme information (Figure 3.3A, entry (h)).
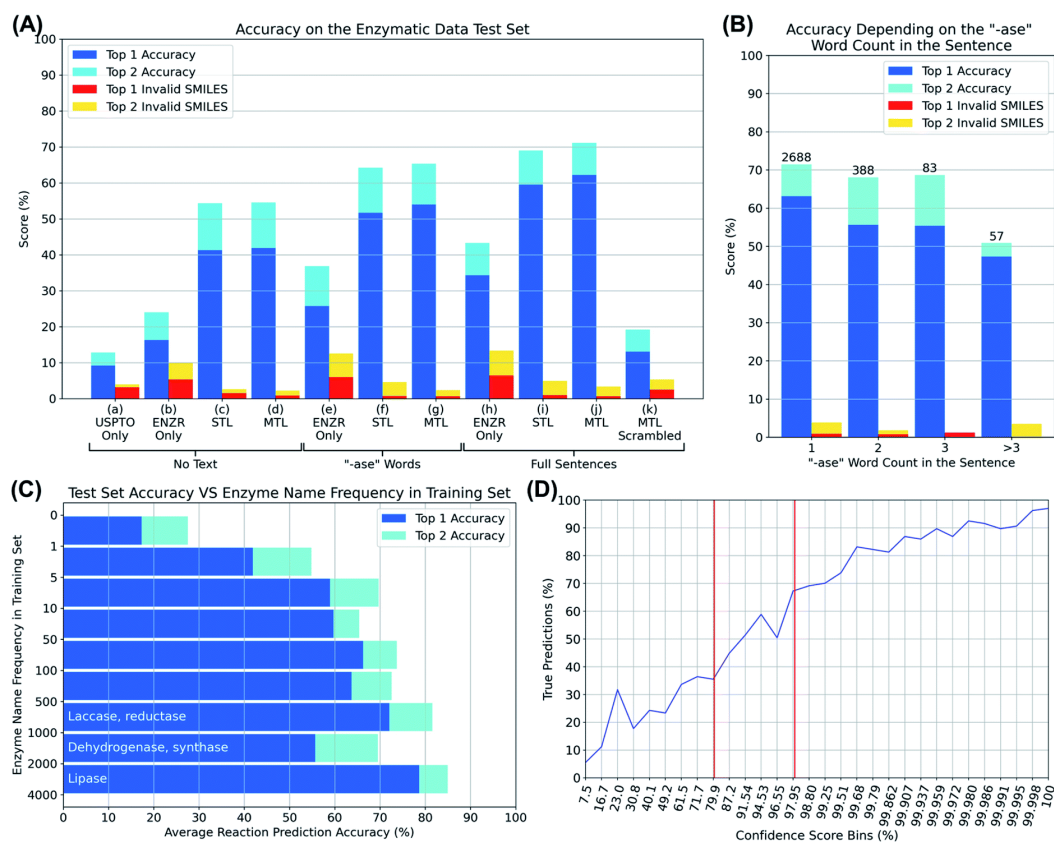
Figure 3.3: **Prediction accuracies (A-D). (A)** Top prediction accuracy and invalid SMILES on the enzyme reaction test set for various models. **(a)** USPTO model from Schwaller *et al.* trained without any enzymatic transfer learning and tested without enzyme sentence. **(b)** Enzymatic DB without USPTO data set. **(c)** USPTO model transfer learned (sequential) to enzymatic DB trained without any enzyme description part. **(d)** USPTO model transfer learned (multi-task) to enzymatic DB trained without any enzyme description part. **(e)** Enzymatic DB without USPTO data set trained with "-ase" words only. **(f)** USPTO model transfer learned (sequential) to enzymatic DB trained with "-ase" words only. **(g)** USPTO model transfer learned (multi-task) to enzymatic DB trained with "-ase" words only. **(h)** Enzymatic DB without USPTO data set trained with enzyme full sentences. **(i)** USPTO model transfer learned (sequential) to enzymatic DB trained with enzyme full sentences. **(j)** USPTO model transfer learned (multi-task) to enzymatic DB trained with enzyme full sentences. **(k)** Best multi-task model tested by swapping enzyme full sentences between reactions of the test set. **(B)** Accuracy on the test set depending on how many "-ase" words are present in the sentence. **(C)** Accuracy on the test set depending on how frequent the "-ase" words combination from the sentences appears in the training set. **(D)** True predictions rate against confidence scores, bins were adjusted to obtain an equal distribution of predictions over the bins. Vertical red bars represent our limits to indicate true or false predictions.
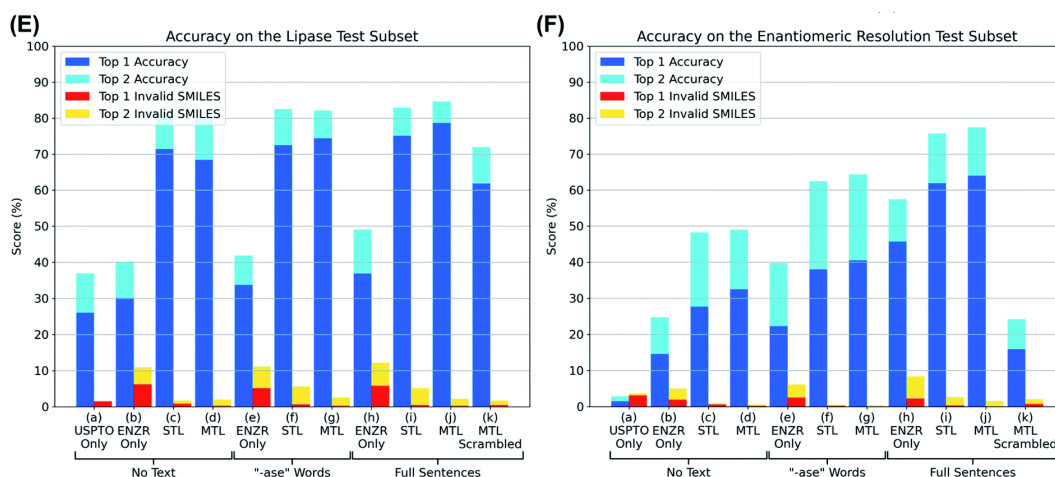
Figure 3.3: **Prediction accuracies (E-F). (E)** Top prediction accuracy and invalid SMILES on lipase re-
actions of the test set only. **(F)** Top prediction accuracy and invalid SMILES on enantiomeric
resolution reactions of the test set only.

Furthermore, testing the MTL with a test set in which the enzyme information was scrambled
between reactions resulted in poor results (~15% accuracy), indicating that the true enzyme in-
formation was required rather than the presence of random text information (Figure 3.3A, entry
(k)). Examples of the added value of enzyme information for predicting the outcome of an enzyme
reaction are provided with the cases of linoleic acid conversion with various oxygenases and dehy-
drogenases, and the conversion of l-tyrosine by a lyase and a tyrosinase. These examples are taken
from the test set and reflect true predictions since they have not been seen by the model during
training or validation (Figure 3.4).

3.2.3 ANALYZING THE PREDICTION PERFORMANCE OF THE ENZYMATIC TRANSFORMER

The comparisons above showed that an excellent prediction performance was reached by the
transformer trained using MTL combining the USPTO and the ENZR dataset using full enzyme
names as enzyme information. Retraining this model with different splits of training, validation
and test sets gave indistinguishable results in terms of prediction accuracy. This model was selected
for further investigation and is referred to as the "enzymatic transformer".

Considering that many reactions in the ENZR dataset contain multiple enzymes, we wondered
if our transformer might be confused in such situations because the main enzyme and the cofactor
regeneration enzyme are not labeled as such. Indeed, the prediction accuracy of the enzymatic
transformer was lower for reactions with multiple enzymes compared to reactions with a single
enzyme (Figure 3.3B). However, in many cases of multi-enzyme reactions including cofactor regen-
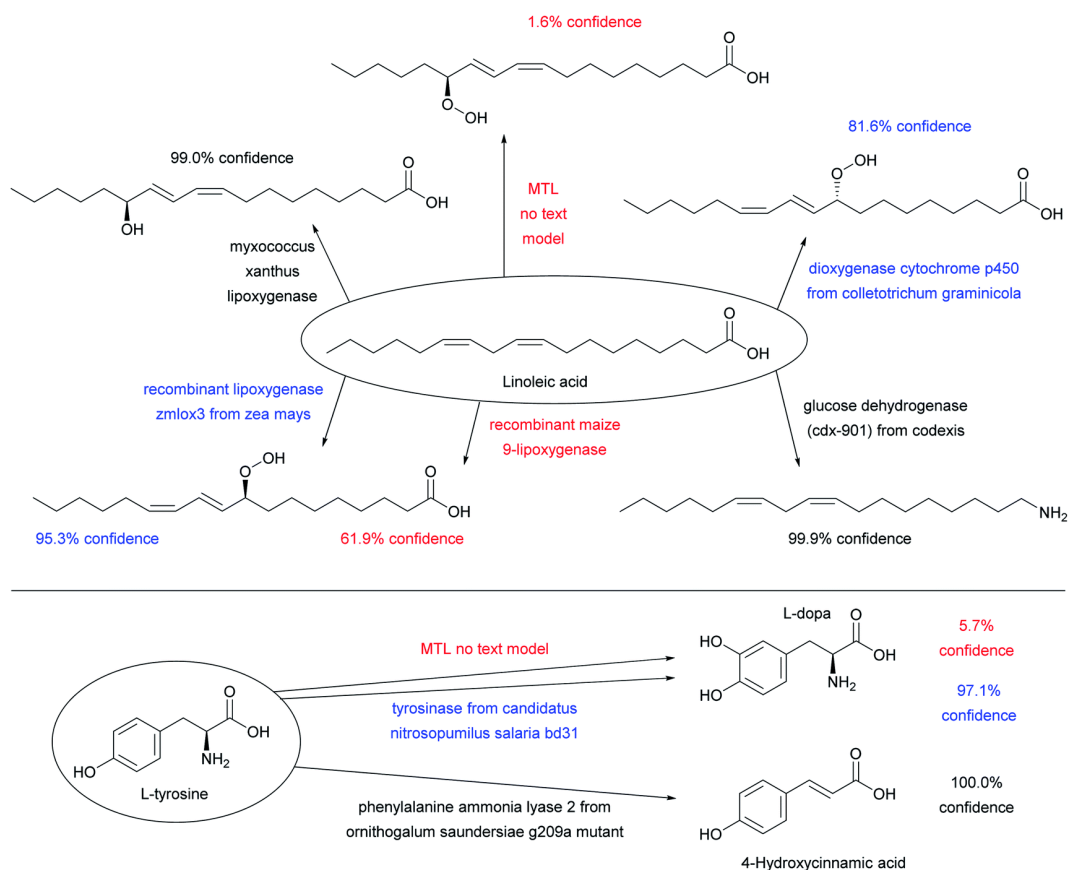
Figure 3.4: **Examples of substrates applied to various enzymes** using the MTL transformer with full sentences, which illustrate predictions of reactions from the test set not seen by the model during training. The color code indicates high confidence predictions (score >98%, black), uncertain predictions (score 80–98%, blue), and low confidence predictions (score <80%), see Figure 3.3D for discussion of confidence scores. All enzymatic reactions are predicted correctly, however the confidence score varies. The predictions of the MTL no text model are shown to illustrate what the transformer predicts when the enzyme information is missing.

eration, the transformer provided the correct prediction when omitting the cofactor regenerating enzyme or swapping it for an equivalent one (glucose dehydrogenase to phosphite dehydrogenase, Figure A.3).

Since transformer models require a large number of examples for good performance, we also tested prediction accuracy as function of the number of occurrences of the enzyme name in the training set. Indeed, a prediction accuracy of almost 80% was reached for lipases, which were the most abundant in the training set (Figure 3.3C). Nevertheless, prediction accuracy reached a good level (~60%) as soon as more than five examples of a particular enzyme were present in the training set.

In the best transformer model using MTL on full sentences, there was a clear association of the prediction confidence score with accuracy, as observed with other transformer models (Figure 3.3D).[180] Overall, 85.5% of the predictions with confidence score >98% were true and 75.6% of the predictions with confidence score <80% were false, suggesting to use confidence score values >98% or <80% as indicators for a true (the reaction is worth testing) or false (the reaction outcome is uncertain) prediction.

Since the subset of the test set containing the word "lipase" performed best (Figure 3.3C), we evaluated this subset exhaustively with all models (Figure 3.3E). While models trained on the USPTO or ENZR dataset without enzyme information performed poorly (Figure 3.3E, entries (a) and (b)), combining both sets with STL (entry (c)) or MTL (entry (d)) reached an excellent accuracy (>70%), indicating that the presence of an ester functional group is sufficient for the model to recognize a lipase biotransformation even in the absence of the enzyme name. However, models trained with ENZR alone using only the "ase" word or the full sentence performed poorly (Figure 3.3E, entries (e) and (h)), showing that this relatively small dataset contained insufficient general chemistry knowledge to training even for the relatively simple lipase reaction. Overall, the model trained on both datasets using STL and the full enzyme description performed best for lipases, as observed in the entire dataset (Figure 3.3E, entry (j)). However, scrambling the enzyme information between different reactions in the lipase only test set did not decrease prediction accuracy as dramatically as for the full set, reflecting the fact that all lipases catalyze very similar reactions. In addition, 36.89% of the lipase test set cases were reactions with *Candida antarctica* lipase B, the most frequently used lipase in biotranformations, in which case swapping the enzyme information does not induce any change.

Enzymatic reactions are often used to perform kinetic resolutions, typically using hydrolase enzymes such as lipases, or to transform achiral substrates into chiral products, typically to produce chiral alcohols or amines from achiral ketone precursors. To evaluate the performance of the transformer on such reactions, we defined enantiomeric resolutions as enzymatic reactions containing chiral centers, identified by the presence of at least one "@" character in the SMILES, in the reaction products only, which corresponded to 6495 reactions in the entire ENZR dataset (20.18%), and 687 reactions in the test set (21.35%). The relative performance of the different transformer models in this subset was comparable to that of the entire dataset, indicating that the transformer model was able to learn the enantiomeric preference of enantioselective enzymes as successfully as the overall enzymatic transformation (Figure 3.3F).

29

3.2.4 EXAMPLES OF CORRECT AND INCORRECT PREDICTIONS BY THE ENZYMATIC TRANSFORMER

The types of enzymatic reactions predicted correctly by the enzymatic transformer are well illustrated by selected cases (Figure 3.5). These include the correct product prediction including chirality for kinetic resolutions using lipases (reactions (1)[192] and (2)),[193] two enantioselective reductions of ketones using alcohol dehydrogenases (reaction (3)[194] and (4)),[195] an enantioselective imine reduction (reaction (5))[196] and reductive amination with a transaminase (reaction (6)).[197]

Considering that none of the products of these reactions have been seen by the model during training, the ability of the enzymatic transformer to predict not only the correct reaction product but also the correct stereochemical outcome of the enantiomeric resolution reactions is remarkable. It must be pointed out that the prediction is always done by analogy to examples, including cases of engineered enzymes. For instance, in reaction (1) with a mutant CALB enzyme, the transformer has learned from the training set that this triple mutant has an altered stereospecificity, and listing the mutation is sufficient for the model to make the correct prediction in the example from the test set. The product structure prediction is still correct but the stereoselectivity is lost when using simply "*Candida antarctica* lipase B" as enzyme description, which corresponds to the experimental result (Figure A.4).

Cytochrome P450 mediated regioselective demethylation (reaction (7))[198] or hydroxylations (reactions (8)[199] and (9))[200] further illustrate the predictive power of the enzymatic transformer. From the 405 cytochrome P450 mediated reactions in ENZR, 316 were used in the training set and 46 in the validation set. The resulting enzymatic transformer correctly predicted the product structure of 17 (40%) of the 43 cytochrome P450 reactions in the test set considering the top 1 predictions and 22 (51%) considering the top 2 predictions. The numbers increased to 21 (49%) correct predictions for the top 1 and 25 (58%) for the top 2 predictions when ignoring stereochemistry. These prediction accuracies are far from perfect but still very remarkable considering that the reaction site and type of cytochrome P450 reactions transformation are difficult to predict for a chemist (Figure A.5 and A.6).

In the above examples, a shorter description of the enzyme often reduces the confidence score and may induce errors in the predicted stereochemistry or product structure (red labels in Figure 3.5 and A.4). Such errors when using short enzyme names are not surprising considering that models trained with only "-ase" words performed worse than models trained with the full enzyme description (Figure 3.3A).

Analyzing unsuccessful predictions by the enzymatic transformer in a random sample of 200 reactions from the test set selected to cover various reaction types and enzymes provides further insights (Figure 3.6). Inaccurate predictions may sometimes simply reflect errors in database en-
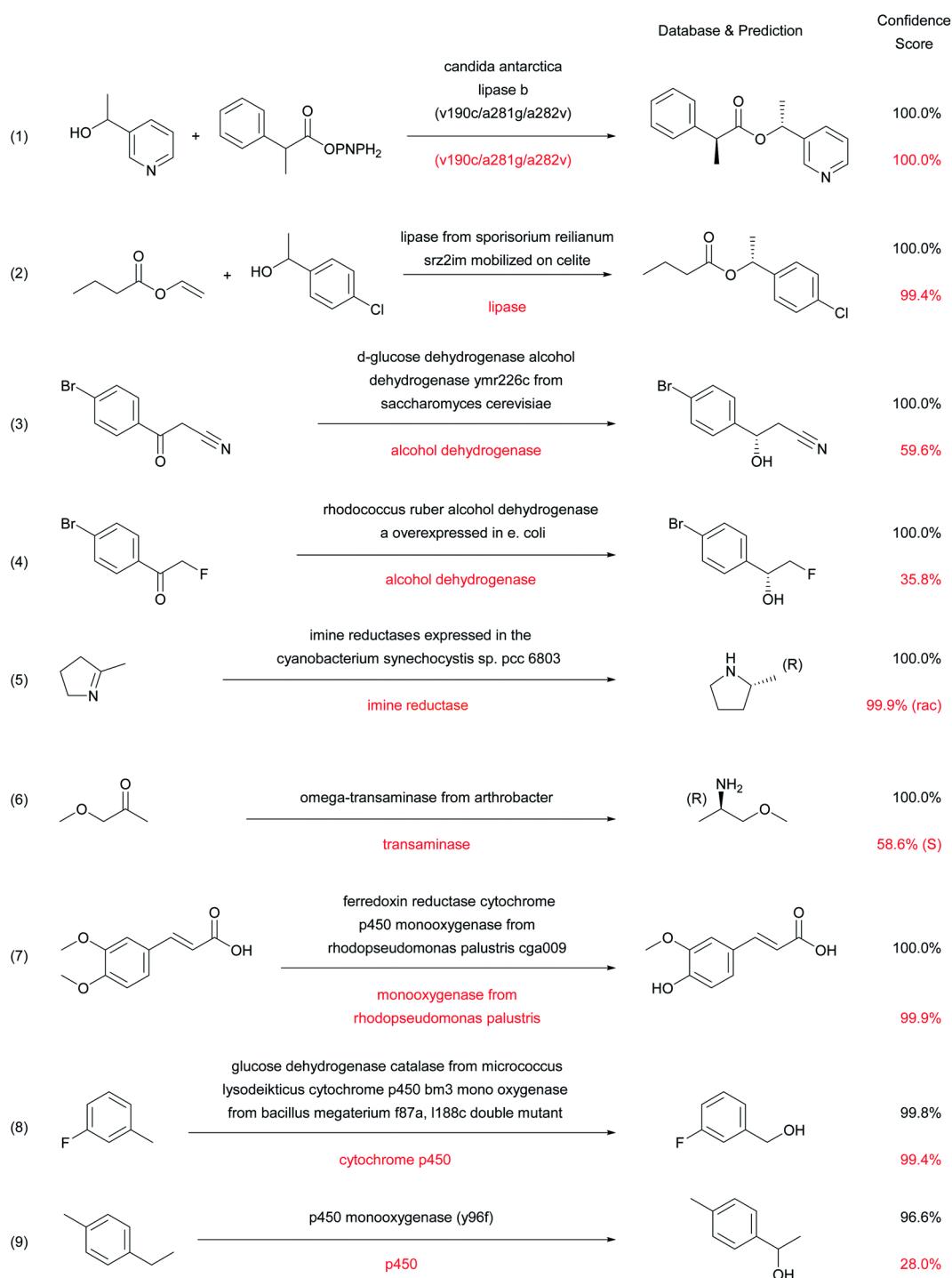
Figure 3.5: **Examples of successful predictions by the enzymatic transformer.**

tries. For instance, the enzymatic transformer correctly predicts, with a high confidence score, the formation of thymine from the hydrolysis of a thymidine nucleoside analog by uridine phosphorylase, however the database entry wrongly recorded the isomeric 6-methyl-uracil as the product (reaction (10)).[201] The model also correctly predicts with high confidence score the alcohol hydrolysis product in the hydrolysis of a β-hydroxysulfone by porcine liver esterase. However, this product is unstable and spontaneously eliminates to form a styrene, which is the product isolated and recorded in the database (reaction (11)).[202] Furthermore, the model correctly predicts that 5-deoxy-b-d-ribofuranose is the product formed by the action of a nucleosidase on the parent adenosine nucleoside, which it writes down in the cyclic hemi-acetal form, while the database entry recorded the open-chain aldehyde form (reaction (12)).[203]

Other examples reflect true limitations of our model, for example errors in the regioselectivity of hydroxylation of 7-methoxy-3,4-dihydronaphthalen-1(2H)-one (reaction (13))[204] and α-naphthol (reaction (17))[205] by cytochrome P450. In the case of the formation of (+)-δ-cadinene from geranyl pyrophosphate by (+) cadinene synthase, our model predicts the correct product structure and stereochemistry, however the deuterium label, which is lost during cyclization, is wrongly incorporated into the predicted product (reaction (14)).[206] The model may also predict the correct product structure but the opposite enantiomer, as illustrated for the benzylic hydroxylation of ethylbenzene by cytochrome P450 (reaction (15)),[207] or with missing stereochemistry, as illustrated for the biotransformation of 4-methyl-cyclohexanol by a sequence of an alcohol dehydrogenase and a cyclohexanone monooxygenase to produce an enantiomerically pure lactone (reaction (16)).[208]

Note that the enzymatic transformer can only predict the structure of reaction products based on what it has learned from examples in the ENZR source database. For example, the reaction rates of 49 different alcohol substrates with a wild-type choline oxidase (WT) and an engineered version with an expanded substrate scope (M) have been reported with a broad range of values.[209] However, the Reaxys entry used for ENZR attributed each reaction only to one of the two enzymes, which was in each case the faster reacting enzyme, even if the rates were almost equal. The enzymatic transformer was trained with a random subset of 32 reactions attributed to M and five reactions attributed to WT (Figure A.7) and validated with five M and two WT cases (Figure A.8). The model then correctly predicts the two WT and three M reactions in the test set, however in each case the same product is predicted with very high confidence for both WT and M enzymes (Figure A.9). This prediction is correct for the two WT cases where the reported rates are almost equal for WT and M, but inaccurate for the three M cases where the activity of WT is much lower, including one case where even the M rate is impractically low, reflecting the fact that the training data does not consider reaction rate information.
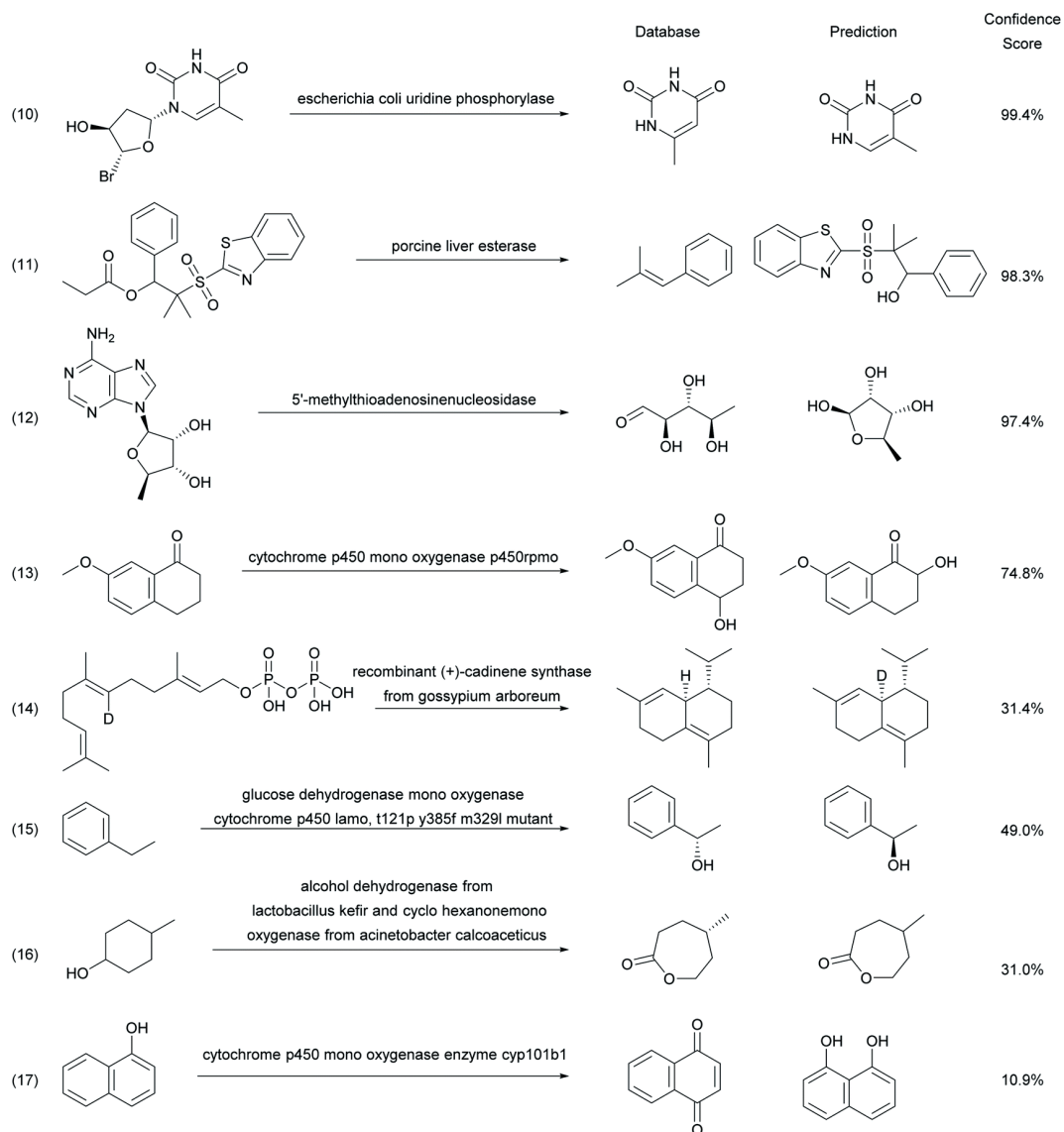
Figure 3.6: **Examples of unsuccessful predictions by the enzymatic transformer.**

3.2.5 How to use the Enzymatic Transformer

The examples discussed above belong to the ENZR test set for which the product molecules have never been seen by the enzymatic transformer during training and validation, but they are recorded cases for which a look-up in the scientific literature will give the answer. In a possible application, one might use the enzymatic transformer to select which enzyme might be best suited for a given biotransformation not yet recorded in the dataset. To carry out such prediction, one would analyze the product structures and confidence scores returned by the model when presented with a given substrate and various enzymes.

As a theoretical example, we consider the reduction of levulinic anilide to either enantiomer of the corresponding chiral alcohol, a reaction which is not present in the training set. We used the enzymatic transformer to predict which product would be formed by exposing this ketone to 163 alcohol dehydrogenases and 60 ketoreductases in the ENZR dataset. In this case, the transformer model predicts with high confidence two experimentally verified cases of two different keto-reductases in the test set forming either the (S) or the (R) enantiomeric alcohol enantioselectively. In addition, the transformer also proposes high confidence reactions to either enantiomers involving other ketoreductase and alcohol dehydrogenases enzymes, which could be considered for experimental testing (Figure 3.7).

One might also use the enzymatic transformer to predict which substrates might be converted by a given enzyme. To illustrate this point, we considered the enzyme "d-glucose dehydrogenase alcohol dehydrogenase ymr226c from Saccharomyces cerevisiae", which is documented in six reactions of the training set to reduce various acetophenones enantioselectively and correctly predicts the product structure and stereochemistry for the 2 examples in the test set (Figure A.10, substrates **D1** and **D2**). One can then challenge the enzymatic transformer to predict which product might be formed with further ketone substrates and the same enzyme. The transformer predicts the probably correct alcohol products with high confidence scores for ketones that are structurally related to the database examples (Figure A.10, substrates **D3–D15**). Among further analogs that are less similar, three cases are predicted with high confidence (Figure A.10, substrates **D16–D18**), and the remaining five cases have much lower confidence scores as well as sometimes unlikely product structure, indicating that the model is uncertain about the possible outcome of these reactions (Figure A.10, substrates **D19–D22**).

3.3 Conclusion

We had previously shown the principle of transfer learning to specialize the general USPTO transformer model at the example of carbohydrate reactions, however this approach used SMILES information only and a limited set of 405 tokens.[180] Here we showed for the first time that

ketoreductase 349 (100% confidence)
ketoreductase from *Geotrichum candidum* (100% confidence)
E.coli alcohol dehydrogenase (100% confidence)
Saccharomyces cerevisiae alcohol dehydrogenase (100% confidence)
ketoreductase enzyme (100% confidence)
ketoreductase p3-b03 (99.9% confidence)

(S)

(R)

quinoprotein membrane-bound *meso*-erythritol dehydrogenase
(100% confidence)
ketoreductase cdx-022 mutant (100% confidence)
ketoreductase kred-nadh-117 (100% confidence)
ketoreductase kred-nadh-126 (100% confidence)
ketoreductase enzyme p1-a04 (100% confidence)
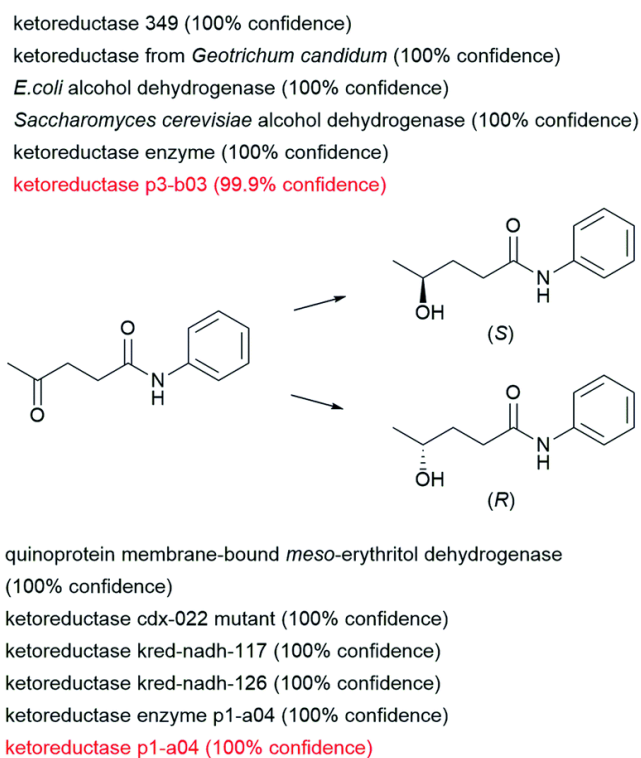ketoreductase p1-a04 (100% confidence)

Figure 3.7: **Examples of usage of the enzymatic prediction model to find suitable enzymes leading to different enantiomers.** Screening sentences were extracted from the entire dataset. Filtering was applied for dehydrogenases and ketoreductases from single enzyme systems and filtered for simple sentences (less than 5 words). Resulting in a total of 223 sentences (163 dehydrogenases and 60 ketoreductases). Are shown the top 5 confidence score sentences leading to both enantiomers. Red colored sentences were present in the test set providing experimental proof.

the general USPTO transformer model can be used as a basis for transfer learning using a more complex language information, here an extended vocabulary of several thousand language and atomic tokens describing enzymatic reactions in text format. Despite of the relatively small size of the ENZR dataset of enzymatic reactions used here, the resulting enzymatic transformer model predicted the outcome of enzymatic transformations including enantioselective reactions with excellent accuracy. This type of approach might be extended in the future to incorporate additional information such as reaction conditions and experimental procedures.

It should be noted that the text descriptions of enzymes used in our ENZR dataset most often represent a rather plain description of the reaction and substrate involved, e.g. "tyrosine decarboxylase", which provides a direct hint for the enzymatic transformer for proposing a product structure. Nevertheless, other descriptions of enzymes such as their EC number,[177] their amino acid sequence or a representation of the sequence produced by an auto-encoder,[210, 211] might also be exploitable for the enzymatic transformer if these would be available since these descriptions in principle contain the same information, even if in a more indirect manner.[212]

Here we demonstrated the feasibility of using a text description of an enzyme to train a transformer model to predict product structure given a substrate and the enzyme. The same data type might be suitable to train a transformer to predict the substrate structure given a product and an enzyme (retro-synthesis) or to predict an enzyme name given a substrate and a product, however to succeed such models might require much larger datasets than the relatively small ENZR dataset used here.

In this study, we obtained the best prediction accuracies when using multi-task transfer learning based on the full description of the enzymes. However, model performance was limited by database size and was lower with enzymes for which only few examples were available. Furthermore, analysis of successes and failures showed that model performance is also limited by the occurrence of database entry errors. Model performance can probably be increased by using larger and higher quality training dataset. Furthermore, the performance of our enzymatic transformer model was highest with the enzymes that are most represented in the ENZR dataset, which were lipases and dehydrogenases due to the historical nature of the data source reflecting which enzymes have been mostly used in the literature. Considering that transformer models learn from example, increasing the performance for other types of biotransformations such as keto-reductases and monooxygenases will critically depend on acquiring training data for such types of enzymes. Provided the availability of experimental training data, the transfer learning approach demonstrated here should be optimally suited to integrate this data into predictive models capable of assisting chemists in implementing biotransformations for chemical synthesis.

3.4 Methods

3.4.1 Data collection

The USPTO data was downloaded from the patent mining work of Lowe.[69] The ENZR data set was downloaded from Reaxys.[64] Enzymatic reactions were found querying "enzymatic reaction" keywords directly in the search field.

3.4.2 Transformer training

The enzymatic transformer model was trained based on the molecular transformer work from Schwaller *et al.*[99] The version 1.1.1 of OpenNMT,[190] freely available on GitHub,[213] were used to preprocess, train and test the models. Minor changes were performed based on the version of Schwaller *et al.*[99] SMILES were also tokenized using the same tokenizer as Schwaller *et al.*[99] The ENZR description sentences were tokenized by the Hugging Face Tokenizers[189] using a byte pair encoding[214] resulting in a vocabulary of 6139 language tokens (top 40 most frequent tokens in Figure A.11) for which the occurrence frequencies follow a power-law distribution shown in Figure A.12. For our model, we used the 3000 most frequent tokens representing 97.4% of tokens found in ENZR sentences. The 3139 remaining tokens only represent 2.6% of occurrences and have less important frequencies going from 7 to 1. The following hyperparameters were used for the multi-task model:

```
preprocess.py -train_ids ENZR ST_sep_aug \
-train_src \$DB/ENZR/src_train.txt \$DB/ST_sep_aug/src-train.txt \
-train_tgt \$DB/ENZR/tgt_train.txt \$DB/ST_sep_aug/tgt-train.txt \
-valid_src \$DB/ENZR/src_val.txt -valid_tgt \$DB/ENZR/tgt_val.txt \
-save_data \$DB/Preprocessed \
-src_seq_length 3000 -tgt_seq_length 3000 \
-src_vocab_size 3000 -tgt_vocab_size 3000 \
-share_vocab -lower \

train.py -data \$DB/Preprocessed \
-save_model ENZR_MTL -seed 42 -train_steps 200000 -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 4 -label_smoothing 0.0 -layers 4 \
-rnn_size 384 -word_vec_size 384 \
-encoder_type transformer -decoder_type transformer \
-dropout 0.1 -position_encoding -global_attention general \
-global_attention_function softmax -self_attn_type scaled-dot \
-heads 8 -transformer_ff 2048 \
-data_ids ENZR ST_sep_aug -data_weights 1 9 \
-valid_steps 5000 -valid_batch_size 4 -early_stopping_criteria accuracy
```

3.4.3 Validation

Canonicalized SMILES were compared to assess the accuracy of the models. Distribution of the training, validation and test set was randomly distributed after being grouped by reaction product multiple time resulting in constant accuracy.

3.4.4 TMAPs

TMAPs were computed using standard parameters.[186] The reaction fingerprint (RXNFP)[19] as well as the molecular substructure fingerprint (MHFP6)[15] was computed with a dimension of 256.

3.5 Availability of Data and Materials

The USPTO data is available from the patent mining work of Lowe.[69] Reactions from Reaxys are accessible with subscription. The modified version of OpenNMT as well as the code for data extraction and preprocessing as well as to tokenize, train and test the model are available from: https://github.com/reymond-group/OpenNMT-py.

4 Multistep Retrosynthesis Combining a Disconnection Aware Triple Transformer Loop with a Route Penalty Score Guided Tree Search

Computer-aided synthesis planning (CASP) aims to automatically learn organic reactivity from literature and perform retrosynthesis of unseen molecules. CASP systems must learn reactions sufficiently precisely to propose realistic disconnections, while avoiding overfitting to leave room for diverse options, and explore possible routes such as to allow short synthetic sequences to emerge. Herein we report an open-source CASP tool proposing original solutions to both challenges. First, we use a triple transformer loop (TTL) predicting starting materials (T1), reagents (T2), and products (T3) to explore various disconnection sites defined by combining systematic, template-based, and transformer-based tagging procedures. Second, we integrate TTL into a multistep tree search algorithm (TTLA) prioritizing sequences using a route penalty score (RPScore) considering the number of steps, their confidence score, and the simplicity of all intermediates along the route. Our approach favours short synthetic routes to commercial starting materials, as exemplified by retrosynthetic analyses of recently approved drugs.

This chapter has been published in Chemical Science:

4.1 Introduction

Retrosynthetic analysis consists in drafting a synthetic sequence to produce a desired product from available starting materials. This analysis is one of the most useful but also difficult tasks in organic chemistry because it requires to integrate the large and complex set of rules that have emerged from millions of reactions reported in almost 200 years of organic synthesis. Computer-aided synthesis planning (CASP), initially conceived by E. J. Corey in the 1960s,[2] aims to harness the power of computers to automate retrosynthesis by exploiting data from experimental reactions collected in databases such as Reaxys[64] or the open-access reaction dataset extracted from US patent office data.[68, 69] These databases list reactions of sets of starting materials (SM) and sets of reagents (R) to form one or several products (P).

While expert systems based on hand-written rules such as Chematica/Synthia™ perform quite well for synthesis planning,[71] CASP ultimately aims to exploit artificial intelligence to automatically learn organic synthesis from reaction examples and propose synthetic routes for new molecules without human intervention.[171, 215, 216, 217, 218] Template-based approaches extract reaction rules in the form of substructure transformations and use machine learning to learn their applicability domain from the structure of P in the training data.[84, 85, 88, 179] On the other hand, transformer-based models use the linear SMILES[25, 26] notation of chemical reactions and learn to translate the character string of P into the character string of SM + R, or *vice versa*.[95, 96, 97, 98, 99, 100, 102, 104, 160, 174, 219] The single-step predictions are then iterated to propose multistep retrosyntheses of target molecules from a selected set of building blocks (BB), which requires prioritizing possible routes using search algorithms such as Monte Carlo tree search,[84, 104, 220] and-or trees,[105, 160] or a multistep graph exploration.[102]

Any CASP system must overcome two critical challenges to propose realistic retrosyntheses. First, the system must learn the context of reactions sufficiently well to propose reactions that make sense, but without overfitting such as to propose diverse retrosynthetic operations on previously unseen molecules. Second, the route-prioritizing algorithm must be designed to allow short sequences to emerge from the multitude of predicted possibilities.[171] Herein, we report a transformer-based retrosynthesis tool which proposes original solutions to both challenges. For single-step retrosynthesis, we use three different transformer models assembled as a triple transformer loop (TTL, Figure 4.1a). To broaden the scope of predicted disconnections on a given target molecule, the TTL explores multiple disconnections by using products with tagged reaction centers (P*) obtained by combining systematic, template-based and transformer-based tagging procedures. Compared to a transformer model trained on predicting SM + R directly from P*, the TTL achieves better round-trip accuracy for single-step retrosynthesis. For multistep retrosynthesis predictions, we integrate the TTL into a multistep tree search algorithm, here named TTLA, which selects

reaction sequences using a new route penalty score (RPScore), which for a route of N steps, is the product of a step-penalty score SPN, the confidence scores of each single-step retrosynthesis (CS), and the simplicity scores[102] of all SM along the route (Figure 4.1b). This selection scheme favours short sequences and is exemplified with the prediction of synthetic routes for recently approved drugs.

4.2 Methods

4.2.1 Dataset

The United States Patent and Trademark Office (USPTO) chemical reaction dataset version shared by Thakkar *et al.*[221] was used for the single-step evaluation as well as for training all transformer models in this study. The dataset is derived from the version of Lowe[68, 69] and has been filtered by these authors to include reactions with a single product (P) and between 2 and 10 starting materials (SM) and reagents (R) only. In the present work, we removed the tagging information, and reactions were remapped and retagged using our new SMILES tagging strategy and syntax. The same dataset split for training, validation, and test (90 : 5 : 5), as shared by Thakkar *et al.*[221] was used across all models resulting in 1 139 608, 63 672 and 63 454 reactions respectively.

4.2.2 Tagging reaction centers

Training the disconnection-aware retrosynthesis model requires a training dataset where all product SMILES have tagged atoms. To tag reacting atoms, we use the atom-mapping tool shared by Schwaller *et al.*[222] to identify the atoms having an environmental change during the reaction, defined as reacting atoms. These reacting atoms are then re-labelled with the atom mapping label "1" while all other atom mapping labels are removed, as described by Byekwaso *et al.*.[223] We then replace the atom tagging syntax by its unmapped SMILES notation, *e.g.* replacing "[C:1]" with "C", and append the atom with another separated tagging token ("!") using RDkit.[30] his modification allows to maintain an invariant SMILES token usage independent of the neighbouring hydrogen count or stereochemistry.

4.2.3 Single-step disconnection aware retrosynthesis (T1)

Being able to identify the reaction center of a given reaction, we apply our reaction tagging algorithm on USPTO to obtain a retrosynthesis-tagged training dataset. We remove reagents, catalysts, and solvents, which are identified as the unmapped species in atom-mapped reactions and train the retrosynthesis model to predict the starting materials given as input the tagged product. We use
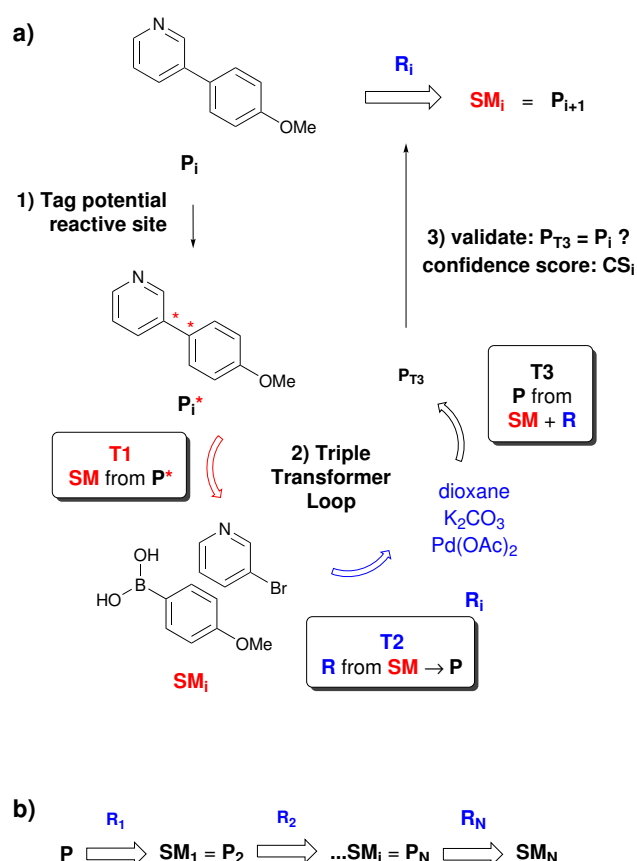
$$RPScore = SP^{N-1} \times \prod_{i=1}^{N} \left(CS_i \times \prod_{\forall \, mol \, \in \, SM_i} Simplicity(mol) \right)$$

$$\text{with } Simplicity(mol) = \begin{cases} 1 & \text{if } mol \in Commercial\_Database, \\ 1 - \dfrac{SCScore(mol) - 1}{4} & \text{otherwise.} \end{cases}$$

Figure 4.1: **Multistep retrosynthesis using TTLA. (a)** Single-step retrosynthesis. At step i, each potential reactive site in P_i is identified systematically, using templates or a tagging transformer, and labelled to produce $P_i{}^*$. Transformer T1 is applied to $P_i{}^*$ to predict SM_i (one or more starting materials), transformer T2 is applied to the top-scoring $SM_i \rightarrow P_i$ to predict R_i (one or more reagents), and finally transformer T3 is applied to the top-scoring $SM_i + R_i$ to produce P_{T3}. The prediction is finally validated if $P_{T3} = P_i$ with confidence score CS_i of T3. Each molecule in the SM_i set is then used as product P_{i+1} for the next iteration. The route branches out if SM_i contains multiple molecules. **(b)** TTLA sequence and route penalty scoring. All molecules in the SM_i set of each step are used in the RPScore calculation of a linear sequence. See text for details.

the transformer architecture[97] and train it using the OpenNMT[190, 213] library with standard previously-reported hyperparameters for this type of task.[99]

4.2.4 Automatic tagging of potentially reactive atoms

We use three complementary methods to maximize the tagging possibilities while maintaining a reasonable number of predictions. First, we systematically tag all possible single atoms, pairs of directly connected atoms, and triplets of adjacent atoms (chain or three-membered ring). Secondly, we use templates for tagging the reactive sets of atoms corresponding to the conditional substructure with a variable radius (typically from 1 to 3). Templates occurring more than once and having between 1 and 10 reactive atoms were identified by analyzing the original USPTO dataset. A given template can contain multiple disconnected sets of reactive atoms. Finally, the transformer model AutoTag reported by Thakkar *et al.*[221] was trained from untagged SMILES to the corresponding tagged molecule to provide additional tagging examples.

4.2.5 Reagent prediction (T2)

Transformer T2 is trained from the untagged USPTO training set to identify reagents (R) from the combination of SM and P using the same hyperparameters as for T1. Note that R often includes actual reagents and solvents.

4.2.6 Forward validation (T3)

The third model of the triple-transformer loop is a forward reaction prediction model trained with untagged reactions (molecular transformer).[99] We give this forward validation model the predicted SM_i (from T1) and the predicted R_i (from T2) as input separated by the ">" token. If T3 predicts the correct P_i as its top-1 prediction, those SM_i and R_i are stored for the tree search. The confidence score CS_i for the T3 prediction is used as confidence score for the reaction. T3 serves to filter down a large number of predictions to retain feasible reactions only.

4.2.7 Single-step TTL tagging strategies study

The performance of individual tagging methods was studied on 500 molecules randomly selected from the USPTO test set for single-step TTL retrosynthesis to which we varied the three strategies over various parameters, changing the template radius from 1 to 3 and the transformer tagging (AutoTag) beam size from 1 to 1000.

4.2.8 ROUTE PENALTY SCORE (RPSCORE)

The RPScore is computed for each predicted linear retrosynthetic sequence of N steps leading from the final product P to starting materials SM_N (Figure 4.1b). To reduce the score of long sequences, we introduce a step penalty SP, with $0 < SP \leq 1$, extended to SP^N for a sequence of N steps. The RPScore is the product of SP^N with the product of all confidence scores CS_i (from the T3 prediction) for each individual step and the Simplicity(mol) for all intermediates along the sequence of N steps. By default, the penalty value SP is set to 0.8, but this could be adapted for every search in the configuration file of the multistep exploration. Simplicity(mol)[102] ranges from 0 for complex to 1 for simple molecules and is derived from the molecular synthetic complexity score (SCScore, ranging from 1 to 5) which describes molecular complexity taking synthetic accessibility into account.[103] Here, we assign a value of 1 if the molecule occurs in the BB set of commercial starting materials. In contrast to Schwaller *et al.*,[102] we exclude reagents R_i from the Simplicity calculation to avoid penalizing steps that use reagents with low calculated Simplicity, which is rarely a measure of their availability or ease of use.

4.2.9 MULTISTEP EXPLORATION STRATEGY

We use a Heuristic Best-First Tree Search algorithm with beam search and iterative expansion to explore retrosynthetic routes as similarly reported for transformer-based retrosynthesis.[102] Once predictions of an iteration are complete, the tree search updates and lists all possible routes, and computes the RPScore. Unsolved routes are sorted by decreasing RPScore. The top 20 unsolved routes, which lead to starting materials absent from the selected set of commercially available building blocks, are selected for expansion by defining them as products P_i and new SM_i are predicted by applying a single-step retrosynthesis using TTL. The resulting set of predicted single-step retrosynthesis is updated back to the tree wherever those SMs were present. The tree is updated for the next iteration. The process stops when a chosen minimum number of solved routes or a maximum number of iterations has been reached.

4.2.10 BUILDING BLOCK (BB) SET

We combined MolPort (https://www.molport.com) and Enamine (https://www.enamine.net) databases to build our database of 534 058 commercially available compounds as the building block (BB) set.

4.3 RESULTS AND DISCUSSION

4.3.1 TRAINING TRANSFORMER T1 FOR SINGLE-STEP RETROSYNTHESIS

Initially, we use the atom-mapping transformer[222] information to annotate reacting atoms in all products P in the training data, which results in a training dataset containing labelled P*. Our code is inspired by the recent report by Byekwaso *et al.*,[223] however with a slightly simplified syntax for tagged atoms. Using the tagged P* data, we then train a transformer model T1 to predict SM from P*, a task which is simpler than predicting both SM and R from P* as done by Byekwaso *et al.*[223]

4.3.2 TAGGING POTENTIAL REACTIVE SITES

To use T1 to predict possible SM_i from a given product P_i at step i, one must first tag potentially reacting atoms in P_i. We do this using complementary methods. First, we tag all single atoms as well as pairs and triplets of adjacent atoms systematically in P_i. Second, we systematically apply templates extracted from tagged P* in the USPTO dataset. These templates with various conditional radiuses (from 1 to 3) are substructures containing up to ten tagged atoms, not necessarily connected. Although the most frequent templates are those with two or three connected atoms, which are tags also obtained in the systematic procedure, the templates also include many tags with disconnected atom pairs and triplets as well as tags with four or more atoms, which are missing from the systematic tagging procedure (Figure 4.2a). As a third tagging option, we use the tagging approach recently reported by Thakkar *et al.*[221] where reacting atoms are identified using a tagging transformer, here named AutoTag, trained to learn the detailed context from the tagged dataset. The number of predicted tags (sorted by confidence score, called beam size) of AutoTag can be varied to generate a given number of possible tags to extend the retrosynthesis options.

   Analyzing the performance of the different tagging methods shows that less restrictive template radius or high AutoTag beam size both lead to an increased number of tagged atoms per molecule (Figure B.1) as well as a much higher number of generated tagged SMILES (Figure B.2) and significantly more single-step starting materials (Figures 4.2b and B.3), but also to a lower number of high confidence predictions (Figure B.4), indicating that most of the additionally obtained tags are less chemically meaningful (Figure B.5). Moreover, the tagging efficiency, evaluated by dividing the number of successful retrosynthetic steps obtained by the number of TTL rounds (number of tags), drops for high AutoTag beam sizes and low radius templates (Figure B.6). To obtain a good number of validated retrosynthetic steps at reasonable computing cost, we combine three strategies: the systematic tagging (1, 2 and 3 atoms), templates with a radius of 2, and the AutoTag transformer with a beam size of 50. A Venn diagram analysis of the number of unique retrosynthetic steps obtained shows that 17% of the steps (37.8% of high confidence steps) are predicted by all three
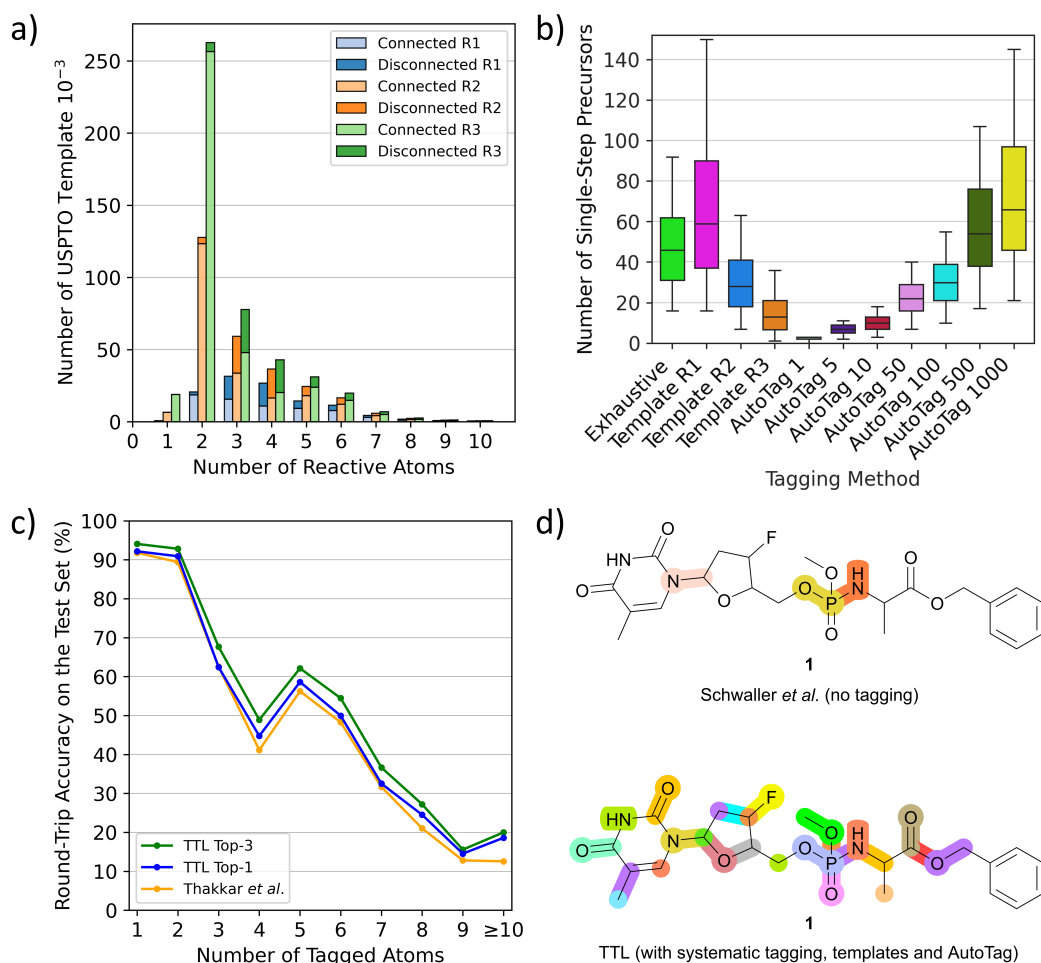
Figure 4.2: **TTL and automatic atom tagging. (a)** Distribution of the number of tagging templates extracted from USPTO depending on the number of atoms it tags, named "reactive atoms". Triple bar plot to show the differences between conditional radiuses beyond tagged atoms from 1 to 3 (R1 to R3). Bars are split into a light-coloured part representing the fraction of templates that tags bond-connected atoms and dark-coloured for disconnected atoms. **(b)** Number of validated single-step starting materials ("precursors") on the TTL generated depending on the automatic tagging strategy, tested over 500 molecules randomly selected from the TTL test set. **(c)** Round-trip accuracies of the TTL using the top-1 SM by T1 and the top-1 or top-3 R predicted by T2, compared to the disconnection-aware dual transformer of Thakkar *et al.*[221] **(d)** Highlighted disconnection sites of the antiviral molecule 1 using the untagged retrosynthesis and forward validation models of Schwaller *et al.*,[102] leading to four unique sets of starting materials among three reactive sites (top) and the TTL augmented by systematic tagging, template-based tagging (radius 2) and AutoTag (beam size 50) after forward validation leading to 231 unique sets of starting materials among 26 reactive sites (bottom).

methods, while 52.6% of the steps (25.5% of high confidence steps) are coming from only one of the three tagging methods, highlighting their complementarity (Figures B.7 and B.8).

4.3.3 TRIPLE TRANSFORMER LOOP (TTL)

To initiate a validated single-step retrosynthesis prediction for product P_i, we run T1 on all P_i^* obtained by the combined selected tagging procedures described above. The transformer outputs a series of possible SM_i, which are sorted in order of the T1 confidence score. For the top-B SM_i (beam size $B = 1$ or more), we then apply a second transformer (T2) trained to predict R from SM→P. For each SM_i, T2 outputs a series of possible R_i, from which we retain the top-B' (beam size $B' = 1$ or more). The TTL is completed with a forward validation[101] transformer (T3) trained to predict P from SM + R using the same training dataset used for T1 and T2. For all combinations of top SM_i predicted by T1 and top R_i predicted by T2, we finally use T3 to predict the most likely product P_{T3}. The TTL prediction is validated if the top-1 predicted P_{T3} is identical to the input product P_i (Figure 4.1a). The T3 confidence scores CS_i of the validated predictions $SM_i + R_i$ are used to select the best R_i if $B' > 1$, and to calculate the route penalty score (RPScore, see below).

4.3.4 PERFORMANCE EVALUATION

The performance of TTL can be compared with previous single-step retrosynthesis models at three different levels. First, transformer T1, which predicts SM from the tagged product P^*, can be compared with other single-step retrosynthesis models predicting SM from P, both transformer-based and template-based.[95, 96, 100, 102, 104, 160, 174, 219] While these models perform between 40% and 55% top-1 accuracy, our tagged T1 achieves 66% top-1 accuracy, which shows that tagging provides a significant advantage for this task.

Second, the performance of the TTL loop can be compared with the disconnection-aware retrosynthesis model of Thakkar *et al.*[221] in terms of single-step round-trip prediction accuracy from the tagged product P^*, which is the accuracy of predicting P from the SM + R initially predicted from P^*. TTL using only the top-1 predictions for T1 and T2 performs comparably to Thakkar's disconnection-aware retrosynthesis model (80.44% *vs.* 79.09% accuracy). The TTL performance increases to 83.04% when considering the top-1 prediction of T1 and the top-3 predictions of T2. Similar to the observation by Thakkar *et al.*,[221] we furthermore find that the prediction accuracy strongly decreases as a function of the number of tagged atoms (Figure 4.2c). Subsequently to our preprinted report, a separate study has investigated the performance of the reagent prediction transformer.[224]

Thirdly, one can compare the single-step round trip accuracy of TTL with that of the non-tagged retrosynthesis model of Schwaller *et al.*,[101, 102] who evaluated if a forward prediction model

predicted the correct product P from the SM + R predicted by their model from the non-tagged P. As discussed by Thakkar *et al.*,[221] the untagged transformer may sometimes choose a different and easier to predict disconnection than that recorded in the test set, and therefore performs slightly better (82.4% top-1 accuracy) than the tagged transformer, which is forced by tagging to apply the retrosynthesis of the test set. Here, we find that the top-1 round-trip prediction accuracy (P→P), obtained by applying our multiple tagging procedure followed by the TTL, reaches 99.9%, which means that our approach is almost always able to propose at least one forward-validated possible retrosynthetic step from any product molecule.

Furthermore, a critical feature of any single-step retrosynthesis model in view of multi-step retrosynthesis concerns the diversity of possible disconnections proposed. We find that this diversity is greatly enhanced by the multiple tagging approach. For instance, when tested on unseen molecules, the TTL combined multiple tagging provides validated disconnections at several possible reactive sites. By contrast, the baseline transformer, trained as reported by Schwaller *et al.*[102] to produce SM directly from P using the unannotated data for training, chooses fewer disconnection points, as exemplified here for the pro-nucleotide **1** (Figure 4.2d).[225]

4.3.5 MULTISTEP RETROSYNTHESIS

By integrating the single-step retrosynthesis TTL into a multistep tree search, we obtain a multistep retrosynthesis algorithm, here named TTLA. In each retrosynthesis iteration, TTLA runs the TTL exhaustively on all SM of the preceding iteration, newly defined as P, and ranks the routes to the newly predicted SM using a composite route penalty score RPScore (Figure 4.1b, see Methods for details).

When prioritizing multiple retrosynthesis options during the tree search, TTLA uses the RPScore to rank the different routes leading to the SM produced in the latest iteration of TTL, and only extends retrosynthesis on a small number (typically 20) of SM taken from the top RPScoring routes. Because each additional step imposes a penalty (usually P = 0.8), lengthy routes and unproductive loops involving protection/deprotection cycles of the same functional group are rapidly falling down the RPScore priority list, which leads the algorithm to explore alternative routes, so that short synthetic sequences are eventually prioritized even if their first retrosynthetic steps were initially not top scoring.

As commonly observed with CASP tools as well as with transformer models in general, the top-scoring outputs of TTLA must be inspected to identify relevant predictions. While the RPScore is used in the tree search, we find relevant routes by inspecting both the top-RPScoring route and the top-CScoring routes (CScore(route) = the product of CSi for all steps) in the TTLA output, as discussed below with examples.

TTLA is exemplified here for predicting the synthesis of two drug molecules approved in 2020, namely fostemsavir (**2**, Figure 4.3), a prodrug which upon phosphatase cleavage releases the antiretroviral agent temsavir as HIV entry inhibitor,[226] and ozanimod (**10**, Figure 4.4), a sphingosine-1-phosphate receptor antagonist used as an immunomodulatory agent to treat multiple sclerosis.[227] The commercial process for both drugs was recently reviewed.[228] None of the synthetic steps involved in these two processes occur in the USPTO dataset used for training TTL, making them a good test case for TTLA. For these examples, we challenged TTLA to predict synthetic routes starting from a list of 534 058 commercially available BB.

The reported commercial process for the antiviral drug fostemsavir (**2**, Figure 4.3a, details in Figure B.9) is a linear sequence involving the sequential *C*-acylation of pyrrolopyridine **3** with oxalyl monochloride **4a** (step a) and benzoylpiperazine (**5a**, step b), followed by coupling of with triazole **6** (step c), *N*-alkylation of the pyrrole with the protected chloromethylphosphate **7a** (step d), and finally deprotection of the *tert*-butyl ester protecting groups (step e).

When challenged with **2**, TTLA proposes many possible routes from similar starting materials as the commercial process, but in a different order. The highest RPScoing route is a linear sequence starting from the double *C*- and *N*-alkylation of oxalyl chloride (**4b**) with pyrrolopyridine (**3**) and 1-boc-piperazine (**5b**) in one pot (step a', Figure 4.3b, details in Figure B.10). The aryl bromide of the resulting intermediate is then substituted with triazole **6** (step b'), and its pyrrole NH group is alkylated with **tert**-butyl chloromethyl phosphate **7a**, similarly to the commercial route (step c'). In the final step, the phosphate and the piperazine groups are deprotected with acid, followed by benzoylation of the free piperazine with benzoylchloride (**8b**) to form fostemsavir **2** (step d').

On the other hand, the highest CScoring route is a convergent sequence starting with alkylation of triazole **6** with pyrrolopyridine **3** on the one hand (step a", Figure 4.3c, details in Figure B.11), and the preparation of the Weinreb amide **9** from boc-oxalylpiperazine **5c** and *N,O*-dimethylhydroxylamine **4c** on the other hand (step b"). The resulting intermediates are then coupled (step c"), and the product is *N*-alkylated on the pyrrole nitrogen with benzyl-protected chloromethyl phosphate **7c** (step d"). Deprotection of the piperazine group allows the acylation with benzoylchloride (**8b**, step e"). Reductive deprotection of the benzyl phosphate esters finally gives the product **2** (step f").

In the second example, the drug ozanimod **10** is synthesized commercially in a convergent sequence of 7 steps from ketone **11a** and benzoic acid **14a** (Figure 4.4a, details in Figure B.12). After initial protection of ketone **11a** as an acetal (step a), its nitrile group is reacted with hydroxylamine **13a** to form the N-hydroxyamidine intermediate **12a** (step b). In parallel, benzoic acid **14a** is activated to the corresponding benzoyl imidazole **15** (step c). Intermediates **12a** and **15** are then condensed to form the oxazole ring (step d). The acetal group of the resulting intermediate is then
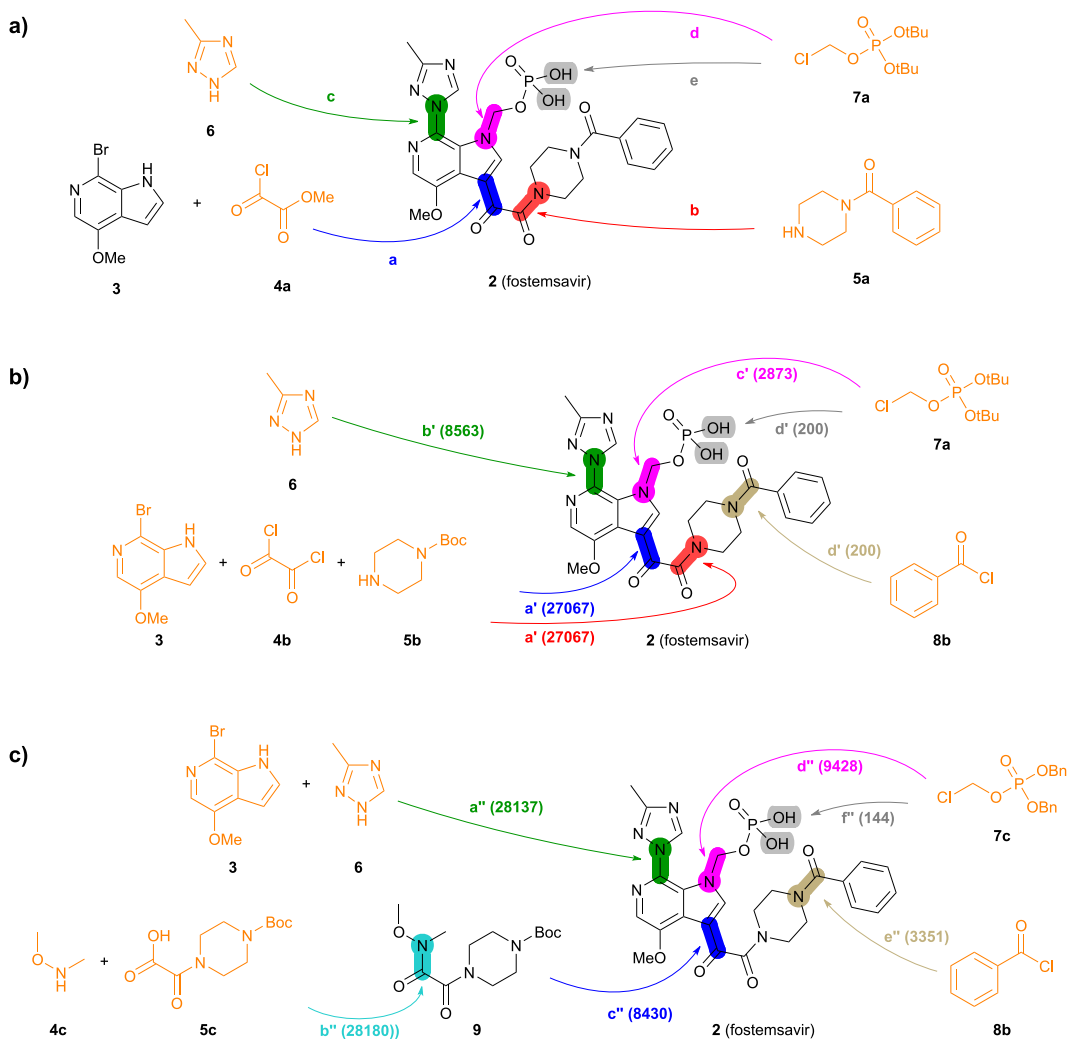
Figure 4.3: **Summary of reported and TTLA predicted routes for fostemsavir 2.** Bonds formed in each step are highlighted in colour. Numbers in parenthesis correspond to the order in which the multistep tree search prioritized predictions. The full retrosynthesis routes are drawn out in the supporting information Figures B.9, B.10 and B.11. **(a)** Commercial process. Reported reagents: a) AlCl₃, Bu₄NHSO₄, CH₂Cl₂, then KOH, then H₃PO₄; b) Ph₂POCl, NMM, NMP; c) KOH, CuI, then KOH, EtOH, LiI; d) Et₄NI, K₂CO₃, CH₃CN/H₂O; e) AcOH, H₂O. **(b)** Highest TTLA RPScoring route. Predicted reagents: a') Et₃N, CH₂Cl₂; b') K₂CO₃, CuI, toluene; c') K₂CO₃, DMF; d') HCl, *N,N*-Diisopropylethylamine, H₂O, dioxane. **(c)** Highest TTLA CScoring route. Predicted reagents: a'') (2S)-pyrrolidine-2-carboxylic acid, K₂CO₃, CuI, EtOAc, DMSO; b'') no reagent predicted; c'') *n*-BuLi, THF; d'') K₂CO₃, DMF; e'') TFA, DMAP, CH₂Cl₂, f'') Pd, EtOH.

deprotected and condensed with ethanolamine (**16a**) to the corresponding imine, which is reduced enantioselectively using a chiral ruthenium catalyst to form **10** (step e).

Many of the high-scoring routes identified with TTLA are extremely short sequences starting with commercially available close analogs of the drug and were removed from the list of top-scoring routes. Interestingly, TTLA also proposes routes that resemble the commercial process but start from chiral starting materials such as aminoindanes **11b** and **11c**, which avoids the enantioselective reaction used for the commercial process. For example, the best RPScoring route is a linear synthesis from **11b** starting with the removal of the Boc and TBS protecting groups of the ethanolamine side chain and conversion of the cyano group to the corresponding N-hydroxyamidine by reaction with TBS-hydroxylamine (**13b**) to form intermediate **12b** (steps a' and b', Figure 4.4b, details in Figure B.13). The third and final step of this short sequence is the condensation of N-hydroxyamidine **12b** with cyanobenzoate **14b** yielding ozanimod **10** (step c').

The best CScoring route is a somewhat longer linear sequence employing the same condensation of **12b** and **14b** as the final step (step e", Figure 4.4c, details in Figure B.14). In this proposed sequence however, intermediate **12b** requires four steps from the chiral aminobromoindane **11c** as follows. First, the cyano group is installed by reaction of the aryl bromide with copper cyanide (step a"). Second, the primary amine reacts with ethylene oxide **16c** to form the N-hydroxyethyl side chain (step b"). Third, the cyano group introduced in step a" reacts with ethanol (**18**) to form an ethyl imidate intermediate (step c"), which further reacts with ethanolamine (**13a**) in a fourth step to form the N-hydroxyamidine group in **12b** (step d").

Analyzing the details of the TTLA collective output shows that, although TTLA did not formulate routes identical to the commercial processes, the set of commercial starting materials used by TTLA are very similar to those used in the reported commercial processes for both drugs (Figure B.15 and B.16). In fact, all starting materials used in the commercial process for fostemsavir are present in the set for this drug.

In terms of individual reaction steps, we find that TTLA explores a large number of single reactions to arrive at the top-scoring short routes proposed in the above retrosynthesis. In the case of fostemsavir, the key retrosynthetic C- and N-acylation of the oxalyl starting material is discovered after 27 067 single predicted steps (Figure 4.3b, step a'), probably because this step is rather complex and unusual. In the case of ozanimod, TTLA performed 7594 individual single-step predictions to arrive at the proposed retrosyntheses, with the best scoring route being discovered after 2700 iterations. Interestingly, the formation of the oxadiazole ring is discovered already at iteration 8 (Figure 4.4b, step c'). It should be noted that the order of iterations and therefore the number of attempts necessary to identify high-scoring routes depends on the scoring function used to prioritize node expansion, here the RPScore, which takes the simplicity and number of steps into account.

The output of TTLA can be visualized by representing the collective predicted single steps in a TMAP[186] computed using the differential reaction fingerprint (DRFP)[229] as a similarity measure. As illustrated for ozanimod, colour-coding by step iteration number indicates that TTLA explores a broad diversity of steps directly from the beginning of the retrosynthesis exploration, which we attribute to our diverse reaction center tagging approach used (Figure 4.5a). This diversity is also visible when colour-coding all steps involving the final product, corresponding to the initial retrosynthesis, which are broadly distributed on the map (Figure 4.5b). A similar pattern is visible in the TMAP of the predicted single steps for fostemsavir (Figure B.17).

4.3.6 COMPARING TTLA WITH OTHER RETROSYNTHESIS TOOLS

Previous retrosynthesis tools, template-based or transformer-based, predict starting material from products by applying the most probable retrosynthetic operation according to a training set. Here we combined exhaustive and template-based methods to label many potential reactive sites, which lead us to test many possible disconnections (Figure 4.2d). These potential reactive sites were then challenged with the TTL, which produced detailed predictions including starting materials and reagents. In the examples discussed above TTLA identified short routes comparable to the reported processes, which were all examples of optimized production routes.

By comparison, a currently available version of AiZynthFinder (v3.7.0),[88] a templated-based retrosynthesis tool, fails to propose a synthesis for fostemsavir due to its inability to find a synthesis for a bis-tert-butyl phosphate starting material (Figure B.18). AizynthFinder furthermore proposes a short route similar to TTLA for ozanimod, although including somewhat less realistic steps, for example, an alkylation of a primary amine with 2-bromoethyl acetate which would probably rather lead to acetyl transfer, and no indication of reagents (Figure B.19). On the other hand, the online portal of IBM RXN for chemistry,[230] which uses a transformer model, predicts essentially the same route as TTLA for fostemsavir (Figure B.20). For ozanimod however, this tool settles on an eight-step route which, although containing realistic steps, is simply much longer than the commercial process or the route proposed by TTLA (Figure B.21). For both of these retrosynthesis tools, whether the routes are part of their training sets is not known.

To statistically evaluate our TTLA, we selected target molecules from the retrosynthesis benchmark dataset shared by Genheden *et al.* which were absent from our training dataset.[231] Due to the high computing time of our method, a random subset of 240 target molecules was selected. Solved routes involving reaction steps present in our training dataset were removed from the evaluation. TTLA proposed retrosyntheses to commercially available starting materials for 97.5% of the target molecules, which is comparable to the performance of other retrosynthetic tools reported in the original paper.46 Selected examples are shown in Figures B.22, B.23, B.24, B.28, B.26, B.27, B.28, B.29, B.30 and B.31.
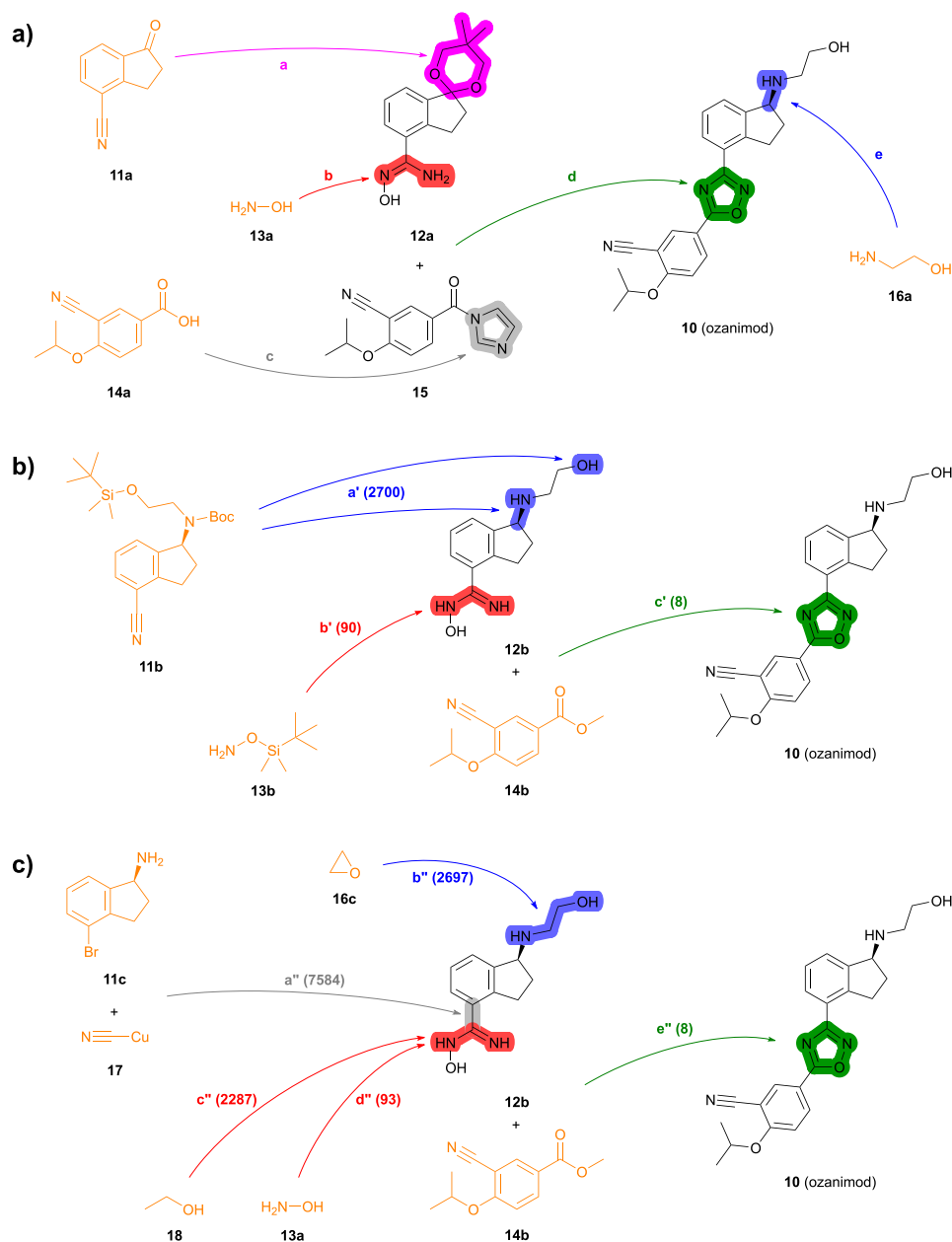
Figure 4.4: **Summary of reported and TTLA predicted routes for ozanimod 10.** Bonds formed in each step are highlighted in colour. Numbers in parenthesis correspond to the order in which the multistep tree search prioritized predictions. The full retrosynthesis routes are drawn out in the supporting information Figures B.12, B.13 and B.14. **(a)** Commercial process. Reported reagents: a) HC(Ome)$_3$, *p*-TsOH, PhCH$_3$; b) NH$_2$OH.HCl, Et$_3$N; c) carbonyl diimidazole; d) NaOH; e) i) *p*-TsOH, acetone, ii) NH$_2$CH$_2$CH$_2$OH, *p*-TsOH, PhCH$_3$, iii) chiral Ru-complex, Et$_3$N/HCO$_2$H. **(b)** Highest TTLA RPScoring route. Predicted reagents: a') HCl, dioxane; b') ZnCl$_2$, AcOEt, toluene; c') HCl, *t*-BuOK, THF. **(c)** Highest TTLA CScoring route. Predicted reagents: a") 1-Methylpyrrolidin-2-one; b") no reagent predicted; c") HCl, Et$_2$O; d") HCl, NaHCO$_3$, EtOH; e") HCl, *t*-BuOK, THF.
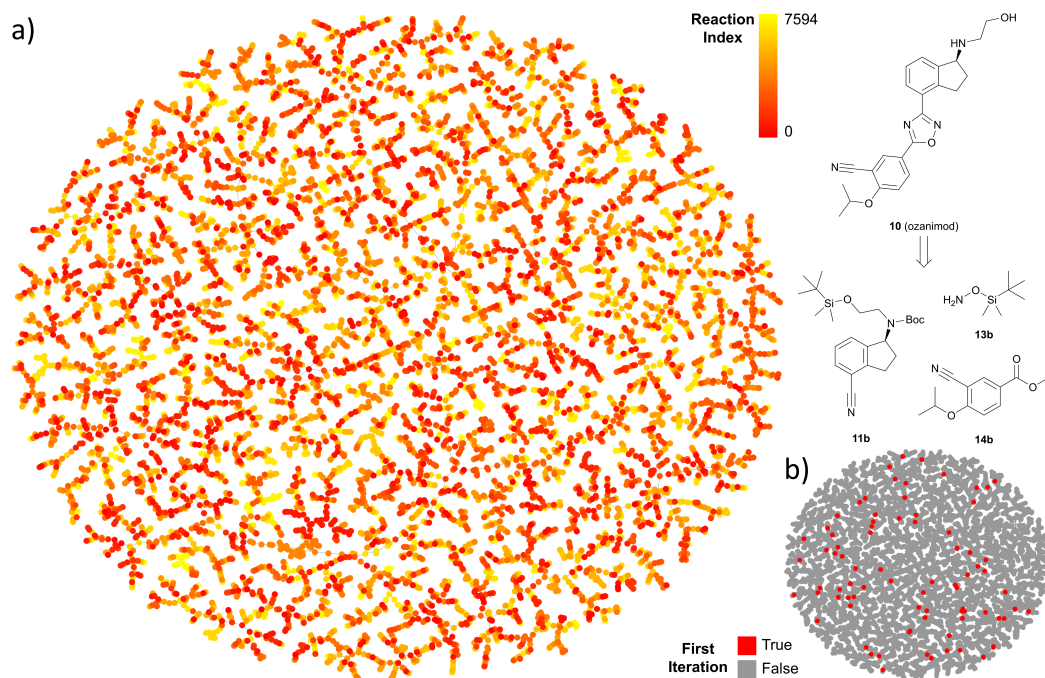
Figure 4.5: **TMAP representation of iterated predictions for the multistep search of ozanimod.** **(a)** Predicted reactions from the target molecule (low indexes) to end nodes. **(b)** Highlighted first iteration of the TTLA search. Interactive map available at https://tm.gdb.tools/TTLA/ozanimod.

4.4 Conclusion

In summary, our data shows that a triple transformer loop (TTL) operating on products with tagged reactive atoms achieves efficient single-step retrosynthesis predictions. TTL was integrated into a tree-exploration strategy using a route penalty scoring scheme to form the multistep retrosynthesis tool TTLA, which can predict short synthetic routes for drug molecules. Since our approach uses transformer models, it should be possible to specialize TTLA for specific reaction classes by transfer learning similar to transformer models for forward prediction.[180] Furthermore, predicting SM from P and R from SM + P separately might be potentially adapted to reactions with more complex reagents such as enzymes[164, 165, 232] and help expand the scope of CASP systems. It should however be noted that the use of multiple transformer models and the detailed analysis of many possible disconnections renders our approach relatively slow, requiring up to several hours of computing time for a full retrosynthetic analysis. Efficiency increases might be possible in the future by fine-tuning the selection of potential disconnections and improving the tree search.

4.5 Data Availability

Code, models and instructions to compute multistep retrosynthesis as well as the code to tag reactive sites can be found on our GitHub repository:

https://github.com/reymond-group/MultiStepRetrosynthesisTTL.

The original USPTO dataset can be found at https://doi.org/10.6084/m9.figshare.5104873.v1. The derived version of USPTO of Thakkar *et al.* could be found in their Zenodo repository.[221, 233]

5   Triple Transformer Loops for Chemoenzymatic Multistep Retrosynthesis

Biocatalysis offers opportunities for enhanced selectivity, efficiency, and greener processes in synthetic chemistry. Therefore, it is important to upgrade computer-assisted synthesis planning (CASP) tools such as to transition towards sustainable chemistry by proposing alternative and greener catalytic methods to chemists. Herein, we describe the expansion of our previously reported open-source multistep retrosynthesis algorithm. We added an enzymatic triple transformer loop (TTL) variant composed of models for retrosynthesis (T1), enzyme prediction as textual descriptions (T2), and forward validation (T3). The enzymatic dataset (ENZR) was extracted from Reaxys, and models were trained using multitask transfer learning on patent reactions (USPTO) with instruction tokens to avoid task ambiguity. The multistep algorithm leverages a heuristic best-first tree search operating both TTL frameworks in parallel, enabling both organic and enzyme catalytic routes, competing by the route penalty score (RPScore). We show that our dual catalysis CASP tool proposes reasonable solutions to drug-like molecules, providing a good starting point for synthesis design.

This chapter contains unpublished work.

5.1   Introduction

Chemical retrosynthetic is a systematic methodology allowing chemists to work in reverse, deconstructing a target molecule into commercially available starting materials and outlining a synthesis pathway.[1] This approach has been harnessed by computer-assisted synthesis planning (CASP) tools exploring the reactivity space automatically.[2] While the field has seen attempts to develop tools employing diverse expert system strategies with limited success,[3, 38, 41, 47] it is only recently that commercial solutions, such as Chematica/Synthia™, have reached a level of effectiveness suitable for use in the industry.[71] More recently, the field of CASP for chemocatalysis has seen

strategies exploiting rules-based approaches combined with decision-making neural networks,[79, 81, 82, 83, 84, 85, 88, 179] but also fully data-driven deep-learning methods.[95, 96, 97, 98, 99, 100, 104, 215, 216, 234, 235, 236]

For instance, our previously reported triple transformer loops (TTL) outperformed single-step transformer-based retrosynthesis models through the integration of tagging strategies within a validation loop. The use of a disconnection-aware model not only prevents dataset bias but also allows tagging at multiple locations for diversity-oriented retrosynthesis, exceeding 99% round-trip accuracy.[237]

Biocatalysis harnesses nature's inherent chemical processes, utilizing enzymes to catalyse various chemical transformations. It emerges as a powerful green chemistry tool, particularly for enantio-selectivity, often challenging with chemocatalysis.[5] Recently, directed enzyme evolution has revolutionized the field by allowing the improvement of an enzyme for a custom substrate,[121, 123] expanding the possibilities of enzymatic transformations and the generalization of biocatalysis in industry.[130, 238, 239, 240, 241] However, current CASP tools still lack full integration of biocatalysis, resulting in missed opportunities.

Enzymatic reaction collections including BRENDA, KEGG, MetaCyc, Rhea, PathBank, MetaNetX, or EzCatDB consist of literature reactions, either text-mined or manually curated, with a primary focus on metabolic pathways.[139, 141, 143, 145, 148, 149, 152] However, capturing enzymatic trans-formations on small molecules is challenging for these databases due to their modified nature and different applicability. Reaxys, with its diverse content, has proven more effective for the machine learning of enzymatic transformations.[232]

The incorporation of biocatalysis into computer synthesis planning has attracted attention, with emerging examples utilizing template-based or machine-learning approaches[156, 160, 164, 165, 166, 167, 242] However, proposed solutions either rely on metabolic data, which is hardly suitable for small molecule synthesis, or exclusively focus on biocatalytic transformations, making them impractical for small molecule total synthesis. Additionally, rule-based approaches often fail to fully capture enzyme-substrate specificity, an aspect where machine learning models have demonstrated superior performance.[160, 232] Notably, none of the above-mentioned works use Reaxys experimental data.

In this study, we tackle the integration of biocatalysis together with organic catalysis using attention-based Transformers, which are presumed to better capture overall substrate structures than templates. We introduce an independent triple transformer loop (TTL) comprising (T1) an enzymatic disconnection-aware retrosynthesis model predicting starting materials (SM), (T2) a model predicting textual descriptions of enzymes given a reaction, and (T3) a forward validation model predicting the product P of an enzymatic reaction given the SM and enzyme description. The resulting ENZR-TTL integrates with our previously reported chemocatalytic TTL, operating
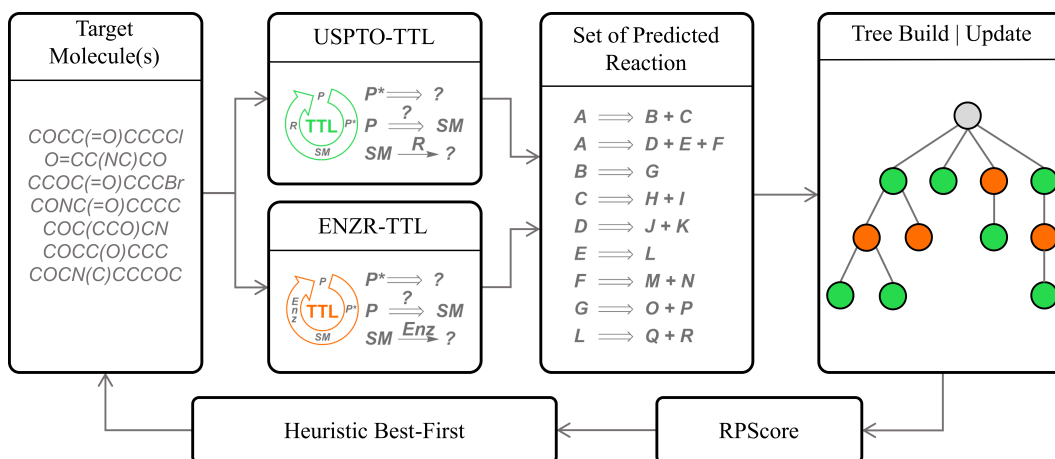
Figure 5.1: **Concept of the dual catalytic multistep search** operating organic (USPTO-TTL) and enzymatic (ENZR-TTL) catalysis in parallel

in parallel, see Figure 5.1. This setup enables the tree builder to select any catalytic type of reaction, competing based on the route penalty score (RPScore) based on SM simplicity scores and confidence scores. It ensures an efficient selection of the appropriate catalytic system, as exemplified in this chapter.

5.2 Methods

5.2.1 Chemocatalysis dataset

The same United States Patent and Trademark Office (USPTO) chemical reaction dataset as in our previous report was used.[237] It is a version curated by Thakkar *et al.*[221] derived from the data mining work of Lowe[68, 69]

5.2.2 Chemocatalysis Triple Transformer Loop models (USPTO-TTL)

The chemocatalytic models trained on the USPTO dataset are identical as in our previous study and available on Zenodo [237, 243] and herein named USPTO-TTL. AutoTag is a tagging model predicting tagged product P* from the target product (P). T1 is a disconnection-aware retrosynthesis model predicting starting materials SM from the target tagged product P*. T2 is a reaction condition model predicting reagent R including catalyst and solvent from the reaction SM→P. T3 is a forward validation model predicting P from SM + R.[101]

5.2.3 Biocatalysis dataset: extraction from Reaxys

The biocatalysis/enzymatic reaction dataset, herein named ENZR was extracted from Reaxys using the API accessible under a commercial license.[64] We first isolated reactions labelled as "enzymatic reaction" in the "other conditions" field ("RXD.COND"). Next, we compiled a list of reagents, catalysts, and solvents typically associated with enzymatic reactions. This involved identifying components with the "ase" suffix in the text fields "RXD.RGT," "RXD.CAT," and "RXD.SOL,". Additionally, we manually selected keywords corresponding to enzymatic transformations, such as "P450," "NADP," "CAL-B," "flavin mononucleotide," and others, from the most frequently occurring reagents and catalysts in the initial data retrieval. Finally, we queried these enzymatic components individually in the Reaxys database and retrieved the associated reactions. This process resulted in a raw dataset consisting of 107 865 biocatalytic reactions.

5.2.4 Biocatalysis dataset: cleaning

The process of cleaning the ENZR dataset involved several steps, wherein the RDKit library was used across various stages.[30] Initially, multistep reactions and those lacking any reactant or product were excluded, leaving 95 389 reactions. Next, reactions were mapped using RxnMapper,[19] for which 1333 reactions failed and were removed. Reactions with unspecified atomic symbols ("*") were also removed. Unmapped reactant molecules were removed for each reaction. A significant number of reactions (32 527) with more than one product were removed. The remaining reactions were tagged its reactive atoms as described in Kreutter *et al.*[237] and reactions with no or more than 10 tagged atoms were removed. This cleaning process results in a final enzymatic dataset of 57 176 unique reactions SMILES[25, 26] associated with textual descriptions of each reagent, catalyst, and solvent.

5.2.5 Biocatalysis Triple Transformer Loop models (ENZR-TTL)

Biocatalysis/Enzymatic models (ENZR-TTL) were trained using the ENZR dataset through multitask transfer learning, similar to our previous Enzymatic Transformer model with identical training hyperparameters.[232] The dataset split was done such as identical products from different reactions belonging to the same set. The split ratio 90:5:5 was applied similarly as in the USPTO dataset resulting in 51 459:2859:2858 reactions in the training, validation, and test set respectively.

While ENZR-AutoTag and ENZR-T1 maintain a SMILES-to-SMILES correspondence, like their USPTO counterparts, a notable difference exists between the second transformer (T2) of the USPTO-TTL and the ENZR-TTL. Specifically, the ENZR-T2 functions as a SMILES-to-text model, predicting textual descriptions of enzymes, while the USPTO-T2 predicts reagents in the

form of SMILES. Consequently, the forward validation model ENZR-T3 is a mixed SMILES-and-text-to-SMILES model.[232]

Additionally, during the multitask transfer learning process for both ENZ-AutoTag and all ENZ-TTL models, we inserted additional instruction tokens: either "ENZYME" for the ENZR dataset or "USPTO" for the USPTO dataset, at both the beginning and end of the SMILES inputs.

5.2.6 Disconnection-aware automatic tagging strategy

In alignment with our previous study,[237] the USPTO-TTL employs a combination of three tagging strategies: (1) a systematic tagging procedure, tagging 1 to 3 neighbouring atoms,(2) tagging templates of reactive sites with a conditional structure radius of 2 atoms, and (3) the AutoTag Transformer model with a beam size of 50.

The ENZR-TTL uses a specific tagging strategy combining only an AutoTag model and templates, excluding the systematic tagging approach. The dedicated ENZR-AutoTag was trained from the ENZR dataset and USPTO by multitask transfer learning. ENZR reactive site templates were extracted from ENZR exclusively with a radius of 2 atoms.

5.2.7 Dual biocatalytic and chemocatalytic multistep tree search algorithm

In parallel to the existing single-step USPTO-TTL, we added the ENZR-TTL which the multistep algorithm uses systematically and independently. The prediction outcomes of both TTLs are provided to the heuristic best-first tree search, elaborating routes mixing both types of catalytic strategies. The RPScore, based on molecular simplicity and confidence scores of T3 distinguishes which routes are the best to explore further.[103, 237]

Our previous report of the Enzymatic Transformer model, herein named ENZR-T3, demonstrated that a confidence score threshold was required to filter unreasonable enzymatic reactions. A similar evaluation using the round-trip evaluation of the ENZ-TTL was performed and a threshold of 90% confidence of ENZR-T3 was defined for considering ENZR-TTL predictions for multistep retrosynthesis search.

5.2.8 Building block (BB) set

We combined MolPort (https://www.molport.com) and Enamine (https://www.enamine.net) databases to build a database of 534 058 commercially available compounds as the building block (BB) set.

5.3 Results and Discussion

5.3.1 Enzymatic reaction dataset from Reaxys

Predicting enzymatic retrosynthesis reactions employing machine learning requires a dataset of enzymatic transformations. We decided to use Reaxys as a source of experimental reaction corpus to train all our Transformer models. The data collection was performed in two phases, first by extracting reactions labelled as enzymatic by Reaxys, followed by querying enzymatic reactants and catalysts to collect additional unlabelled enzymatic reactions. Only 38 173 reactions of the enzymatic dataset (ENZR) have the "enzymatic reaction" label from Reaxys, indicating a nearly 50% increase in the dataset size using our two phases strategy, resulting in an ENZR dataset of 57 176 enzymatic reactions with annotated enzyme textual descriptions.

5.3.2 ENZR for small molecule synthesis

Assessing the potential and applicability of the ENZR dataset for small molecule synthesis involves estimating the overlapping molecular structures in the molecular space with those found in traditional chemocatalytic reactions, represented by the USPTO dataset. Furthermore, a comparative analysis could be performed between our ENZR dataset and other notable biocatalytic reaction datasets. For instance, "ECREACT," a dataset meticulously compiled by Probst *et al.*[164] aggregates reactions from Rhea, BRENDA, PathBank, and MetaNetX.[139, 145, 148, 149] This compilation contains 62 222 reactions, each associated with its respective enzyme commission (EC) number. To visually facilitate such an analysis encompassing the three selected datasets, we plot a TMAP [186] employing the MinHashed atom-pair fingerprint (MAP4),[16] which was computed on 10 000 randomly selected molecules from each dataset, either starting materials or products of reactions (Figure 5.2a).

When analysing the TMAP, a clear distinction could be observed between ENZR and ECRE-ACT, mostly due to the overpopulation of lipid compounds in ECREACT, forming a cluster in the top left of the TMAP (blue). Moreover, only a small fraction of the ECREACT compounds overlap with USPTO, demonstrating its difficult integration for small molecule synthesis. On the other hand, ENZR compounds form smaller clusters which spread over the TMAP, showing a better overlap with USPTO.

A detailed analysis of the dataset molecules reveals key distinctions. Specifically, when examining the heavy atom count distribution, ENZR aligns closely to USPTO, in contrast to ECREACT, which predominantly features heavy moieties (Figure 5.2b). Furthermore, the fraction of aromatic atoms highlights that ENZR reassembles more USPTO than ECREACT (Figure 5.2c). Following a similar pattern, it is worth noting that 47.9% of ECREACT molecules contain a phosphate functional group, whereas only 0.5% of USPTO molecules include this functional group. In

a)
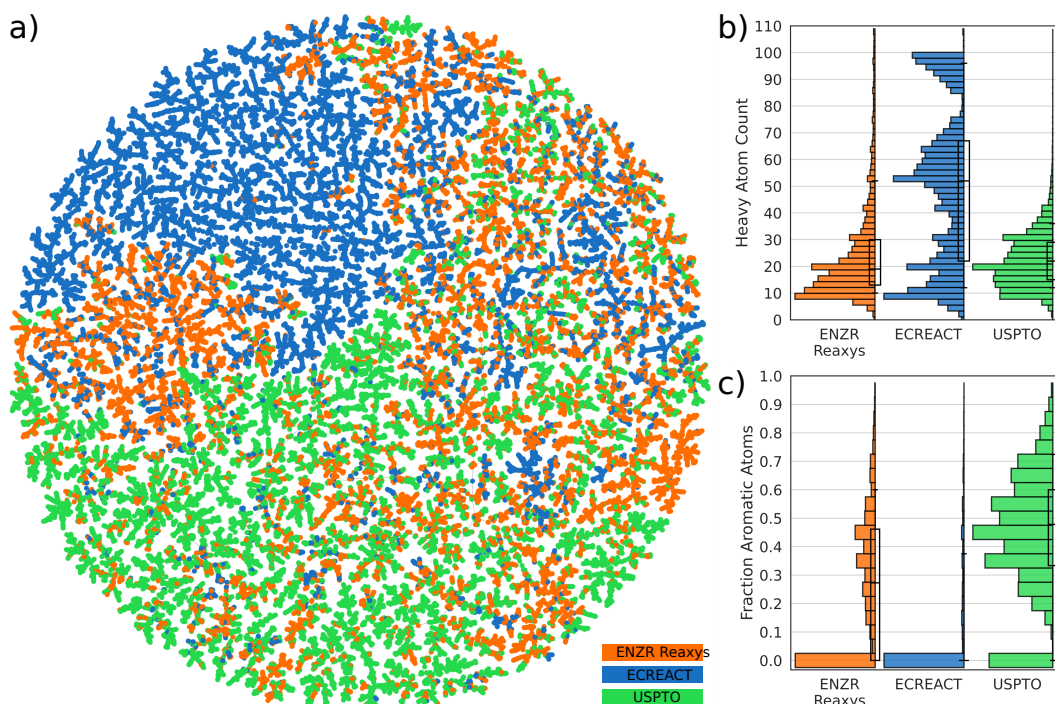
Figure 5.2: **(a)** TMAP of molecules (starting materials or products) present in our Reaxys ENZR, the ECRE-ACT enzymatic dataset of Probst et al.48 and USPTO computed with the MAP4 fingerprint. A selection of 10 000 molecules were randomly chosen from each reaction dataset. The interactive map is available at https://tm.gdb.tools/TTLA/EnzymeDB.html. **(b)** Number of heavy atoms distribution for all materials or products as function of the dataset. **(c)** Fraction of aromatic atoms distribution for all materials or products as function of the dataset.

comparison, our ENZR dataset comprises a substantial but lower fraction of 8.2% of molecules containing phosphates.

While the diverse nature of the datasets from which ECREACT is composed may be valuable for enzyme catalysis involving lipids and metabolism predictions, this analysis reveals that it has limited applicability for retrosynthesis. In contrast, our ENZR dataset, extracted from Reaxys, seems to be better suited for the synthesis of small organic molecules.

5.3.3 Training of models and instruction strategy for training ENZR-TTL models

Multitask transfer learning could encounter task ambiguity when using unbalanced dataset sizes, especially when similar inputs yield different outputs across various tasks. This issue emerges in the case of the enzyme prediction model ENZR-T2, as it predicts textual descriptions of enzymes capable of catalysing reactions from an input SMILES (SMILES-to-Text). However, this task significantly differs from the SMILES-to-SMILES reagent conditions prediction dataset of USPTO-
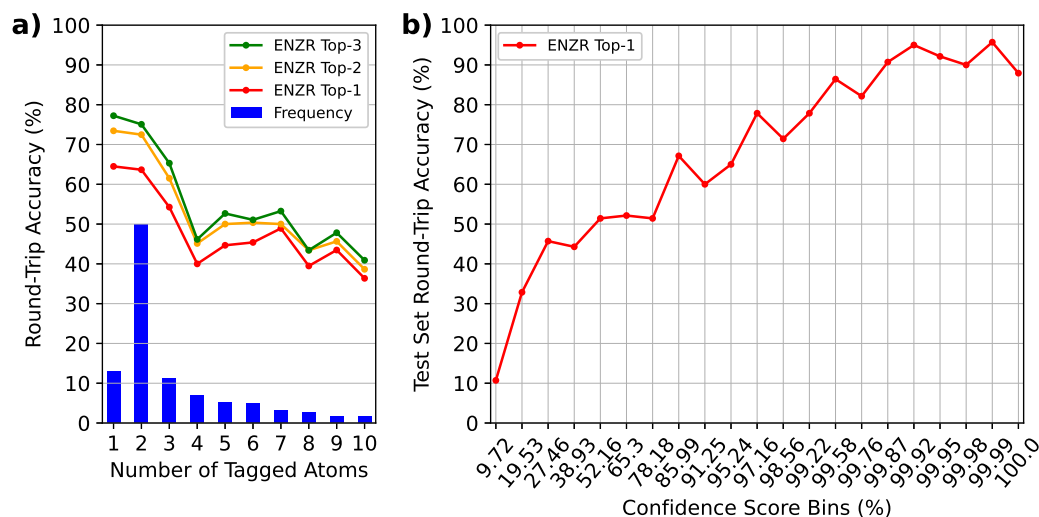
Figure 5.3: **(a)** Round-trip accuracy as function of the number of tagged atoms on the target molecules from the ENZR test set. The top-N represents the round-trip accuracy considering multiple examples of enzyme textual descriptions predicted by ENZR-T2. The blue bar plot shows the frequency fraction as function of the number of tagged atoms. **(b)** Round-trip accuracy as function of confidence scores bins of ENZR-T3. Bins were selected to equally distribute predictions.

T2 which ENZR-T2 leverage as a transfer learning dataset. The ambiguity arises when ENZR-T2 is employed aiming to predict an enzyme, but the model instead predicts reagent conditions in the form of SMILES. To prevent our enzymatic ENZR-T2 from predicting chemocatalytic SMILES, we labelled the datasets with "ENZYME" and "USPTO" for ENZR and USPTO datasets, respectively. These instructions ensure that the model outputs a textual description of an enzyme when the intention is to predict a biocatalytic step using the ENZR-TTL. In fact, these instruction tokens improved the fraction of textual enzyme description produced by ENZR-T2 from 85.3% to 99.7%.

5.3.4 SINGLE-STEP RETROSYNTHESIS PERFORMANCE OF THE ENZR-TTL

The ENZR-TTL could be benchmarked and evaluated in many aspects. For instance, the assessing the capability of ENZR-T1 to predict the expected starting materials (ground-truth accuracy) or execute disconnections at the anticipated tagged atoms (disconnection accuracy). However, a more crucial metric for multistep retrosynthesis is the ability of ENZR-TTL to propose feasible reactions. This can be evaluated through the round-trip evaluation,[101] achieving an overall 57.2% top-1 performance on the ENZR test set. A detailed examination reveals a decreasing accuracy as the number of tagged atoms increases, as illustrated in Figure 5.3a.

Examining the round-trip accuracy depending on the confidence score reveals a clear correlation, as shown in Figure 5.3b. This highlights that the confidence score of ENZR-T3 serves as a reliable metric for determining the realism of a predicted reaction, further stating its significance in the RPScore calculation.

5.3.5 Dual catalytic multistep retrosynthesis

Through the parallel integration of ENZR-TTL alongside the previously reported USPTO-TTL, we have developed a dual single-step retrosynthesis prediction system. The tree search algorithm utilizes both catalytic single-step TTLs, allowing it to selectively combine reactions and construct mixed-catalytic retrosynthesis routes, as illustrated in Figure 5.1.

To ensure the reliability of the enzymatic steps selected by TTLA, a confidence score filter of 90% was applied to ENZR-T3. This step aims to prevent the inclusion of highly effective enzymatic steps in terms of molecular simplification but with chemical infeasibility. Many unrealistic enzymatic reactions were initially predicted, and the simplicity score in the RPScore was compensating for the low confidence score of ENZR-T3.

The dual TTLA can be used in chemo-bio-catalytic mode, generating mixed synthesis routes, as exemplified in Figure 5.4 and Figure 5.5 with routes predicted without human intervention. Notably, these predicted routes do not have any single-step reaction present in any of the training datasets, but the enzymatic reactions shown were present in the test set, demonstrating the ability of the dual-catalytic multistep search to assign high scores to the most realistic routes.

For the example of compound **1**, the search stopped after exceeding the defined solved route target with 2518 different solved routes including a biocatalysis step. The best-scoring route incorporating one enzymatic step is a linear sequence of three steps. It starts with the condensation of vinylglycine **2** with acetyl chloride **3** and ethanol to form **4**. It further reacts with ethyl methylphosphinate **5** to form **6** which is finally enzymatically reacting with ethyl methionine **7** by alkaline phosphatase to form the product **1**.

Similarly, the TTLA completed 3050 different solved routes for compound **8**. The best RP-Scoring route starts with 2-hexenal **9** which is converted to the corresponding nitrile **10** which condensates with cyanoacetic acid **11** to form the dinitrile intermediate **12**. It then undergoes an enantioselective hydrolysis using a nitrilase, acting on a single nitrile resulting in the corresponding nitrile carboxylic acid **8** (Figure 5.5).

5.4 Conclusion

In summary, our work integrates biocatalysis in a computer-assisted synthesis planning (CASP) system, going towards greener and more sustainable chemistry. We achieved this by introducing a
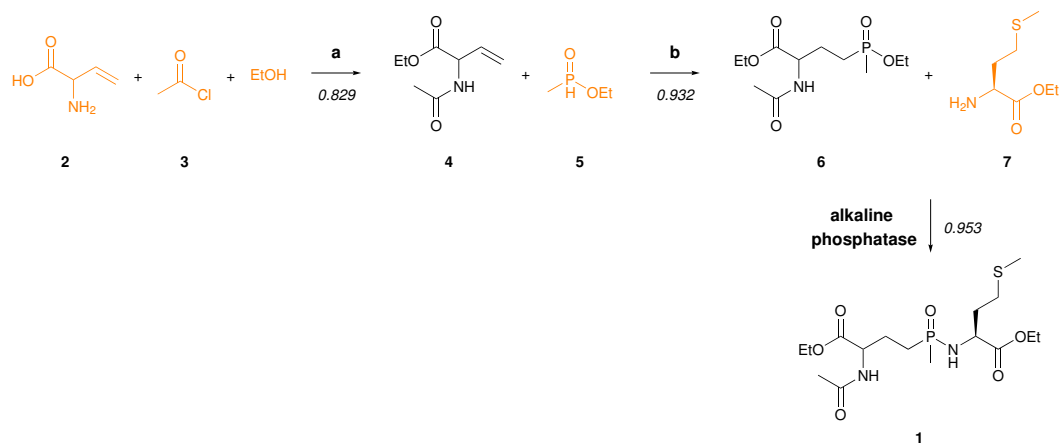
Figure 5.4: Best RPScoring retrosynthesis route of **1** having an enzymatic step predicted by the dual-catalytic TTLA. Forward prediction confidence scores are shown under synthesis arrows. Detailed reaction conditions: (a) HCl; DCM; Et₃N; (b): ammonium acetate; EtOH.
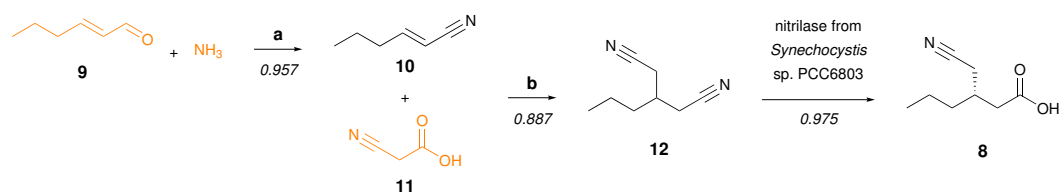
Figure 5.5: Best RPScoring retrosynthesis route of **8** having an enzymatic step predicted by the dual-catalytic TTLA. Forward prediction confidence scores are shown under synthesis arrows. Detailed reaction conditions: (a): THF; EtOAc; sodium carbonate; POCl₃, (b): ammonium acetate; DMF.

dual-catalytic multistep retrosynthesis prediction system, integrating both organic and enzyme catalytic routes. Trained on experimental enzymatic reactions from Reaxys, the enzymatic triple transformer loop operates in parallel to the chemocatalytic loop. The competitive framework, driven by the route penalty score (RPScore), drives the selection of optimal steps by our best-first tree search, incorporating both catalytic steps to generate mixed synthesis routes. Our results not only showcase the tool's capabilities in proposing viable solutions for drug-like molecules but also establish it as a valuable resource for synthesis design. Furthermore, the continuous enrichment of data in Reaxys promises ongoing enhancements in enzymatic capabilities, progressively going towards enzymatic synthesis.

5.5 Availability of Data and Materials

Code and instructions to compute multistep retrosynthesis as well as the code to tag reactive sites will be available on our GitHub repository once this chapter becomes published: https://github.com/reymond-group/MultiStepRetrosynthesisTTL

The original USPTO dataset can be found at https://doi.org/10.6084/m9.figshare.5104873.v1. The derived version of the USPTO dataset of Thakkar *et al.* can be found in their preprint.[221] The Reaxys enzymatic dataset is a licensed commercial database that cannot be made available.

6   Conclusion and Outlook

The objective of this thesis was to explore the application of deep learning to enzymatic reactions and to close the gap between Computer-Aided Synthesis Planning (CASP) tools and biocatalysis by introducing a dual catalysis retrosynthesis software, which was absent at the time. In this chapter, I will summarize the work presented in this thesis and provide an outlook on future challenges.

6.1 Summary

In this thesis, I explored the potential of natural language processing (NLP) models to learn from biocatalysis experimental databases for predicting the outcomes of enzymatic reactions. Additionally, I designed a novel CASP software leveraging Transformer models with disconnection-tagging strategies for diversity exploration and synthesis route optimization. Subsequently, the CASP tool was implemented in a dual mode, encompassing both biocatalysis and chemocatalysis, to predict mixed catalytic reaction pathways.

Firstly, in Chapter 3, I pioneered the application of the Transformer architecture, coupled with transfer learning strategies, to enzymatic transformations. Drawing inspiration from how humans learn biocatalysis, where recognition and prediction of enzyme functions are often based on classification and common names rather than structural understanding, the Transformer model demonstrated the ability to learn similarly. By integrating molecular linear structure with textual descriptions of the associated enzymes, the resulting Enzymatic Transformer achieved a top-2 accuracy higher than 70% in predicting reaction outcomes, including correct stereochemistry — an essential aspect for enzymatic transformations. I introduced the concept of using confidence scores as a metric for the trustworthiness of predicted enzymatic reactions and established defined threshold values. The Enzymatic Transformer was showcased as a valuable tool for searching enzymes for desired enantioselective reactions, with the potential to link back to literature reports. In this chapter, I highlighted the potential to condense biocatalysis knowledge into a deep-learning model directly from experimental data, paving the way for further applications in retrosynthesis.

In the context of integrating biocatalysis into dual catalytic retrosynthesis, I initiated the development of a novel CASP software designed for chemocatalysis, anticipating the integration of multiple catalysis types. I detailed the software development and performance in Chapter 4. The

core part is a Triple Transformer Loop (TTL) for retrosynthesis prediction, operating based on the principle of round-trip validation.[102] The first Transformer is a disconnection-aware model predicting starting materials (SM), augmented by tagging strategies for exploring disconnections in the broader possible sense. The second Transformer predicts reagents, solvents, and catalysts for each disconnection. The third Transformer is a forward validation model, assessing the feasibility of the predicted retrosynthetic step. I demonstrated that the resulting TTL significantly enhances the diversity of retrosynthetic steps while being critical of the validity of the generated reactions. Subsequently, I integrated the Triple Transformer Loop (TTL) into a multistep framework using a heuristic best-first tree search. This search is guided by a route penalty score (RPScore), which takes into account factors such as the number of synthetic steps, the simplicity of the generated starting materials (SM), their commercial availability, and the chemical steps feasibility. Finally, I showcased its ability to find efficient synthesis routes for drug molecules. I have made the CASP tool and models freely accessible as an open-source Python package. Future work could involve the development of a Graphical User Interface (GUI) to enhance accessibility for chemists less familiar with command lines.

Following the demonstration of the Transformer model's capability to learn biocatalysis and the development of a CASP tool, I detailed in Chapter 5 the incorporation of an independent biocatalysis Triple Transformer Loop (TTL) into the previously established CASP tool. The resulting CASP software stands as the first and only Transformer-based multistep retrosynthesis tool, capable of exploring chemical spaces in both chemocatalysis and biocatalysis concurrently, constructing synthesis routes that involve a combination of catalytic approaches. I illustrated the dual TTL ability to identify viable synthesis routes incorporating enzymatic transformations on unseen molecules. Future implementations could involve the integration of additional catalytic modes or specialized chemistry with transfer learning, such as photocatalysis, electrocatalysis, carbohydrate, or natural product synthesis.

Overall, the thesis initiated with an exploration of language models for biocatalysis transformations, followed by the development of a CASP tool for chemocatalysis, and concluded with the integration of biocatalysis, based on experimental data. The resulting software stands as the first and only Transformer-based dual catalytic retrosynthesis tool, capable of exploring both chemical spaces for mixed synthesis routes. Furthermore, the Enzymatic Transformer prediction models, along with CASP tool developed, were implemented and tested internally at Novartis. All code already is or will be upon publication, open-source and freely accessible on GitHub, encouraging further community contributions and improvements.

6.2 Outlook

The work presented in this thesis owes its feasibility to recent advances in deep learning and the accessibility of large datasets. Despite the promising applications, the integration of deep learning into chemistry has numerous challenges awaiting to be addressed. In this section, I will discuss the forthcoming challenges and opportunities associated with the application of deep learning to retrosynthesis and biocatalysis. This discussion will address issues related to dataset limitations, considerations for multistep retrosynthesis, and the inherent challenges in benchmarking methods.

6.2.1 Datasets

Regardless of whether methods are template/rule-based or deep-learning-based, the quality of data is of utmost importance for data-driven approaches. The initial CASP tool, LHASA,[3] was renowned for its proficiency in suggesting rearrangement reactions frequently overlooked by chemists. Present CASP tools should similarly be able to offer non-obvious ideas to chemists. However, an inherent imbalance in datasets inevitably results in biased models, whether they are Transformers or policy networks. Infrequently occurring reactions, like rearrangements, are often neglected and not recommended, even though they could significantly contribute to solving a synthesis route prediction and introduce originality. Initiatives should be undertaken to prioritize such efficient steps in balancing training data or adapting model architectures.

Another consideration in historical patent reaction datasets is that they may reflect older chemistry, which chemists might wish to avoid due to concerns about toxicity, cost, or environmental impact, opting for greener alternatives. Once again, achieving a proper balance or labeling of reactions could enhance the quality of suggested synthesis routes.

Efforts should also be directed towards reporting negative data, which could substantially enhance the quality of models. It is frequently ignored in the literature because defining a negative reaction can be challenging and could be due to various reasons, such as unfavorable reaction conditions (temperature, pH, catalyst, solvent, etc.). Robotics could play a pivotal role in this regard by conducting reactions in a reproducible manner, expanding the coverage of chemical space, and offering deeper insights into chemistry.

Concerning biocatalysis, the convergence of directed evolution with robotic advancements holds the potential to significantly enhance our comprehension of enzymatic transformations and contribute additional data for training models. Moreover, such platforms could offer more detailed enzyme information than textual descriptions used in this thesis, such as the amino acid sequences. This detailed information could potentially lead to improved biocatalysis prediction models by integrating structural information, which can be predicted using recent advancements in protein structural prediction methods.[244]

6.2.2 MULTISTEP RETROSYNTHESIS

The synthesis planning software discussed in this thesis primarily relies on single-step retrosynthesis models, whether template/rule-based or deep-learning-based, re-iterating uncommercial predicted molecules. However, the simulation of all conceivable moves is not feasible due to the combinatorial explosion, and it does not align with the way a chemist or a Go player would approach tasks that involve multiple moves. In contrast, there is a need for more intelligent retrosynthesis planning software that can envision and plan beyond individual steps, resembling the strategic thinking observed in games like Go or Chess. This concept involves "smart synthesis", incorporating rearrangements through multiple steps in advance, akin to navigating a tree of thoughts.[245]

Synthesis planning tools encoded by experts may exhibit improved performance, especially when equipped with meticulously curated rules. Future work could focus on advancing deep-learning approaches that seamlessly integrate both expert-encoded knowledge and data-driven strategies, harnessing the synergies between these two methodologies.[246]

6.2.3 BENCHMARKING SYNTHESIS PLANNING TOOLS

Assessing the performance of CASP tools poses a significant challenge due to the intricate nature of evaluating predicted synthesis route quality automatically. While one approach involves comparing the ability to reproduce known synthesis routes, I believe that the evaluation of a CASP tool should lie not in its capacity to replicate its training dataset but rather in its capability to present chemists with diverse and innovative options that are chemically feasible in the laboratory.

Furthermore, the diverse parameters and settings of each tool, coupled with variations in training datasets and limited accessibility to certain tools, obstruct reproducibility and direct comparisons. Additionally, defining a synthesis route as "solved" or not is complex and directly depends on the chosen dataset of commercially available starting materials, which varies among tools.

Synthetic routes can be evaluated based on various factors, including route length, chemical feasibility, safety, starting materials, cost, etc., aspects that only synthetic experts can effectively appraise. The ideal solution for evaluating CASP tool performance would involve individual route assessments by human experts. Although challenging to implement, efforts have been made toward establishing guidelines for benchmarking multistep retrosynthesis approaches, marking an initial step in conceiving evaluation metrics.[231]

While benchmarking platforms are an initial stride toward clearer evaluation of CASP tools, the field could benefit immensely from a community-wide initiative similar to the Critical Assessment of Protein Structure Prediction experiment. Such an effort would involve using the same training dataset for all participants, with the results evaluated by expert judges in a blind fashion.

A APPENDIX: PREDICTING ENZYMATIC REACTIONS WITH A MOLECULAR TRANSFORMER

A.1 DEHYDROGENASE FREQUENCY ANALYSIS
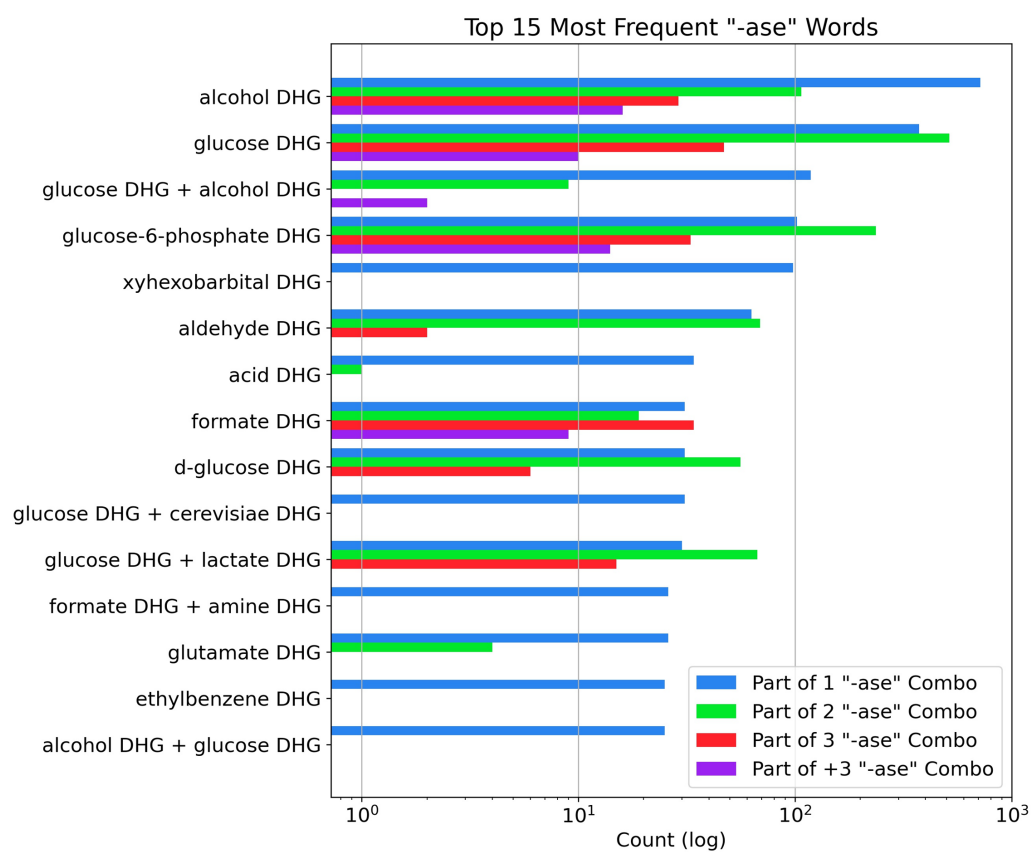
Figure A.1: **Analysis of the dehydrogenase ("DHG") diversity in the entire ENZR dataset.**

A.2 Tmap of the Enzr dataset by substrate similarity
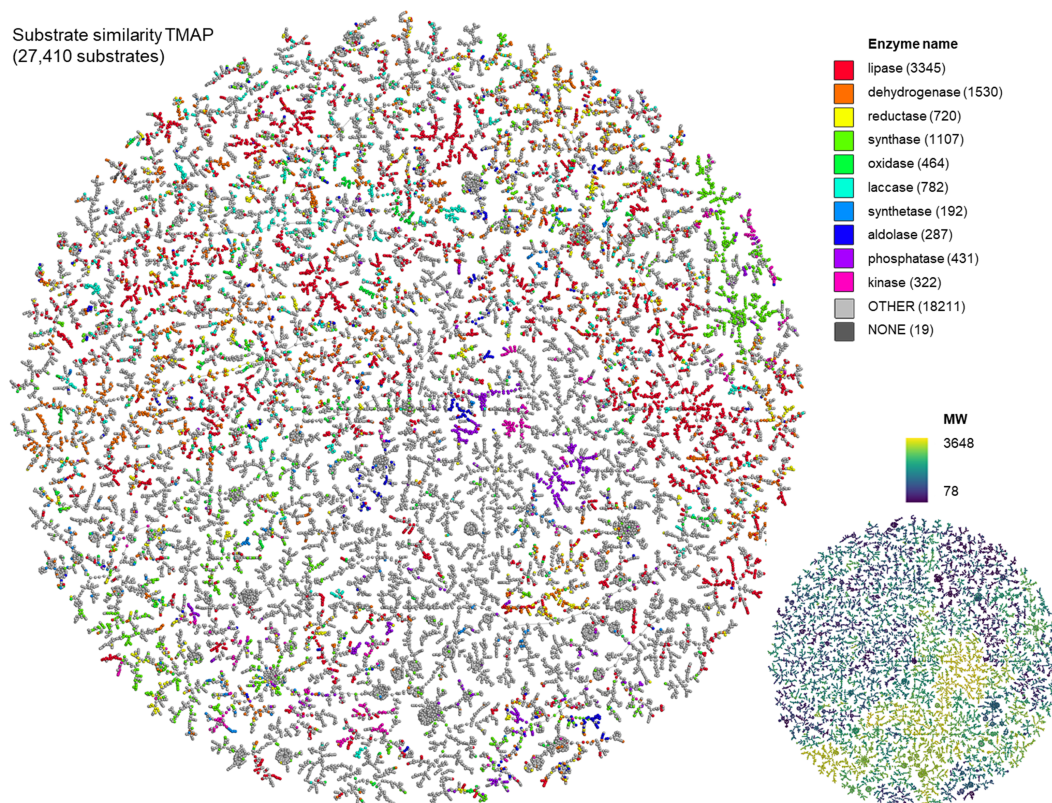
Figure A.2: **TMAP of the ENZR dataset** analyzed by substrate similarity and color-coded by "-ase" word combinations. Insert: TMAP color-coded by substrate molecular weight.

A.3 Cofactor importance in the prediction
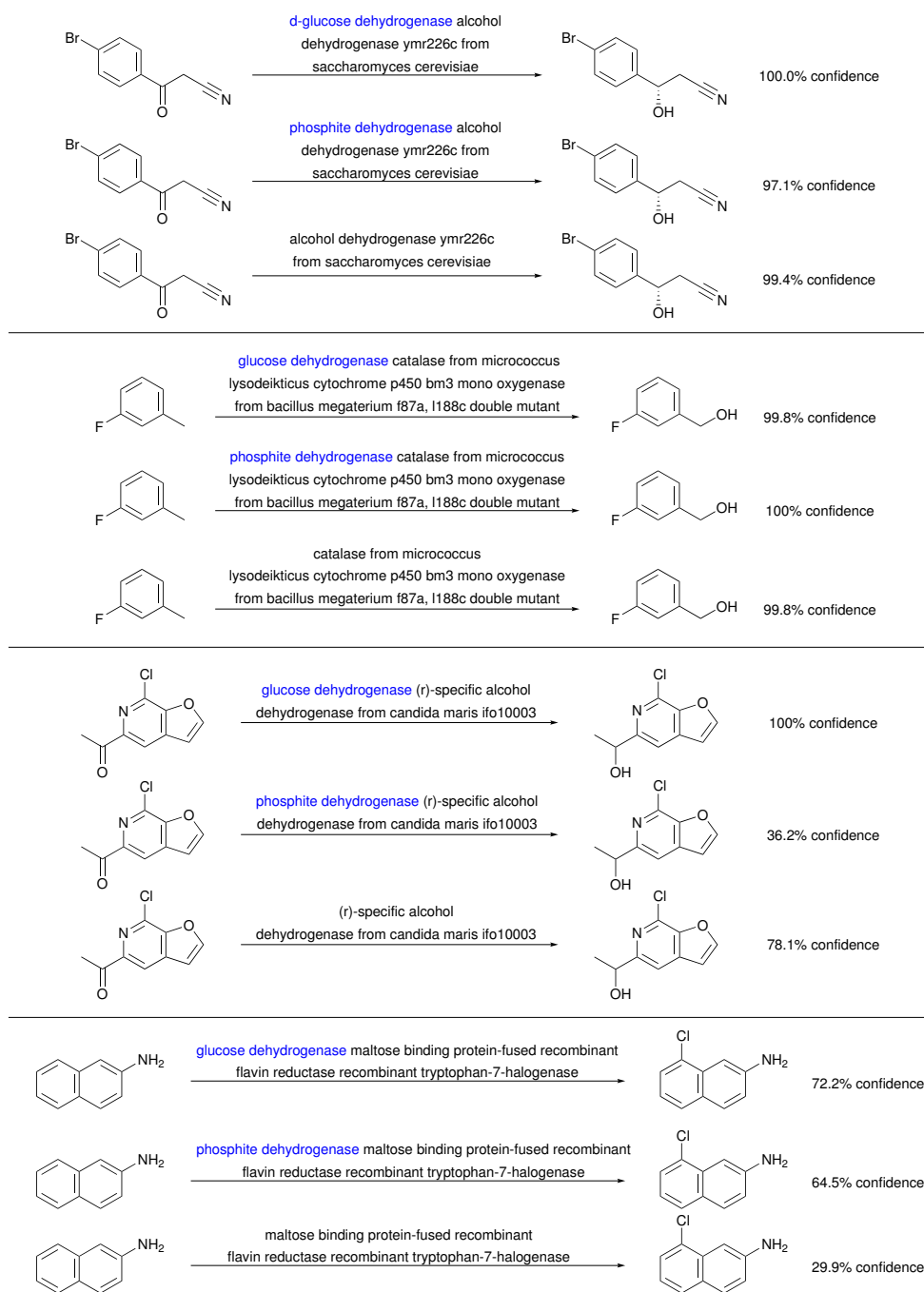
Figure A.3: **Examples of cofactor generator swapped or removed.**

A.4 Effect of word on the prediction
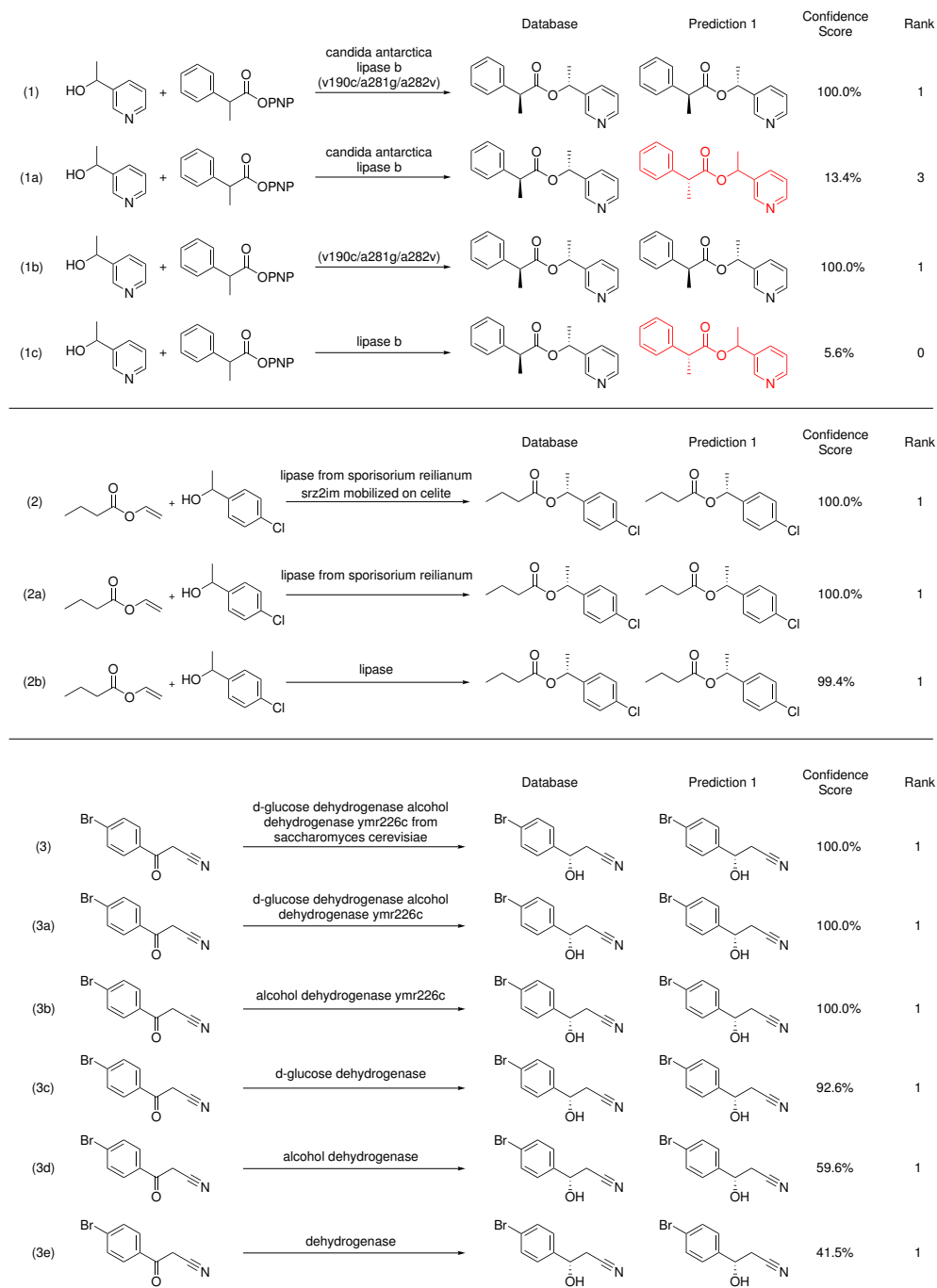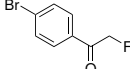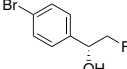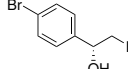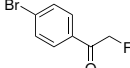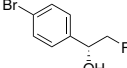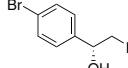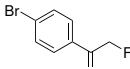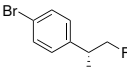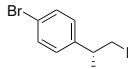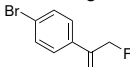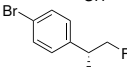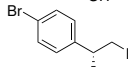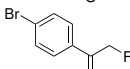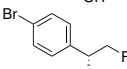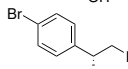
Figure A.4: **(Part A) - Examples of predictions from success examples from figure 3.5** with a variety of truncated sentences. "Rank" represent the top position prediction containing the correct product.

Figure A.4: **(Part B) - Examples of predictions from success examples from figure 3.5** with a variety of truncated sentences. "Rank" represent the top position prediction containing the correct product.

		Database	Prediction	Confidence Score	Rank
(7)	ferredoxin reductase cytochrome p450 monooxygenase from rhodopseudomonas palustris cga009			100.0%	1
(7a)	ferredoxin reductase cytochrome p450 mono oxygenase from rhodopseudomonas palustris			99.8%	1
(7b)	cytochrome p450 mono oxygenase from rhodopseudomonas palustris			99.8%	1
(7c)	p450 monooxygenase from rhodopseudomonas palustris cga009			99.8%	1
(7d)	monooxygenase from rhodopseudomonas palustris			99.9%	1
(7e)	reductase cytochrome p450 monooxygenase			43.0%	0
(7f)	rhodopseudomonas palustris cga009			100.0%	1
(7g)	cytochrome p450			16.4%	0
(7h)	p450			31.5%	0

		Database	Prediction	Confidence Score	Rank
(8)	glucose dehydrogenase catalase from micrococcus lysodeikticus cytochrome p450 bm3 mono oxygenase from bacillus megaterium f87a, l188c double mutant			99.8%	1
(8a)	cytochrome p450 bm3 mono oxygenase from bacillus megaterium f87a, l188c double mutant			99.6%	1
(8b)	glucose dehydrogenase catalase cytochrome p450 mono oxygenase			100.0%	1
(8c)	cytochrome p450 bm3 mono oxygenase			37.8%	1
(8d)	glucose dehydrogenase catalase			97.0%	1
(8e)	glucose dehydrogenase			61.2%	1
(8f)	cytochrome p450			99.4%	1

Figure A.4: **(Part C) - Examples of predictions from success examples from figure 3.5** with a variety of truncated sentences. "Rank" represent the top position prediction containing the correct product.

78

Figure A.4: **(Part D) - Examples of predictions from success examples from figure** **3.5** with a variety of truncated sentences. "Rank" represent the top position prediction containing the correct product.

A.5 All P450 reactions from the test set

Figure A.5: **(Part A) - Every reaction from the test set containing "p450" in the sentence correctly predicted by the full sentence model.** Reactions sorted by decreasing confidence score.

Figure A.5: **(Part B) - Every reaction from the test set containing "p450" in the sentence correctly predicted by the full sentence model.** Reactions sorted by decreasing confidence score.
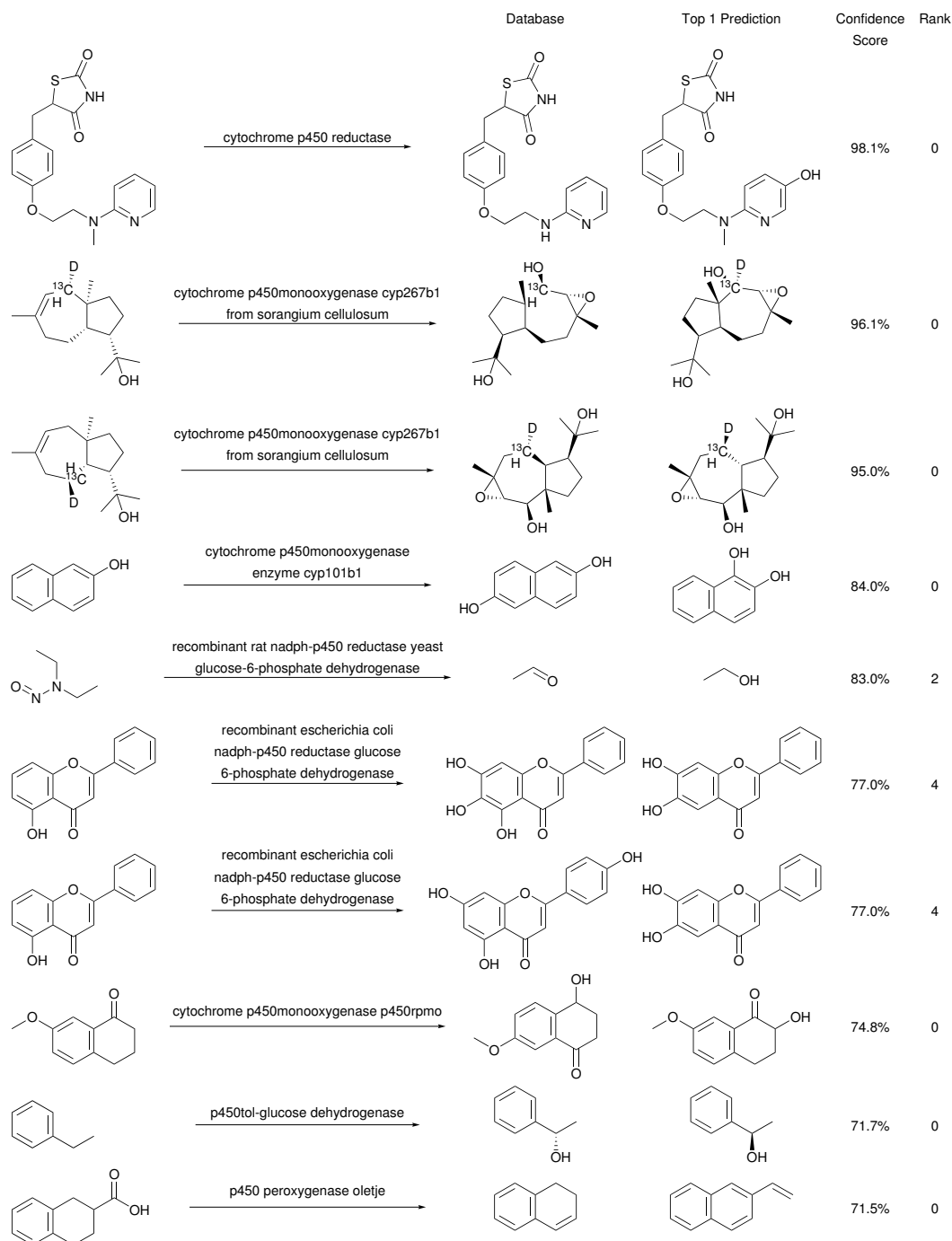
Figure A.6: **(Part A) - Every reaction from the test set containing "p450" in the sentence incorrectly predicted by the full sentence model.** "Rank" showing the rank of the correct prediction assigned by the model, "0" meaning that the model did not predict the correct product within the 5 first predictions. Reactions are sorted by decreasing confidence score.

Figure A.6: **(Part B) - Every reaction from the test set containing "p450" in the sentence incorrectly predicted by the full sentence model.** "Rank" showing the rank of the correct prediction assigned by the model, "0" meaning that the model did not predict the correct product within the 5 first predictions. Reactions are sorted by decreasing confidence score.
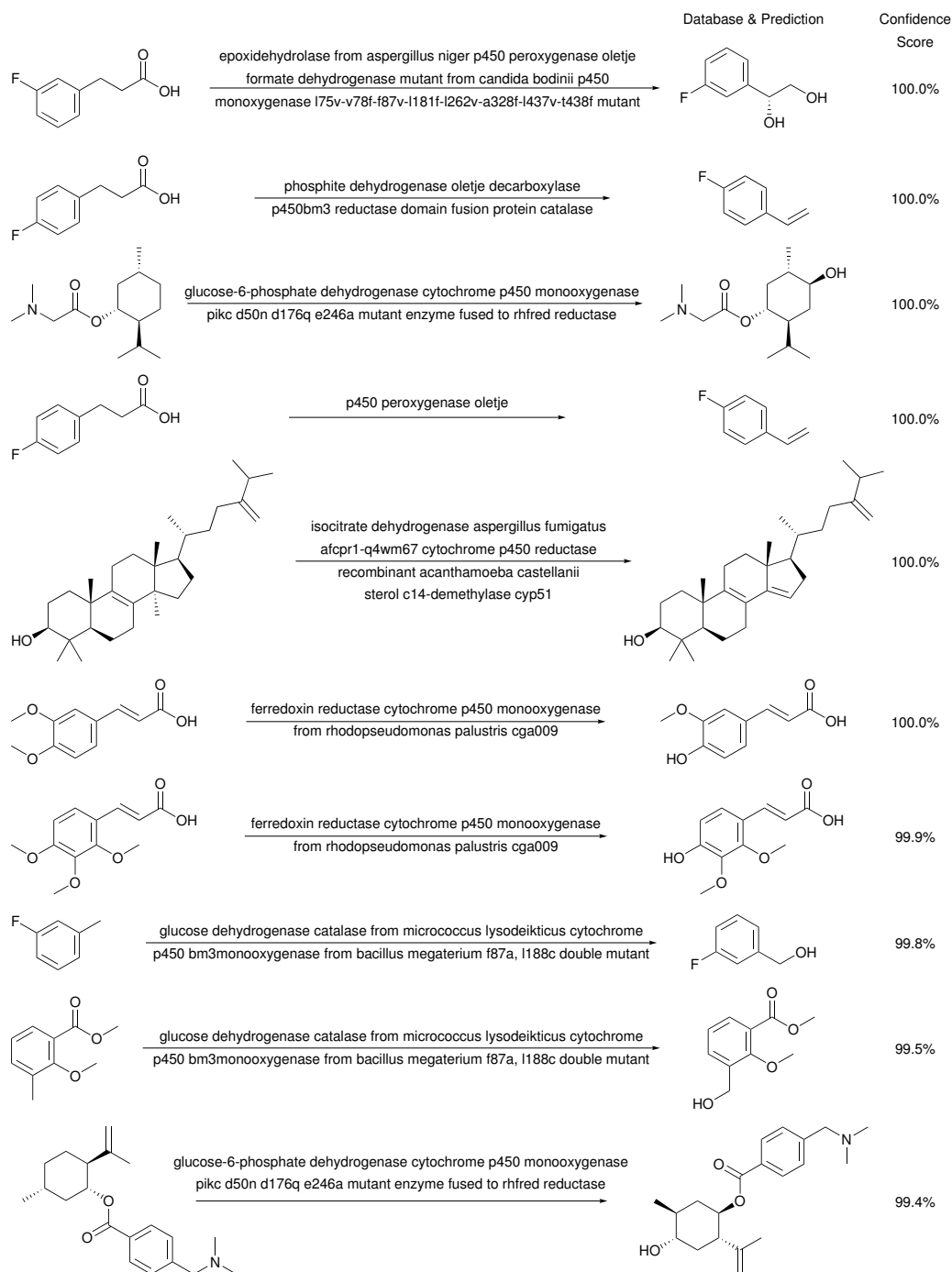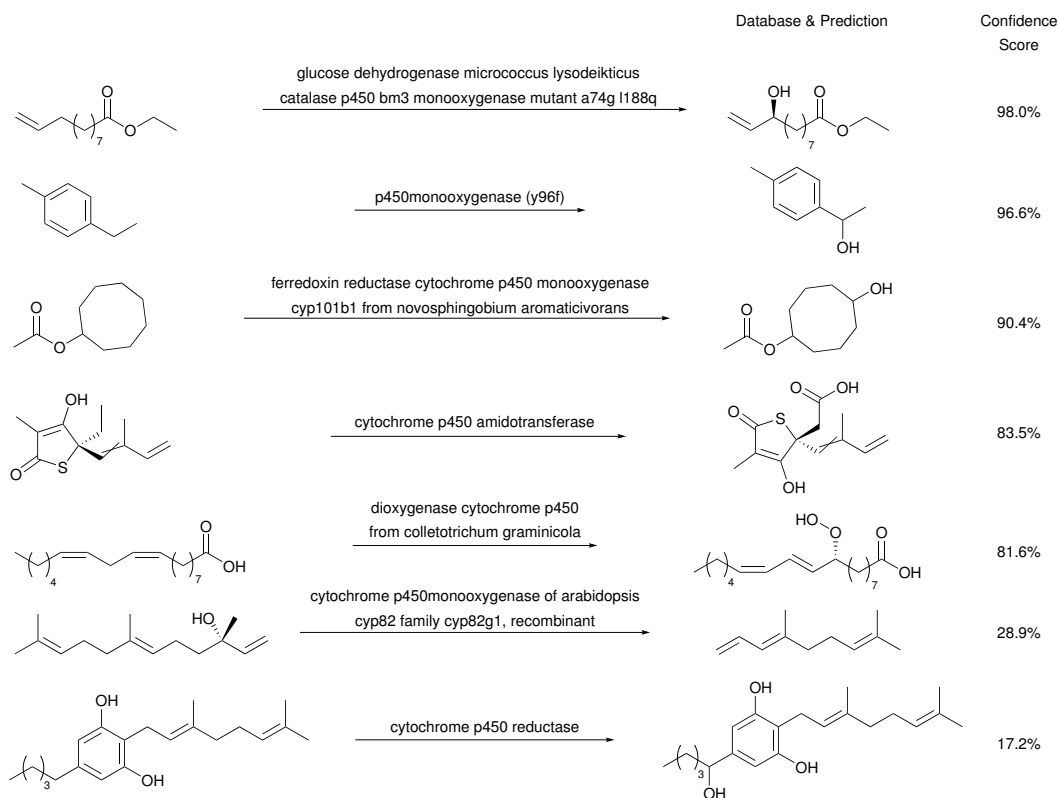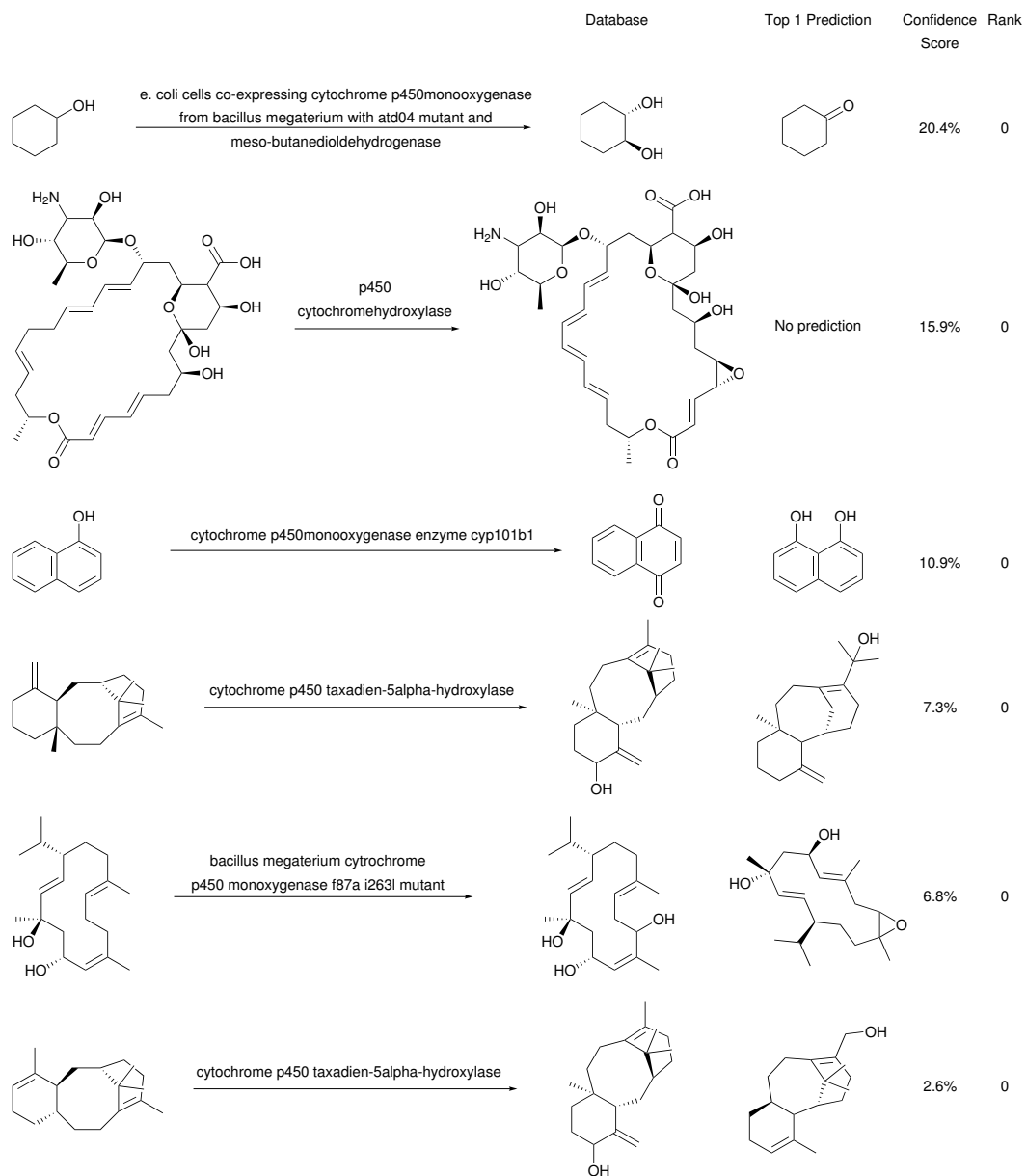
Figure A.6: **(Part C) - Every reaction from the test set containing "p450" in the sentence incorrectly predicted by the full sentence model.** "Rank" showing the rank of the correct prediction assigned by the model, "0" meaning that the model did not predict the correct product within the 5 first predictions. Reactions are sorted by decreasing confidence score.

A.6 Oxidase wild type (WT) and mutant (M)
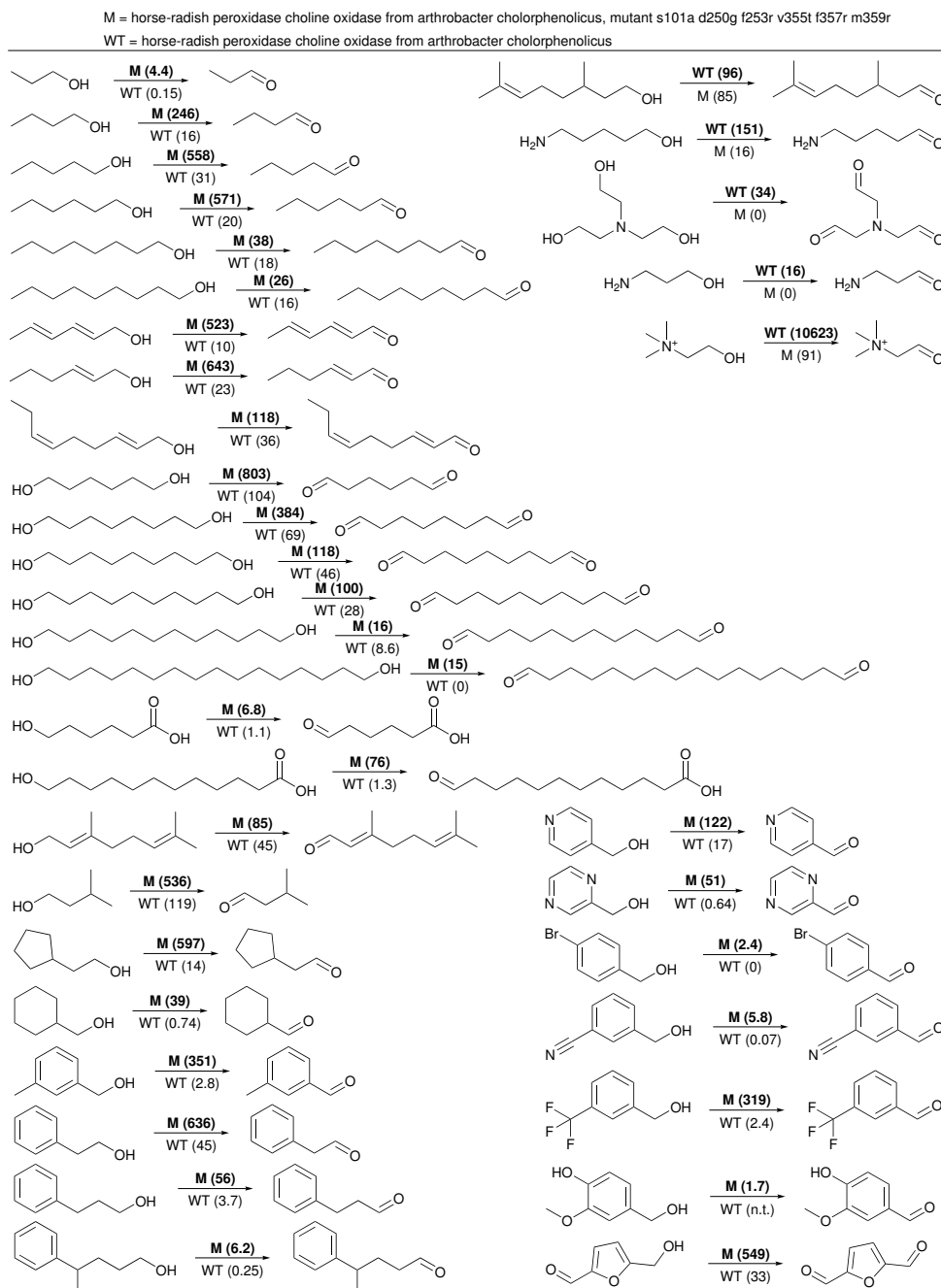
Figure A.7: **Reactions using the choline oxidase wild type (WT) and mutant (M)** from Heath *et al.*[209] that were assigned to the training set. The numbers in parenthesis correspond to the specific activity of either the mutant or the wild type enzyme express in mU.mg^{-1}. (n.t. = not tested).

M = horse-radish peroxidase choline oxidase from arthrobacter cholorphenolicus, mutant s101a d250g f253r v355t f357r m359r
WT = horse-radish peroxidase choline oxidase from arthrobacter cholorphenolicus
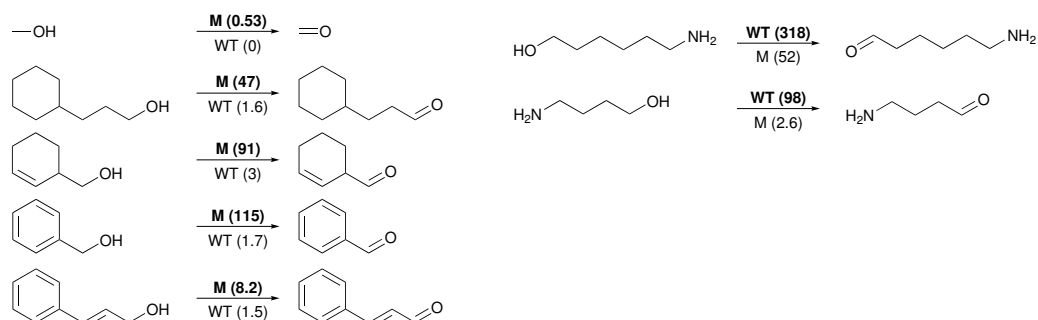
Figure A.8: **Reactions using the choline oxidase wild type (WT) and mutant (M)** from Heath *et al.*[209] that were assigned to the validation set. The numbers in parenthesis correspond to the specific activity of either the mutant or the wild type enzyme express in mU.mg^{-1}.

M = horse-radish peroxidase choline oxidase from arthrobacter cholorphenolicus, mutant s101a d250g f253r v355t f357r m359r
WT = horse-radish peroxidase choline oxidase from arthrobacter cholorphenolicus
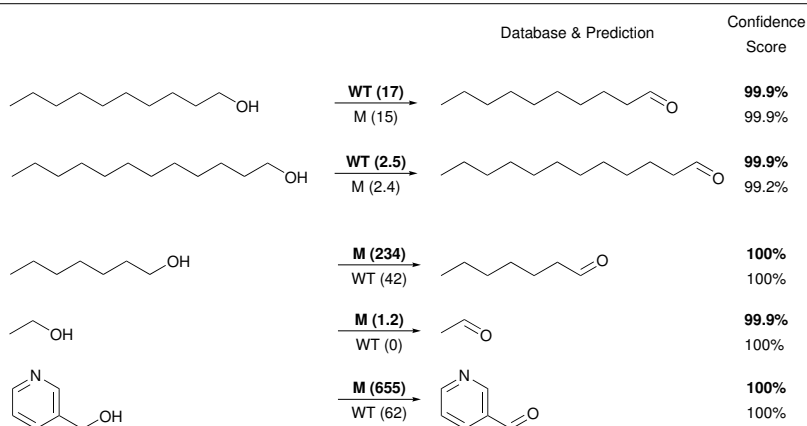
Figure A.9: **Reactions using the choline oxidase wild type (WT) and mutant (M)** from Heath *et al.*[209] that were assigned to the test set. All reactions were predicted correctly. The numbers in parenthesis correspond to the specific activity of either the mutant or the wild type enzyme express in mU.mg^{-1}.

A.7 Screening of various substrates for the same sentences
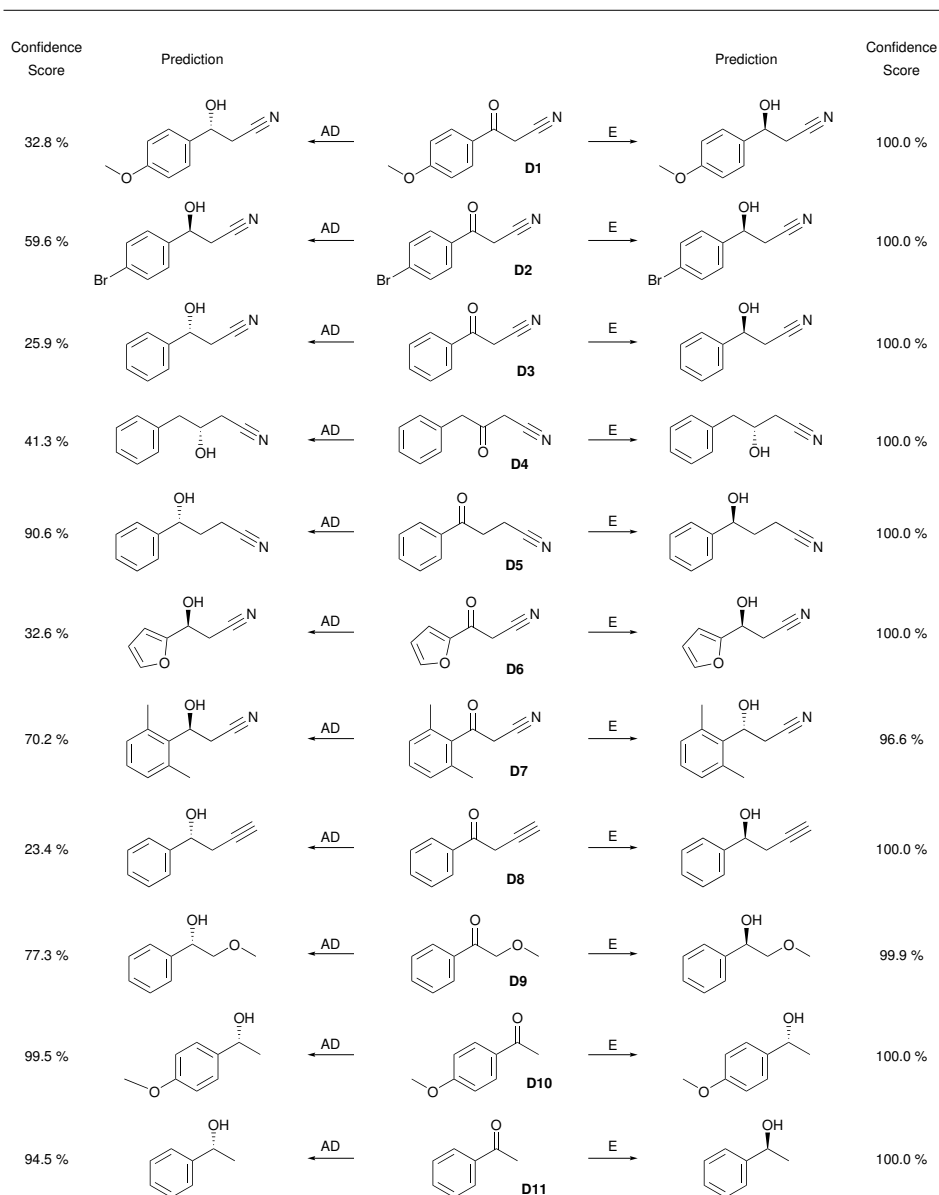
Figure A.10: **(Part A) - Various substrates tested on two sentences**, a simple "alcohol dehydrogenase" (AD) and the "d-glucose dehydrogenase alcohol dehydrogenase ymr226c from *Saccharomyces cerevisiae*" (E). All substrates were derivatives from **D1** and **D2** which were present in the test set[194] and predicted correctly. Even though products from substrates **D16** and **D19** using enzyme "E" are not chiral, the model gave those chiral centers in the output SMILES ("CC[C@H](O)CC" for **D16**, "O[C@H]1CCCCC1" for **D19**).

Figure A.10: **(Part B) - Various substrates tested on two sentences**, a simple "alcohol dehydroge-nase" (AD) and the "d-glucose dehydrogenase alcohol dehydrogenase ymr226c from *Sac-charomyces cerevisiae*" (E). All substrates were derivatives from **D1** and **D2** which were present in the test set[194] and predicted correctly. Even though products from substrates **D16** and **D19** using enzyme "E" are not chiral, the model gave those chiral centers in the output SMILES ("CC[C@H](O)CC" for **D16**, "O[C@H]1CCCCC1" for **D19**).

A.8 Token frequencies analysis
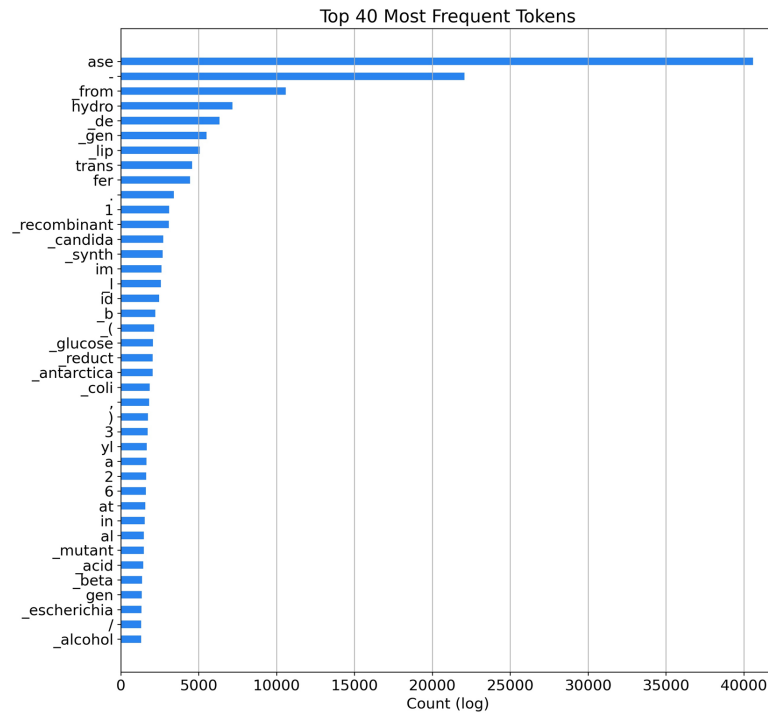
Figure A.11: **Top 40 most frequent tokens from the entire ENZR dataset.**
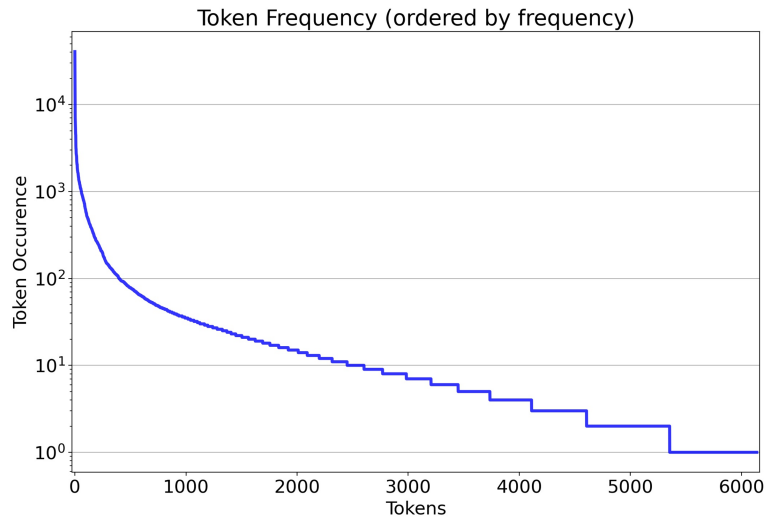
Figure A.12: **Power law distribution of the occurrence frequencies of all tokens in the ENZR sentences sorted by frequency (total of** 6139 **tokens).**

B Appendix: Multistep Retrosynthesis Combining a Disconnection Aware Triple Transformer Loop with a Route Penalty Score Guided Tree Search

B.1 Single-step tagging strategies study
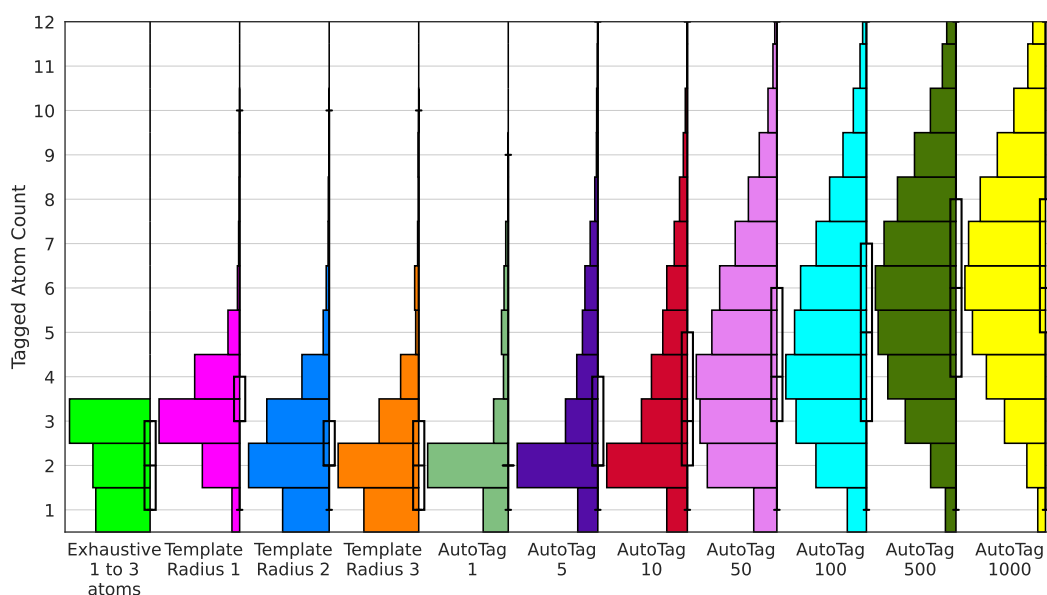
Figure B.1: **Number of tagged atoms per molecule as function of the tagging method.** The relative number of molecules (horizontal bar length) is plotted as function of the number of atoms tagged (vertical axis) by different tagging methods (horizontal categories), tested over 500 molecules (randomly selected from the test set). The exhaustive tagging was performed together for tags containing 1, 2 and 3 atoms. The template tagging was performed separately for templates of radius of 1, 2 or 3 bonds. The AutoTag model was tested using the top-N predictions using N = 1, 5, 10, 50, 100, 500 and 1000.
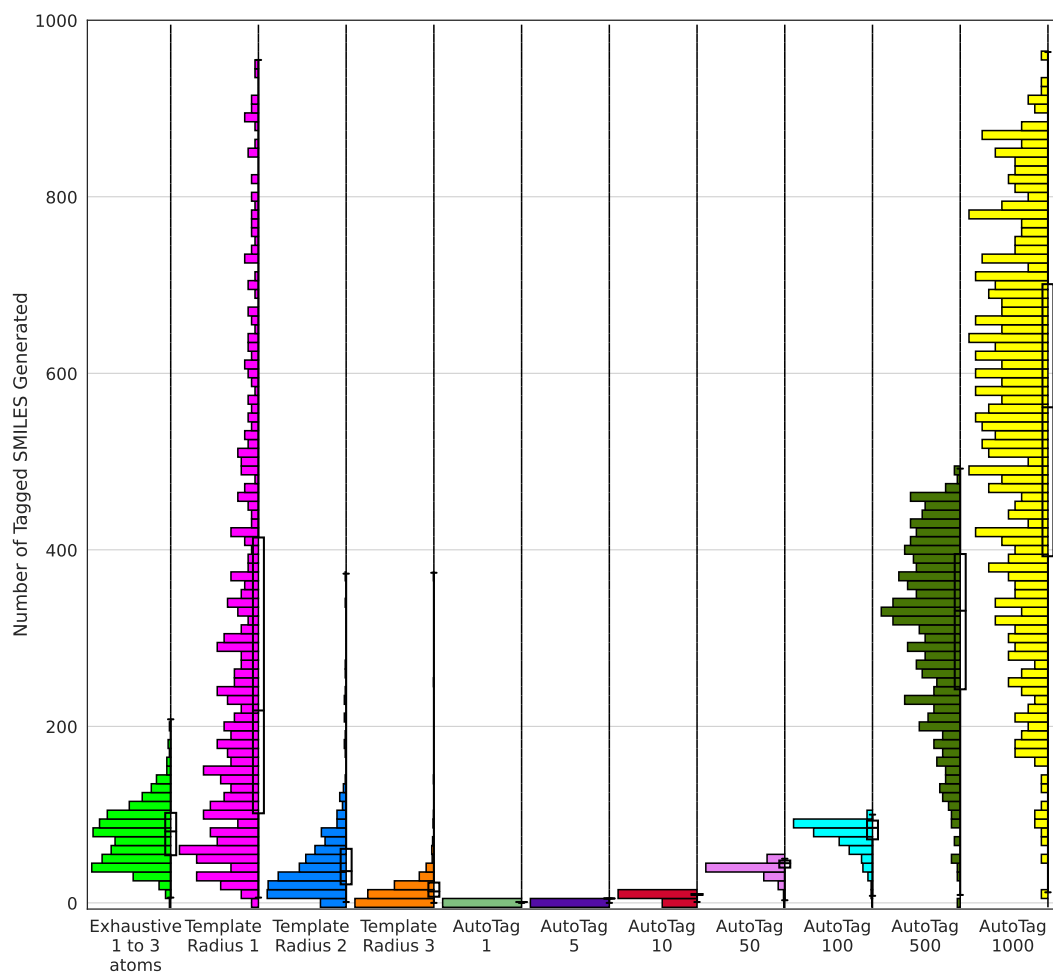
Figure B.2: **Number of tagged SMILES per molecule as function of the tagging method.** The relative number of molecules (horizontal bar length) is plotted as function of the number of valid tagged SMILES per molecule (vertical axis) produced by different tagging methods (horizontal categories), tested over 500 molecules (randomly selected from the test set). A higher number of tags corresponds to a higher computational cost as each tagged starting material must be processed by the TTL.
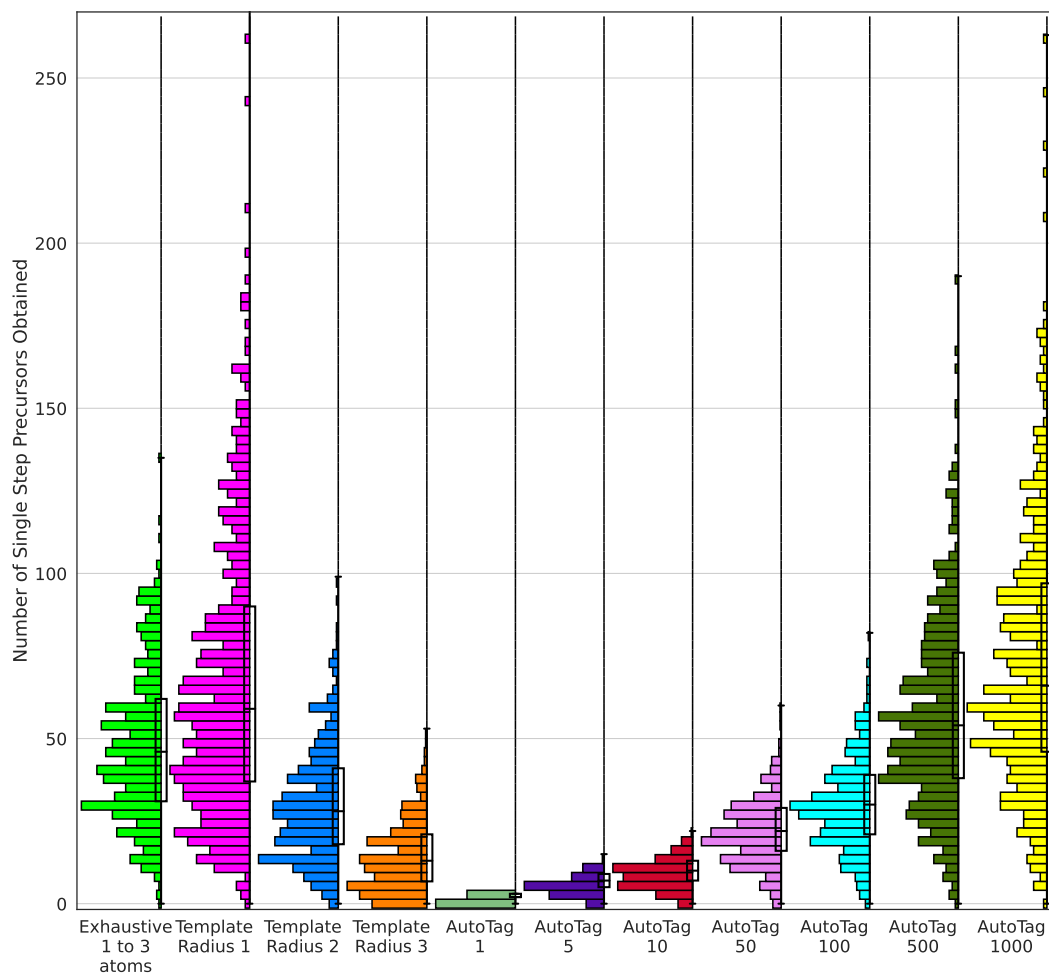
Figure B.3: **Number of starting materials per molecule from TTL as function of the tagging method.** The relative number of molecules (horizontal bar length) is plotted a function of the number of starting materials per molecule (vertical axis, "single step precursors") produced by applying TTL to the tagged SMILES resulting from the indicated tagging method (horizontal categories), tested on 500 molecules (randomly selected from the test set) across multiple tagging strategies.
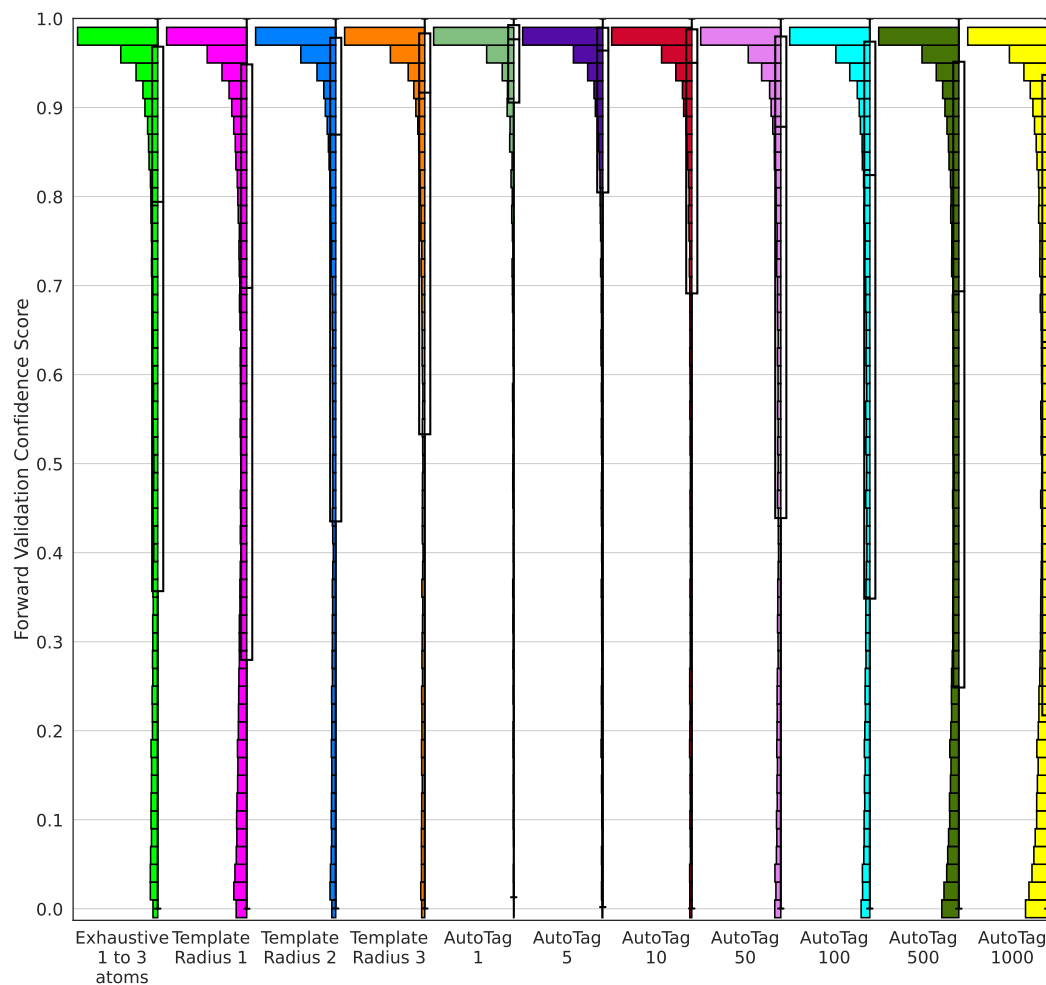
Figure B.4: **Distribution of forward validation confidence scores for validated TTL steps a function of the tagging method.** The relative number of forward validated steps (horizontal bar length) is plotted as function of the confidence score of the forward validation transformer T3 (vertical axis) for steps predicted from SMILES tagged with different tagging methods (horizontal categories), tested over 500 molecules (randomly selected from the test set).
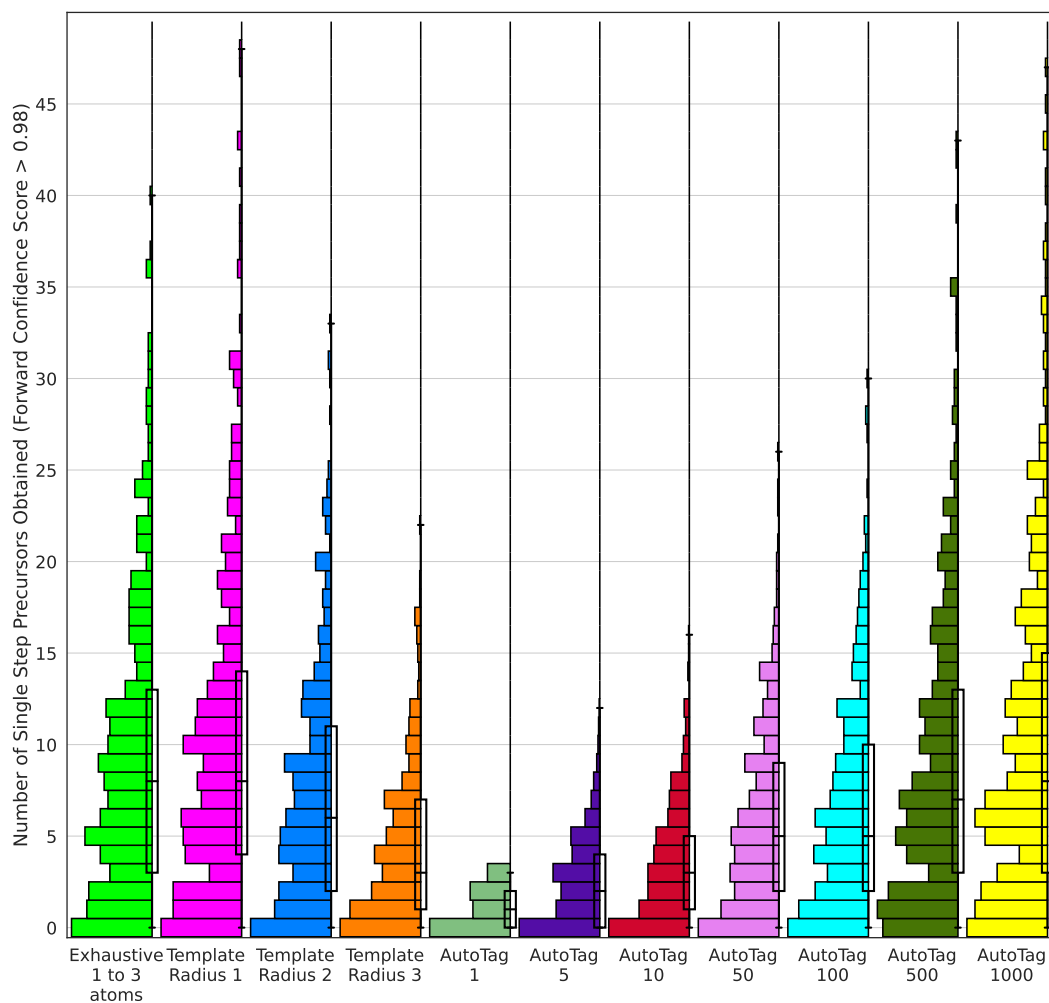
Figure B.5: **Number of single step precursors produced by TTL as function of the tagging method.** The relative number of molecules (horizontal bar length) is plotted as function of the number of precursors obtained from validated TTL predicted single retrosynthetic steps per molecule (vertical axis) using different tagging methods (horizontal categories), tested on 500 molecules (randomly selected from the test set).
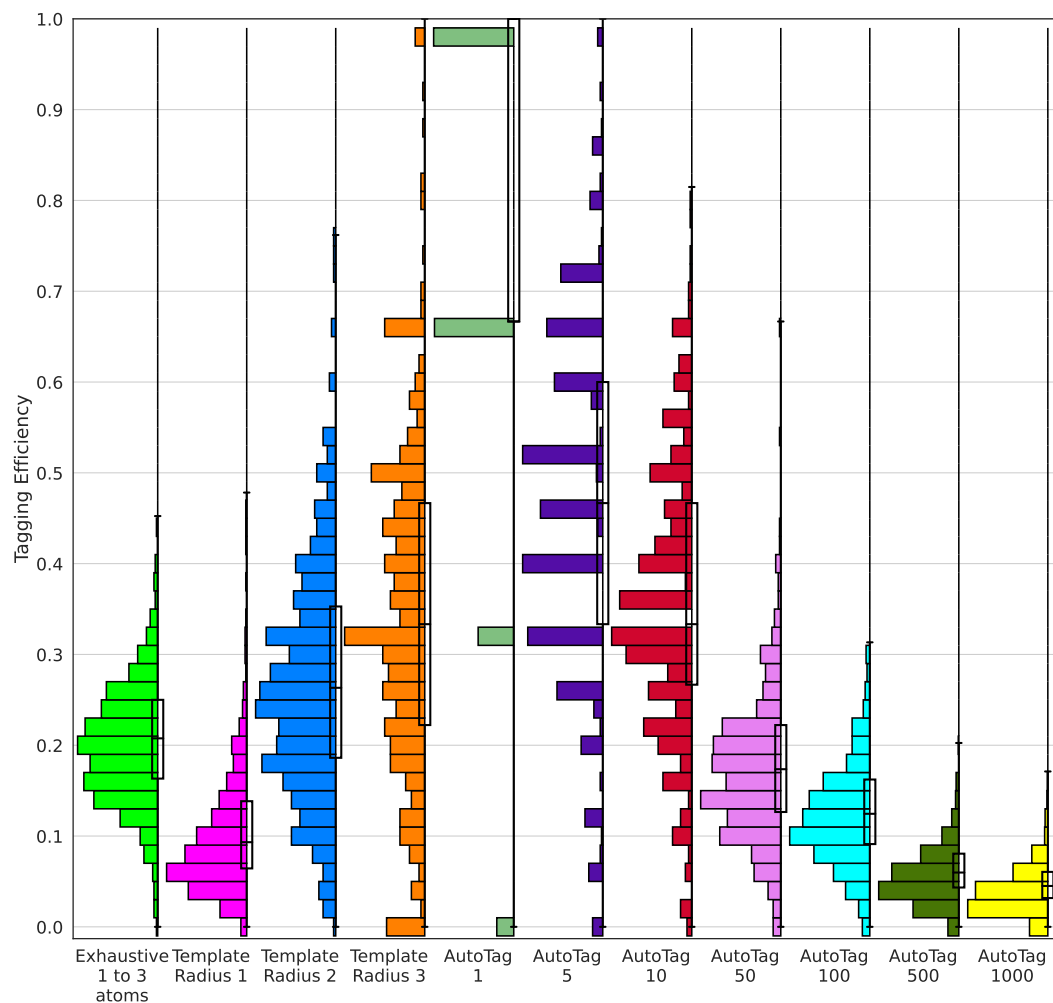
Figure B.6: **Tagging efficiency as function of the tagging method.** The number of molecules (horizontal bar length) is plotted as function of the fraction of tags leading to a TTL validated retrosynthetic step (vertical axis) using different tagging methods (horizontal categories), tested over 500 molecules (randomly selected from the test set). The tagging efficiency was computed by dividing the number of TTL validated retrosyntheses obtained by the number of generated tagged SMILES. Values are normalized, predictions were obtained with a beam size of 3 for T2 (reagent prediction), all tested on the forward validation model T3.
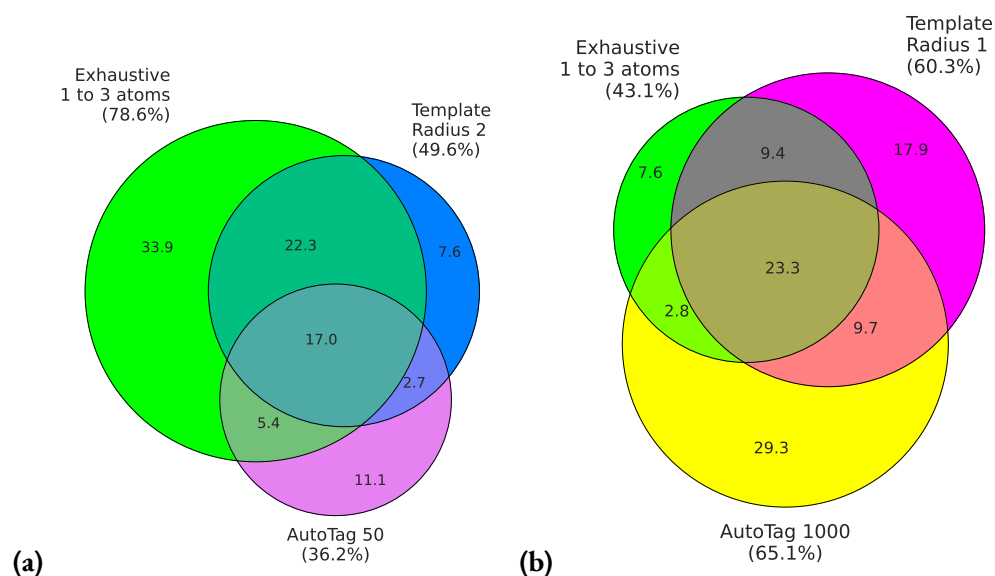
Figure B.7: **Overlap of retrosynthetic steps predicted by TTL using different tagging methods.** The Venn diagram shows the percentage of TTL predicted steps distributed across three different tagging methods choses as **(a)** the selected set of reasonable tagging methods that avoids excessive number of tags, and **(b)** the three least restrictive tagging methods generating large number of tags (computationally expensive), tested over 500 molecules (randomly selected from the test set). Selection **(a)** is subsequently used for the multistep predictions in TTLA.
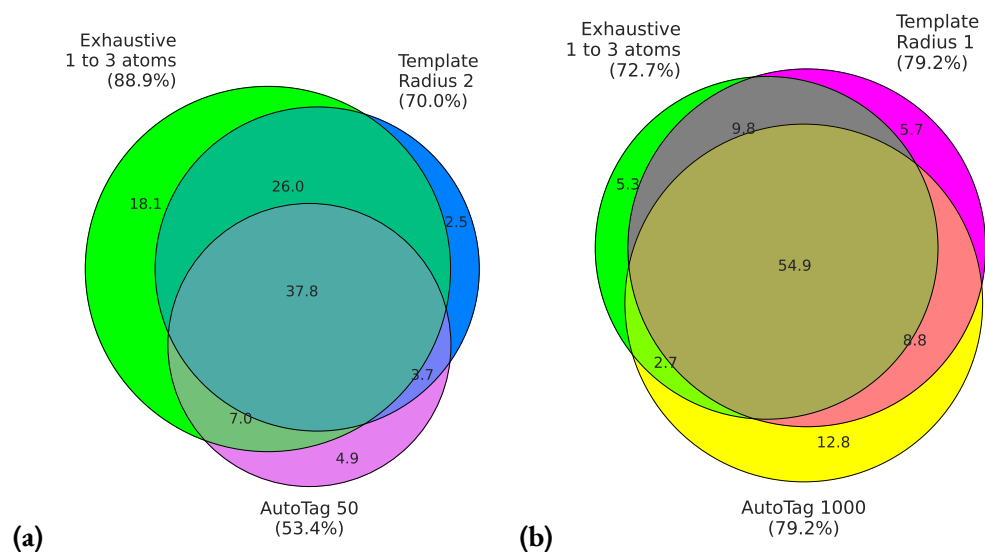
Figure B.8: **Overlap of high confidence retrosynthetic steps predicted by TTL using different tagging methods.** Same analysis as Figure B.7 for the subset of validated step having a confidence score than >98% for forward validation transformer T3.

B.2 Multistep predictions
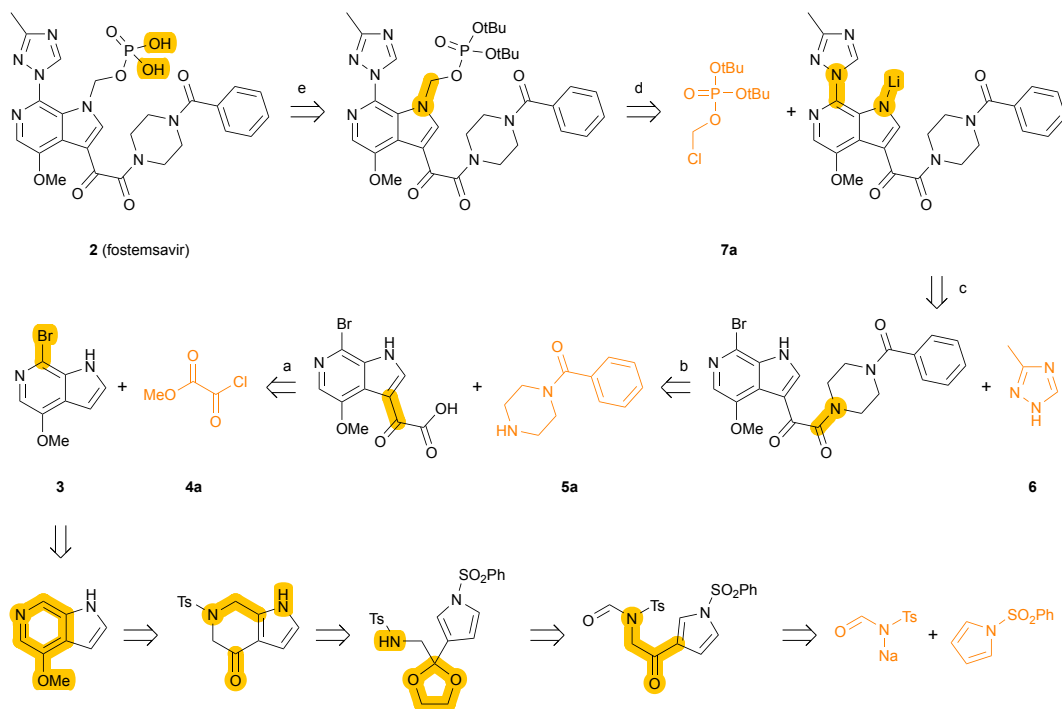
Figure B.9: **Literature reported retrosynthesis for fostemsavir.**[228] Orange-coloured compounds are commercially available. Reported reagents: a) AlCl$_3$, Bu$_4$NHSO$_4$, CH$_2$Cl$_2$, then KOH, then H$_3$PO$_4$; b) Ph$_2$POCl, NMM, NMP; c) KOH, CuI, then KOH, EtOH, LiI; d) Et$_4$NI, K$_2$CO$_3$, CH$_3$CN/H$_2$O; e) AcOH, H$_2$O.
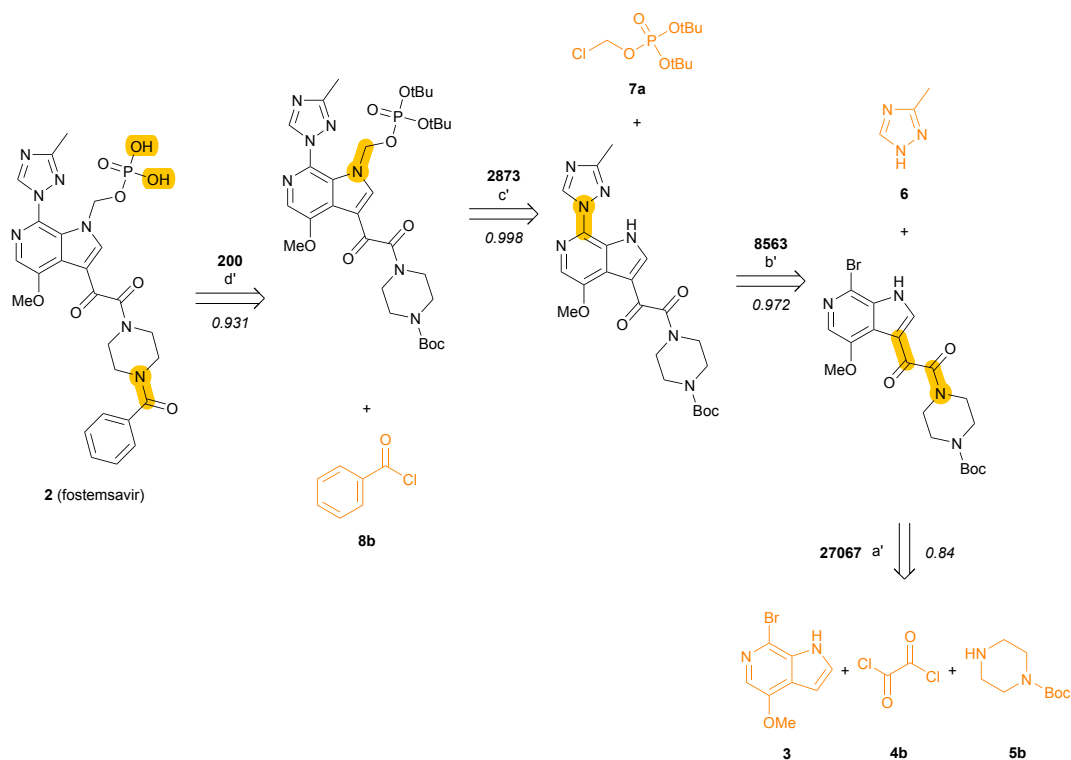
Figure B.10: **Best RPScoring predicted retrosynthesis route for fostemsavir.** Orange-coloured compounds are commercially available. Except for some of the commercial precursors that were present but involved in different reactions, none of the intermediate compounds were present in the training dataset. The reaction prediction numbers in bold on retrosynthesis arrows correspond to the order in which the multistep tree search prioritized the prediction. Forward prediction confidence scores are shown under retrosynthesis arrows. Predicted reaction conditions: a') Et$_3$N, CH$_2$Cl$_2$; b') K$_2$CO$_3$, CuI, toluene; c') K$_2$CO$_3$, DMF; d') HCl, N,N-Diisopropylethylamine, H$_2$O, dioxane.
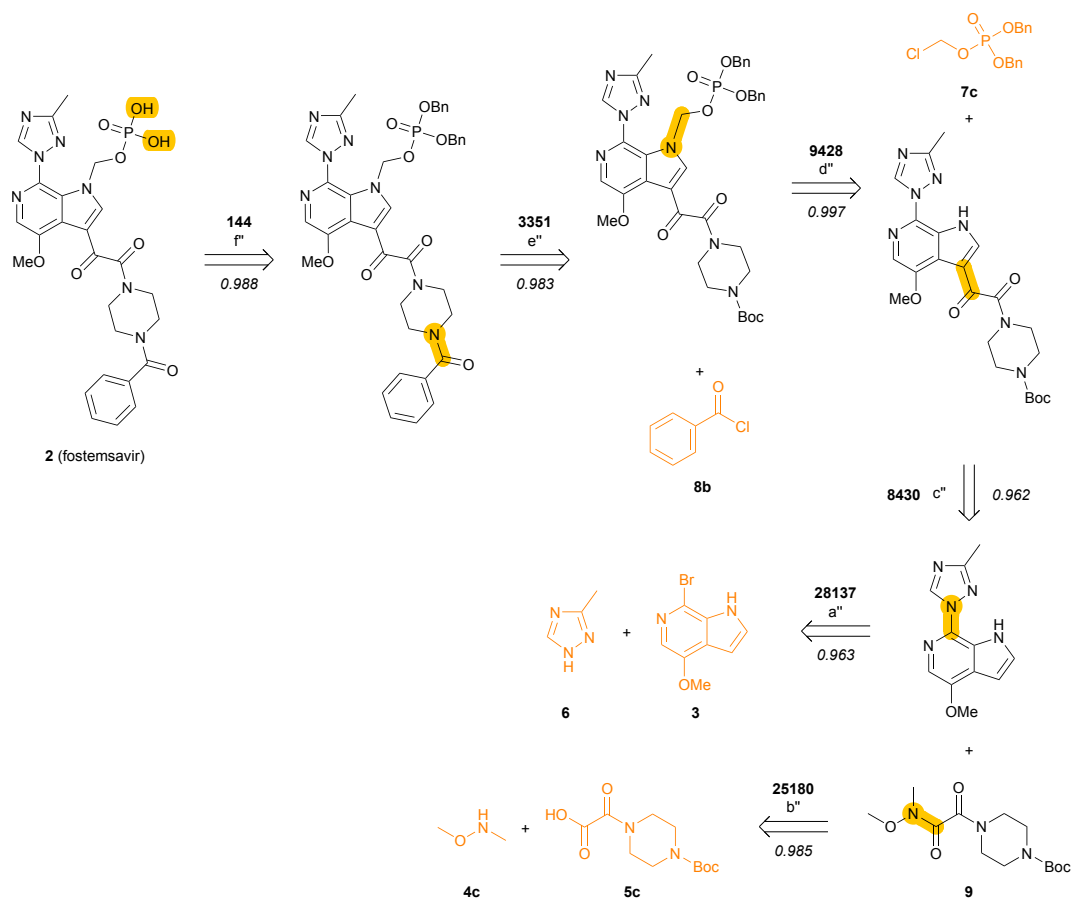
Figure B.11: **Best overall confidence score predicted retrosynthesis route for fostemsavir.** Orange-coloured compounds are commercially available. Except for some of the commercial precursors that were present but involved in different reactions, none of the intermediate compounds were present in the training dataset. The reaction prediction numbers in bold on retrosynthesis arrows correspond to the order in which the multistep tree search prioritized the prediction. Forward prediction confidence scores are shown under retrosynthesis arrows. Predicted reagents: a") (2S)-pyrrolidine-2-carboxylic acid, K_2CO_3, CuI, EtOAc, DMSO; b") no reagent predicted; c") n-BuLi, THF; d") K_2CO_3, DMF; e") TFA, DMAP, CH_2Cl_2, f") Pd, EtOH.

Figure B.12: **Literature reported retrosynthesis for ozanimod.**[228] Orange-coloured compounds are commercially available. Reported reagents: a) HC(Ome)$_3$, p-TsOH, PhCH$_3$; b) NH$_2$OH.HCl, Et$_3$N; c) carbonyl diimidazole; d) NaOH; e) i) p-TsOH, acetone, ii) NH$_2$CH$_2$CH$_2$OH, p-TsOH, PhCH$_3$, iii) Chiral Ru-complex, Et$_3$N/HCO$_2$H.
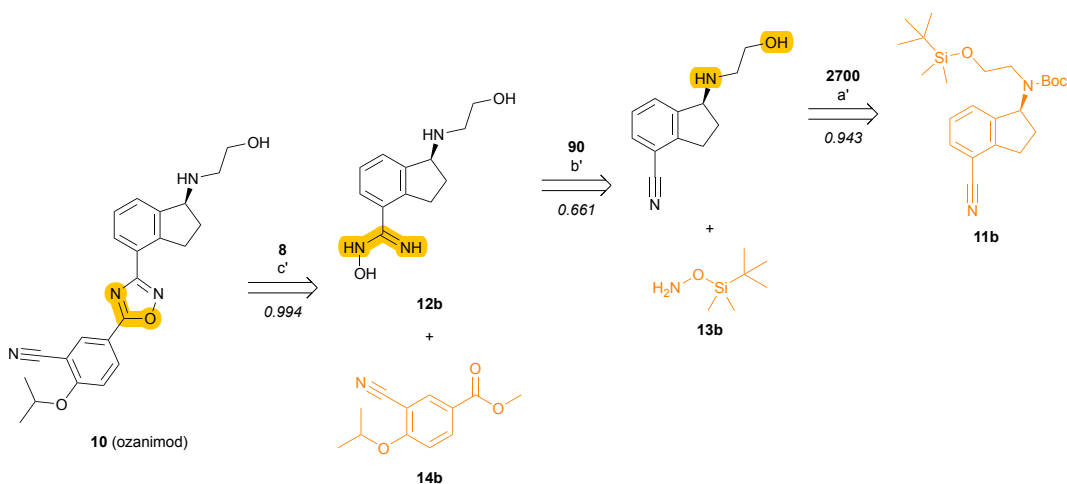
Figure B.13: **Best RPScoring predicted retrosynthesis route for ozanimod.** Orange-coloured compounds are commercially available. Except for some of the commercial precursors that were present but involved in different reactions, none of the intermediate compounds were present in the training dataset. The reaction prediction numbers in bold on retrosynthesis arrows correspond to the order in which the multistep tree search prioritized the prediction. Forward prediction confidence scores are shown under retrosynthesis arrows. Predicted reagents: a') HCl, dioxane; b') ZnCl$_2$, AcOEt, toluene; c') HCl, *t*-BuOK, THF.

Figure B.14: **Best overall confidence score predicted retrosynthesis route for ozanimod.** Orange-coloured compounds are commercially available. Except for some of the commercial precursors that were present but involved in different reactions, none of the intermediate compounds were present in the training dataset. The reaction prediction numbers in bold on retrosynthesis arrows correspond to the order in which the multistep tree search prioritized the prediction. Forward prediction confidence scores are shown under retrosynthesis arrows. Predicted reagents: a") 1-Methylpyrrolidin-2-one; b") no reagent predicted; c") HCl, Et$_2$O; d") HCl, NaHCO$_3$, EtOH; e") HCl, *t*-BuOK, THF.

Figure B.15: **Set of commercially available precursors of all solved routes for fostemsavir.** All building blocks of the literature reported retrosynthesis are highlighted in blue.
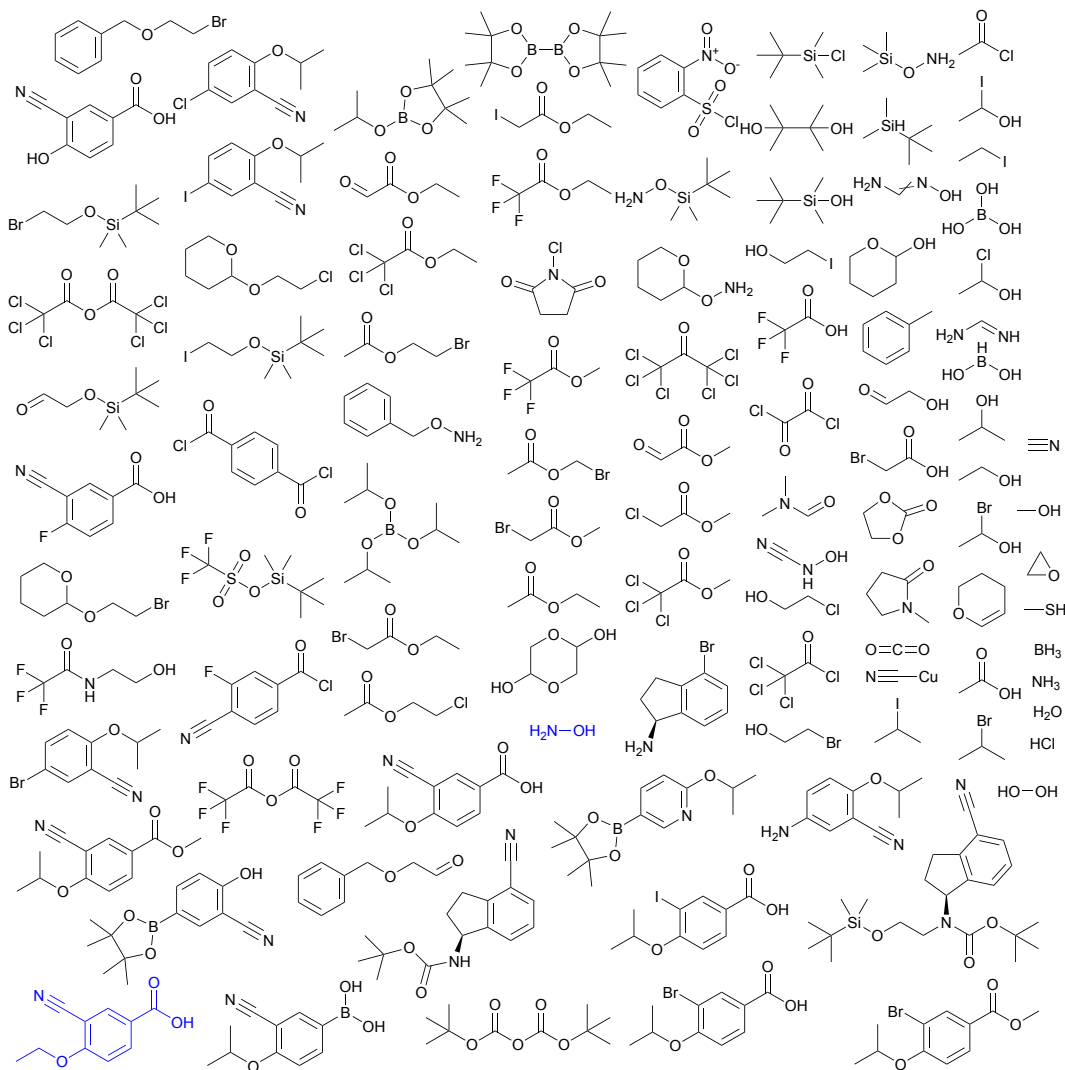
Figure B.16: **Set of commercially available precursors of all solved routes for ozanimod.** Some of the building blocks of the literature reported retrosynthesis are highlighted in blue.
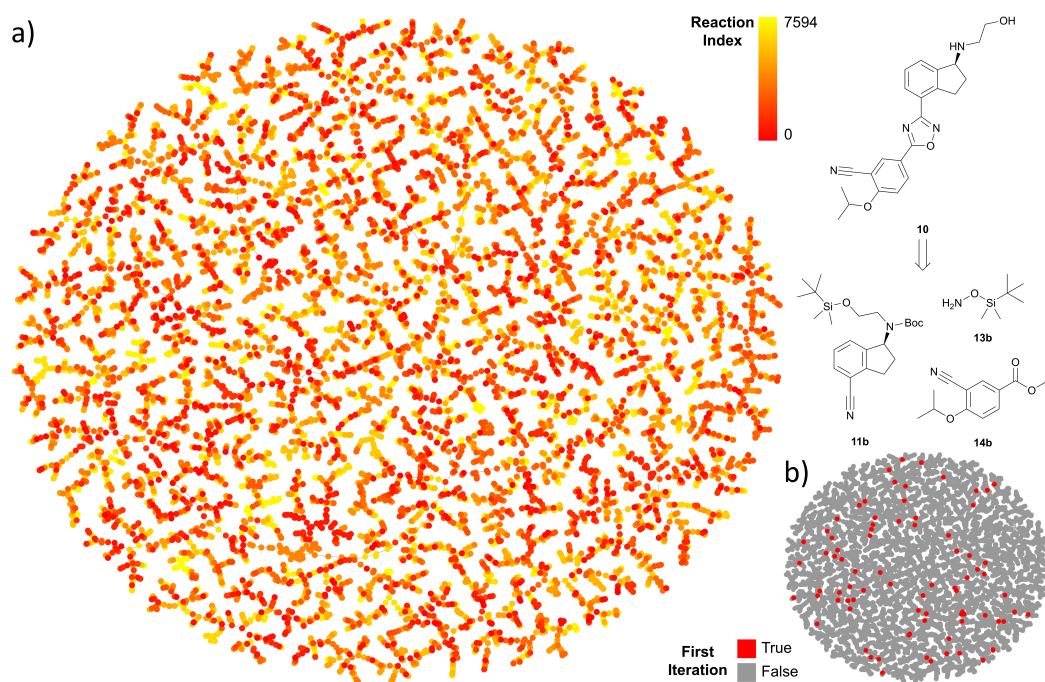
Figure B.17: **TMAP representation of iterated predictions for the multistep search of fostemsavir.** **(a)** Predicted reactions from the target molecule (low indexes) to end nodes. **(b)** Highlighted first iteration of the TTLA search. Interactive map available at https://tm.gdb.tools/TTLA/fostemsavir.

B.3 Route predicted by AiZynthFinder
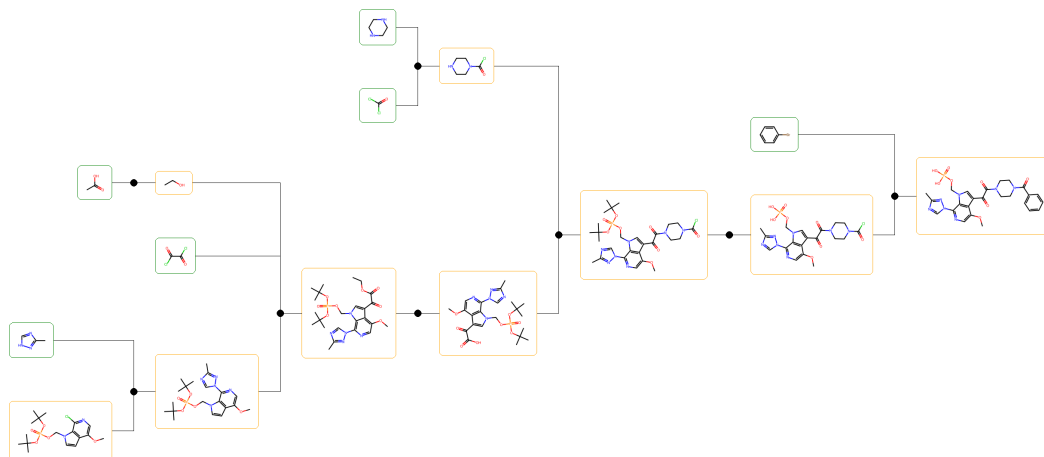
Figure B.18: **Fostemsavir retrosynthesis route predicted by AiZynthFinder (v3.7.0).**
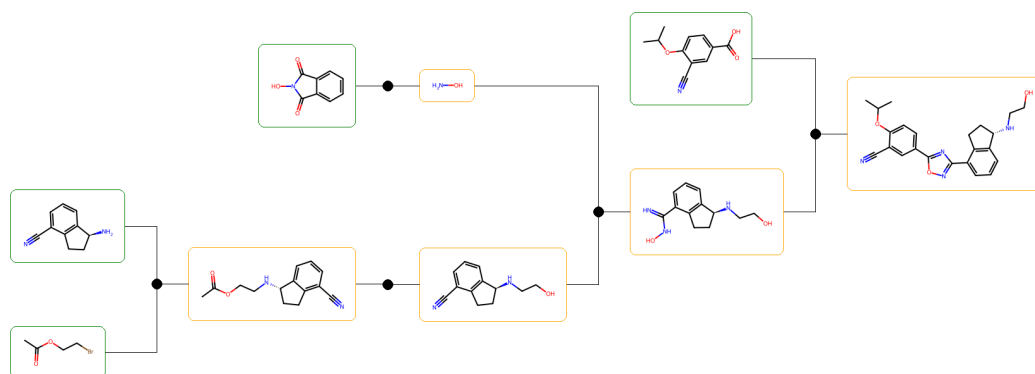
Figure B.19: **Ozanimod retrosynthesis route predicted by AiZynthFinder (v3.7.0).**

B.4 Route predicted by IBM RXN for Chemistry
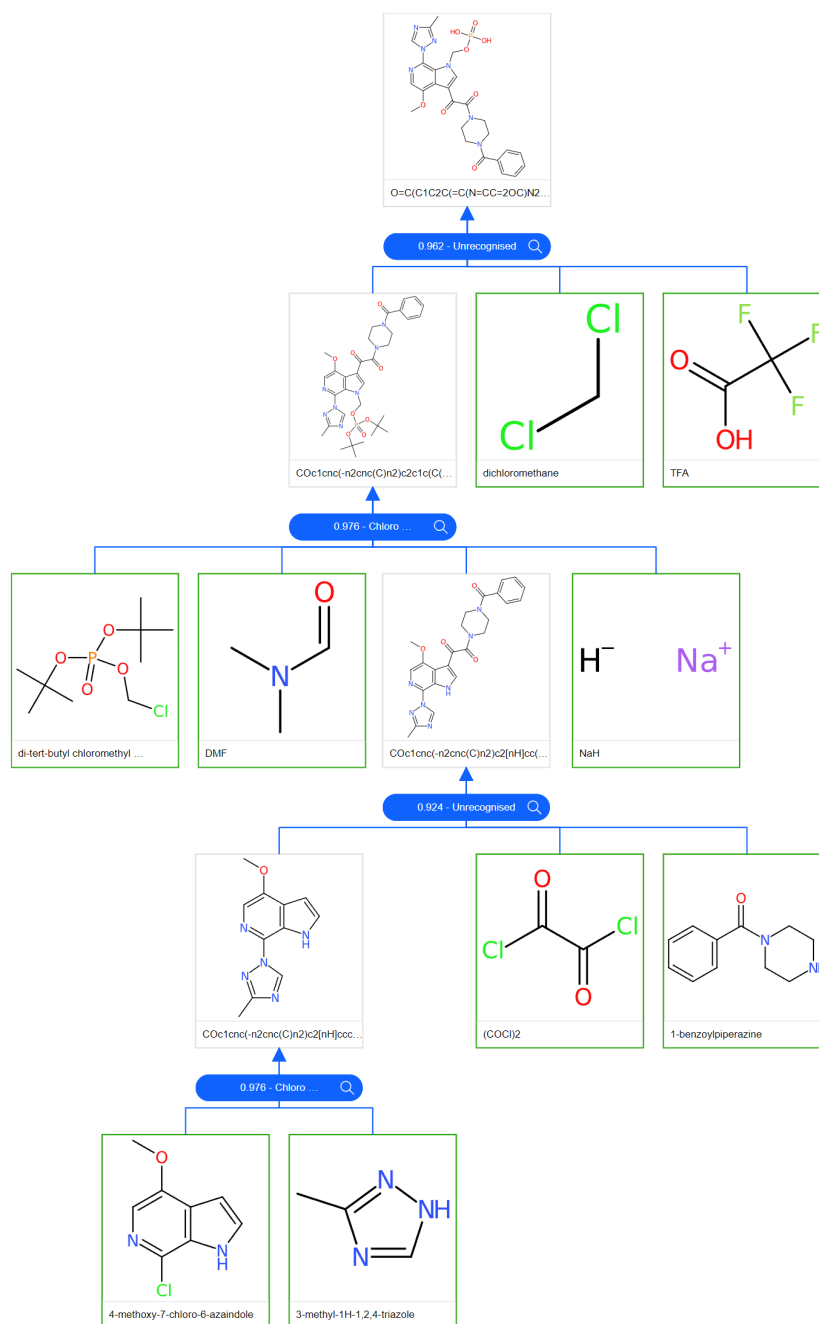
Figure B.20: **Fostemsavir retrosynthesis route predicted by IBM RXN for Chemistry user interface** using the default "12class-tokens-2021-05-14" models, with highest quality tuning, and excluding commercially similar compounds as in our route prediction settings.
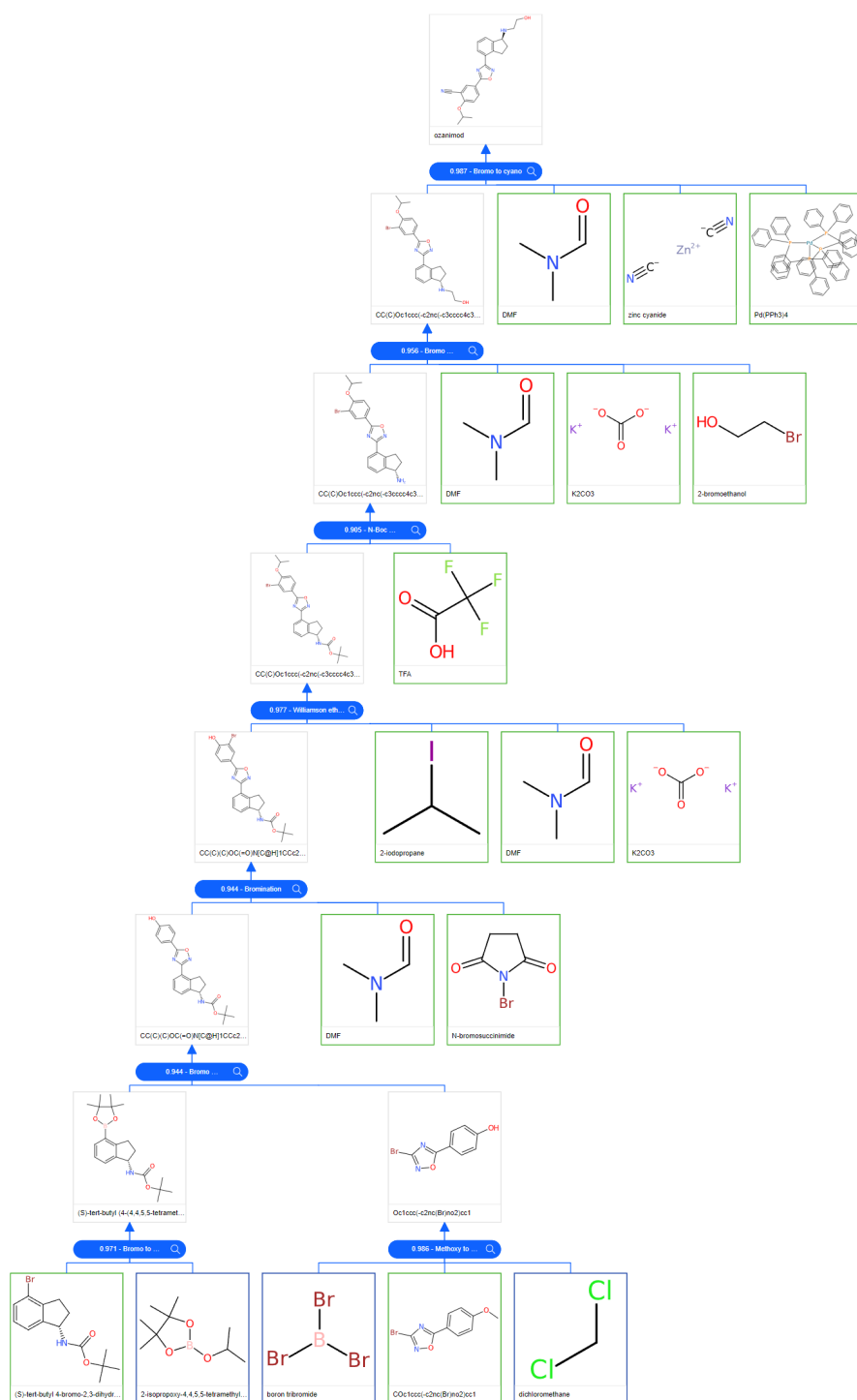
Figure B.21: **Ozanimod retrosynthesis route predicted by IBM RXN for Chemistry user interface** using the default "12class-tokens-2021-05-14" models, with highest quality tuning, and excluding commercially similar compounds as in our route prediction settings.

B.5 Benchmark routes

Target SMILES: CCCCc1nn(-c2cc(NC(=O)CC)ccc2Cl)c(C#N)c1Cc1ccc(Br)cc1F
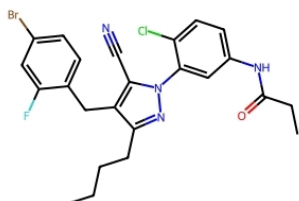
Overall forward confidence score = 0.6502
Overall Guiding RPScore = 0.025
Overall Penalties = 0.0938
Number of steps = 5

Best RPScore route:
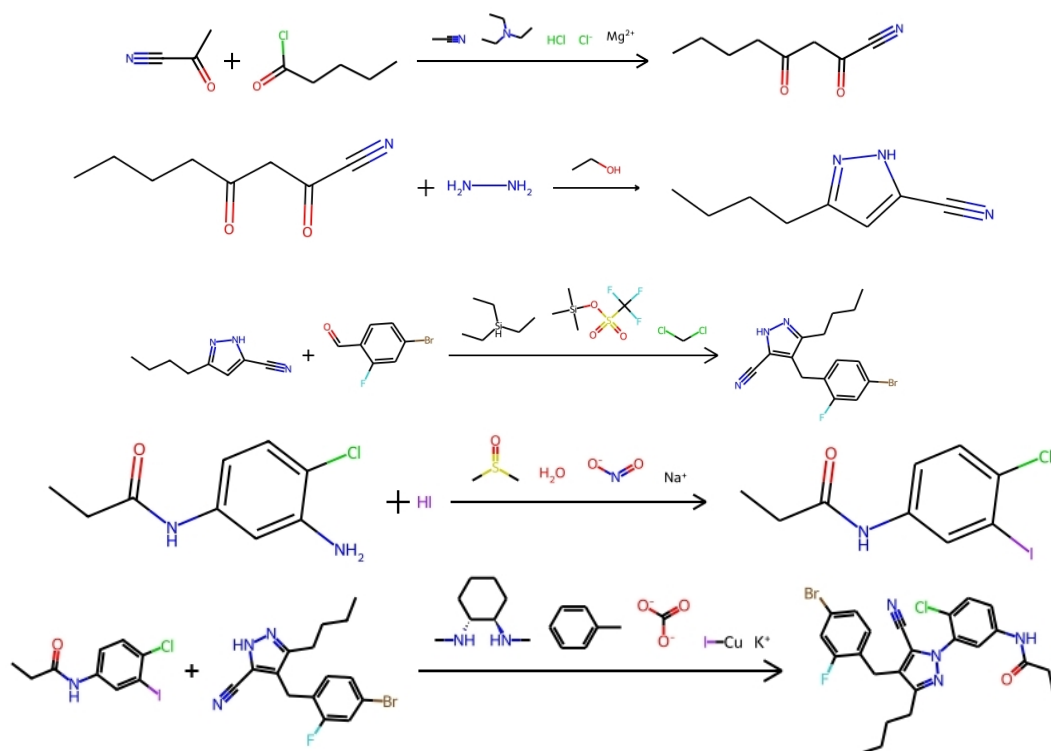
Figure B.22: **Best RPScoring predicted route by our TTLA, example 1.** Target molecule selected from the benchmark of Genheden *et al.*,[231] see main text.

Target SMILES: CCOC(=O)N1CCc2c(c3cccc(N)c3n2C)C1

Overall forward confidence score = 0.7725
Overall Guiding RPScore = 0.2858
Overall Penalties = 0.4625
Number of steps = 2
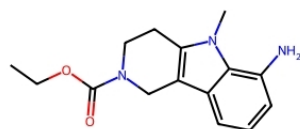
Best RPScore route:
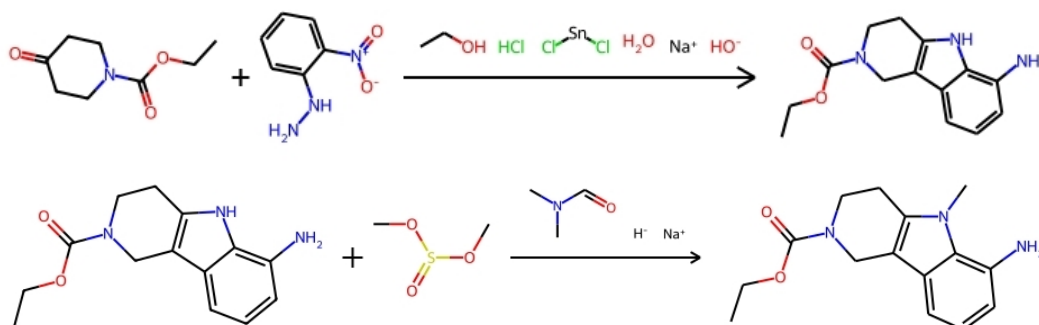
Figure B.23: **Best RPScoring predicted route by our TTLA, example 2.** Target molecule selected from the benchmark of Genheden *et al.*,[231] see main text.

Target SMILES: COc1ccc(CCNc2nccc(-c3cccc(NC(C)=O)c3)n2)cc1OC

Overall forward confidence score = 0.8665
Overall Guiding RPScore = 0.4439
Overall Penalties = 0.6403
Number of steps = 2
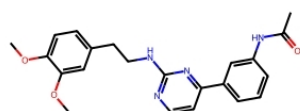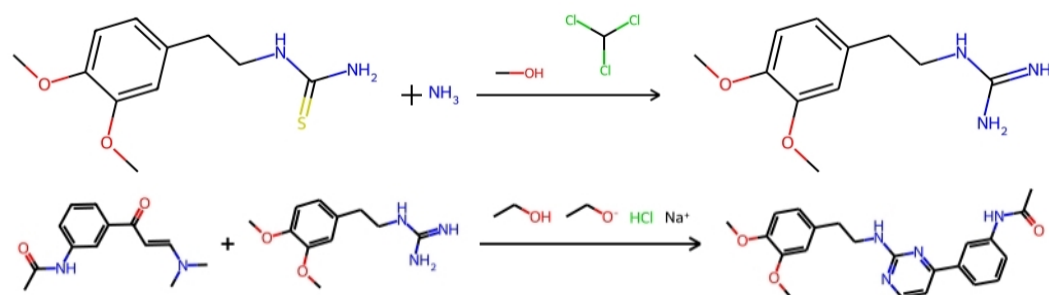
Best RPScore route:

Figure B.24: **Best RPScoring predicted route by our TTLA, example 3.** Target molecule selected from the benchmark of Genheden *et al.*,[231] see main text.

Target SMILES: CCCCc1nc(-c2ccc(F)cc2)c(C(=O)N(CC)CC)n1Cc1ccc(-c2ccccc2C#N)cc1
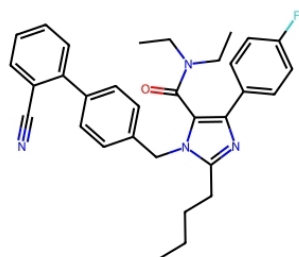
Overall forward confidence score = 0.8949
Overall Guiding RPScore = 0.0795
Overall Penalties = 0.1389
Number of steps = 3

Best RPScore route:
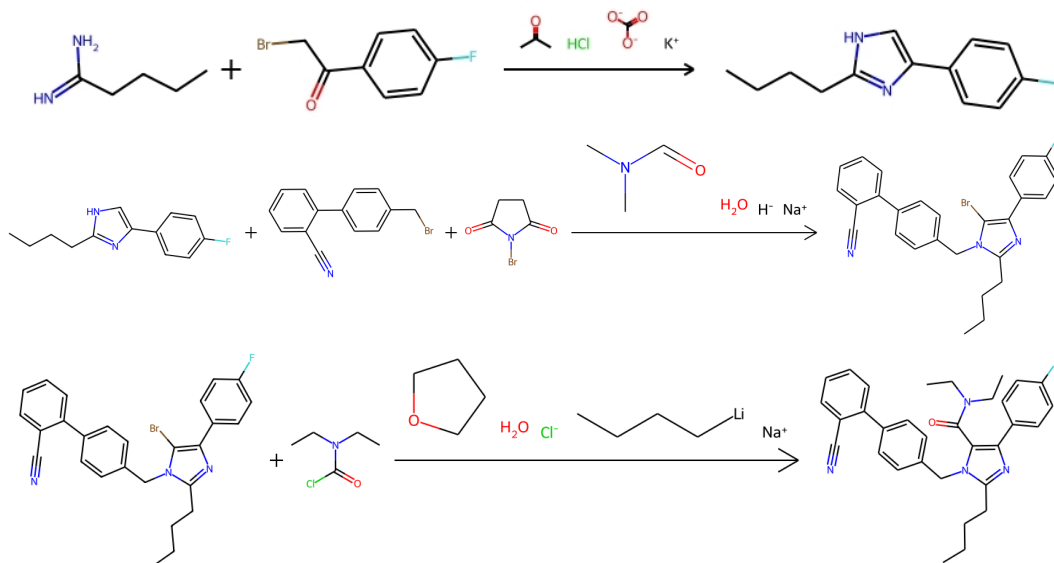
Figure B.25: **Best RPScoring predicted route by our TTLA, example 4.** Target molecule selected from the benchmark of Genheden *et al.*,[231] see main text.

Target SMILES: O=C(O)C1CN(Cc2ccc(-c3noc(CCC4(c5ccccc5)CCCCC4)n3)cc2)C1

Overall forward confidence score = 0.7397
Overall Guiding RPScore = 0.1437
Overall Penalties = 0.3035
Number of steps = 3
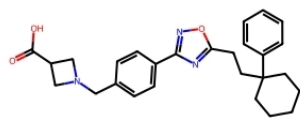
Best RPScore route:
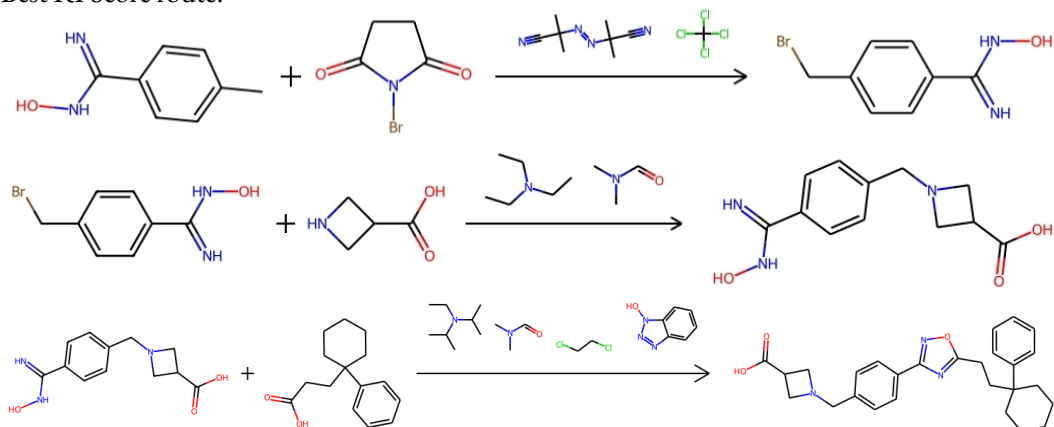
Figure B.26: **Best RPScoring predicted route by our TTLA, example 5.** Target molecule selected from the benchmark of Genheden *et al.*,[231] see main text.

Target SMILES: Cc1ccc(-n2ccc(C(F)(F)F)n2)nc1Cl

Overall forward confidence score = 0.9783
Overall Guiding RPScore = 0.6513
Overall Penalties = 0.8323
Number of steps = 2
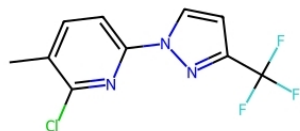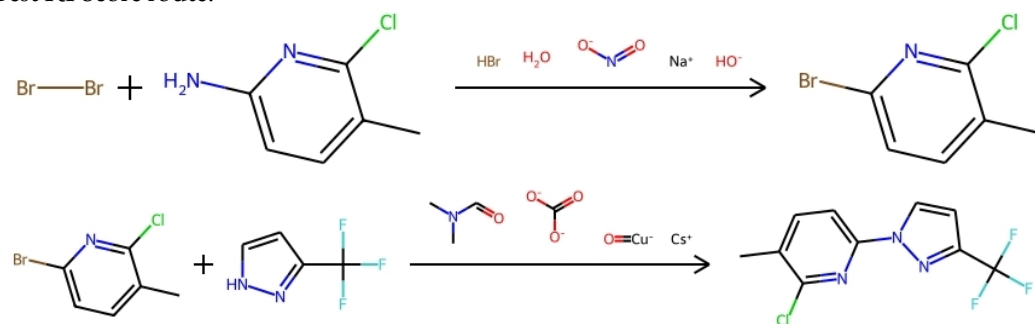
Best RPScore route:

Figure B.27: **Best RPScoring predicted route by our TTLA, example 6.** Target molecule selected from the benchmark of Genheden *et al.*,[231] see main text.

Target SMILES: O=C(Nc1ccc2c(CN3CCCC3)cn(Cc3c(Cl)cccc3Cl)c2c1)Oc1ccccc1

Overall forward confidence score = 0.5652
Overall Guiding RPScore = 0.1587
Overall Penalties = 0.351
Number of steps = 2
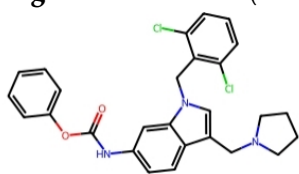
Best RPScore route:
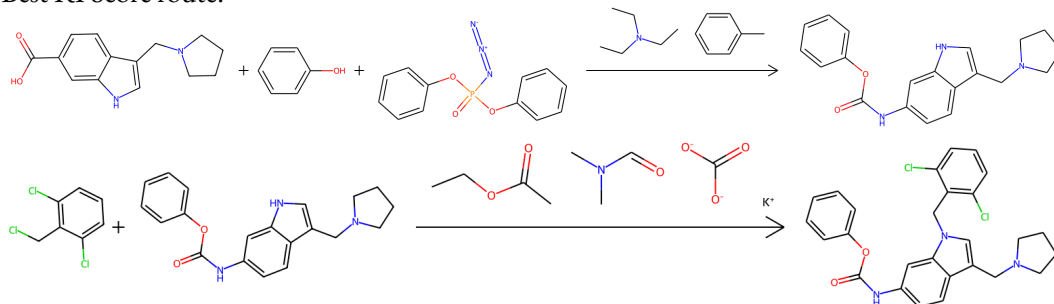
Figure B.28: **Best RPScoring predicted route by our TTLA, example 7.** Target molecule selected from the benchmark of Genheden *et al.*,[231] see main text.

Target SMILES: c1ccc2oc(-c3ccc4c(c3)[nH]c3ccccc34)nc2c1

Overall forward confidence score = 0.8932
Overall Guiding RPScore = 0.8932
Overall Penalties = 1.0
Number of steps = 1
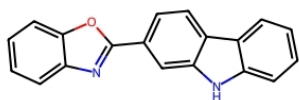
Best RPScore route:
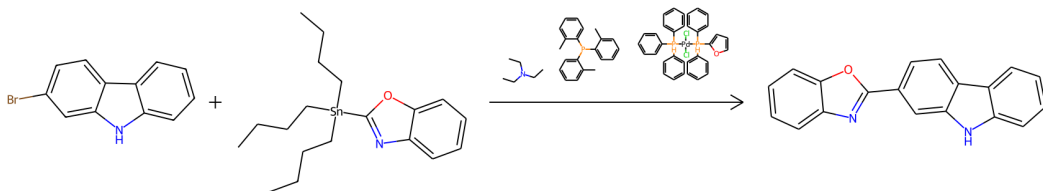
Figure B.29: **Best RPScoring predicted route by our TTLA, example 8.** Target molecule selected from the benchmark of Genheden *et al.*,[231] see main text.

Target SMILES: COc1ccc(C2(c3cccc(OCCCF)c3)N=C(N)c3ncccc32)cc1C(F)(F)F
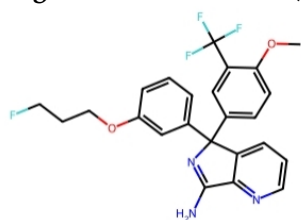
Overall forward confidence score = 0.4259
Overall Guiding RPScore = 0.0892
Overall Penalties = 0.3272
Number of steps = 3

Best RPScore route:
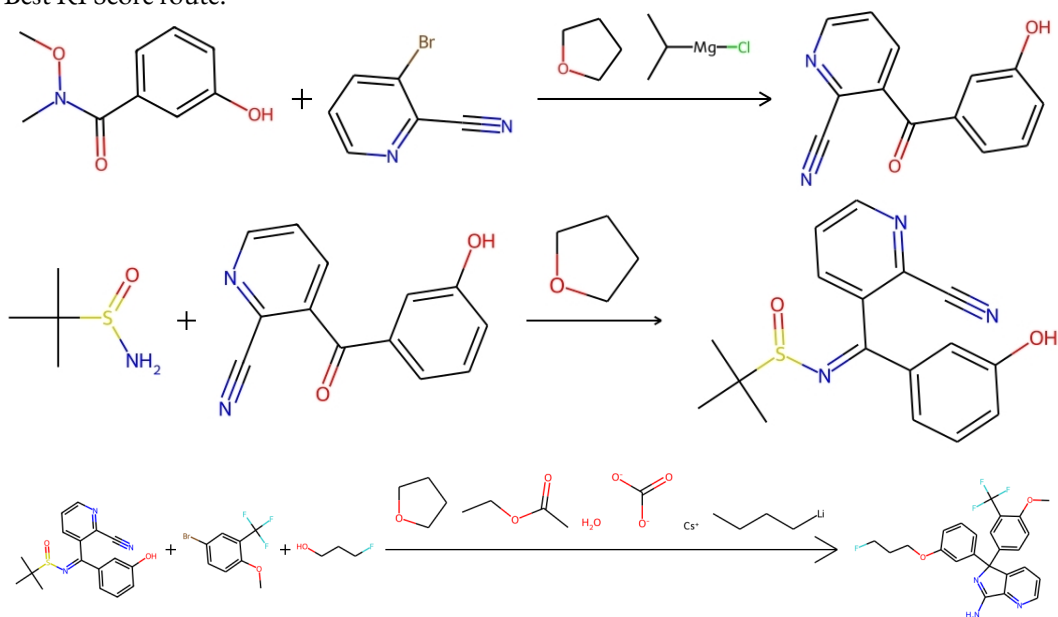
Figure B.30: **Best RPScoring predicted route by our TTLA, example 9.** Target molecule selected from the benchmark of Genheden *et al.*,[231] see main text.

Target SMILES: CS(=O)(=O)c1ccc(C2=C(c3ccc(-c4ccsc4)cc3)C(=O)OC2)cc1
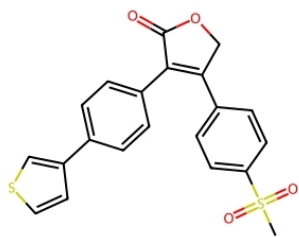
Overall forward confidence score = 0.7994
Overall Guiding RPScore = 0.3674
Overall Penalties = 0.5744
Number of steps = 2

Best RPScore route:
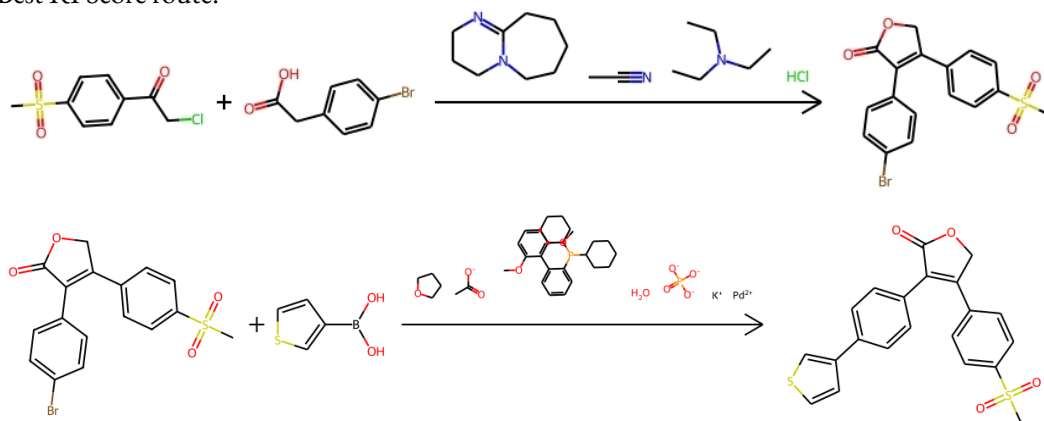
Figure B.31: **Best RPScoring predicted route by our TTLA, example 10.** Target molecule selected from the benchmark of Genheden *et al.*,[231] see main text.

Abbreviations

AI	Artificial intelligence
CALB	*Candida antarctica* lipase B
CASP	Computer-aided synthesis planning
CHMTRN	CHeMistry TRaNslator
EC	Enzyme commission
GUI	Graphical User Interface
LHASA	Logic and Heuristics Applied to Synthetic Analysis
MCTS	Monte Carlo tree search
NLP	Natural language processing
OCSS	Organic Chemical Simulation of Synthesis
P	Product
PATRAN	PAttern TRANslator
R	Reagent
SCScore	Synthetic complexity score
SECS	Simulation and evaluation of chemical synthesis
SM	Starting material
SMILES	Simplified molecular-input line-entry system
TMAP	Tree Map
TTL	Triple Transformer Loop
TTLA	Multistep Triple Transformer Loop Algorithm
USPTO	United States Patent and Trademark Office

Bibliography

1. E. J. Corey. "General Methods for the Construction of Complex Molecules". *Pure and Applied Chemistry* 14:1, 1967, pp. 19–38.

2. E. J. Corey and W. T. Wipke. "Computer-Assisted Design of Complex Organic Syntheses". *Science (New York, N.Y.)* 166:3902, 1969, pp. 178–192.

3. D. A. PENSAK and E. J. COREY. "LHASA—Logic and Heuristics Applied to Synthetic Analysis". In: *Computer-Assisted Organic Synthesis*. Vol. 61. 0 vols. ACS Symposium Series 61. AMERICAN CHEMICAL SOCIETY, 1977, pp. 1–32.

4. F. H. Arnold. "Design by Directed Evolution". *Accounts of Chemical Research* 31:3, 1998, pp. 125–131.

5. E. L. Bell, W. Finnigan, S. P. France, A. P. Green, M. A. Hayes, L. J. Hepworth, S. L. Lovelock, H. Niikura, S. Osuna, E. Romero, K. S. Ryan, N. J. Turner, and S. L. Flitsch. "Biocatalysis". *Nature Reviews Methods Primers* 1:1, 1 2021, pp. 1–21.

6. F. Gallou, H. Gröger, and B. H. Lipshutz. "Status Check: Biocatalysis; It's Use with and without Chemocatalysis. How Does the Fine Chemicals Industry View This Area?" *Green Chemistry* 25:16, 2023, pp. 6092–6107.

7. A. Kekulé. "Sur La Constitution Des Substances Aromatiques". *Bulletin de la Société Chimique de Paris* 3:2, 1865, pp. 98–110.

8. A. S. Couper. "Sur Une Nouvelle Théorie Chimique". *Comptes rendus hebdomadaires des séances de l'Académie des sciences* 46, 1858, pp. 1157–1160.

9. J. Cribb. *Earth Detox: How and Why We Must Clean Up Our Planet*. Cambridge University Press, 2021. 331 pp. Google Books: `hfIuEAAAQBAJ`.

10. J. H. Clark. "Green Chemistry: Challenges and Opportunities". *Green Chemistry* 1:1, 1999, pp. 1–8.

11. J. B. Zimmerman, P. T. Anastas, H. C. Erythropel, and W. Leitner. "Designing for a Green Chemistry Future". *Science* 367:6476, 2020, pp. 397–400.

12. R. A. Sheldon and D. Brady. "Broadening the Scope of Biocatalysis in Sustainable Organic Synthesis". *ChemSusChem* 12:13, 2019, pp. 2859–2881.

13. S. Wu, R. Snajdrova, J. C. Moore, K. Baldenius, and U. T. Bornscheuer. "Biocatalysis: Enzymatic Synthesis for Industrial Applications". *Angewandte Chemie International Edition* 60:1, 2021, pp. 88–119.

14. D. Rogers and M. Hahn. "Extended-Connectivity Fingerprints". *Journal of Chemical Information and Modeling* 50:5, 2010, pp. 742–754.

15. D. Probst and J.-L. Reymond. "A Probabilistic Molecular Fingerprint for Big Data Settings". *Journal of Cheminformatics* 10:1, 2018, p. 66.

16. A. Capecchi, D. Probst, and J.-L. Reymond. "One Molecular Fingerprint to Rule Them All: Drugs, Biomolecules, and the Metabolome". *Journal of Cheminformatics* 12:1, 2020, p. 43.

17. R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, and A. Aspuru-Guzik. "Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules". *ACS Central Science* 4:2, 2018, pp. 268–276.

18. D. P. Kingma and M. Welling. "Auto-Encoding Variational Bayes. Paper Presented at the 2nd International Conference on Learning Representations, ICLR 2014–Conference Track Proceedings", 2014.

19. 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". *Nature Machine Intelligence* 3, 2, 2 2021, pp. 144–152.

20. *THE WISWESSER LINE-NOTATION: AN INTRODUCTION.*

21. S. Heller, A. McNaught, S. Stein, D. Tchekhovskoi, and I. Pletnev. "InChI - the Worldwide Chemical Structure Identifier Standard". *Journal of Cheminformatics* 5:1, 2013, p. 7.

22. A. Dalby, J. G. Nourse, W. D. Hounshell, A. K. I. 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". *Journal of Chemical Information and Computer Sciences* 32:3, 1992, pp. 244–255.

23. L. David, A. Thakkar, R. Mercado, and O. Engkvist. "Molecular Representations in AI-driven Drug Discovery: A Review and Practical Guide". *Journal of Cheminformatics* 12:1, 2020, p. 56.

24. D. S. Wigh, J. M. Goodman, and A. A. Lapkin. "A Review of Molecular Representation in the Age of Machine Learning". *WIREs Computational Molecular Science* 12:5, 2022, e1603.

25. D. Weininger. "SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules". *Journal of Chemical Information and Computer Sciences* 28:1, 1988, pp. 31–36.

26. D. Weininger, A. Weininger, and J. L. Weininger. "SMILES. 2. Algorithm for Generation of Unique SMILES Notation". *Journal of Chemical Information and Computer Sciences* 29:2, 1989, pp. 97–101.

27. E. J. Bjerrum. "SMILES Enumeration as Data Augmentation for Neural Network Modeling of Molecules". *CoRR* abs/1703.07076, 2017. arXiv: `1703.07076`.

28. T. B. Kimber, S. Engelke, I. V. Tetko, E. Bruno, and G. Godin. "Synergy Effect between Convolutional Neural Networks and the Multiplicity of SMILES for Improvement of Molecular Prediction". *CoRR* abs/1812.04439, 2018. arXiv: `1812.04439`.

29. N. M. O'Boyle. "Towards a Universal SMILES Representation - A Standard Method to Generate Canonical SMILES Based on the InChI". *Journal of Cheminformatics* 4:1, 2012, p. 22.

30. G. Landrum. "RDKit: Open-source Cheminformatics", 2006.

31. A. H. Thomas. "Directness and Convenience for the Reading Approach". *The German Quarterly* 9:3, 1936, pp. 109–120. JSTOR: `400358`.

32. *Name Reactions*. Springer, Berlin, Heidelberg, 2006.

33. K. C. Nicolaou, Z. Yang, J. J. Liu, H. Ueno, P. G. Nantermet, R. K. Guy, C. F. Claiborne, J. Renaud, E. A. Couladouros, K. Paulvannan, and E. J. Sorensen. "Total Synthesis of Taxol". *Nature* 367:6464, 6464 1994, pp. 630–634.

34. R. B. Woodward. "The total synthesis of vitamin B12". *Pure and Applied Chemistry* 33:1, 1973, pp. 145–178.

35. K. C. Nicolaou, D. Vourloumis, N. Winssinger, and P. S. Baran. "The Art and Science of Total Synthesis at the Dawn of the Twenty-First Century". *Angewandte Chemie International Edition* 39:1, 2000, pp. 44–122.

36. 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". *Nature* 529:7587, 7587 2016, pp. 484–489.

37. E. J. Corey, W. T. Wipke, R. D. I. Cramer, and W. J. Howe. "Computer-Assisted Synthetic Analysis. Facile Man-Machine Communication of Chemical Structure by Interactive Computer Graphics". *Journal of the American Chemical Society* 94:2, 1972, pp. 421–430.

38. E. J. Corey, A. K. Long, and S. D. Rubenstein. "Computer-Assisted Analysis in Organic Synthesis". *Science (New York, N.Y.)* 228:4698, 1985, pp. 408–418.

39. E. J. Corey. *The Logic of Chemical Synthesis*. 1991. 447 pp. Google Books: `0YIVAwAAQBAJ`.

40. W. T. Wipke, G. I. Ouchi, and S. Krishnan. "Simulation and Evaluation of Chemical Synthesis—SECS: An Application of Artificial Intelligence Techniques". *Artificial Intelligence*. Applications to the Sciences and Medicine 11:1, 1978, pp. 173–193.

41. P. Y. Johnson, I. Burnstein, J. Crary, M. Evans, and T. Wang. "Designing an Expert System for Organic Synthesis". In: *Expert System Applications in Chemistry*. Vol. 408. 0 vols. ACS Symposium Series 408. American Chemical Society, 1989, pp. 102–123.

42. 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". *Science* 197:4308, 1977, pp. 1041–1049.

43. K. K. Agarwal, T. D. L. Larsen, and H. L. Gelernter. "Application of Chemical Transforms in Synchem2, a Computer Program for Organic Synthesis Route Discovery". *Computers & Chemistry* 2:2, 1978, pp. 75–84.

44. J. B. Hendrickson and A. G. Toczko. "SYNGEN Program for Synthesis Design: Basic Computing Techniques". *Journal of Chemical Information and Computer Sciences* 29:3, 1989, pp. 137–145.

45. G. Mehta, P. Azario, R. Barone, and M. Chanon. "How to Search Original Key Steps in a Synthesis of Complex Structures by Using a Microcomputer as a Chemical Pocket-like Calculator". *Tetrahedron* 45:7, 1989, pp. 1985–1994.

46. F. Barberis, R. Barone, M. Arbelot, A. Baldy, and M. Chanon. "CONAN (CONnectivity ANalysis): A Simple Approach in the Field of Computer-Aided Organic Synthesis. Example of the Taxane Framework". *Journal of Chemical Information and Computer Sciences* 35:3, 1995, pp. 467–471.

47. W.-D. Ihlenfeldt and J. Gasteiger. "Computer-Assisted Planning of Organic Syntheses: The Second Generation of Programs". *Angewandte Chemie International Edition in English* 34:23-24, 1996, pp. 2613–2633.

48. J. Blair, J. Gasteiger, C. Gillespie, P. D. Gillespie, and I. Ugi. "CICLOPS: A Computer Program for the Design of Syntheses on the Basis of a Mathematical Model". *Computer Representation and Manipulation of Chemical Information. New York, NY: John Wiley and Sons* 1974, 1974, pp. 137–139.

49. J. Gasteiger and C. Jochum. "EROS A Computer Program for Generating Sequences of Reactions". In: *Organic Compunds*. Topics in Current Chemistry. Springer, Berlin, Heidelberg, 1978, pp. 93–126.

50. J. Gasteiger and W. D. Ihlenfeldt. "The WODCA System". In: *Software Development in Chemistry 4*. Ed. by J. Gasteiger. Springer, Berlin, Heidelberg, 1990, pp. 57–65.

51. 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 and Applied Chemistry* 62:10, 1990, pp. 1921–1932.

52. H. Satoh and K. Funatsu. "SOPHIA, a Knowledge Base-Guided Reaction Prediction System - Utilization of a Knowledge Base Derived from a Reaction Database". *Journal of Chemical Information and Computer Sciences* 35:1, 1995, pp. 34–44.

53. P. Judson. "Knowledge-Based Expert Systems in Chemistry : Not Counting on Computers". In: 2009.

54. 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*. Ed. by W. A. Warr. Springer, Berlin, Heidelberg, 1988, pp. 303–313.

55. A. P. Johnson, K. Burt, A. P. F. Cook, K. M. Higgins, G. A. Hopkinson, and G. Singh. "Integration and Standards: Use of a Host Language Interface". In: *Chemical Structure Information Systems*. Vol. 400. 0 vols. ACS Symposium Series 400. American Chemical Society, 1989, pp. 50–58.

56. D. Chodosh and W. L. Mendelson. "SYNthesis LIBrary — Graphics at the Bench". *Drug Information Journal* 17:4, 1983, pp. 231–238.

57. A. Barth. "Status and Future Developments of Reaction Databases and Online Retrieval Systems". *Journal of Chemical Information and Computer Sciences* 30:4, 1990, pp. 384–393.

58. J. E. Blake and R. C. Dana. "CASREACT: More than a Million Reactions". *Journal of Chemical Information and Computer Sciences* 30:4, 1990, pp. 394–399.

59. A. J. Lawson. "The Beilstein Database". In: *Handbook of Chemoinformatics*. John Wiley & Sons, Ltd, 2003, pp. 608–628.

60. A. Parlow, C. Weiske, and J. Gasteiger. "ChemInform - an Integrated Information System on Chemical Reactions". *Journal of Chemical Information and Computer Sciences* 30:4, 1990, pp. 400–402.

61. W. A. Warr. "A Short Review of Chemical Reaction Database Systems, Computer-Aided Synthesis Design, Reaction Prediction and Synthetic Feasibility". *Molecular Informatics* 33:6-7, 2014, pp. 469–476.

62. "SciFinder: A New Generation of Research Tool",

63. S. W. Gabrielson. "SciFinder". *Journal of the Medical Library Association* 106:4, 4 2018, pp. 588–590.

64. A. J. Lawson, J. Swienty-Busch, T. Géoui, and D. Evans. "The Making of Reaxys—Towards Unobstructed Access to Relevant Chemistry Information". In: *The Future of the History of Chemical Information*. Vol. 1164. ACS Symposium Series 1164. American Chemical Society, 2014, pp. 127–148.

65. M. Rubacha, A. K. Rattan, and S. C. Hosselet. "A Review of Electronic Laboratory Notebooks Available in the Market Today". *JALA: Journal of the Association for Laboratory Automation* 16:1, 2011, pp. 90–98.

66. W. Heyndrickx, L. Mervin, T. Morawietz, N. Sturm, L. Friedrich, A. Zalewski, A. Pentina, L. Humbeck, M. Oldenhof, R. Niwayama, P. Schmidtke, N. Fechner, J. Simm, A. Arany, N. Drizard, R. Jabal, A. Afanasyeva, R. Loeb, S. Verma, S. Harnqvist, M. Holmes, B. Pejo, M. Telenczuk, N. Holway, A. Dieckmann, N. Rieke, F. Zumsande, D.-A. Clevert, M. Krug, C. Luscombe, D. Green, P. Ertl, P. Antal, D. Marcus, N. Do Huu, H. Fuji, S. Pickett, G. Acs, E. Boniface, B. Beck, Y. Sun, A. Gohier, F. Rippmann, O. Engkvist, A. H. Göller, Y. Moreau, M. N. Galtier, A. Schuffenhauer, and H. Ceulemans. "MELLODDY: Cross-pharma Federated Learning at Unprecedented Scale Unlocks Benefits in QSAR without Compromising Proprietary Information". *Journal of Chemical Information and Modeling*, 2023.

67. L. C. Ray and R. A. Kirsch. "Finding Chemical Records by Digital Computers". *Science (New York, N.Y.)* 126:3278, 1957, pp. 814–819.

68. D. M. Lowe. "Extraction of Chemical Structures and Reactions from the Literature". Thesis. University of Cambridge, 2012.

69. D. M. Lowe. *Chemical Reactions from US Patents (1976-Sep2016)*. 2017.

70. "NextMove Software | Pistachio",

71. 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". *Angewandte Chemie International Edition* 55:20, 2016, pp. 5904–5937.

72. 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.

73. M. Fialkowski, K. J. M. Bishop, V. A. Chubukov, C. J. Campbell, and B. A. Grzybowski. "Architecture and Evolution of Organic Chemistry". *Angewandte Chemie International Edition* 44:44, 2005, pp. 7263–7269.

74. M. Kowalik, C. M. Gothard, A. M. Drews, N. A. Gothard, A. Weckiewicz, P. E. Fuller, B. A. Grzybowski, and K. J. M. Bishop. "Parallel Optimization of Synthetic Pathways within the Network of Organic Chemistry". *Angewandte Chemie International Edition* 51:32, 2012, pp. 7928–7932.

75. 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.

76. K. Molga, E. P. Gajewska, S. Szymkuć, and B. A. Grzybowski. "The Logic of Translating Chemical Knowledge into Machine-Processable Forms: A Modern Playground for Physical-Organic Chemistry". *Reaction Chemistry & Engineering* 4:9, 2019, pp. 1506–1521.

77. E. S. Blurock. "Computer-Aided Synthesis Design at RISC-Linz: Automatic Extraction and Use of Reaction Classes". *Journal of Chemical Information and Computer Sciences* 30:4, 1990, pp. 505–510.

78. K. Satoh and K. Funatsu. "A Novel Approach to Retrosynthetic Analysis Using Knowledge Bases Derived from Reaction Databases". *Journal of Chemical Information and Computer Sciences* 39:2, 1999, pp. 316–325.

79. C. D. Christ, M. Zentgraf, and J. M. Kriegl. "Mining Electronic Laboratory Notebooks: Analysis, Retrosynthesis, and Reaction Based Enumeration". *Journal of Chemical Information and Modeling* 52:7, 2012, pp. 1745–1756.

80. P. P. Plehiers, G. B. Marin, C. V. Stevens, and K. M. Van Geem. "Automated Reaction Database and Reaction Network Analysis: Extraction of Reaction Templates Using Chem-informatics". *Journal of Cheminformatics* 10:1, 2018, p. 11.

81. 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". *Journal of Chemical Information and Modeling* 49:3, 2009, pp. 593–602.

82. A. Bøgevig, H.-J. Federsel, F. Huerta, M. G. Hutchings, H. Kraut, T. Langer, P. Löw, C. Oppawsky, T. Rein, and H. Saller. "Route Design in the 21st Century: The ICSYNTH Software Tool as an Idea Generator for Synthesis Prediction". *Organic Process Research & Development* 19:2, 2015, pp. 357–368.

83. M. H. S. Segler and M. P. Waller. "Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction". *Chemistry – A European Journal* 23:25, 2017, pp. 5966–5971.

84. M. H. S. Segler, M. Preuss, and M. P. Waller. "Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI". *Nature* 555, 7698, 7698 2018, pp. 604–610.

85. C. W. Coley, R. Barzilay, T. S. Jaakkola, W. H. Green, and K. F. Jensen. "Prediction of Organic Reaction Outcomes Using Machine Learning". *ACS Central Science* 3:5, 2017, pp. 434–443.

86. 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". *Chemical Science* 10:2, 2019, pp. 370–377.

87. C. W. Coley, D. A. Thomas, J. A. M. Lummiss, J. N. Jaworski, C. P. Breen, V. Schultz, T. Hart, J. S. Fishman, L. Rogers, H. Gao, R. W. Hicklin, P. P. Plehiers, J. Byington, J. S. Piotti, W. H. Green, A. J. Hart, T. F. Jamison, and K. F. Jensen. "A Robotic Platform for Flow Synthesis of Organic Compounds Informed by AI Planning". *Science* 365:6453, 2019, eaax1566.

88. 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". *Journal of Cheminformatics* 12:1, 2020, p. 70.

89. H. Dai, C. Li, C. W. Coley, B. Dai, and L. Song. "Retrosynthesis Prediction with Conditional Graph Logic Network". arxiv: `2001.01408` (cs, stat), 2020.

90. J. Bauer, E. Fontain, D. Forstmeyer, and I. Ugi. "Interactive Generation of Organic Reactions by IGOR 2 and the PC-assisted Discovery of a New Reaction". *Tetrahedron Computer Methodology* 1:2, 1988, pp. 129–132.

91. P. Röse and J. Gasteiger. "Automated Derivation of Reaction Rules for the EROS 6.0 System for Reaction Prediction". *Analytica Chimica Acta* 235, 1990, pp. 163–168.

92. I. Ugi, J. Bauer, C. Blomberger, J. Brandt, A. Dietz, E. Fontain, B. Gruber, A. v. Scholley-Pfab, A. Senff, and N. Stein. "Models, Concepts, Theories, and Formal Languages in Chemistry and Their Use as a Basis for Computer Assistance in Chemistry". *Journal of Chemical Information and Computer Sciences* 34:1, 1994, pp. 3–16.

93. M. A. Kayala and P. Baldi. "ReactionPredictor: Prediction of Complex Chemical Reactions at the Mechanistic Level Using Machine Learning". *Journal of Chemical Information and Modeling* 52:10, 2012, pp. 2526–2540.

94. J. N. Wei, D. Duvenaud, and A. Aspuru-Guzik. "Neural Networks for the Prediction of Organic Chemistry Reactions". *ACS Central Science* 2:10, 2016, pp. 725–732.

95. J. Nam and J. Kim. "Linking the Neural Machine Translation and the Prediction of Organic Chemistry Reactions". arxiv: `1612.09529` (cs), 2016.

96. 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 Central Science* 3:10, 2017, pp. 1103–1113.

97. 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 30*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Curran Associates, Inc., 2017, pp. 5998–6008.

98. P. Schwaller, T. Gaudin, D. Lányi, C. Bekas, and T. Laino. ""Found in Translation": Predicting Outcomes of Complex Organic Chemistry Reactions Using Neural Sequence-to-Sequence Models". *Chemical Science* 9:28, 2018, pp. 6091–6098.

99. 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 Central Science* 5:9, 2019, pp. 1572–1583.

100. P. Karpov, G. Godin, and I. V. Tetko. "A Transformer Model for Retrosynthesis". In: *Artificial Neural Networks and Machine Learning – ICANN 2019: Workshop and Special Sessions*. Ed. by I. V. Tetko, V. Kůrková, P. Karpov, and F. Theis. Lecture Notes in Computer Science. Springer International Publishing, Cham, 2019, pp. 817–830.

101. P. Schwaller, R. Petraglia, V. H. Nair, and T. Laino. "Evaluation Metrics for Single-Step Retrosynthetic Models". *Second Workshop on Machine Learning and the Physical Sciences (NeurIPS 2019)*, 2019.

102.   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". *Chemical Science* 11:12, 2020, pp. 3316–3325.

103.   C. W. Coley, L. Rogers, W. H. Green, and K. F. Jensen. "SCScore: Synthetic Complexity Learned from a Reaction Corpus". *Journal of Chemical Information and Modeling* 58:2, 2018, pp. 252–261.

104.   K. Lin, Y. Xu, J. Pei, and L. Lai. "Automatic Retrosynthetic Route Planning Using Template-Free Models". *Chemical Science* 11:12, 2020, pp. 3355–3364.

105.   B. Chen, C. Li, H. Dai, and L. Song. "Retro*: Learning Retrosynthetic Planning with Neural Guided A* Search". In: *Proceedings of the 37th International Conference on Machine Learning*. International Conference on Machine Learning. PMLR, 2020, pp. 1608–1616.

106.   P. E. McGovern, J. Zhang, J. Tang, Z. Zhang, G. R. Hall, R. A. Moreau, A. Nuñez, E. D. Butrym, M. P. Richards, C.-s. Wang, G. Cheng, Z. Zhao, and C. Wang. "Fermented Beverages of Pre- and Proto-Historic China". *Proceedings of the National Academy of Sciences* 101:51, 2004, pp. 17593–17598.

107.   W. Kühne. "Ueber Das Verhalten Verschiedener Organisirter Und Sog. Ungeformter Fermente". *Verhandlungen des Naturhistorisch-medicinischen Vereins zu Heidelberg* 1, 1877, pp. 190–193.

108.   J. Wisniak. "The History of Catalysis. From the Beginning to Nobel Prizes". *Educación Química* 21:1, 2010, pp. 60–69.

109.   C. Chothia. "One Thousand Families for the Molecular Biologist". *Nature* 357:6379, 6379 1992, pp. 543–544.

110.   J. A. Gerlt and P. C. Babbitt. "Divergent Evolution of Enzymatic Function: Mechanistically Diverse Superfamilies and Functionally Distinct Suprafamilies". *Annual Review of Biochemistry* 70:1, 2001, pp. 209–246. pmid: 11395407.

111.   S. D. Brown and P. C. Babbitt. "New Insights about Enzyme Evolution from Large Scale Studies of Sequence and Structure Relationships*". *Journal of Biological Chemistry* 289:44, 2014, pp. 30221–30228.

112.   D. G. Herries. "Enzyme Structure and Mechanism (Second Edition), by Alan Fersht. Pp 475. W H Freeman, New York. 1984. £28.95 or £14.95 (Paperback) ISBN 0–7167–1614–3 or ISBN 0–7167–1615–1 (Pbk)". *Biochemical Education* 13:3, 1985, pp. 146–146.

113.   A. Warshel, P. K. Sharma, M. Kato, Y. Xiang, H. Liu, and M. H. M. Olsson. "Electrostatic Basis for Enzyme Catalysis". *Chemical Reviews* 106:8, 2006, pp. 3210–3235.

114. "Chapter 6: Enzyme Principles and Biotechnological Applications",

115. D. P. Pantaleone. "Biotransformations: "Green" Processes for the Synthesis of Chiral Fine Chemicals". In: *Handbook Of Chiral Chemicals*. CRC Press, 1999.

116. R. A. Sheldon. "The E Factor 25 Years on: The Rise of Green Chemistry and Sustainability". *Green Chemistry* 19:1, 2017, pp. 18–43.

117. J. D. Rozzell. "Commercial Scale Biocatalysis: Myths and Realities". *Bioorganic & Medicinal Chemistry* 7:10, 1999, pp. 2253–2261.

118. M. Hönig, P. Sondermann, N. J. Turner, and E. M. Carreira. "Enantioselective Chemo- and Biocatalysis: Partners in Retrosynthesis". *Angewandte Chemie International Edition* 56:31, 2017, pp. 8942–8973.

119. P. D. de María, G. de Gonzalo, and A. R. Alcántara. "Biocatalysis as Useful Tool in Asymmetric Synthesis: An Assessment of Recently Granted Patents (2014–2019)". *Catalysts* 9:10, 10 2019, p. 802.

120. F. H. Arnold and G. Georgiou. *Directed Evolution Library Creation*. Vol. 231. Humana Press, New Jersey, 2003.

121. F. H. Arnold. "Directed Evolution: Bringing New Chemistry to Life". *Angewandte Chemie International Edition* 57:16, 2018, pp. 4143–4148.

122. J. C. Moore and F. H. Arnold. "Directed Evolution of a Para-Nitrobenzyl Esterase for Aqueous-Organic Solvents". *Nature Biotechnology* 14:4, 4 1996, pp. 458–467.

123. S. Soni. "Trends in Lipase Engineering for Enhanced Biocatalysis". *Biotechnology and Applied Biochemistry* 69:1, 2022, pp. 265–272.

124. L. Giver, A. Gershenson, P.-O. Freskgard, and F. H. Arnold. "Directed Evolution of a Thermostable Esterase". *Proceedings of the National Academy of Sciences* 95:22, 1998, pp. 12809–12813.

125. M. Petersen and A. Kiener. "Biocatalysis". *Green Chemistry* 1:2, 1999, pp. 99–106.

126. U. T. Bornscheuer, G. W. Huisman, R. J. Kazlauskas, S. Lutz, J. C. Moore, and K. Robins. "Engineering the Third Wave of Biocatalysis". *Nature* 485:7397, 7397 2012, pp. 185–194.

127. P. Jacques, M. Béchet, M. Bigan, D. Caly, G. Chataigné, F. Coutte, C. Flahaut, E. Heuson, V. Leclère, D. Lecouturier, V. Phalip, R. Ravallec, P. Dhulster, and R. Froidevaux. "High-Throughput Strategies for the Discovery and Engineering of Enzymes for Biocatalysis". *Bioprocess and Biosystems Engineering* 40:2, 2017, pp. 161–180.

128. X. Gu, J. Zhao, L. Chen, Y. Li, B. Yu, X. Tian, Z. Min, S. Xu, H. Gu, J. Sun, X. Lu, M. Chang, X. Wang, L. Zhao, S. Ye, H. Yang, Y. Tian, F. Gao, Y. Gai, G. Jia, J. Wu, Y. Wang, J. Zhang, X. Zhang, W. Liu, X. Gu, X. Luo, H. Dong, H. Wang, B. Schenkel, F. Venturoni, P. Filipponi, B. Guelat, T. Allmendinger, B. Wietfeld, P. Hoehn, N. Kovacic, L. Hermann, T. Schlama, T. Ruch, N. Derrien, P. Piechon, and F. Kleinbeck. "Application of Transition-Metal Catalysis, Biocatalysis, and Flow Chemistry as State-of-the-Art Technologies in the Synthesis of LCZ696". *The Journal of Organic Chemistry* 85:11, 2020, pp. 6844–6853.

129. L. A. Hardegger, P. Beney, D. Bixel, C. Fleury, F. Gao, A. G.-G. Perrenoud, X. Gu, J. Haber, T. Hong, R. Humair, A. Kaegi, M. Kibiger, F. Kleinbeck, V. T. Luu, L. Padeste, F. A. Rampf, T. Ruch, T. Schlama, E. Sidler, A. Udvarhelyi, B. Wietfeld, and Y. Yang. "Toward a Scalable Synthesis and Process for EMA401, Part III: Using an Engineered Phenylalanine Ammonia Lyase Enzyme to Synthesize a Non-natural Phenylalanine Derivative". *Organic Process Research & Development* 24:9, 2020, pp. 1763–1771.

130. S. J. Novick, N. Dellas, R. Garcia, C. Ching, A. Bautista, D. Homan, O. Alvizo, D. Entwistle, F. Kleinbeck, T. Schlama, and T. Ruch. "Engineering an Amine Transaminase for the Efficient Production of a Chiral Sacubitril Precursor". *ACS Catalysis* 11:6, 2021, pp. 3762–3770.

131. J. B. Pyser. "State-of-the-Art Biocatalysis". *ACS Central Science* 7, 2021, pp. 1105–1116.

132. N. J. Turner and E. O'Reilly. "Biocatalytic Retrosynthesis". *Nature Chemical Biology* 9:5, 5 2013, pp. 285–288.

133. R. O. M. A. de Souza, L. S. M. Miranda, and U. T. Bornscheuer. "A Retrosynthesis Approach for Biocatalysis in Organic Synthesis". *Chemistry – A European Journal* 23:50, 2017, pp. 12040–12063.

134. B. P. Dwivedee, S. Soni, M. Sharma, J. Bhaumik, J. K. Laha, and U. C. Banerjee. "Promiscuity of Lipase-Catalyzed Reactions for Organic Synthesis: A Recent Update". *ChemistrySelect* 3:9, 2018, pp. 2441–2466.

135. F. Rudroff, M. D. Mihovilovic, H. Gröger, R. Snajdrova, H. Iding, and U. T. Bornscheuer. "Opportunities and Challenges for Combining Chemo- and Biocatalysis". *Nature Catalysis* 1:1, 1 2018, pp. 12–22.

136. K.-E. Jaeger and T. Eggert. "Enantioselective Biocatalysis Optimized by Directed Evolution". *Current Opinion in Biotechnology* 15:4, 2004, pp. 305–313.

137.   C. K. Chung, P. G. Bulger, B. Kosjek, K. M. Belyk, N. Rivera, M. E. Scott, G. R. Humphrey, J. Limanto, D. C. Bachert, and K. M. Emerson. "Process Development of C–N Cross-Coupling and Enantioselective Biocatalytic Reactions for the Asymmetric Synthesis of Niraparib". *Organic Process Research & Development* 18:1, 2014, pp. 215–227.

138.   N. Wamser, H. Wu, F. Buono, A. Brundage, F. Ricci, J. C. Lorenz, J. Wang, N. Haddad, J. Paolillo, J. C. Leung, H. Lee, and A. Hossain. "Discovery and Process Development of a Scalable Biocatalytic Kinetic Resolution toward Synthesis of a Sterically Hindered Chiral Ketone". *Organic Process Research & Development* 26:6, 2022, pp. 1820–1830.

139.   I. Schomburg, A. Chang, and D. Schomburg. "BRENDA, Enzyme Data and Metabolic Information". *Nucleic Acids Research* 30:1, 2002, pp. 47–49. pmid: 11752250.

140.   A. Chang, L. Jeske, S. Ulbrich, J. Hofmann, J. Koblitz, I. Schomburg, M. Neumann-Schaal, D. Jahn, and D. Schomburg. "BRENDA, the ELIXIR Core Data Resource in 2021: New Developments and Updates". *Nucleic Acids Research* 49:D1, 2021, pp. D498–D508.

141.   M. Kanehisa. "The KEGG Database". In: *'In Silico' Simulation of Biological Processes*. John Wiley & Sons, Ltd, 2002, pp. 91–103.

142.   M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe. "KEGG for Integration and Interpretation of Large-Scale Molecular Data Sets". *Nucleic Acids Research* 40:D1, 2012, pp. D109–D114.

143.   P. D. Karp, M. Riley, S. M. Paley, and A. Pellegrini-Toole. "The MetaCyc Database". *Nucleic Acids Research* 30:1, 2002, pp. 59–61.

144.   C. J. Krieger, P. Zhang, L. A. Mueller, A. Wang, S. Paley, M. Arnaud, J. Pick, S. Y. Rhee, and P. D. Karp. "MetaCyc: A Multiorganism Database of Metabolic Pathways and Enzymes". *Nucleic Acids Research* 32, suppl_1 2004, pp. D438–D442.

145.   R. Alcántara, K. B. Axelsen, A. Morgat, E. Belda, E. Coudert, A. Bridge, H. Cao, P. de Matos, M. Ennis, S. Turner, G. Owen, L. Bougueleret, I. Xenarios, and C. Steinbeck. "Rhea–a Manually Curated Resource of Biochemical Reactions". *Nucleic Acids Research* 40, Database issue 2012, pp. D754–760. pmid: 22135291.

146.   P. Bansal, A. Morgat, K. B. Axelsen, V. Muthukrishnan, E. Coudert, L. Aimo, N. Hyka-Nouspikel, E. Gasteiger, A. Kerhornou, T. B. Neto, M. Pozzato, M.-C. Blatter, A. Ignatchenko, N. Redaschi, and A. Bridge. "Rhea, the Reaction Knowledgebase in 2022". *Nucleic Acids Research* 50:D1, 2022, pp. D693–D700.

147.   J. W. K. Ho, T. Manwaring, S.-H. Hong, U. Roehm, D. C. Y. Fung, K. Xu, T. Kraska, and D. Hart. "PathBank: Web-Based Querying and Visualziation of an Integrated Biological Pathway Database". In: *International Conference on Computer Graphics, Imaging and Visualisation (CGIV'06)*. International Conference on Computer Graphics, Imaging and Visualisation (CGIV'06). 2006, pp. 84–89.

148.   D. S. Wishart, C. Li, A. Marcu, H. Badran, A. Pon, Z. Budinski, J. Patron, D. Lipton, X. Cao, E. Oler, K. Li, M. Paccoud, C. Hong, A. C. Guo, C. Chan, W. Wei, and M. Ramirez-Gaona. "PathBank: A Comprehensive Pathway Database for Model Organisms". *Nucleic Acids Research* 48:D1, 2020, pp. D470–D478.

149.   M. Ganter, T. Bernard, S. Moretti, J. Stelling, and M. Pagni. "MetaNetX.Org: A Website and Repository for Accessing, Analysing and Manipulating Metabolic Networks". *Bioinformatics* 29:6, 2013, pp. 815–816.

150.   S. Moretti, O. Martin, T. Van Du Tran, A. Bridge, A. Morgat, and M. Pagni. "MetaNetX/MNXref – Reconciliation of Metabolites and Biochemical Reactions to Bring Together Genome-Scale Metabolic Networks". *Nucleic Acids Research* 44:D1, 2016, pp. D523–D526.

151.   S. Moretti, V. D. T. Tran, F. Mehl, M. Ibberson, and M. Pagni. "MetaNetX/MNXref: Unified Namespace for Metabolites and Biochemical Reactions in the Context of Metabolic Models". *Nucleic Acids Research* 49:D1, 2021, pp. D570–D574.

152.   N. Nagano. "EzCatDB: The Enzyme Catalytic-mechanism Database". *Nucleic Acids Research* 33, suppl_1 2005, pp. D407–D412.

153.   N. Nagano, N. Nakayama, K. Ikeda, M. Fukuie, K. Yokota, T. Doi, T. Kato, and K. Tomii. "EzCatDB: The Enzyme Reaction Database, 2015 Update". *Nucleic Acids Research* 43:D1, 2015, pp. D453–D458.

154.   E. Boutet, D. Lieberherr, M. Tognolli, M. Schneider, P. Bansal, A. J. Bridge, S. Poux, L. Bougueleret, and I. Xenarios. "UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase: How to Use the Entry View". In: *Plant Bioinformatics: Methods and Protocols*. Ed. by D. Edwards. Methods in Molecular Biology. Springer, New York, NY, 2016, pp. 23–54.

155.   N. Hadadi and V. Hatzimanikatis. "Design of Computational Retrobiosynthesis Tools for the Design of de Novo Synthetic Pathways". *Current Opinion in Chemical Biology*. Synthetic Biology • Synthetic Biomolecules 28, 2015, pp. 99–104.

156.   B. Delépine, T. Duigou, P. Carbonell, and J.-L. Faulon. "RetroPath2.0: A Retrosynthesis Workflow for Metabolic Engineers". *Metabolic Engineering* 45, 2018, pp. 158–170.

157. M. Koch, T. Duigou, and J.-L. Faulon. "Reinforcement Learning for Bioretrosynthesis". *ACS Synthetic Biology* 9:1, 2020, pp. 157–168.

158. C. E. Lawson, J. M. Martí, T. Radivojevic, S. V. R. Jonnalagadda, R. Gentz, N. J. Hillson, S. Peisert, J. Kim, B. A. Simmons, C. J. Petzold, S. W. Singer, A. Mukhopadhyay, D. Tanjore, J. G. Dunn, and H. Garcia Martin. "Machine Learning for Metabolic Engineering: A Review". *Metabolic Engineering*. Tools and Strategies of Metabolic Engineering 63, 2021, pp. 34–60.

159. S. Ranganathan, S. Mahesh, S. Suresh, A. Nagarajan, T. Z. Sen, and R. M.Yennamalli. "Experimental and Computational Studies of Cellulases as Bioethanol Enzymes". *Bioengineered* 13:5, 2022, pp. 14028–14046. pmid: 35730402.

160. S. Zheng, T. Zeng, C. Li, B. Chen, C. W. Coley, Y. Yang, and R. Wu. "Deep Learning Driven Biosynthetic Pathways Navigation for Natural Products with BioNavi-NP". *Nature Communications* 13, 1, 1 2022, p. 3342.

161. H.-H. Cheng and L.-M. Whang. "Applying Metabolic Flux Analysis to Hydrogen Fermentation Using a Metabolic Network Constructed for Anaerobic Mixed Cultures". *Environmental Research* 235, 2023, p. 116636.

162. B. F.-L. Sieow, R. De Sotto, Z. R. D. Seet, I. Y. Hwang, and M. W. Chang. "Synthetic Biology Meets Machine Learning". In: *Computational Biology and Machine Learning for Metabolic Engineering and Synthetic Biology*. Ed. by K. Selvarajoo. Methods in Molecular Biology. Springer US, New York, NY, 2023, pp. 21–39.

163. D. Walther. "Specifics of Metabolite-Protein Interactions and Their Computational Analysis and Prediction". In: *Cell-Wide Identification of Metabolite-Protein Interactions*. Ed. by A. Skirycz, M. Luzarowski, and J. C. Ewald. Methods in Molecular Biology. Springer US, New York, NY, 2023, pp. 179–197.

164. D. Probst, M. Manica, Y. G. Nana Teukam, A. Castrogiovanni, F. Paratore, and T. Laino. "Biocatalysed Synthesis Planning Using Data-Driven Learning". *Nature Communications* 13, 1, 1 2022, p. 964.

165. W. Finnigan, L. J. Hepworth, S. L. Flitsch, and N. J. Turner. "RetroBioCat as a Computer-Aided Synthesis Planning Tool for Biocatalytic Reactions and Cascades". *Nature Catalysis* 4, 2, 2 2021, pp. 98–104.

166. K. Sankaranarayanan and K. F. Jensen. "Computer-Assisted Multistep Chemoenzymatic Retrosynthesis Using a Chemical Synthesis Planner". *Chemical Science* 14:23, 2023, pp. 6467–6475.

167. I. Levin, M. Liu, C. A. Voigt, and C. W. Coley. "Merging Enzymatic and Synthetic Chemistry with Computational Synthesis Planning". *Nature Communications* 13:1, 1 2022, p. 7747.

168. M. Lang, M. Stelzer, and D. Schomburg. "BKM-react, an Integrated Biochemical Reaction Database". *BMC Biochemistry* 12:1, 2011, p. 42.

169. M. Ali, H. M. Ishqi, and Q. Husain. "Enzyme Engineering: Reshaping the Biocatalytic Functions". *Biotechnology and Bioengineering* 117:6, 2020, pp. 1877–1894.

170. R. A. Sheldon and J. M. Woodley. "Role of Biocatalysis in Sustainable Chemistry". *Chemical Reviews* 118:2, 2018, pp. 801–838.

171. C. W. Coley, W. H. Green, and K. F. Jensen. "Machine Learning in Computer-Aided Synthesis Planning". *Accounts of Chemical Research* 51:5, 2018, pp. 1281–1289.

172. V. H. Nair, P. Schwaller, and T. Laino. "Data-Driven Chemical Reaction Prediction and Retrosynthesis". *CHIMIA* 73, 12, 12 2019, pp. 997–997.

173. S. Johansson, A. Thakkar, T. Kogej, E. Bjerrum, S. Genheden, T. Bastys, C. Kannas, A. Schliep, H. Chen, and O. Engkvist. "AI-assisted Synthesis Prediction". *Drug Discovery Today: Technologies*. Artificial Intelligence 32–33, 2019, pp. 65–72.

174. 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". *Nature Communications* 11, 1, 1 2020, p. 5575.

175. W. W. Qian, N. T. Russell, C. L. W. Simons, Y. Luo, M. D. Burke, and J. Peng. "Integrating Deep Neural Networks and Symbolic Inference for Organic Reactivity Prediction", 2020.

176. Y. Cai, H. Yang, W. Li, G. Liu, P. W. Lee, and Y. Tang. "Multiclassification Prediction of Enzymatic Reactions for Oxidoreductases and Hydrolases Using Reaction Fingerprints and Machine Learning Methods". *Journal of Chemical Information and Modeling* 58:6, 2018, pp. 1169–1181.

177. N. Hadadi, H. MohammadiPeyhani, L. Miskovic, M. Seijo, and V. Hatzimanikatis. "Enzyme Annotation for Orphan and Novel Reactions Using Knowledge of Substrate Reactive Sites". *Proceedings of the National Academy of Sciences* 116:15, 2019, pp. 7298–7307.

178. E. E. Litsa, P. Das, and L. E. Kavraki. "Prediction of Drug Metabolites Using Neural Machine Translation". *Chemical Science* 11:47, 2020, pp. 12777–12788.

179. 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". *Chemical Science* 11:1, 2019, pp. 154–168.

180. G. Pesciullesi, P. Schwaller, T. Laino, and J.-L. Reymond. "Transfer Learning Enables the Molecular Transformer to Predict Regio- and Stereoselective Reactions on Carbohydrates". *Nature Communications* 11, 1, 1 2020, p. 4874.

181. S. Ferri, K. Kojima, and K. Sode. "Review of Glucose Oxidases and Glucose Dehydrogenases: A Bird's Eye View of Glucose Sensing Enzymes". *Journal of Diabetes Science and Technology* 5:5, 2011, pp. 1068–1076.

182. O. K. Tawfik and D. S. "Enzyme Promiscuity: A Mechanistic and Evolutionary Perspective". *Annual Review of Biochemistry* 79:1, 2010, pp. 471–505. pmid: 20235827.

183. K. Hult and P. Berglund. "Enzyme Promiscuity: Mechanism and Applications". *Trends in Biotechnology* 25:5, 2007, pp. 231–238.

184. S. Velikogne, W. B. Breukelaar, F. Hamm, R. A. Glabonjat, and W. Kroutil. "C=C-Ene-Reductases Reduce the C=N Bond of Oximes". *ACS Catalysis* 10:22, 2020, pp. 13377–13382.

185. M. Kanehisa. "Enzyme Annotation and Metabolic Reconstruction Using KEGG". In: *Protein Function Prediction: Methods and Protocols*. Ed. by D. Kihara. Springer New York, New York, NY, 2017, pp. 135–145.

186. D. Probst and J.-L. Reymond. "Visualization of Very Large High-Dimensional Data Sets as Minimum Spanning Trees". *Journal of Cheminformatics* 12:1, 2020, p. 12.

187. C. W. Coley, W. H. Green, and K. F. Jensen. "RDChiral: An RDKit Wrapper for Handling Stereochemistry in Retrosynthetic Template Extraction and Application". *Journal of Chemical Information and Modeling* 59:6, 2019, pp. 2529–2537.

188. G. Landrum et al. *RDKit: Open-Source Cheminformatics Software*. Version 2020_03_4 (Q1 2020) Release. 2020_03_4 (Q1 2020) Release.

189. T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush. "HuggingFace's Transformers: State-of-the-art Natural Language Processing". 2020. arXiv: 1910.03771 [cs].

190. G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. Rush. "OpenNMT: Open-Source Toolkit for Neural Machine Translation". In: *Proceedings of ACL 2017, System Demonstrations*. Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 67–72.

191. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019.

192. J. Xu, Y. Cen, W. Singh, J. Fan, L. Wu, X. Lin, J. Zhou, M. Huang, M. T. Reetz, and Q. Wu. "Stereodivergent Protein Engineering of a Lipase To Access All Possible Stereoisomers of Chiral Esters with Two Stereocenters". *Journal of the American Chemical Society* 141:19, 2019, pp. 7934–7945.

193. Y.-H. Kim and S. Park. "Surveying Enantioselectivity of Two Candida Antarctica-Lipase-B Homologs Towards Chiral Sec-Alcohols". *Bulletin of the Korean Chemical Society* 38:11, 2017, pp. 1358–1361.

194. H. Ankati, D. Zhu, Y. Yang, E. R. Biehl, and L. Hua. "Asymmetric Synthesis of Both Antipodes of β-Hydroxy Nitriles and β-Hydroxy Carboxylic Acids via Enzymatic Reduction or Sequential Reduction/Hydrolysis". *The Journal of Organic Chemistry* 74:4, 2009, pp. 1658–1662.

195. W. Borzęcka, I. Lavandera, and V. Gotor. "Synthesis of Enantiopure Fluorohydrins Using Alcohol Dehydrogenases at High Substrate Concentrations". *The Journal of Organic Chemistry* 78:14, 2013, pp. 7312–7317.

196. H. C. Büchsenschütz, V. Vidimce-Risteski, B. Eggbauer, S. Schmidt, C. K. Winkler, J. H. Schrittwieser, W. Kroutil, and R. Kourist. "Stereoselective Biotransformations of Cyclic Imines in Recombinant Cells of Synechocystis Sp. PCC 6803". *ChemCatChem* 12:3, 2020, pp. 726–730.

197. F. G. Mutti and W. Kroutil. "Asymmetric Bio-amination of Ketones in Organic Solvents". *Advanced Synthesis & Catalysis* 354:18, 2012, pp. 3409–3413.

198. R. R. Chao, J. J. D. Voss, and S. G. Bell. "The Efficient and Selective Catalytic Oxidation of Para-Substituted Cinnamic Acid Derivatives by the Cytochrome P450 Monooxygenase, CYP199A4". *RSC Advances* 6:60, 2016, pp. 55286–55297.

199. K. Neufeld, J. Marienhagen, U. Schwaneberg, and J. Pietruszka. "Benzylic Hydroxylation of Aromatic Compounds by P450 BM3". *Green Chemistry* 15:9, 2013, pp. 2408–2421.

200. P. Both, H. Busch, P. P. Kelly, F. G. Mutti, N. J. Turner, and S. L. Flitsch. "Whole-Cell Biocatalysts for Stereoselective C-H Amination Reactions". *Angewandte Chemie International Edition* 55:4, 2016, pp. 1511–1513.

201.   C. S. Alexeev, G. G. Sivets, T. N. Safonova, and S. N. Mikhailov. "Substrate Specificity of E. Coli Uridine Phosphorylase. Further Evidences of High-Syn Conformation of the Substrate in Uridine Phosphorolysis". *Nucleosides, Nucleotides & Nucleic Acids* 36:2, 2017, pp. 107–121. pmid: 27846376.

202.   W. Wang and B. Wang. "Esterase-Sensitive Sulfur Dioxide Prodrugs Inspired by Modified Julia Olefination". *Chemical Communications* 53:73, 2017, pp. 10124–10127.

203.   H. A. Namanja-Magliano, C. F. Stratton, and V. L. Schramm. "Transition State Structure and Inhibition of Rv0091, a 5'-Deoxyadenosine/5'-Methylthioadenosine Nucleosidase from Mycobacterium Tuberculosis". *ACS Chemical Biology* 11:6, 2016, pp. 1669–1676.

204.   R.-J. Li, J.-H. Xu, Y.-C. Yin, N. Wirth, J.-M. Ren, B.-B. Zeng, and H.-L. Yu. "Rapid Probing of the Reactivity of P450 Monooxygenases from the CYP116B Subfamily Using a Substrate-Based Method". *New Journal of Chemistry* 40:10, 2016, pp. 8928–8934.

205.   E. A. Hall, M. R. Sarkar, and S. G. Bell. "The Selective Oxidation of Substituted Aromatic Hydrocarbons and the Observation of Uncoupling via Redox Cycling during Naphthalene Oxidation by the CYP101B1 System". *Catalysis Science & Technology* 7:7, 2017, pp. 1537–1548.

206.   J. A. Faraldos, D. J. Miller, V. González, Z. Yoosuf-Aly, O. Cascón, A. Li, and R. K. Allemann. "A 1,6-Ring Closure Mechanism for (+)-δ-Cadinene Synthase?" *Journal of the American Chemical Society* 134:13, 2012, pp. 5900–5908.

207.   R.-J. Li, A. Li, J. Zhao, Q. Chen, N. Li, H.-L. Yu, and J.-H. Xu. "Engineering P450LaMO Stereospecificity and Product Selectivity for Selective C–H Oxidation of Tetralin-like Alkylbenzenes". *Catalysis Science & Technology* 8:18, 2018, pp. 4638–4644.

208.   S. Schmidt, H. C. Büchsenschütz, C. Scherkus, A. Liese, H. Gröger, and U. T. Bornscheuer. "Biocatalytic Access to Chiral Polyesters by an Artificial Enzyme Cascade Synthesis". *ChemCatChem* 7:23, 2015, pp. 3951–3955.

209.   R. S. Heath, W. R. Birmingham, M. P. Thompson, A. Taglieber, L. Daviet, and N. J. Turner. "An Engineered Alcohol Oxidase for the Oxidation of Primary Alcohols". *Chembiochem : a European journal of chemical biology* 20:2, 2019, pp. 276–281.

210.   J. Wang, L. Zhang, L. Jia, Y. Ren, and G. Yu. "Protein-Protein Interactions Prediction Using a Novel Local Conjoint Triad Descriptor of Amino Acid Sequences". *International Journal of Molecular Sciences* 18, 11, 11 2017, p. 2373.

211.   V. Gligorijević, M. Barot, and R. Bonneau. "deepNF: Deep Network Fusion for Protein Function Prediction". *Bioinformatics (Oxford, England)* 34:22, 2018, pp. 3873–3881.

212.   D. Probst, M. Manica, Y. G. N. Teukam, A. Castrogiovanni, F. Paratore, and T. Laino. "Molecular Transformer-aided Biocatalysed Synthesis Planning", 2021.

213.   "OpenNMT/OpenNMT-py",

214.   R. Sennrich, B. Haddow, and A. Birch. "Neural Machine Translation of Rare Words with Subword Units". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, 2016, pp. 1715–1725.

215.   F. Strieth-Kalthoff, F. Sandfort, M. H. S. Segler, and F. Glorius. "Machine Learning the Ropes: Principles, Applications and Directions in Synthetic Chemistry". *Chemical Society Reviews* 49:17, 2020, pp. 6154–6168.

216.   A. Thakkar, S. Johansson, K. Jorner, D. Buttar, J.-L. Reymond, and O. Engkvist. "Artificial Intelligence and Automation in Computer Aided Synthesis Planning". *Reaction Chemistry & Engineering* 6:1, 2021, pp. 27–51.

217.   K. Molga, S. Szymkuć, and B. A. Grzybowski. "Chemist Ex Machina: Advanced Synthesis Planning by Computers". *Accounts of Chemical Research* 54:5, 2021, pp. 1094–1106.

218.   P. Schwaller, A. C. Vaucher, R. Laplaza, C. Bunne, A. Krause, C. Corminboeuf, and T. Laino. "Machine Intelligence for Chemical Reaction Space". *WIREs Computational Molecular Science* 12:5, 2022, e1604.

219.   R. Irwin, S. Dimitriadis, J. He, and E. J. Bjerrum. "Chemformer: A Pre-Trained Transformer for Computational Chemistry". *Machine Learning: Science and Technology* 3:1, 2022, p. 015022.

220.   X. Wang, Y. Qian, H. Gao, C. W. Coley, Y. Mo, R. Barzilay, and K. F. Jensen. "Towards Efficient Discovery of Green Synthetic Pathways with Monte Carlo Tree Search and Reinforcement Learning". *Chemical Science* 11:40, 2020, pp. 10959–10972.

221.   A. Thakkar, A. C. Vaucher, A. Byekwaso, P. Schwaller, A. Toniato, and T. Laino. "Unbiasing Retrosynthesis Language Models with Disconnection Prompts". *ACS Central Science* 9:7, 2023, pp. 1488–1498.

222.   P. Schwaller, B. Hoover, J.-L. Reymond, H. Strobelt, and T. Laino. "Extraction of Organic Chemistry Grammar from Unsupervised Learning of Chemical Reactions". *Science Advances* 7:15, 2021, eabe4166.

223.   A. Byekwaso, A. C. Vaucher, P. Schwaller, A. Toniato, and T. Laino. "A Sequence-to-Sequence Transformer Model for Disconnection Aware Retrosynthesis", 2021.

224. M. Andronov, V. Voinarovska, N. Andronova, M. Wand, D.-A. Clevert, and J. Schmidhuber. "Reagent Prediction with a Molecular Transformer Improves Reaction Data Quality". *Chemical Science* 14:12, 2023, pp. 3235–3246.

225. W. Velanguparackel, N. Hamon, J. Balzarini, C. McGuigan, and A. D. Westwell. "Synthesis, Anti-HIV and Cytostatic Evaluation of 3'-Deoxy-3'-Fluorothymidine (FLT) pro-Nucleotides". *Bioorganic & Medicinal Chemistry Letters* 24:10, 2014, pp. 2240–2243.

226. T. Wang, Y. Ueda, Z. Zhang, Z. Yin, J. Matiskella, B. C. Pearce, Z. Yang, M. Zheng, D. D. Parker, G. A. Yamanaka, Y.-F. Gong, H.-T. Ho, R. J. Colonno, D. R. Langley, P.-F. Lin, N. A. Meanwell, and J. F. Kadow. "Discovery of the Human Immunodeficiency Virus Type 1 (HIV-1) Attachment Inhibitor Temsavir and Its Phosphonooxymethyl Prodrug Fostemsavir". *Journal of Medicinal Chemistry* 61:14, 2018, pp. 6308–6327.

227. F. L. Scott, B. Clemons, J. Brooks, E. Brahmachary, R. Powell, H. Dedman, H. G. Desale, G. A. Timony, E. Martinborough, H. Rosen, E. Roberts, M. F. Boehm, and R. J. Peach. "Ozanimod (RPC1063) Is a Potent Sphingosine-1-Phosphate Receptor-1 (S1P1) and Receptor-5 (S1P5) Agonist with Autoimmune Disease-Modifying Activity". *British Journal of Pharmacology* 173:11, 2016, pp. 1778–1792.

228. A. C. Flick, C. A. Leverett, H. X. Ding, E. L. McInturff, S. J. Fink, S. Mahapatra, D. W. Carney, E. A. Lindsey, J. C. DeForest, S. P. France, S. Berritt, S. V. Bigi-Botterill, T. S. Gibson, R. B. Watson, Y. Liu, and C. J. O'Donnell. "Synthetic Approaches to the New Drugs Approved During 2020". *Journal of Medicinal Chemistry* 65:14, 2022, pp. 9607–9661.

229. D. Probst, P. Schwaller, and J.-L. Reymond. "Reaction Classification and Yield Prediction Using the Differential Reaction Fingerprint DRFP". *Digital Discovery* 1:2, 2022, pp. 91–97.

230. *IBM RXN for Chemistry*.

231. S. Genheden and E. Bjerrum. "PaRoutes: Towards a Framework for Benchmarking Retrosynthesis Route Predictions". *Digital Discovery* 1:4, 2022, pp. 527–539.

232. D. Kreutter, P. Schwaller, and J.-L. Reymond. "Predicting Enzymatic Reactions with a Molecular Transformer". *Chemical Science* 12:25, 2021, pp. 8648–8659.

233. A. Thakkar, A. Vaucher, A. Byekwaso, P. Schwaller, A. Toniato, and T. Laino. *Disconnection Labelled Reaction Data*. Version 1.0. Zenodo, 2022.

234. 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". *Chemical Communications* 55:81, 2019, pp. 12152–12155.

235. S. Zheng, J. Rao, Z. Zhang, J. Xu, and Y. Yang. "Predicting Retrosynthetic Reactions Using Self-Corrected Transformer Neural Networks". *Journal of Chemical Information and Modeling* 60:1, 2020, pp. 47–55.

236. H. Duan, L. Wang, C. Zhang, L. Guo, and J. Li. "Retrosynthesis with Attention-Based NMT Model and Chemical Analysis of "Wrong" Predictions". *RSC Advances* 10:3, 2020, pp. 1371–1378.

237. D. Kreutter and J.-L. Reymond. "Multistep Retrosynthesis Combining a Disconnection Aware Triple Transformer Loop with a Route Penalty Score Guided Tree Search". *Chemical Science* 14:36, 2023, pp. 9959–9969.

238. K. Honda. "Chapter 16 - Industrial Applications of Multistep Enzyme Reactions". In: *Biotechnology of Microbial Enzymes*. Ed. by G. Brahmachari. Academic Press, 2017, pp. 433–450.

239. M. Arroyo, I. de la Mata, J.-L. García, and J.-L. Barredo. "Chapter 17 - Biocatalysis for Industrial Production of Active Pharmaceutical Ingredients (APIs)". In: *Biotechnology of Microbial Enzymes*. Ed. by G. Brahmachari. Academic Press, 2017, pp. 451–473.

240. S. A. Kelly, S. Pohle, S. Wharry, S. Mix, C. C. Allen, T. S. Moody, and B. F. Gilmore. "Application of ω-Transaminases in the Pharmaceutical Industry". *Chemical Reviews* 118:1, 2018, pp. 349–367.

241. J. P. Adams, M. J. B. Brown, A. Diaz-Rodriguez, R. C. Lloyd, and G.-D. Roiban. "Biocatalysis: A Pharma Perspective". *Advanced Synthesis & Catalysis* 361:11, 2019, pp. 2421–2432.

242. P. Carbonell, T. Fehér, I. Grigoras, and J.-L. Faulon. "RetroPath: Retrosynthesis Design of Metabolic Pathways". *advances in Systems and Synthetic Biology*, 2014.

243. D. Kreutter and J.-L. Reymond. *Transformer Models for Disconnection-Aware Triple Transformer Loop*. Version 1.1. Zenodo, 2023.

244. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis. "Highly Accurate Protein Structure Prediction with AlphaFold". *Nature* 596:7873, 7873 2021, pp. 583–589.

245. S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan. "Tree of Thoughts: Deliberate Problem Solving with Large Language Models". arxiv: `2305.10601` (cs), 2023.

246. T. Badowski, E. P. Gajewska, K. Molga, and B. A. Grzybowski. "Synergy Between Expert and Machine-Learning Approaches Allows for Improved Retrosynthetic Planning". *Angewandte Chemie International Edition* 59:2, 2020, pp. 725–730.

During the preparation of this work, I used the following tools:

- ChatGPT v3.5 — Coding support, rewording, synonym.

- Grammarly — Grammar and spelling checking.

After using these tools, I reviewed and edited the content as needed and I take full responsibility for the content of the publication. I am aware that in case of dis-compliance, the generated text is considered plagiarism with its legal consequences.