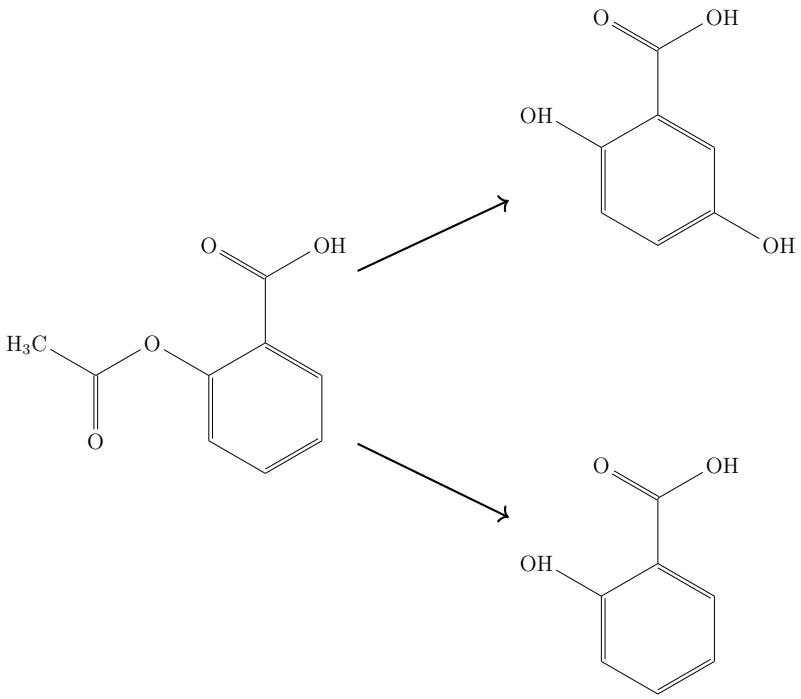
CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG



Optimising a Transformer-Based Model for Metabolite Prediction in Drug Discovery

Master's Thesis in Computer Science and Engineering

MIRANDA CARLSSON, SOFIA LARSSON

Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2025

Master's thesis 2025

Optimising a Transformer-Based Model for Metabolite Prediction in Drug Discovery

MIRANDA CARLSSON, SOFIA LARSSON

UNIVERSITY OF
GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Computer Science and Engineering
Division of Data Science and AI
AI Laboratory for Molecular Engineering
Chalmers University of Technology
University of Gothenburg
Gothenburg, Sweden 2025

Optimising a Transformer-Based Model for Metabolite Prediction in Drug Discovery

MIRANDA CARLSSON, SOFIA LARSSON

Supervisor: Richard Beckmann, Department of Computer Science and Engineering, Chalmers
Advisor: Filip Miljković, AstraZeneca
Examiner: Rocío Mercado Oropeza, Department of Computer Science and Engineering, Chalmers

Master's Thesis 2025
Department of Computer Science and Engineering
Division of Data Science and AI
AI Laboratory for Molecular Engineering
Chalmers University of Technology and University of Gothenburg
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: The chemical structures of the drug aspirin (to the left) and its metabolites gentisic acid and salicylic acid (to the right).

Typeset in LaTeX
Gothenburg, Sweden 2025

Optimising a Transformer-Based Model for Metabolite Prediction in Drug Discovery

MIRANDA CARLSSON, SOFIA LARSSON
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg

Abstract

Drug metabolism plays a crucial role in drug discovery, impacting the safety and efficacy of medications. Experimental methods of predicting metabolic reactions of potential drugs have long suffered from high costs in terms of both time and resources. Computational methods have emerged as more cost-effective and time-efficient approaches for predicting drug metabolites, but are often hindered by reliance on rigid rules. Large language models provide a more adaptable and rule-free alternative. In this project, the transformer-based language model Chemformer, previously employed in various chemical tasks, was optimised for drug metabolism prediction. To achieve this, a dataset of drugs and their corresponding metabolites was compiled and preprocessed. The Chemformer model was fine-tuned and evaluated using this dataset. Further optimisation methods to enhance the model's performance involved incorporating an additional pre-training of the Chemformer model, randomisation of the input SMILES (Simplified Molecular Input Line Entry System) strings, augmenting the dataset, annotating the data with chemical information, employing ensemble models, and optimising prediction space by pre-training the model further. The most promising optimisation methods attempted were the additional pre-trainings and the randomisation of the SMILES strings, both of which showed a significant increase in performance. Benchmarking the best-performing model resulted in an outperformance in precision and F_1 score compared to the existing models GLORYx and SyGMa. These results suggest that Chemformer is a promising tool for metabolite prediction.

Keywords: drug discovery, metabolism, language model, transformer, computational chemistry, drugs, metabolites, optimisation

Acknowledgements

First and foremost, we would like to thank our academic supervisor, Richard Beckmann, and company advisor, Filip Miljković. Your insightful discussions, valuable feedback, and support throughout this master's thesis have been greatly appreciated.

We would also like to thank our examiner, Rocío Mercado Oropeza, for guiding us in the right direction and supporting us throughout this project.

Additionally, we would like to show gratitude to the team of AI Laboratory for Molecular Engineering (AIME) at Chalmers University of Technology and the Computational Chemistry team at AstraZeneca for your warm welcome and supportive environment that greatly enhanced our research experience.

We would also like to acknowledge the developers of the Chemformer model for making this project possible. A special thanks to Annie Westerlund for your invaluable support with Chemformer throughout the project.

Finally, we would like to thank our friends and families for your support during this master's thesis and throughout our studies. We would not have made it without you.

Miranda Carlsson and Sofia Larsson, Gothenburg, 2025-06-11

Contents

List of Figures

List of Tables

1

Introduction

The development of new drugs is a key part of modern medicine [1]. While drug discovery was a matter of luck in the past, systematic approaches to this field emerged in the 20th century and permitted many discoveries to be made [2]. Drug discovery usually begins with the identification of an unmet medical need for a disease, followed by selection of potential molecular targets that can be addressed with drugs, and then screening chemical compound libraries to find promising candidates to these targets. Yet, the size of the sampled chemical space is small compared to the space of possibly biologically active compounds. For this reason, the methods of drug discovery are constantly undergoing optimisation.

Metabolism plays a crucial role in drug discovery, with approximately 75% of all drugs undergoing metabolic transformation before being excreted from the body [3]. This transformation is thus a key phenomenon to explore. Metabolism is a complex essential system in living organisms that sustains life [4]. In metabolism, substrates are the starting molecules in chemical reactions. They are transformed by enzymes, a type of protein, into products called metabolites, which often have altered chemical properties [4].

It is critical to consider the implications of drug metabolism, as the resulting metabolites can have vastly different properties compared to the original drug, potentially being toxic even if the substrate molecule is not. This risk is particularly evident regarding liver toxicity, as the liver is the primary site of metabolic reactions involving foreign substances [5]. Additionally, it is important to consider potential drug-drug interactions, which refers to when a drug or component of food affects the free concentration of another drug, potentially leading to unforeseen effects or loss of efficacy [6]. Consequently, understanding drug metabolism is essential for optimising drug discovery as it directly influences drug efficacy and safety.

Traditional experimental methodologies for investigating drug metabolism are both time-consuming and resource-intensive, creating a need for more efficient approaches [6]. Computational methods have emerged as a powerful alternative, offering the potential for accurate metabolic predictions while significantly reducing resource demands [5], [7]–[9]. As early as in the 1990s, knowledge-based computer systems emerged that were able to predict the metabolic fate of chemical compounds [10]. However, computational techniques often rely on predefined transformation rules tied to specific enzyme families, i.e., classification groups, which limits scalability and often requires expert input [7]. Moreover, these rules rely on extreme simplifica-

tions, which prevent them from fully capturing the chemical context of the system. Simultaneously, more rules could increase the prediction of false positives [5]. In the context of the drug discovery pipeline, these false positives represent molecules that require time and resources for investigation without yielding meaningful results. To overcome these limitations, deep learning approaches provide a more adaptable and scalable solution, and could potentially enable accurate predictions of metabolites and metabolic pathways without relying on rigid rules. These rule-free methods have the potential of being more comprehensive and generalisable, thus being suitable for less common metabolic occurrences.

There are multiple research studies that have explored the possibility of predicting drug metabolites [5], [7]–[9]. Some are rule-based [7]–[9], whereas others are rule-free [5], [8]. Most of the developed tools can predict at least one metabolite per drug, but they struggle when accounting for all possible metabolite structures [5]. It is thus evident that more research in the field is necessary to reach the desired reliability for a wider range of metabolites.

A transformer-based model named Chemformer was previously developed by Irwin *et al.* [11] to tackle diverse tasks within computational chemistry. One of its successful applications was retrosynthesis prediction, i.e., conversion of final products of synthesis into its constituent building blocks. For the predictions with the highest probability of accuracy for each product, 54.3% correctly identified one substrate [11]. The retrosynthesis task is analogous to substrate-to-product conversions during metabolic reactions, often converting a large molecule into a set of smaller metabolites that also retain similarity with its substrate. Therefore, the Chemformer model has the potential of also performing well on the metabolic prediction task.

This project is a continuation of the master's thesis *Prediction of Drug Metabolites Using a Deep Learning Language Model* (2024) by Dehlén and Aronsson [12], conducted within the same research groups at Chalmers and AstraZeneca. Dehlén and Aronsson [12] utilised the Focused Transformer (FoT) for metabolite prediction. They found that their model's predictions more often resembled the parent molecule rather than the intended metabolite and it struggled to handle multiple metabolites per parent molecule. This limitation may have stemmed from the FoT being a decoder-only model, which processes one input at a time. Exploring a transformer architecture with both an encoder and a decoder, such as Chemformer, could provide more comprehensive insights and potentially yield improved prediction results.

1.1 Aim

The aim of this master's thesis is to optimise the Chemformer model to reliably predict drug metabolites. A dataset of drugs and their metabolic reactions will be gathered and preprocessed to fine-tune the Chemformer model, and thereafter predict metabolites from given drugs. To optimise the performance of the model, established techniques will be explored, such as additional pre-training, ensemble models and augmentation of the data. The final step involves evaluating the performance of the model by comparing it to prior work using an external dataset.

2

Theory

This chapter introduces the project's background, starting with the importance of pharmacokinetics in drug discovery, specifically emphasising drug metabolism. Following this, the field of cheminformatics is presented, highlighting the chemical language SMILES and how it can be used in machine learning. Lastly, an overview of transformers and their role in large language models is provided, focusing on the specific transformer used in this project called Chemformer.

2.1 Pharmacokinetics

Pharmacokinetics is the study of how a foreign substance, a drug, is processed in the body over time. It consists of four steps: absorption, distribution, metabolism and excretion (also known as ADME) [13]. Understanding this process is important in drug discovery and can contribute to finding the most efficient drug with minimal side effects.

Prior to a drug exhibiting its systemic effect, that is within the body, it must be absorbed. *Absorption* is the process in which a drug goes from its administrated form, e.g., a pill, into the blood stream [13]. Medication absorption can directly influence bioavailability, which is the fraction of the drug that reaches the systemic circulation. It is affected by both the quality of the substance, and how it is administrated, e.g., orally or intravenously [13].

Distribution is the process in which a drug or a substance travels throughout the body, typically via the bloodstream to its effector site [13]. A molecule can enter a cell from the bloodstream via either passive or active transport. In passive transport, the molecule moves through the membrane via diffusion, which is the random movement of particles from high to low concentrations [4]. Conversely, in active transport, a molecule moves against the concentration gradient through a membrane protein, a process requiring energy and a transmembrane protein facilitating this transport [4]. Factors such as polarity, basicity and size of the drug can affect its rate of distribution.

The next step in the process is *metabolism*. In this stage, the drug undergoes chemical modification. Figure 2.1 illustrates a drug and its metabolites, highlighting the chemical modifications while preserving structural similarity.

A common chemical modification in drug metabolism is the transformation into a

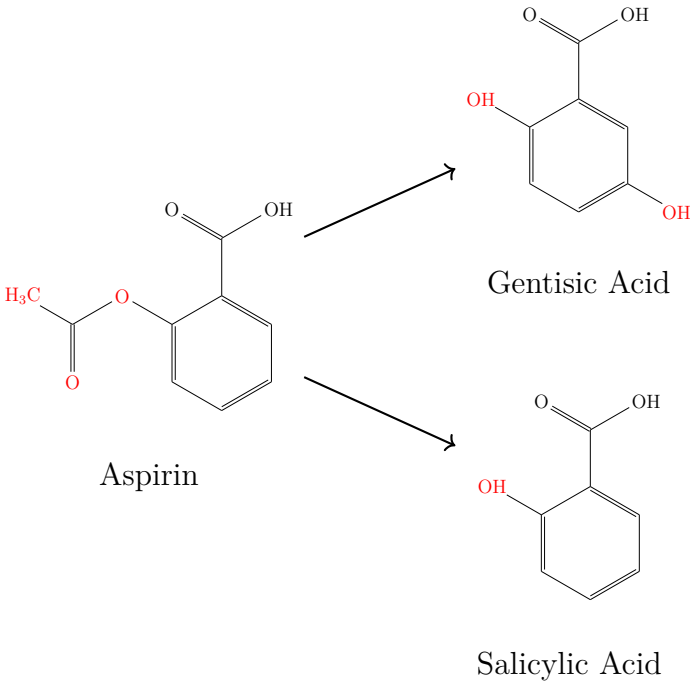


Figure 2.1: The chemical structures of the drug aspirin and its metabolites gentisic acid and salicylic acid.

substance with high water solubility, making it easier to excrete [13]. The partition coefficient P is a measure that describes a substance's tendency to dissolve in either organic solvents (e.g., fats and oils) or water [14]. It is calculated as

$$P = \frac{[\text{organic}]}{[\text{aqueous}]}, \tag{2.1}$$

where [ ] denotes the concentration of the substance in the respective partitions. In simple terms, the partition coefficient measures how much of a substance dissolves in the organic phase versus the aqueous phase in chemical equilibrium. If the substance dissolves mainly in the organic phase, it is called lipophilic (fat-liking) and if it mainly dissolves in water, it is called hydrophilic (water-liking) [14]. A more manageable value to work with is logP, which is simply defined as

$$LogP = log_{10}(P). \tag{2.2}$$

A value of logP = 0 indicates that the concentration in the organic and aqueous phases are equal. In this case, a negative logP value means that the substance is more hydrophilic, whereas a positive value means that it is more lipophilic [14]. A drug that is to be orally administered, which is the easiest and most common form of administration, should have a logP value between zero and five [15].

The reactivity of a drug also influences its metabolic pathway and efficiency. A potential element to consider in this context is the nature of chemical bonds, particularly

carbon-carbon bonds, as all organic molecules consist of carbon. Hybridisation is the concept in which atomic orbitals, functions describing the distribution of electrons around the atomic nucleus, are combined to form new hybrid orbitals. Different types of orbitals exhibit distinct shapes and behaviours. For instance, carbon atoms can undergo hybridisation through linear combination of an s-orbital, shaped like a sphere, and one, two or three p-orbitals, shaped like peanuts, to form sp, sp^2 or sp^3 hybrid orbitals, respectively. Bonds formed between two sp^3-hybridised carbon atoms are exceptionally stable and are rarely cleaved under normal biological conditions [16]. Thus, a drug with a higher fraction of sp^3-hybridised carbon atoms is more likely to change less during metabolism [16].

Lastly, *excretion* is the process of eliminating the drug from the body. Most commonly this is taken care of by the kidneys [13]. A hydrophobic (water-fearing, often synonymous to lipophilic) drug is often metabolised into a more water-soluble compound for easier elimination via urine [17]. Additionally, various other factors can influence the excretion process. This includes molecular factors such as size and pK_a. pK_a represents the pH value at which the concentration of the molecule's acid-form is equal to its concentration in base-form [17], [18]. A drug that is weakly acidic will be excreted more easily in basic urine, while drugs that are weakly basic will be excreted easier in urine that is acidic [17]. Improper excretion of drugs can lead to accumulation in the body with potential toxicity.

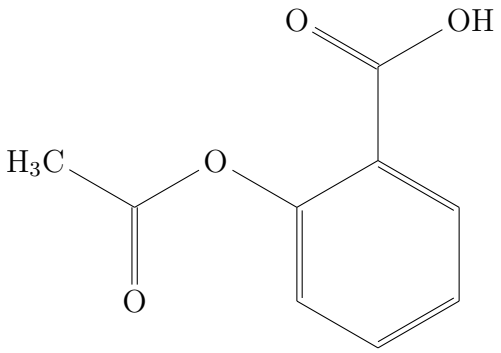
2.2 Cheminformatics

Cheminformatics is an interdisciplinary field of study between chemistry and computational and informational science. It uses *in silico* techniques to solve problems in the chemistry field, meaning to perform experiments with the help of computers and software. A fundamental concept within cheminformatics is the *in silico* mapping of chemical space, referring to the exploration of the set of all possible molecules and chemical compounds. This mapping is crucial in the early stages of drug discovery, as it helps identify promising candidates for further development. RDKit [19] is an open-source cheminformatics and machine learning software, that enables working with molecular data.

2.2.1 Simplified Molecular Input Line Entry System

Simplified Molecular Input Line Entry System (SMILES) is a language specifically developed for chemical information processing [20]. Using SMILES, a molecule graphically represented in two dimensions is written as a single string of characters, which is well-adapted for machine processing. An example is shown in Figure 2.2.

Elements in the organic subset are simply represented as their atomic symbol, e.g., "O" for oxygen and "Cl" for chlorine. Non-organic elements are also represented as their atomic symbol but with square brackets, e.g., "[Ag]" for silver. The organic subset should also be expressed with squared brackets if the number of attached hydrogens does not align with the minimum standard valence that is consistent with the explicit bonds. An atom in an aromatic ring is expressed with a lower case letter;



SMILES: CC(=O)Oc1ccccc1C(=O)O

Figure 2.2: Two-dimensional molecular structure and one-dimensional SMILES of the drug aspirin.

an aromatic carbon is thus written as "c" whereas an aliphatic carbon is expressed as an upper case "C". An aromatic ring is a cyclic molecule with electrons that are delocalised over the ring, typically resulting in increased stability. In contrast, an aliphatic atom is part of an organic molecule joined by single bonds and can be arranged in linear, branched, or cyclic structures, lacking the delocalised electrons characteristic of aromatic compounds.

In SMILES, bonds are expressed as "-", "=", "#", and ":", representing single, double, triple and aromatic bonds respectively. However, single and aromatic bonds can be omitted. Branches are enclosed in normal parentheses, e.g., "CC(C)C(=O)O" for isobutyric acid. When expressing a cyclic structure in SMILES, the two atoms closing the cycle receive a number to mark the beginning and end of the cyclic structure. For example cyclohexane, a cycle of six carbons connected with single bonds, can be written as "C1CCCCC1".

There are often multiple ways to express the same molecule as a SMILES string, which are all equally valid. This can cause significant issues when training a machine learning model. Because of this, canonical SMILES have been defined; a molecule can solely be represented by one specific canonical SMILES string [21]. SMILES strings can easily be canonicalised using chemical standardisation procedures offered by software packages such as RDKit [19]. To form a canonicalised SMILES, the atoms in the molecule need to be numbered distinctly and consistently [22]. Thus, the SMILES can be generated in that same order every time. Table 2.1 shows the same molecule expressed in two different ways in the left column. When canonicalised, it is evident that the SMILES strings are identical.

Table 2.1: Non-canonical and canonical SMILES of the molecular structure shown in Figure 2.2.

Non-canonical SMILES	Canonical SMILES
c1(C(O)=O)c(OC(=O)C)cccc1	CC(=O)Oc1ccccc1C(=O)O
O=C(Oc1c(C(O)=O)cccc1)C	CC(=O)Oc1ccccc1C(=O)O

As mentioned previously, a single molecule can have several different non-canonical SMILES representations. This can be used as an advantage when more data is

required for machine learning purposes, e.g., for data augmentation purposes [21]. To form a randomised SMILES, the ordering of the atoms can be randomised. This will affect the starting atom of the SMILES and in what order the branches are written [23]. A randomised SMILES can be generated by unrestricted or restricted atom order. Restricted ordering avoids forming substrings in the SMILES, e.g., c1cc(c(cc1)), by including side-chains in the order as they come when traversing a path [23]. Augmenting data by randomising SMILES has shown to improve model performance [23]. Using restricted rather than unrestricted randomisation shows to be slightly more ideal. Training on randomised SMILES is expected to create a better and more generalisable model for several reasons [23]. For example, using solely canonicalised SMILES implies that the model does not only have to learn to output valid SMILES strings but also how to produce a canonicalised SMILES. Parallels can also be drawn to image classification [23]. An image recognition task, a machine learning model, which should predict whether something is a cat, will be more reliable if trained on pictures of different angles of cats and not just the ones from the front [23]. With similar reasoning, it is advantageous to train a model on molecules from different angles, i.e., differently randomised SMILES strings.

2.2.2 Molecular Fingerprints

Molecular fingerprints are used to represent chemical structures of molecules for data science applications, either as binary or numerical vectors [24]. The Morgan fingerprint [25] is a popular choice of fingerprint and is a binary vector based on the molecule's topology, with each bit representing a specific substructure of the molecule [25]. The two key parameters that can be adjusted are the radius and the vector length. The radius determines the size of the atom groups considered, with a larger radius capturing larger structural features. The vector length is related to the hashing process. Hashing is used because the total size of molecular features scales with the chemical space considered, requiring compression into a fixed number of bits. This process can lead to a single bit representing multiple atom groups due to possible collisions.

The aim of using fingerprints is to compare molecules, i.e., estimate chemical similarity. This can be achieved using the Tanimoto similarity score which quantifies the overlap of features shared between two molecules [26]. The Tanimoto coefficient can be defined as

$$\textit{Tanimoto coefficient} = \frac{c}{a + b - c}, \tag{2.3}$$

where a and b represent the number of bits set to 1 (meaning that the corresponding feature is present) in each molecule's fingerprint bit-string, and c denotes the number of bits set to 1 in *both* molecules. The coefficient measures the ratio of shared features to the total numbers of features present, giving a score between 0, when no bits are the same, and 1, when all bits are the same, thus reflecting their similarity [26]. Tanimoto scores can be used to cluster molecules based on their similarity, e.g.,

using the Butina algorithm [27]. A distance matrix and a distance threshold are used in the algorithm for this purpose.

2.3 Transformers

The transformer is a type of neural network architecture that can be applied for various machine learning tasks [28], [29]. Transformers typically consist of two main components: an encoder and a decoder. The encoder processes an input and generates a fixed-size vector, encapsulating the essential features of the data. This vector then serves as input for the decoder, which produces the output. The defining feature of the transformer architecture is the reliance on only attention mechanisms [28], unlike previous architectures which relied heavily on recurrence and convolutions [30], [31].

Transformers use an attention mechanism known as self-attention to establish relationships between the tokens within an input vector, thereby enabling it to compute a representation of that input [28]. Tokens, the units that compose the input, vary depending on the task. For instance, a token can be a character, subword or word [32]. The attention function defines a process that maps a query with a collection of key-value pairs to produce an output [28]. In this context, the query, key, and value are all vector representations of tokens from the input. Each token has a corresponding query, key, and value vector that are used to determine the relationships between tokens. The output is determined by a weighted sum of the values, where the weights indicate how much a query resembles a key [28].

Introduced in 2017 by Vaswani *et al.* [28], the transformer achieved state-of-the-art performance as a sequence-to-sequence model. Sequence-to-sequence models follow the concept of translating one sequence (an input) into another (an output) and are a type of large language model (LLM). This revolutionised the field of natural language processing (NLP) [32], [33], which uses computational methods to represent and analyse texts sourced from real-world usage, aiming to perform tasks with human-like language understanding [34].

Inspired by the success of the transformer architecture in large language models, these models have been adapted for chemical structure notations, such as SMILES (Section 2.2.1). This adaptation makes it possible to undertake tasks like molecular property prediction, *de novo* molecular design, and synthesis prediction [35]–[38]. The approach is particularly suitable as SMILES represent two-dimensional molecular structures as linear strings, similar to natural language [20]. The simplicity and compactness of SMILES make them particularly advantageous for the application of LLMs in chemical contexts.

A typical LLM is usually pre-trained on generic data. After this initial pre-training, the model can be further trained on data specific to a particular task. This process, known as fine-tuning, involves adjusting the pre-trained model's weights to enhance its performance on the specific task.

During both pre-training and fine-tuning, the learning rate is an important pa-

rameter that adjusts the pace of learning. A popular optimiser that determines the optimal learning rate is Adam, introduced in 2014 by Kingma and Ba [39]. This stochastic optimisation method is memory-efficient and adaptively computes the learning rates from estimates of the mean and variance of the gradients of the stochastic function. This has shown to be a robust way of optimising the learning rate [39].

To monitor training progress, evaluation is performed on a validation set. A common metric in machine learning is cross entropy loss [40]. This measure estimates the difference between the true and the predicted outcomes, and adjusts the weights of the machine learning model accordingly. A lower cross entropy loss typically indicates better model performance.

For making predictions in large language models, a frequent choice is to use beam search. As a heuristic search algorithm, it follows the most promising paths without exploring every possible option [41]. While there is no guarantee of finding the most optimal solution, beam search offers significant advantages in comparison to exhaustive searches by improving efficiency in large search spaces through limiting node expansion. Its reduced memory requirements make it ideal for tasks with complex and extensive solution paths. The method is broadly used in fields such as natural language processing [42], and has shown success in chemical prediction tasks [11], [43]. Beam search works by initialising a root node and generating B successor nodes, where B is the branching factor. Central to beam search is the heuristic function, which is used to evaluate the potential of each node, selecting the top W, known as the beam width. Node expansion and selection are iterated until no more nodes are available or when the the goal is attained. The beam search algorithm is illustrated in Figure 2.3.

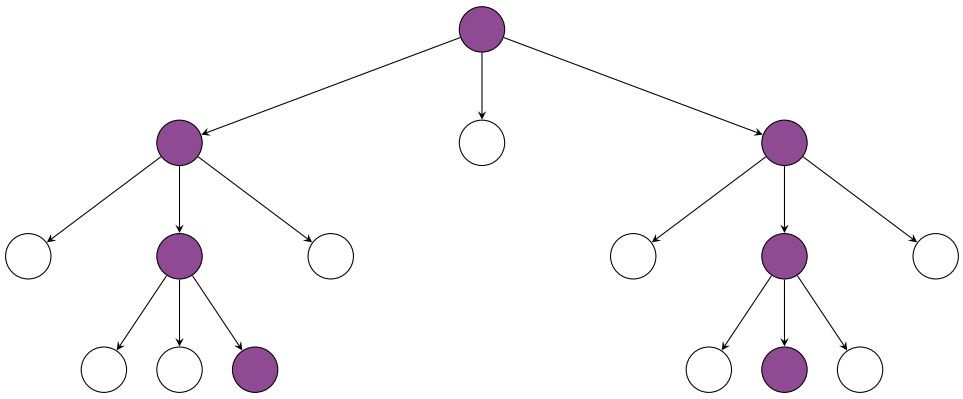


Figure 2.3: An illustration of the beam search algorithm where the branching factor (B) is 3, and the width of the beam (W) is 2. The coloured nodes are the ones selected for further expansion. The algorithm continues until the goal is met or when there are no more nodes available.

One of the most crucial factors determining the success of a large language model is the data it is trained on. In principle, larger [44] and more carefully curated [45] datasets typically lead to better model performance. Utilising diverse data is crucial for reducing the risk of over-fitting. One technique used to improve generalisation

is called masking. In this approach, random tokens within the data are replaced with a masking token, increasing the importance for the model to learn to rely on neighbouring tokens. This encourages the model to generalise better, thereby enhancing its overall performance.

A major challenge in applying LLMs for chemistry is specifically the lack of qualitative and quantitative data [46]. Chemical data, typically obtained in laboratory experiments, is varying in terms of quality and quantity due to the influence of human and environmental factors, such as temperature and humidity. Moreover, this data is often securely held within private companies and academic institutions, which results in limited availability of publicly accessible datasets. Consequently, accessing comprehensive and reliable chemical data for research purposes is a significant challenge.

2.3.1 Chemformer

Chemformer is a large language model that has a transformer-based architecture with both an encoder and a decoder and is well-suited for chemical sequence-to-sequence tasks [11]. It uses SMILES as inputs and outputs, i.e., taking an input in the form of a SMILES string from which a new SMILES string is generated. The Chemformer model [11] was pre-trained on around 100 million SMILES from the ZINC-15 database [47], where the task was to predict the same molecule as the input. The intention of this approach was for the model to learn and understand the chemical language of SMILES. The dataset was split into train, validation and test set, with each set being assigned 99%, 0.5%, and 0.5% of the data points respectively. Two different SMILES modification techniques were implemented in order to improve the model's performance, namely randomisation and masking. Three models were trained using SMILES that were augmented with randomisation, masking, or a combination of both. In Chemformer, the default randomisation was implemented in a restricted way (see Section 2.2.1). This randomisation is based on a manually set hyperparameter that determines the probability of each SMILES string being randomised. The default masking in Chemformer masks a span of tokens, where the span length is determined by a Poisson distribution [48]. The masking probability determines the likelihood of each token being masked.

After pre-training, the model was fine-tuned on tasks such as direct reaction prediction, retrosynthesis prediction and molecular optimisation. For direct reaction prediction, the input is a substrate and the model is intended to predict its product. Retrosynthesis is the reverse action, i.e., starting with the product and predicting its substrates. For molecular optimisation, the input molecule is tagged with an optimisation token, for instance "increase solubility", and the model should output an optimised molecule in consideration to the token. In both direct reaction and retrosynthesis, Chemformer outperformed other existing models in top-1 prediction. The molecular optimisation task was difficult to compare to other models due to different evaluation criteria, but generated more desirable molecules than other approaches.

In Chemformer, the Adam optimiser is used to adaptively adjust the learning rate.

Validation loss is estimated through cross entropy loss to evaluate model performance during training. Additionally, beam search is integrated for predictions.

Irwin *et al.* [11] concluded that for both direct reaction prediction and retrosynthesis prediction, the pre-trained Chemformer model with both masking and randomisation performed the best, while the model using only masking performed better on the molecular optimisation task. Since metabolite prediction is a task analogous to retrosynthesis prediction, the combined pre-trained model was chosen for this project.

3

Methods

This section describes the data sources used in this project, along with the curation process of the dataset. Furthermore, a description of the experimental pipeline is presented as well as the scores used to evaluate performance. Additionally, optimisation strategies implemented to enhance model performance are explained.

The code for this project can be found at https://github.com/tsofiac/LAGOM.

3.1 Data Sources

Data from MetXBioDB [8] and DrugBank [49] were utilised for fine-tuning, while a test dataset from GLORYx [7] was used for external evaluation. All datasets are publicly available.

The MetXBioDB and the Drugbank datasets each contained one column with the drug SMILES (input) and one with the metabolite SMILES (output). If a drug had multiple metabolites, it appeared multiple times in the dataset. The relationship between the drug and metabolite is analogous to parent (input) and child (output). Metabolic reactions typically occur in multiple steps, which is illustrated in Figure 3.1, where all the metabolites in a multistep reaction share the same originating drug. This implies that some metabolites can act as drugs themselves, but with another drug as their origin. To keep track of the multistep reactions, each metabolic reaction was labelled with its originating drug. See an example of the dataset with repeated drug SMILES and assigned origins in Table 3.1.

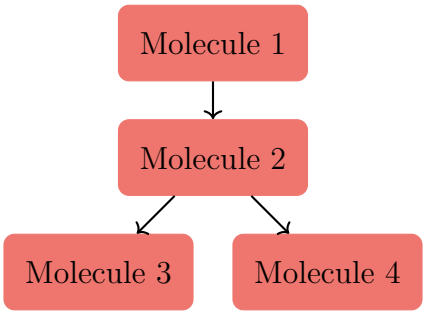


Figure 3.1: Example of a multistep reaction. All three reactions have Molecule 1 as their originating drug.

Table 3.1: Example of the dataset format with assigned origins. SMILES 1-4 represent the molecules in Figure 3.1.

Drug	Metabolite	Origin
SMILES 1	SMILES 2	SMILES 1
SMILES 2	SMILES 3	SMILES 1
SMILES 2	SMILES 4	SMILES 1
SMILES 5	SMILES 6	SMILES 5
SMILES 5	SMILES 7	SMILES 5
SMILES 8	SMILES 9	SMILES 8

3.1.1 MetXBioDB

The biotransformation database MetXBioDB, used for fine-tuning, contains information on metabolite reactions and is extracted from the database behind BioTransformer [8]. The purpose of the database is to assist in the development of metabolite prediction models [8], and is therefore well suited for metabolite prediction. It contains over 2000 unique and experimentally confirmed biotransformations that are extracted from six publicly available datasets including DrugBank [49] and SuperCYP [50], as well as from scientific articles [8].

The latest version of MetXBioDB, accessed in January 2025, was used in this project. The InChI (International Chemical Identifier), a structure-based chemical identifier [51], of each molecule in the dataset were converted to SMILES using the Python library RDKit [19]. In total, MetXBioDB contains around 2100 reaction pairs.

3.1.2 DrugBank

DrugBank, also used for fine-tuning, is a public online database that contains information about drugs and drug targets [52]. It originated in 2006 at the University of Alberta, as an initiative aimed at providing researchers with detailed and organised information on drugs, and is now a part of The Metabolomics Innovation Center (TMIC). DrugBank is publicly available, but accessing the data files requires a licence that can be obtained for free when used for academic purposes.

Metabolic reaction data was extracted from DrugBank (version 5.1.13) alongside drug and metabolite structures. Subsequently, the metabolic reactions were matched with their corresponding structures, written as SMILES, if available. In total, the DrugBank dataset contains about 3500 reaction pairs.

3.1.3 Test Dataset from GLORYx

The test dataset from Bruyn Kops *et al.* [7] is a manually assembled dataset that stems from scientific literature, and is based on the top 100 best-selling drugs in 2018 [7]. The data was assembled to be used for evaluation of the metabolite prediction tool GLORYx. To be included in the dataset, the metabolite must have been detected in the human body, and the dataset only includes drugs that are solely

composed of the elements hydrogen (H), carbon (C), nitrogen (N), sulphur (S), oxygen (O), fluorine (F), chlorine (Cl), bromine (Br), iodine (I) and phosphorus (P) [7]. In total it includes 37 unique drug molecules, and 136 first-generation metabolites. In this project, the test dataset from GLORYx was used as an external test set for benchmarking the best-performing model.

3.2 Data Preprocessing

The extracted data for fine-tuning was curated to ensure that it suited the task of drug metabolism prediction and to create a more homogeneous dataset. The following steps were conducted to filter the data. A summary is shown in Figure 3.2.

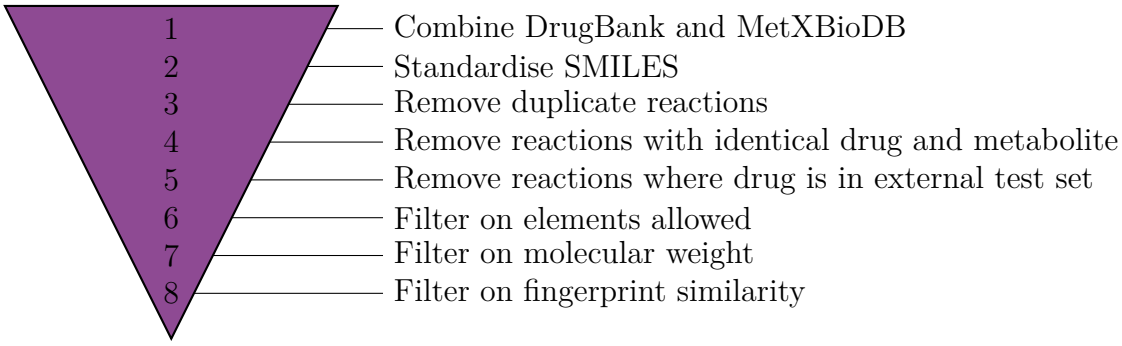


Figure 3.2: A summary of the data curation process for the metabolic dataset, containing the MetXBioDB and DrugBank datasets.

As a first step, the parsed DrugBank dataset and MetXBioDB dataset were combined into one file. This file contained 5619 data points and will be called the metabolic dataset from now on.

As a second step, all SMILES in the metabolic dataset were standardised. To better reflect the chemical reality in the body, any salt components were removed from each molecule. If this process resulted in disconnected fragments, only the largest fragment was retained. Next, stereochemistry, referring to the three-dimensional arrangement of atoms, was removed from the molecules to further simplify the SMILES strings. This was done because two nearly identical molecules that differ only in stereochemistry can produce different metabolites, creating additional complexity for the model. The resulting SMILES were written in their canonical form.

As a third step, duplicate rows were removed, i.e., if both the parent SMILES and child SMILES were identical across multiple data points, only one instance was kept. The appearance of identical rows resulted from existing duplicates in the raw data, due to the combining of two data sources, as well as from the standardisation process, as this could make certain molecules identical. The majority of the duplicates emerged when stereochemistry was removed, while a smaller portion emerged from the removal of salts and the canonicalisation process. Consequently, 1101 data points were removed due to duplicate rows.

As a fourth step, data points with identical parent and child SMILES were removed. This was done as equal molecules can indicate that the parent has not been metabolised, that there is a flaw in the data, or that the standardisation process has made the SMILES identical. In this step, 56 data points were removed.

As a fifth step, the metabolic dataset was compared with the GLORYx test dataset, and reactions with overlapping drug molecules were removed from the metabolic dataset. This was essential to ensure an unbiased benchmarking of the results against the external test dataset. In this step 93 data points were removed.

Up to this point, the curation process was solely based on binary criteria. The remaining filtering process was decided based on cut-off criteria from the data analysis process.

In the next step, both parent and child molecules were filtered on allowed elements, namely C, N, S, O, H, F, I, P, Cl, and Br, as these are the most frequent elements in organic compounds. The distribution of the elements in the metabolic dataset, where H is excluded, can be seen in Figure 3.3. The same ten elements were also seen in the GLORYx test dataset [7]. In contrast to Sícho *et al.* [53], who curated data for their tool FAME 3 for prediction of sites of metabolism (SoMs), and Dehlén and Aronsson [12], we excluded the elements silicon (Si) and boron (B) since these are less common in drug molecules. As seen in Figure 3.3, excluding silicon did not have an effect on the dataset as none of the molecules included that element. Boron only had 11 occurrences in the metabolic dataset, and considering the low frequency, this confirmed the decision to exclude it. Beyond boron, the dataset contained some metal elements with even less occurrences that also were removed. In total, 36 data points were removed.

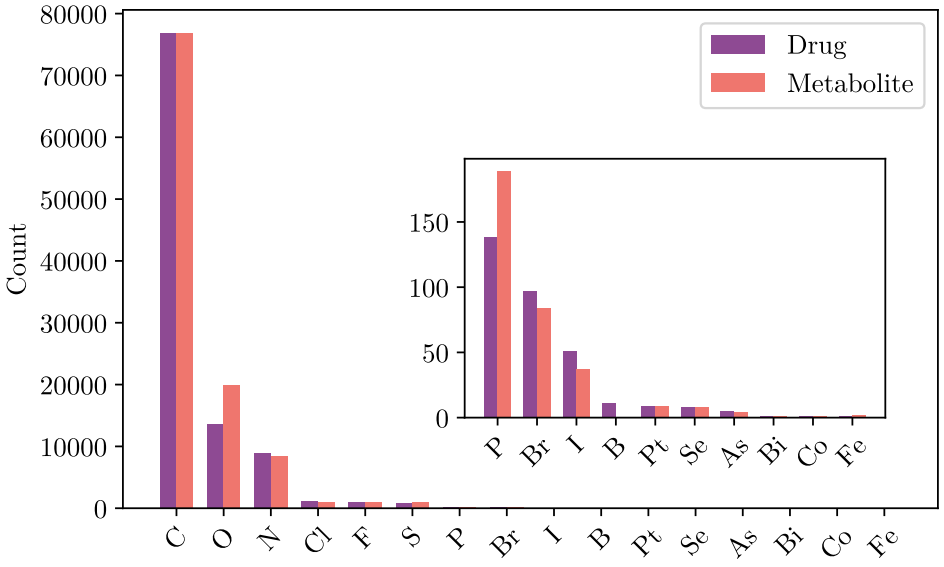


Figure 3.3: Distribution of elements in the metabolic dataset, in descending order with respect to the drugs, before the filtering process was conducted. A close-up is provided of the elements with the fewest counts.

Thereafter, the dataset was filtered based on molecular weight of the parents. Only drug molecules that weighed over 100 Da and under 750 Da were kept. Sícho *et al.* [53], who predicted sites of metabolism, kept molecules in the range 100 Da to 1000 Da. The decision to have a lower upper limit in this project is supported by the molecular weight distribution of the molecules as seen in Figure 3.4. The figure shows that a majority of the data points, more precisely 97% at this stage of the filtering, were located in the chosen weight span. Note that Figure 3.4 shows the distribution of the dataset before and after the full filtering process. The dataset was not filtered based on the molecular weight of the metabolites. The motivation for this is that some metabolic reactions lead to an increase of molecular weight rather than a decrease, and an ideal weight span for metabolites is therefore complicated to determine. An example of this is the common biosynthetic reaction glucuronidation, which is when glucuronic acid is attached to the molecule [54]. This adds about 200 Da to the compound, which justifies not specifying a weight span for metabolites. Regardless, the weight distribution of the metabolites also became more homogeneous after filtering. In this step 143 data points were removed.

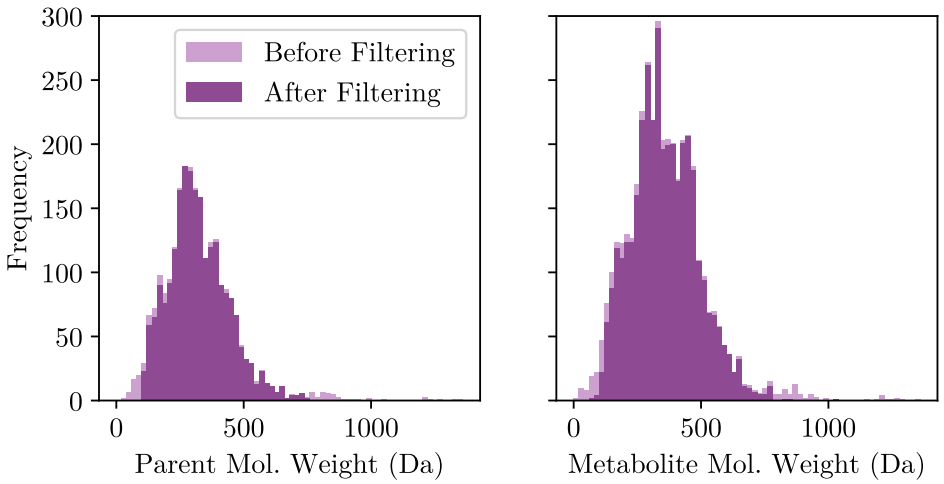


Figure 3.4: Comparison of the molecular weight distribution of the metabolic dataset, before and after all filtering steps. Weights of the unique drug molecules are shown to the left, and unique metabolite weights to the right. Note that there is a drug with molecular weight > 4000 Da that is not included in the graph.

As a last step, fingerprint similarity between each parent and its respective child molecule was analysed. The Morgan fingerprints, as mentioned in Section 2.2.2, were calculated and the Tanimoto similarity score was estimated. This score ranges from 0 to 1, where a higher value indicates greater structural similarity. The distribution of the fingerprint similarity score before and after the complete filtering process is shown in Figure 3.5. Based on the distribution, reactions with a similarity score lower than 0.20 were excluded as these may have constrained artefactual metabolic transformations, introduced by the original sources, and would be too difficult for models to confidently predict. In this step 135, data points were removed.

The final curated metabolic dataset contained 4055 reactions and its partition of

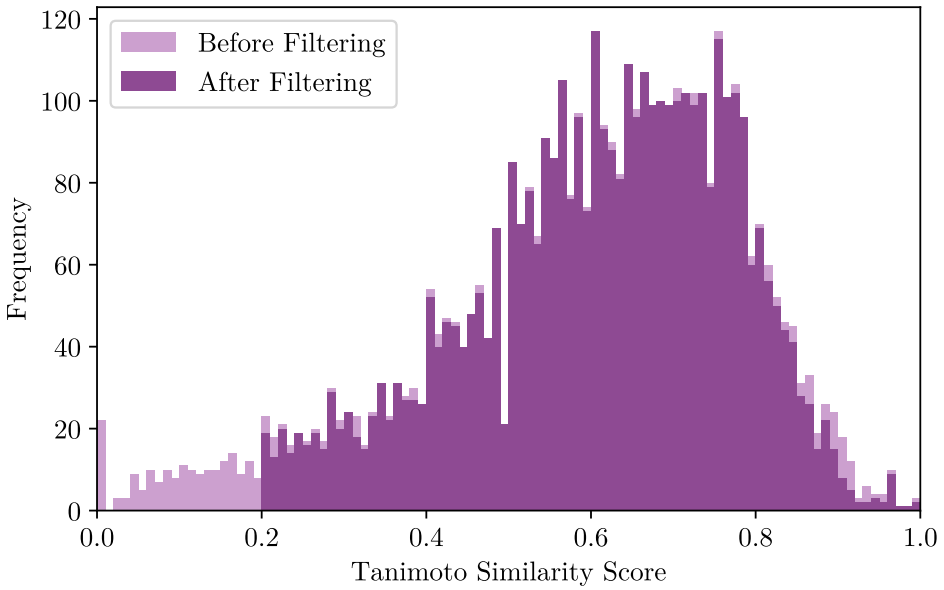


Figure 3.5: Tanimoto similarity score based on Morgan fingerprints, comparing each metabolite with its parent. Distribution of the metabolic dataset from DrugBank and MetXBioDB, before and after all filtering steps.

DrugBank and MetXBioDB data can be seen in Figure 3.6. A portion of the resulting metabolic dataset, specifically 13 %, contained reactions that were present in both DrugBank and MetXBioDB.

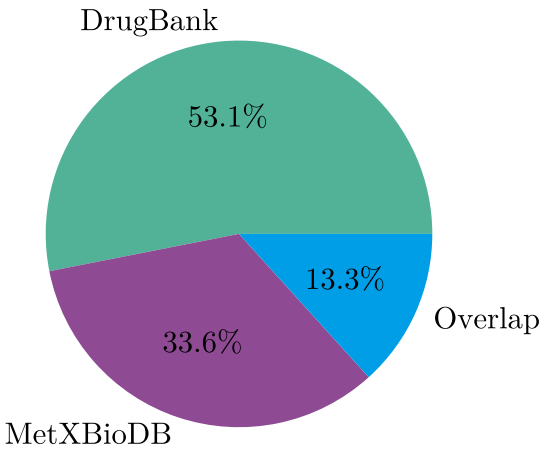


Figure 3.6: Pie chart of the fully curated metabolic dataset.

The distribution of metabolites per drugs can be seen in Figure 3.7. This shows that a majority of the drugs have only one metabolite.

3.2.1 Data Splitting

The metabolic dataset was split into a training set, validation set and test set. Since metabolic reactions can occur in multiple steps, beginning with one origin drug that

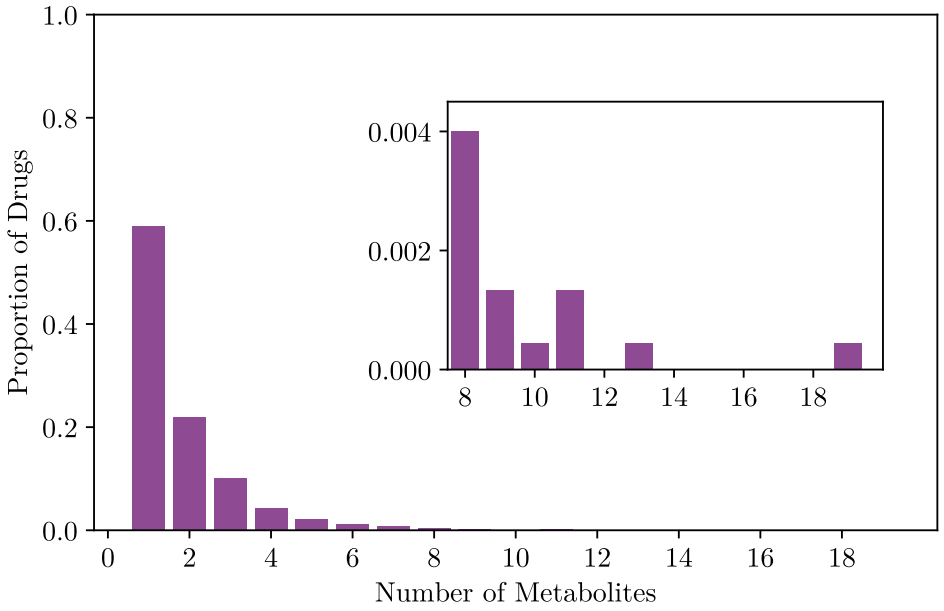


Figure 3.7: Distribution of drugs by number of metabolites. The drug with the most metabolites has 19 of them. A close-up view is provided of the less frequent number of metabolites.

is metabolised into a metabolite that can metabolise further, the division was based on the origin of each drug. The origin refers to the first drug in a multistep reaction, ensuring that all drugs with the same origin were in the same set. Specifically, 85 % of the data was allocated for training, 10 % for validation, and 5 % for testing. The validation set was used for scoring during the fine-tuning, whereas the test set was used for testing the performance afterwards. These sets were kept consistent for all different setups of the model in the project.

3.3 Experimental Pipeline

After completing the data preparation, the Chemformer architecture was utilised. This included choosing the best-performing model during fine-tuning, followed by postprocessing and scoring of the predictions.

3.3.1 Scoring During Fine-Tuning

When fine-tuning in the Chemformer architecture, a beam search may be conducted to predict metabolites in the validation set. This was utilised, with a beam width of 5. Given that drugs often have multiple metabolites, the true metabolite in a specific reaction is not necessarily the only correct one. To address this, a scoring function was implemented to account for all possible metabolites when evaluating predictions as well as a column listing all true metabolites for each drug was added for comparison.

The scoring function was implemented as follows. For each drug, the beam search

yielded predicted metabolites. These were canonicalised for fair comparison. The score was calculated as the ratio of the number of true metabolites found in the predictions out of all true metabolites. This score is called recall and is described in more detail in Section 3.3.3. The scoring function, together with validation loss, was used to monitor the progress during fine-tuning that was run for 200 epochs. The three models with the highest scores were saved. From these, the model from the lowest epoch was selected for evaluation, provided that the validation loss had reached a minimum. The lowest epoch was chosen because, beyond this point, the validation loss tends to increase, indicating over-fitting. This means that the epoch chosen is closest to the validation loss minimum among the three. This strategy ensured that the selected model had one of the highest scores while also maintaining a low validation loss, which is indicative of better performance. Figure 3.8 shows an example of the change of the scoring function (metabolite recall) and validation loss during a fine-tuning. Table 3.2 displays the specific values for the three epochs saved based on metabolite recall score. Here, epoch 38 was chosen for evaluation.

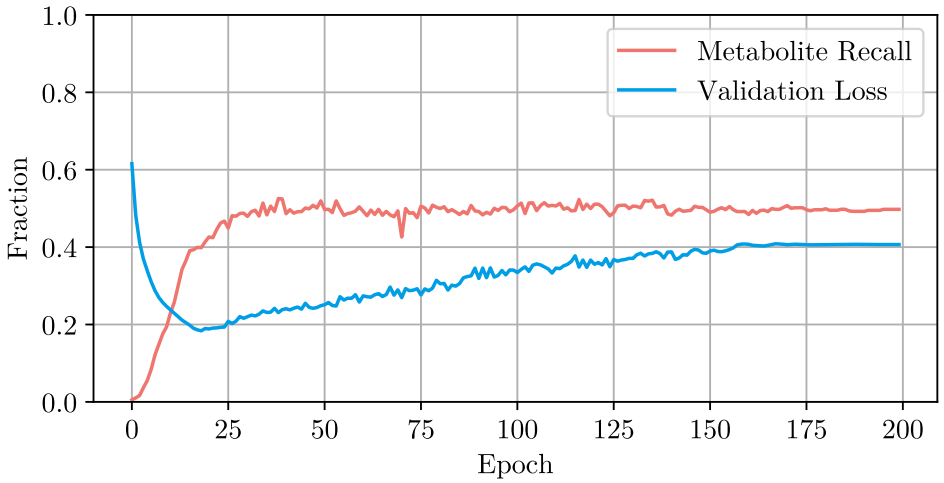


Figure 3.8: Metabolite recall score and validation loss during a fine-tuning of the metabolic dataset.

Table 3.2: Metrics values for the fine-tuning shown in Figure 3.8, showing specific values for the three epochs that were saved based on metabolite recall score.

Epoch	Validation Loss	Metabolite Recall
38	0.23	0.53
39	0.24	0.52
116	0.36	0.52

Fine-tuning was conducted on the pre-trained Chemformer model, mentioned previously in Section 2.3.1 as the best-performing model for the retrosynthesis task, analogue to metabolite prediction.

3.3.2 Data Postprocessing

After a model was fine-tuned, it was set to predict 20 metabolites per drug from the test dataset. In order to evaluate the predictions fairly, postprocessing of the predicted SMILES was needed. First of all, the predicted SMILES were standardised, ensuring the predictions were written in the same format as the true metabolites. This means that invalid SMILES were simultaneously removed from the predictions. Thereafter, the predicted SMILES were compared to one another to remove duplicates, as well as compared with their respective parent to remove the predictions identical to the parent. The remaining SMILES were therefore all valid and unique.

3.3.3 Performance Metrics

Various scoring functions were implemented to evaluate the models and to use benchmarks for comparison of the performance. The metabolic test set was divided into four equally sized batches, with a size of approximately 38, and scores for each batch were calculated for more reliable results. For the GLORYx test dataset, the same batch size of 38 was used, resulting in just one batch for the external test. Note that different scoring functions were used depending on the comparison.

The first score implemented was SMILES string validity. This score calculated the percentage of valid SMILES predictions out of all generated predictions.

The second and third scores used were accuracy based on whether at least one true metabolite was correctly predicted for each drug, and whether all true metabolites were correctly predicted for each drug. This was assessed based on the top-n predictions, where n is an integer greater than zero, meaning that only the top n predicted unique and valid SMILES were considered when evaluating accuracy. These accuracy scores demonstrated how well the model predicted metabolites for each drug and indicated whether the correct predictions were highly ranked by the beam search.

The fourth and fifth scores used were precision and recall. These scores vary between zero and one, where a higher score indicates better performance, and were also evaluated based on the top-n predictions. In comparison to the previous score, these scores demonstrate the overall performance to predict metabolites and rank them high by the beam search. High precision indicates a low number of incorrect predictions, while high recall reflects a high proportion of correctly predicted metabolites among all true metabolites in the dataset. These were defined as follows:

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN}, \tag{3.1}$$

where TP denotes *true positives*, i.e., the correctly predicted metabolites, FP denotes *false positives*, i.e., the valid but incorrect predictions, and FN denotes *false negatives*, i.e., the true metabolites not identified in the predictions.

As seen by the definitions, precision and recall coordinate with each other, meaning maximising one of them typically leads to a loss of the other and vice versa. To take

both scores into consideration, a sixth score was also used, called F_1. This score is a harmonic mean of precision and recall and varies between zero and one, where higher is better, and is estimated as follows:

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}. \tag{3.2}$$

3.4 Performance Optimisation

With the aim of enhancing model performance by improving predictive accuracy and robust generalisation, several methods were investigated. This chapter presents five strategies: pre-training, data augmentation, data annotation, the implementation of ensemble models and optimising the prediction space.

3.4.1 Pre-Training

Pre-training is an initial training on generic data, where a model is able to learn generic features and patterns in the dataset. Despite being the earliest stage, pre-training can significantly affect a model's performance as it establishes the foundational knowledge. Adapting the pre-training process therefore has the potential of optimising overall model performance. In addition of evaluating the existing pre-trained Chemformer model, the usage of another dataset, namely the virtual analogue dataset, was explored.

To discover novel compounds, a method was developed by Dimova and Bajorath [55] to create virtual compound analogous to known ones. This method used known compounds found in the ChEMBL database [56] which contained bioactive drug-like small molecules. In essence, the process began with the extraction of active compounds known as matched molecular pairs [55]. These are compounds that differ by a chemical modification at a single site. Next, these pairs were fragmented into cores, referring to the part of the compound that remains mainly unchanged, and R-groups, which are the chemical modifications at the single site. The resulting fragments were then recombined to generate new virtual analogue compounds. These new virtual analogues underwent retrosynthetic analysis to verify their synthetic accessibility. In this dataset [57] the virtual analogues, presented as SMILES, were matched with the corresponding known compound, presented with ChEMBL-ID. The conversions between the virtual analogues and their known compounds resemble metabolic transformations as they maintain each compound's core structure while sampling a multitude of different modifications.

This dataset [57] contains approximately 1.3 million virtual analogues, each associated with around 150 000 known unique compounds, resulting in a total of around 11 million reaction pairs. In this project, CHEMBL35 (December 2024) [58] was used to add the SMILES of the ChEMBL-ID to the dataset. This dataset was used to pre-train models in this project.

The virtual analogues dataset was preprocessed in the same way as the metabolic

dataset. Beyond removing the reactions with parents identical to any of those in the GLORYx test dataset, the overlapping reactions with the metabolic dataset were also removed. This resulted in a dataset of about 10.8 million reactions.

The virtual analogue (VA) dataset was split into a training set and a validation set randomly. The ratio was 99.5% for training and 0.5% for validation, which is the same ratio as for the pre-trained Chemformer model [11], when excluding the test set. The validation set was used to monitor the validation loss during training. No test set was needed since the test set from the metabolic dataset was used for testing the performance of the pre-training.

As an initial pre-trained model in this project, the previously mentioned Chemformer model by Irwin *et al.* [11] was used. This model used both randomisation with a probability of 0.5 and masking with a probability of 0.1, and was trained for two days. In order to explore the possibility of enhancing the performance of a metabolite prediction model, an additional pre-training was explored using the VA dataset. This second model, the ChemVA model, was built upon the existing pre-trained Chemformer model, undergoing further pre-training with the VA dataset for an additional two days. The pre-training with this dataset was developed in four variants: one using only randomisation at a probability of 0.5 (Rand), another using only masking at a probability of 0.1 (Mask), a third combining both approaches (Comb) and a baseline model with no augmentation (Base). Anytime that masking or randomisation was implemented in this project, the mentioned probabilities were used. To evaluate the effectiveness of the VA dataset independently, pre-trainings not based on the Chemformer model were conducted mirroring the previous pre-training process on the dataset. All the pre-trained models are summarised in Table 3.3.

Table 3.3: Pre-trained models. The numerical values under "Augmentation" represent randomisation and masking probabilities.

Model Name	Pre-Trained	Augmentation
Chemformer	None	Rand. 0.5 and Mask. 0.1
ChemVA Base	Chemformer	None
ChemVA Rand	Chemformer	Randomisation 0.5
ChemVA Mask	Chemformer	Masking 0.1
ChemVA Comb	Chemformer	Rand. 0.5 and Mask. 0.1
VA Base	None	None
VA Rand	None	Randomisation 0.5
VA Mask	None	Masking 0.1
VA Comb	None	Rand. 0.5 and Mask. 0.1

A comparison of the different pre-trained models was conducted based on evaluation on the test set from the metabolic dataset of MetXBioDB and DrugBank, without any fine-tuning, using the Chemformer model as a reference. In order to determine the best-performing model, an investigation was done on SMILES string validity. A high validity means that the model is at least generating valid SMILES strings

which is a good indication for predicting reasonable metabolites. Since the Chemformer model used both randomisation and masking, a comparison of SMILES string validity was made on each model's corresponding augmentation (Comb).

Thereafter, each variant of the best-performing model was investigated. This was again done by comparing the validity. The best-performing model variant was then selected to proceed with. As a verification step, the selected model was fine-tuned and compared to the initial fine-tuned model, and the best-performing fine-tuned model was selected to continue with.

3.4.2 Data Randomisation

As a first fine-tuning performance optimisation method, the built-in data randomisation in the Chemformer architecture was explored. The best-performing pre-trained model was fine-tuned with a randomisation probability of 0.5 and compared to the model without randomisation. The best-performing fine-tuning approach was chosen to continue with.

3.4.3 Data Augmentation

Assuming the validity of the data, increasing the size of the dataset is a fundamental strategy for enhancing model performance. Given the relatively small dataset of approximately 4000 data points curated for this project, augmentation techniques were employed to expand the dataset with the aim of enhancing model performance. In this project, two augmentation techniques were explored.

The first technique involved extending the dataset by generating new reactions from the existing ones. Since metabolic reactions typically occur in multiple steps, every reaction can be connected to an originating drug, illustrated in Figure 3.1. Due to this, new reactions were generated by connecting the origin drug to all of its metabolites' metabolites. These new reactions were then curated in a manner consistent with the original dataset to maintain coherence. This generated 796 new reactions to the training dataset.

The second technique involved generating new reactions by linking each parent to itself, which is representative of a drug that does not undergo metabolism. Given that a metabolite typically resembles its parent, this approach enhances the model's ability to capture these similarities more effectively. This generated 1923 new unique reactions to the training dataset.

Furthermore, the newly generated reactions were added to the metabolic dataset, both separately and together. The best-performing pre-trained model was then fine-tuned with each of the extended datasets and their performances were evaluated on the metabolic test set.

3.4.4 Data Annotations

Enhancing the input data with descriptive annotations provides the model with additional information, enabling it to find new patterns to be able to distinguish

between the inputs, with the potential of optimising model performance.

The descriptive annotations used in this project were drug properties important for drug metabolism. The properties explored were LogP (hydrophilicity) and carbon sp^3 fraction as these properties may affect ADME, as described in Section 2.1. This was done by calculating these properties for each drug molecule and sorting them into groups depending on their respective value. These blocks were codified as tokens and subsequently appended to the drug molecule. LogP annotations, carbon sp^3 fraction (csp3) annotations, or both simultaneously, were added to the input SMILES strings. An example of an annotated SMILES string is shown in Table 3.4.

Table 3.4: Example of a SMILES string with LogP and carbon sp^3 fraction annotations.

SMILES	Annotated SMILES
Oc1ccc(O)cc1	[LogP_[1.0, 2.0)][csp3_[0.0, 0.1)]Oc1ccc(O)cc1

The best-performing pre-trained model was then fine-tuned using the metabolic dataset annotated in one of the three specified methods. Note that randomisation was not combined with annotations, since this resulted in invalid SMILES. The performance of these models was assessed on the metabolic test set, which was annotated in the corresponding manner.

3.4.5 Ensemble Models

With the aim of correctly predicting a greater range of metabolites, the concept of ensemble models was explored. An ensemble model was produced by combining four models that were fine-tuned on different splits of the metabolic dataset. Three different approaches to determine the data splits were examined.

The models may face challenges in accurately predicting all metabolites of a drug with multiple metabolites due to the varying number of metabolites per drug in the dataset, see Figure 3.7. Thus, one of the explored approaches to split the data was to decrease the number of metabolites per drug in each model. If a drug had more than one metabolite, these were put into different splits. If it had fewer metabolites than the number of splits, the drug and a metabolite was also put in the remaining split(s). For drugs with only one metabolite, it was put into all splits. Consequently, all splits had at least one occurrence of each drug. This approach was named "Random Split".

Another approach was to split the data based on similarities between either the drug molecules or the metabolites, with the intention of creating models with different expertise. For splitting based on similarity, Morgan fingerprints and Tanimoto similarity scores between the molecules were calculated with RDKit. These were then used in RDKit's Butina clustering algorithm to divide the molecules into different clusters. A threshold of 0.8 in the clustering was used, implying that molecules with a similarity of 0.2 or higher were clustered together. When dividing the data into splits, the largest clusters were put into different splits and the remaining ones were filled up to make the splits even in size. These approaches, aimed at splitting the

training data for ensemble models into parent- and child-based clusters, were named "Parent Split" and "Child Split", respectively.

Thereafter, the best-performing pre-trained model was fine-tuned on each data split separately, and each model was set to predict 20 metabolites per drug each for the test dataset. As mentioned previously, this number decreases as metabolite predictions that are identical to its parent are removed, as well as duplicates. Thereafter, the same number of metabolites from each model were combined and duplicates were removed. The number of metabolites per model was determined based on the aim of obtaining an average number of metabolites as close to 10 as possible to make its performance validations comparable to top-10 scoring of other models, similarly to what was done by Litsa *et al.* [5].

3.4.6 Optimise Prediction Space

For the last performance optimisation technique, additional pre-training was explored in order to guide the best-performing pre-trained model towards the relevant prediction space prior to using the metabolic dataset for fine-tuning. A dataset from Litsa *et al.* [5], who also used a transformer architecture, called MetaTrans, for drug metabolite prediction, was used for this purpose. Litsa *et al.* [5] compiled a dataset consisting of pairs of parent molecules and human metabolites, similar to the metabolic dataset used in this project. The dataset consists of metabolites for both foreign substances, such as drugs, and compounds originating from within the human body [5]. From here on, this dataset will be called the MetaTrans dataset. The test dataset and validation dataset were combined as one for usage in this project, and they were not treated any differently.

Standardisation and filtering of the MetaTrans dataset was conducted in the same manner as for the previously used data, as described in section Section 3.2. Any reactions with drugs already existing in the metabolic dataset were also removed. The complete cleaning and filtering process decreased the dataset from 11 666 data points to 4243 data points. A majority were removed due to duplicates, molecular weight, and overlap with the metabolic dataset. The dataset was split into training and validation sets in the same ratio as the previous pre-training data, that is 99.5% and 0.5% respectively.

The best-performing pre-trained model was further pre-trained on the filtered MetaTrans dataset, with the same settings as used for pre-training on the VA-dataset. However, the best epoch was chosen differently due to the much smaller size of the dataset. The model ran for 100 epochs, and the epoch with the lowest validation loss was chosen. This new pre-trained will be referred to as the ChemVA-Met model.

Thereafter, the ChemVA-Met model was fine-tuned using the already established best-performing fine-tuning methods. The models were then evaluated and compared to previous fine-tuning models.

4

Results

This section provides the results of the experiments outlined in the previous section. Initially, outcomes from the initial fine-tuning are presented. Subsequent experiments aim to improve these first results. Finally, the best-performing model is compared with other existing models for metabolite prediction.

4.1 Initial Fine-Tuned Model

As an initial experiment, the pre-trained Chemformer model was fine-tuned on the metabolic dataset without any adjustments. Precision and recall scores, which are defined in eq. (3.1), of this initial fine-tuned model were estimated based on the predictions on the metabolic test set. The results are shown in Figure 4.1. Further experiments were aimed at enhancing this performance.

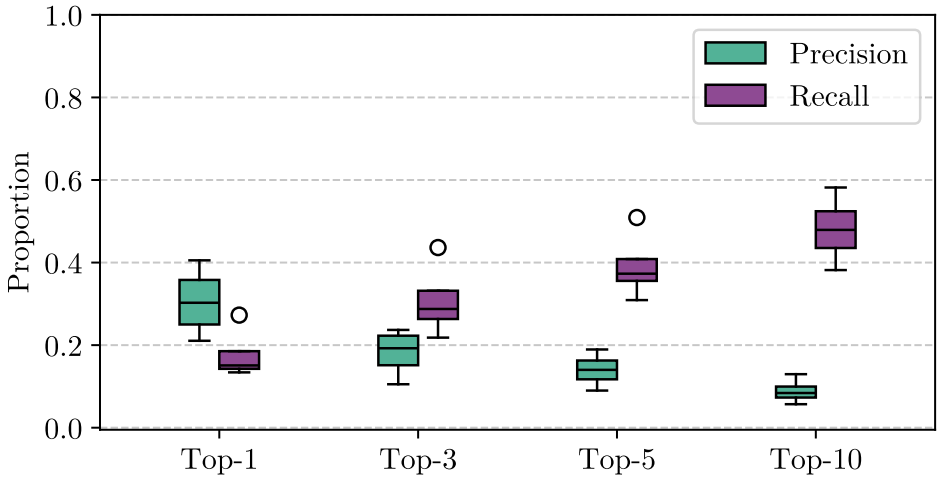


Figure 4.1: The precision and recall on predictions of the metabolic test set obtained using Chemformer Fine-Tuned, the initial fine-tuned model. The box represents the middle 50% of the data points, with lines extending to the data's full range. Outliers are shown beyond. The line inside the box indicates the median.

4.2 Pre-Trained Models

The first approach to optimise performance was to evaluate different pre-trained models, summarised in Table 3.3 in Section 3.4.1. As seen in Figure 4.2, the ChemVA model outperforms both the Chemformer model and VA model in terms of validity. This suggests further investigation into the various ChemVA variants to determine the best-performing model.

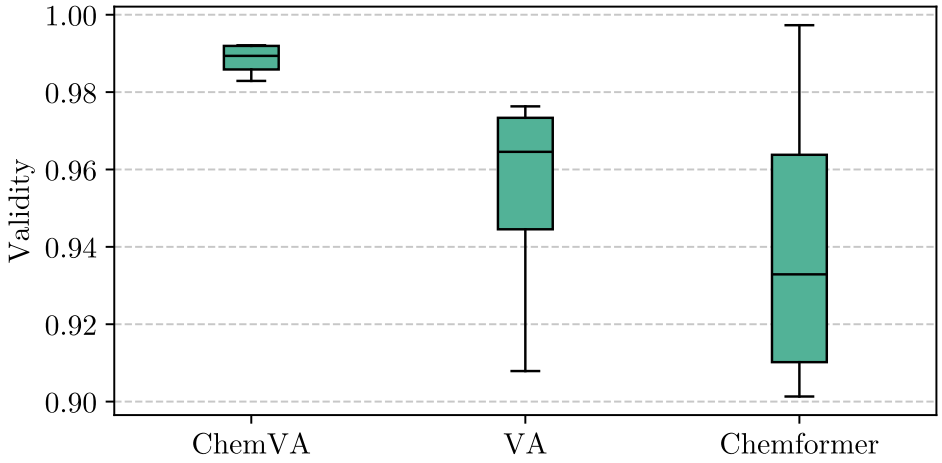


Figure 4.2: SMILES validity of the predictions of the pre-trained models. All models in the figure have a randomisation probability of 0.5 and a masking probability of 0.1. The box represents the middle 50% of the data points, with lines extending to the data's full range. Outliers are shown beyond. The line inside the box indicates the median.

Figure 4.3 shows the comparison of the variants of ChemVA. Here, the highest validity is given by ChemVA Rand, but with a greater variance than ChemVA Comb. The median is still slightly higher for ChemVA Comb. This suggests that the best performing model is ChemVA Comb, indicating that this model should be used for further fine-tuning and will henceforth be referred to simply as "ChemVA". The same investigation was conducted for the VA variants and the results can be seen in Figure A.1 in Appendix A.

To validate choosing ChemVA, the recall and precision of this model, after being fine-tuned on the metabolic dataset, were evaluated. These results, together with a comparison to the initial fine-tuned model (Chemformer Fine-Tuned), are provided in Figure 4.4.

The improvement in recall and precision when using ChemVA compared to the initial model is minimal. However, given that there is no decline and that ChemVA shows the best SMILES string validity, this pre-trained model was selected for further optimisation.

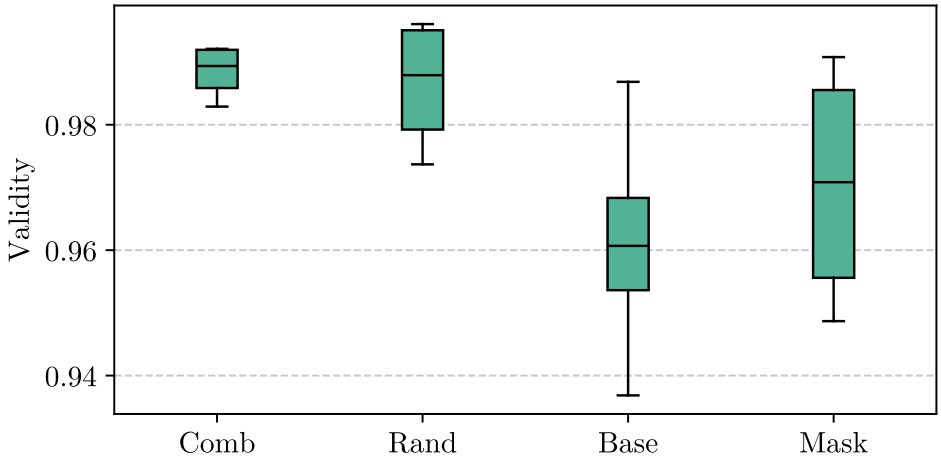


Figure 4.3: SMILES string validity of the predictions of the variants of ChemVA. The box represents the middle 50% of the data points, with lines extending to the data's full range. Outliers are shown beyond. The line inside the box indicates the median.

4.3 Randomised Data Model

After choosing the best-performing pre-trained model, the investigation continued to optimising the fine-tuning of the metabolic dataset. As a first optimisation technique, data randomisation during fine-tuning was introduced. The ChemVA fine-tuned model with randomisation was compared to the model without randomisation, and the precision and recall of the predictions of the metabolic dataset are shown in Figure 4.5. A slight outperformance in the majority of scores is seen for the model with randomisation, and it is therefore chosen for further experiments and comparisons. This model will henceforth be referred to as "ChemVA Fine-Tuned Rand".

4.4 Augmented Data Models

Thereafter, different experiments with data augmentation were explored. The augmented models also included randomisation as this gave the best results in the previous fine-tuning. The results of the three augmented models: parent-grandchild (PG), parent-parent (PP) and both (PG-PP), are shown in Figure 4.6. The results of ChemVA Fine-Tuned Rand without addition of data (None) is included for comparison.

None of the augmentation methods show a significant improvement. The PG-PP model shows a slightly worse performance than the other models in almost all cases for both precision and recall. There is a slight outperformance of the PP model seen in Figure 4.6b, especially in the top-5 and top-10. The model with no addition of data, ChemVA Fine-Tuned Rand, was chosen for further comparison as the observed improvement of the PP model was minimal. Adding parent-parent reactions could

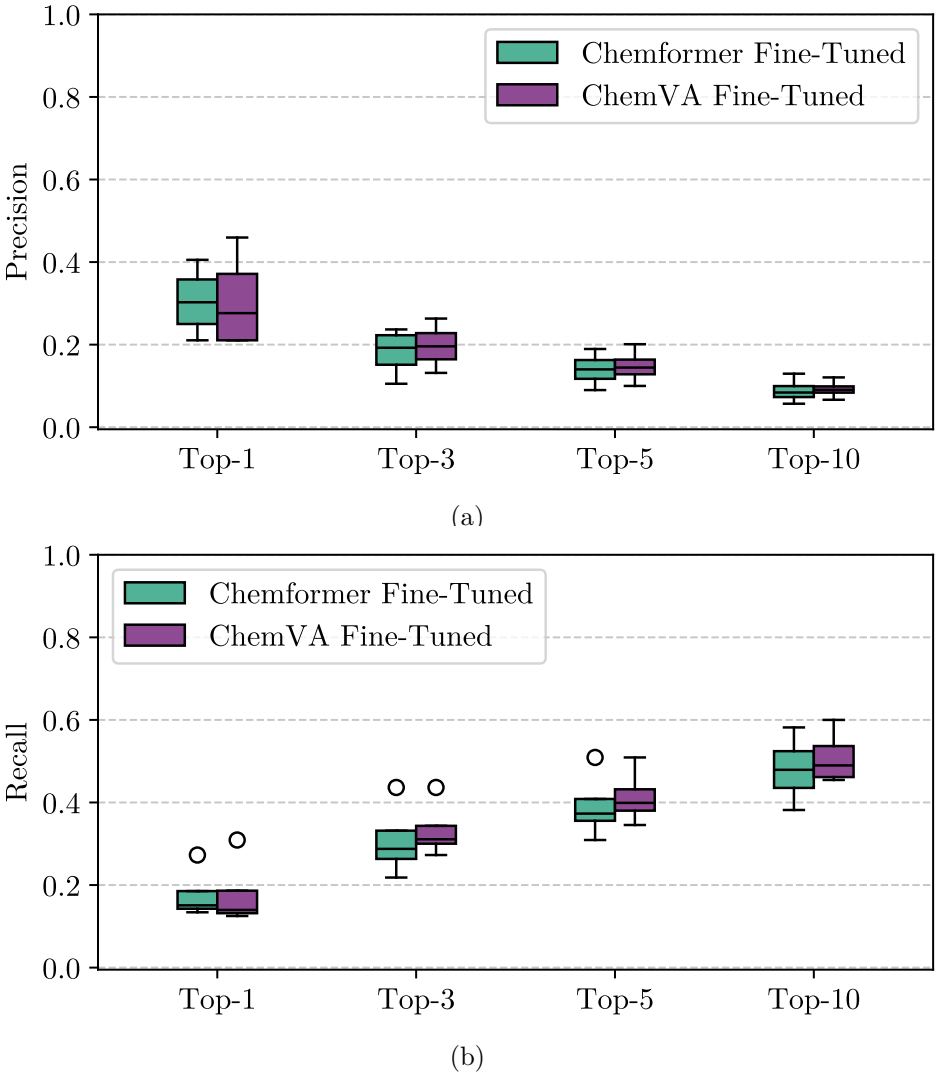


Figure 4.4: The precision (a) and recall (b) scores of predictions on the metabolic test set, obtained by Chemformer and ChemVA fine-tuned models. The box encapsulates the middle 50% of the data points. The extended lines illustrate the full reach of the data, with potential outliers shown individually beyond this range. The line within the box marks the median.

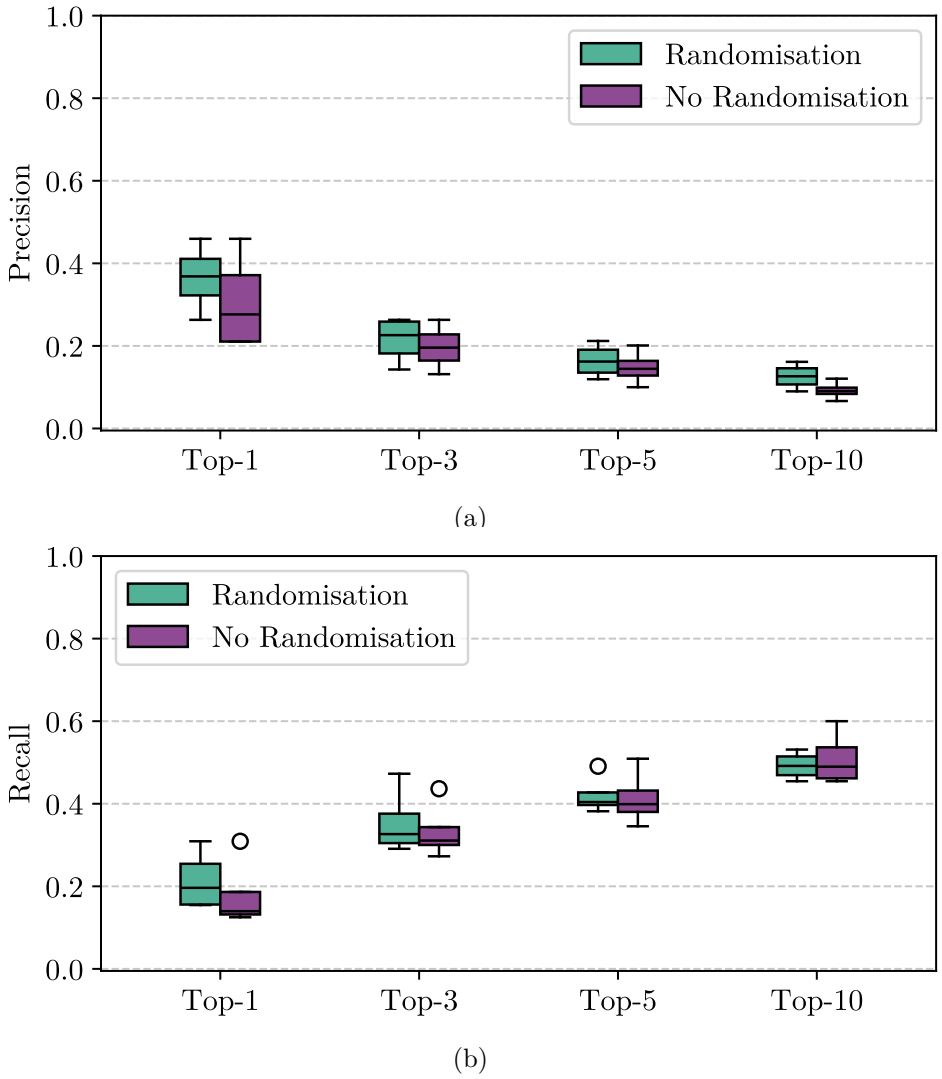


Figure 4.5: The precision (a) and recall (b) scores of predictions on the metabolic test set, obtained by fine-tuning either with or without randomisation. The box represents the middle 50% of the data points, with lines extending to the data's full range. Outliers are shown beyond. The line inside the box indicates the median.

also potentially reduce the chemical space, thereby limiting the model's applicability to new data.

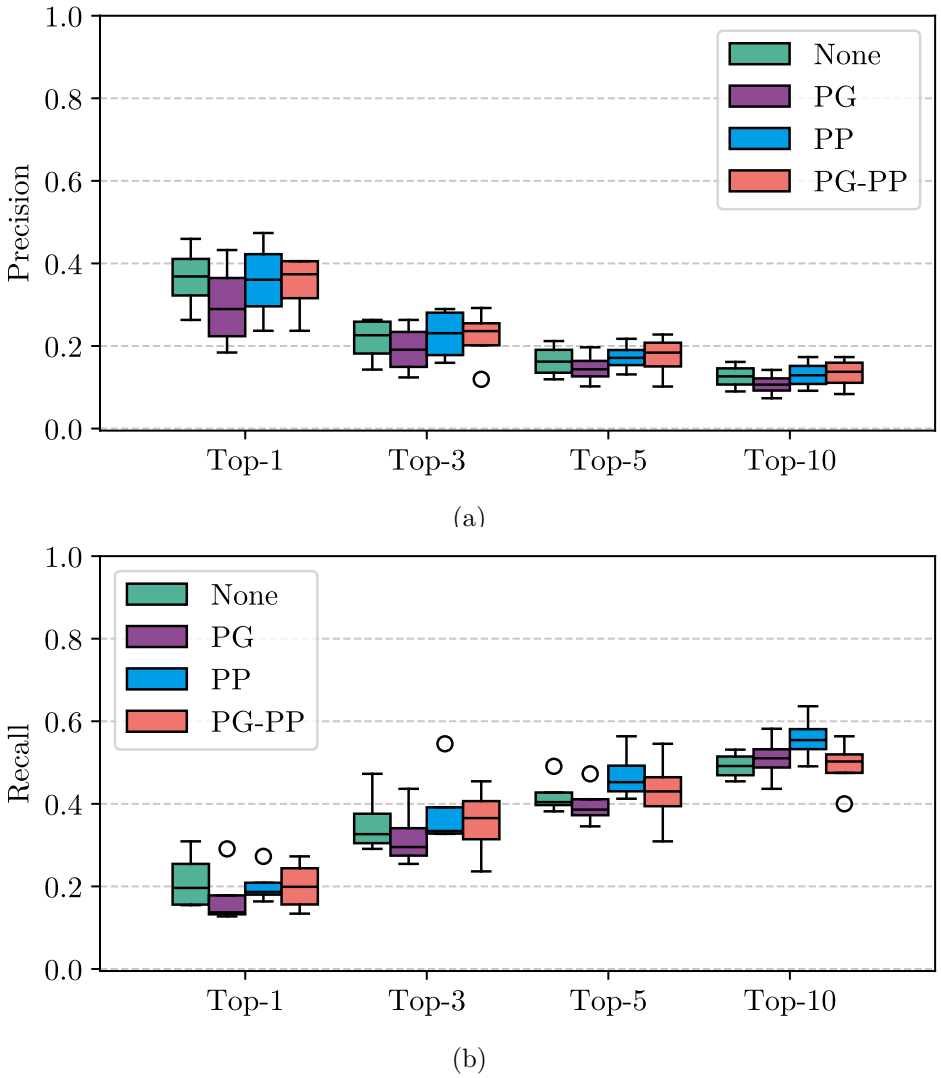


Figure 4.6: The precision (a) and recall (b) scores of predictions on the metabolic dataset, comparing results between the augmented models: parent-grandchild (PG), parent-parent (PP) and both (PG-PP), which are all randomised, and ChemVA Fined-Tuned Rand. The box represents the middle 50% of the data points, with lines extending to the data's full range. Outliers are shown beyond. The line inside the box indicates the median.

4.5 Annotated Data Models

The next experiment involved annotating the data with logP and Csp3 labels (see Section 2.1). Figure 4.7 show the result of the annotations, as well as ChemVA Fine-Tuned Rand (labelled as "None") for comparison. Note that the annotated models are not trained with randomisation as this generated invalid SMILES strings. Despite being randomised, ChemVA Fine-Tuned Rand is used for comparison since

this, so far, is the best-performing model. It is evident that there are no significant differences in the performance of the models. In some cases, especially in Figure 4.7a, ChemVA Fine-Tuned Rand (None) has a higher median than all other models. Consequently, further experiments with data annotations are unjustified.

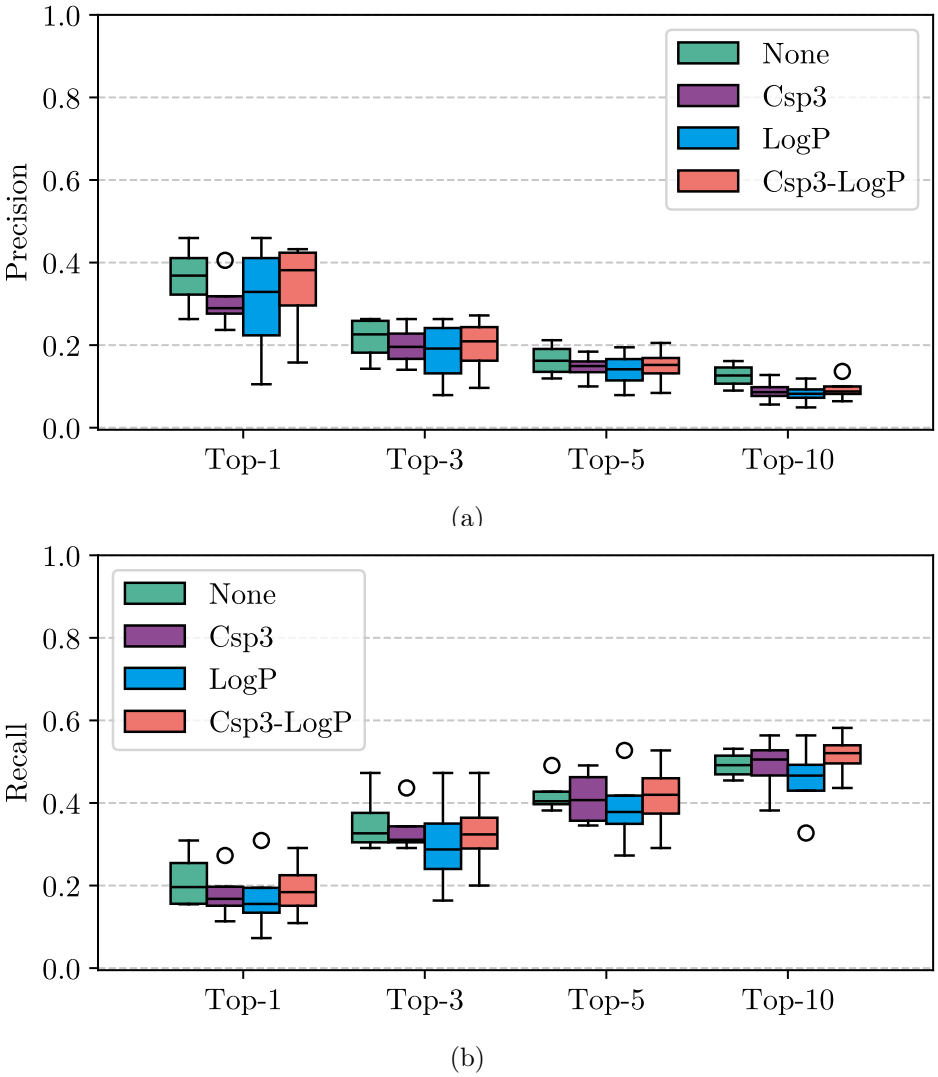


Figure 4.7: The precision (a) and recall (b) scores of predictions on the metabolic dataset, comparing results between the different annotated models (fine-tuned without randomisation) and ChemVA Fine-Tuned Rand. The box represents the middle 50% of the data points, with lines extending to the data's full range. Outliers are shown beyond. The line inside the box indicates the median.

4.6 Ensemble Models

Following this, ensemble models were investigated. Figure 4.8 presents the results from different ensemble models, showing the performance of the separate splitting approaches as well as a comparison to ChemVA Fine-Tuned Rand (labelled as "Single") with top-10 predictions. "Random Split" is the approach of splitting the data

at random, while the "Child Split" and "Parent Split" approaches are based on similarity clusters of the children or parents, respectively. The result shows that there is no significant difference between the splitting approaches. Additionally, they show slightly better recall than the single model, whereas the precisions are slightly worse.

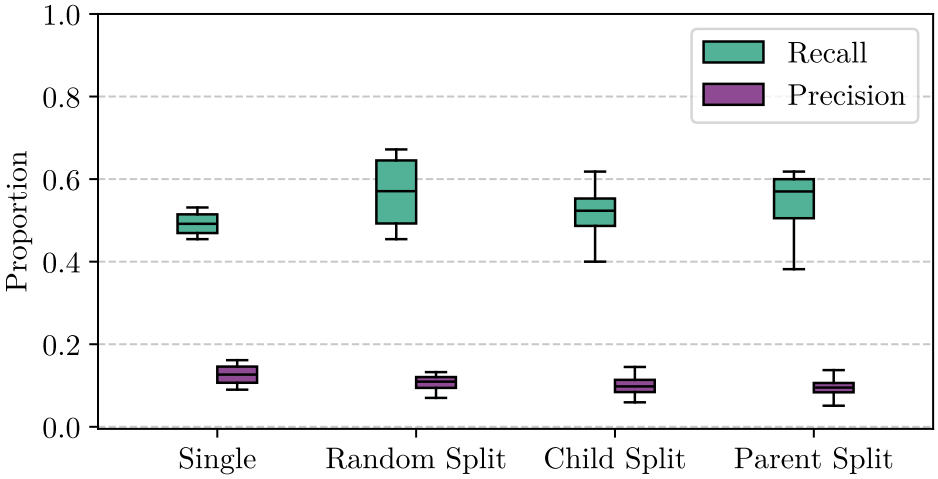


Figure 4.8: Recall and precision scores of predictions on the metabolic dataset. The results are compared between the single model, Chemformer Fine-Tuned Rand, and different data splitting approaches to form ensemble models. The box represents the middle 50% of the data points, with lines extending to the data's full range. Outliers are shown beyond. The line inside the box indicates the median.

Table 4.1 shows the specifications of the predictions for each model, as well as their number of identified true metabolites. The ensemble model using the random splitting approach identified the highest number of metabolites in the metabolic test dataset. This is therefore deemed the best splitting approach for an ensemble model.

Table 4.1: Quantified predictions for the ensemble models and the single model ChemVA Fine-Tuned Rand. Note that the predictions per drug per split is chosen to achieve a value as close to 10 as possible. Top-10 predictions are shown for the single model.

	Single	Random Split	Child Split	Parent Split
Model Type	Single	Ensemble	Ensemble	Ensemble
Predictions/Drug/Split	-	5	3	4
Predictions/Drug	7.05	9.52	9.36	10.3
True Met. out of 271	133	152	140	146

In Table 4.1, it can be noted that the single model generated fewer than 10 predictions per drug, despite its scoring being based on the top-10 predictions. To increase the average number of predicted metabolites per drug, a larger beam width was attempted. However, the increase in number of metabolites was negligible. Therefore,

the original beam width was kept. Since these are two methodologically different approaches, both the best ensemble model and the single model were chosen to proceeded with.

4.7 Models with Optimised Prediction Space

The last experiments performed included additional pre-training of the ChemVA model with the MetaTrans dataset [5] in order to optimise the prediction space prior to fine-tuning. This new pre-trained model is referred to as the "ChemVA-Met" model. Note that the pre-training on the MetaTrans dataset also utilised randomisation and masking.

ChemVA-Met was fine-tuned in the same way as the two best-performing ChemVA fine-tuned models; that is the single model ChemVA Fine-Tuned Rand and the ChemVA ensemble model with a random split. The new single ChemVA-Met fine-tuned model is named "ChemVA-Met Fine-Tuned Rand" and the ensemble model is accordingly named "ChemVA-Met Random Split". Figure 4.9 presents a comparison in top-10 of the ChemVA and ChemVA-Met single models, and the ChemVA and ChemVA-Met ensemble models with the randomised splitting approach. As before, all models incorporate randomisation. The figure shows that ChemVA-Met Random Split performs similarly to ChemVA Random Split but with a smaller variance, especially for recall. For the single models, the recall for the ChemVA-Met model is slightly higher than for the ChemVA model, whereas the precisions are similar. Since the single ChemVA-Met model exhibits higher precision and comparable recall to both ensemble models, it suggests that this model (ChemVA-Met Fine-Tuned Rand) is the best-performing model. Table 4.2 verifies this by showing that ChemVA-Met Rand identified the highest number of true metabolites. Therefore, ChemVA-Met Fine-Tuned Rand is chosen as the model to continue with.

Table 4.2: Quantified predictions of the single models ChemVA Fine-Tuned Rand and ChemVA-Met Fine-Tuned Rand, as well as the ensemble models with randomised splitting. Note that the predictions per drug per split is chosen to achieve a value as close to 10 as possible.

	ChemVA Single	ChemVA-Met Single	ChemVA Random Split	ChemVA-Met Random Split
Model Type	Single	Single	Ensemble	Ensemble
Pred./Drug/Split	-	-	5	5
Pred./Drug	7.05	8.26	9.52	9.6
True Met. of 271	133	153	152	149

4.8 Best-Performing Model

Based on the results, the best-performing model is the ChemVA-Met Fine-Tuned Rand model. It was pre-trained with both the Chemformer model and the vir-

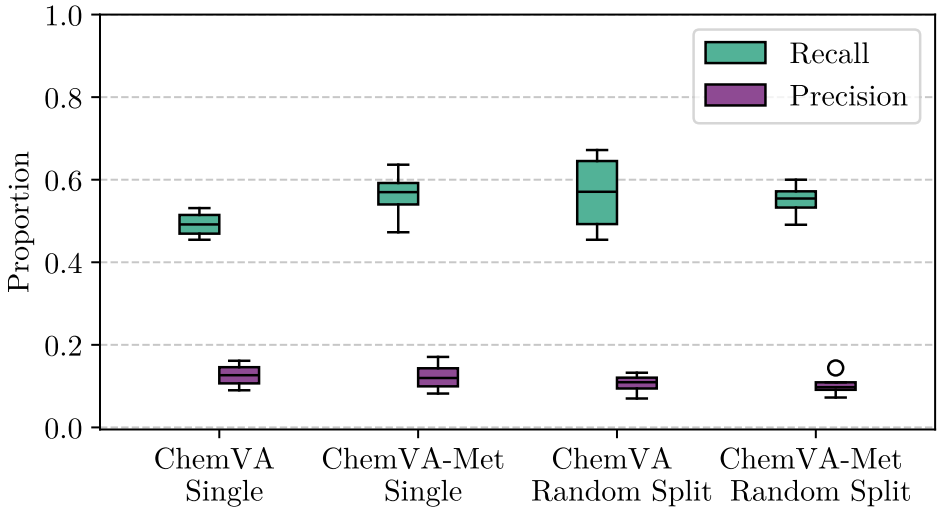


Figure 4.9: Recall and precision scores of predictions on the metabolic test set. The predictions were obtained using the single models ChemVA Fine-Tuned Rand and ChemVA-Met Fine-Tuned Rand, and the ensemble models ChemVA Random Split and ChemVA-Met Random Split. The box represents the middle 50% of the data points, with lines extending to the data's full range. Outliers are shown beyond. The line inside the box indicates the median.

tual analogues dataset, and then additionally pre-trained on the MetaTrans dataset for optimised prediction space. All pre-trainings used augmentation and masking. Thereafter, it was fine-tuned on the metabolic dataset with randomisation. Figure 4.10 presents the result in top-10 of the best-performing model, ChemVA-Met Fine-Tuned Rand, and the initial fine-tuned model (Chemformer Fine-Tuned) on different scores. "At Least One Met." refers to the proportion of drugs that have obtained at least one correctly predicted metabolite, whereas "All Met." refers to the proportion of drugs that have had all of its metabolites correctly predicted. The choice of showing only the top-10 evaluation was motivated by the occurrence of drugs with large numbers of metabolites, for which "Recall" and "All Met." are poor measures of their actual performance for small values of n. It is evident that the optimisation increased the performance of the model significantly, receiving a higher performance in all scores.

Up to this point, the scoring has been based on the success of predicting the exact true metabolites. To gain insight in how close the model is to predict the true metabolites, even when it does not succeed, the constraint can be relaxed by using fingerprint similarity instead. This has been used by both Litsa *et al.* [5], with a fingerprint of 1, and by Multari *et al.* [59], with fingerprint of 0.8. Figure 4.11 shows the result in top-10 of scoring ChemVA-Met Fine-Tuned Rand on the metabolic test dataset with the different types of constraints. "Exact" means that the prediction was fully correct, "Similarity = 1" means that the Tanimoto similarity score (based on Morgan fingerprints) between the true metabolite and the prediction was equal to 1, and "Similarity ≥ 0.8" refers to a fingerprint similarity greater or equal to 0.8. As seen in the figure, both precision and recall increases with a relaxed constraint.

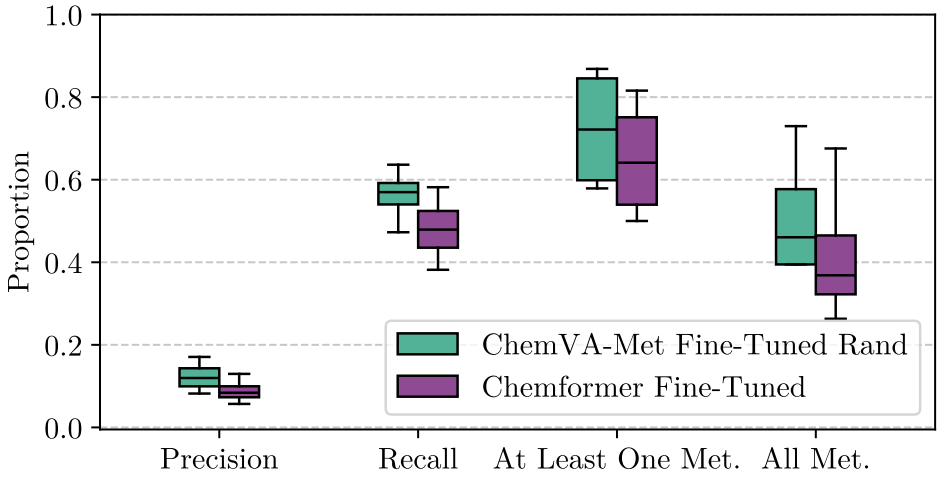


Figure 4.10: Results comparing the best-performing model, ChemVA-Met Fine-Tuned Rand, with the initial fine-tuned model, Chemformer Fine-Tuned. The results are based on the top-10 predictions for each drug in the metabolic test set. The box represents the middle 50% of the data points, with lines extending to the data's full range. Outliers are shown beyond. The line inside the box indicates the median.

This indicates that the model is actually close to predicting the true metabolites, but misplaces some tokens.

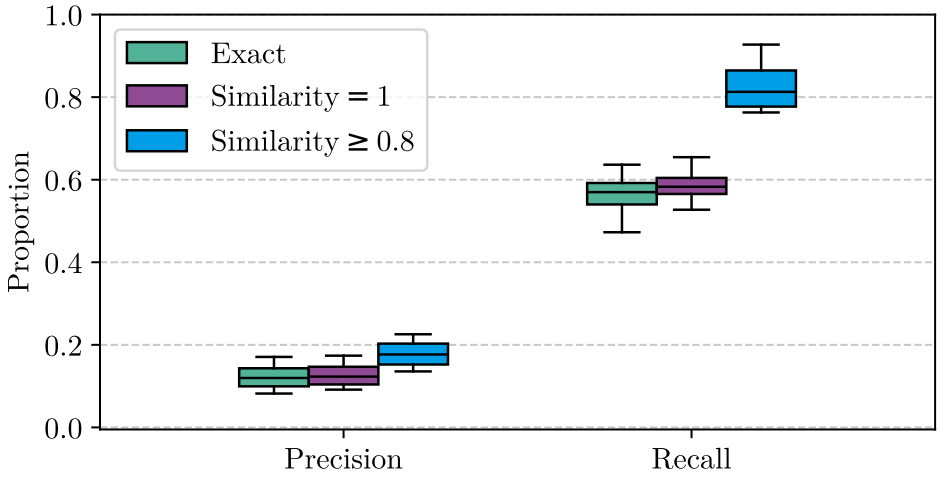


Figure 4.11: Recall and precision on top-10 scoring of the ChemVA-Met Rand model on the metabolic test set using different constrains. Similarity refers to Tanimoto similarity score based on Morgan fingerprints, comparing predictions with true metabolites. The box represents the middle 50% of the data points, with lines extending to the data's full range. Outliers are shown beyond. The line inside the box indicates the median.

4.9 Benchmark

A comparison of the best-performing model in this project to previous work is shown in Table 4.3. This was done on the GLORYx test dataset [7]. The table shows the result of the top-10 predictions in the best-performing model and the initially fine-tuned model. By comparing the result, it can be seen that both Chemformer Fine-Tuned and ChemVA-Met Fine-Tuned Rand have a significantly lower number of total predictions in comparison to both SyGMa and GLORYx. Additionally, ChemVA-Met Fine-Tuned Rand shows the highest precision score as well as F_1 score out of the models. The table shows the result of ChemVA-Met Fine-Tuned Rand Exact, meaning scoring with the predicted SMILES strings that are fully correct, ChemVA-Met Fine-Tuned Rand 1 and ChemVA-Met Rand 0.8, meaning with scoring based on fingerprint similarity of one and greater or equal to 0.8, respectively. It can be seen that changing the scoring to a similarity of one does not affect the performance, whereas relaxing it to 0.8 improves the performance significantly, achieving a higher recall than SyGMa.

Table 4.3: Benchmarks of the predictions in top-10 of the best-performing model ChemVA-Met Fine-Tuned Rand and the initial model Chemformer Fine-Tuned, on the GLORYx test set, against benchmarks of GLORYx [7] and SyGMa [9]. The number of true metabolites is out of 136. The specification next to the model name represents that the predictions are scored based on either exact SMILES strings or the corresponding Tanimoto fingerprint similarity score.

Model	Recall	Precision	F_1	True met.	Total pred.
Chemformer Fine-Tuned	0.37	0.14	0.20	50	358
ChemVA-Met Fine-Tuned Rand Exact	0.43	0.18	0.25	58	328
ChemVA-Met Fine-Tuned Rand 1	0.43	0.18	0.25	58	328
ChemVA-Met Fine-Tuned Rand 0.8	0.69	0.29	0.41	94	328
SyGMa	0.68*	0.12*	0.20**	93*	800*
GLORYx	0.77*	0.061*	0.11**	105*	1724*

**values obtained from Bruyn Kops* et al. *[7].*
***scores calculated from obtained values according to eq.* (3.1) *and eq.* (3.2).

5

Discussion

The initial fine-tuning results of the Chemformer model, presented in Figure 4.4, served as the project's starting point. As expected, an increase in the top-n value led to a decrease in precision and an increase in recall, due to the way these two metrics are defined. From this, further optimisation techniques were explored.

The first technique was to explore the potential of pre-training. The performance of the different pre-trained models was examined based on their ability to generate chemically valid SMILES strings, with the main goal of enabling the model to comprehend the SMILES language prior to task-specific fine-tuning. Among the tested models, the best-performing model by this metric was the ChemVA model, which was the Chemformer model further pre-trained on the virtual analogues [55] dataset. This outcome was expected, as further task-specific training typically enhances performance by increasing familiarity with the language's structures and rules. The VA model, which was pre-trained solely on the virtual analogues dataset, showed worse performance than the ChemVA model but similar performance as the Chemformer model. This was unsurprising given that both VA and Chemformer only consist of one pre-training each, and thus have been less exposed to the SMILES language in comparison to ChemVA with its two pre-trainings.

Upon examining various ChemVA variants, the expected superiority of the Comb variant (trained with both masking and randomisation) was confirmed, as seen by the validity in Figure 4.3. Introducing randomisation of SMILES strings showed an increase in SMILES validity by eliminating the need to learn SMILES canonicalisation. This improvement was evident in the Rand variant and motivates the increased performance when incorporating randomisation during fine-tuning. Although masking is typically used to enhance generalisation, the results indicated that the Mask variant underperformed compared to the Base variant in most cases. Despite this, the Comb variant, which integrates both masking and randomisation, demonstrated superior performance compared to using either approaches alone.

After this, augmentation and annotation of the data were attempted. Neither of these techniques showed any particular increase in performance, as seen in Figure 4.6 for the augmented data, and Figure 4.7 for the annotated data.

The only augmented model that showed a slight enhancement involved the addition of parent-parent reactions (PP). This may be due to the large increase in training data, which can benefit large language models, or because the model improved

its ability to predict metabolites closely resembling the parent structure. In contrast, the addition of parent-grandchild (PG) reactions resulted in slightly worse performance compared to the baseline. This could be due to the greater structural dissimilarity between parent and grandchild molecules compared to parent-child reactions, increasing the complexity of accurately predicting the metabolites. These results might suggest that the majority of the metabolic dataset only introduces minor modifications since guiding the model to make fewer or smaller alterations to the input SMILES strings enhances overall performance, while guiding it towards greater changes worsens performance. For future research it would be worthwhile to investigate the diversity of the transformations further, as this could affect the choice of optimisation methods.

As mentioned, annotating the data did not yield any significant increase in performance. This might be because the model struggled with interpreting annotations, given its lack of pre-training on SMILES strings with annotations. Additionally, the selection of the molecular properties logP and Csp3 may not align with the models ability to detect patterns. While these properties are considered important in drug development, they may not be as useful to the model. Exploring alternative molecular properties could be beneficial in future research.

Ensemble models were one of the more successful optimisation techniques, with randomised splitting identified as the best strategy. The intention behind splitting based on metabolite (child) or drug (parent) similarity was to create distinct expert models, with the anticipation that at least one model would predict true metabolites for each drug. Although these attempts did not improve performance, they generated a larger variety between the different splits, seen in Table 4.1. Achieving approximately the same number of total unique drugs per molecule with fewer molecules required from each split, implies fewer duplicates and increased variety among the splits. Splitting based on metabolite similarity showed the greatest variety. Nevertheless, the performance difference between the child and parent split was too miniscule to draw any further conclusions.

Since increased variety was achieved by splitting based on similarity, one of the intended goals was partly achieved. However, this did not lead to improved performance. Potentially, splitting data based on metabolite or drug similarity would show enhanced performance for a dataset that could be divided into more clearly defined and balanced categories. For the metabolic dataset, there was a significant amount of reactions that were put into their own individual clusters using the specified fingerprint cut-off. However, altering the cut-off either made some clusters too general and large for four even splits, or resulted in too many small groups that would not justify such a splitting strategy. Onward, it would be interesting to explore other ways to categorise the data for the potential of training expert models that also would increase performance.

In the final attempt to improve the performance, the focus was on optimising the model's prediction space before fine-tuning. As seen in Figure 4.9, an additional pre-training with the MetaTrans dataset showed an increase in recall for the single model (ChemVA-Met Single) in comparison to the ChemVA Single model. This was

anticipated, as previous additional pre-training with the virtual analogues dataset also showed improved results. However, there was no improvement for the ChemVA-Met ensemble model compared to the ChemVA ensemble model. The main difference between the ChemVA Random Split and ChemVA-Met Random Split is the smaller recall variance for the ChemVA-Met model. This smaller variance makes the results slightly more reliable and could be explained by the optimised search space of the ChemVA-Met model. Nevertheless, the lack of improvement could also be explained by this, meaning that the model has already limited the search space by overfitting the data. This is likely due to the limited amount of data used for training.

This leads to the best-performing model, which was found to be the single model ChemVA-Met Fine-Tuned Rand. Compared to the first fine-tuned model Chemformer Fine-Tuned, it is evident that ChemVA-Met Fine-Tuned Rand excels on both the Metabolic test set and on the external test set, as seen in Figure 4.10 and Table 4.3. These results confirm the effectiveness of the optimisation methods employed.

Analysing the best-performing model's predictions using similarity scores instead of exact SMILES produced expected results. As seen in Figure 4.11, there is a significant increase when using similarity of 0.8, whereas a similarity of one shows a slight increase. This outcome is reasonable since a similarity of 0.8 might still indicate significant differences between the true and predicted metabolites. In contrast, a similarity of one suggests minor modifications, such as the displacement of a hydroxyl group (-OH) or the addition of a carbon, which do not significantly alter the overall structure. Employing these less strict criteria shows that the model, while not always perfectly accurate, frequently predicts molecules that closely resemble the true metabolites. This insight underscores the model's promising potential. Although employing a fingerprint similarity cut-off may benefit model performance, it is important to recognise that this may yield incorrect predictions. Such errors could result in the misidentification of potential sites of metabolism, incorrect classification of metabolic transformations, or unintended alterations to the original drug structures that are not related to drug metabolism. Even minor molecular differences can lead to vastly different effects on an organism. Therefore, the ultimate aim is still to obtain predictions that match exact SMILES representations.

At first glance, when comparing the benchmarks seen in Table 4.3, it can be assumed that GLORYx is the most suitable model for metabolite prediction due to its high recall. However, upon closer inspection, GLORYx makes twice the number of predictions compared to SyGMa and five times the number of ChemVA-Met Fine-Tuned Rand, which could explain its high recall value and correspondingly low precision. An increase in predictions that leads to more false positives poses a significant issue within a drug discovery pipeline. False positives involves molecules that require further investigation yet fail to yield positive outcomes, thereby consuming substantial time and resources. It is crucial to minimise these occurrences as much as possible. Therefore, the more informative metric to consider is the F_1-score, as it accounts for both recall and precision, aiming to maximize both. ChemVA-Met Fine-Tuned Rand achieved the highest F_1-score, indicating that this model successfully balanced both metrics.

6

Conclusion

The aim of this project was to optimise the transformer-based Chemformer model for predicting drug metabolites. Several techniques for performance optimisation were explored, including additional pre-training, data randomisation, data augmentation, adding annotations with chemical information and using an ensemble model approach. Of these, SMILES string randomisation and additional pre-trainings, with both the virtual analogues and MetaTrans dataset, were used in the best-performing model. While the ensemble model approach also showed potential, it was not a part of the best-performing model. The results indicate that further training and more qualitative data should enhance model performance.

The best-performing model proved comparable to existing metabolite prediction models. In the benchmarking, our model revealed the highest precision and F_1 score, outperforming the existing models. Additionally, using a fingerprint similarity threshold of 0.8, representing a less strict criterion to assess the accuracy of predicted metabolites, significantly improved the results. This shows that our model often predicts molecules that closely approximate true metabolites with only minor deviations from the true structure. It can be asserted that this method for determining success rate lacks fairness, as even minor variations in molecular structure can have significantly different effects on the body. Consequently, predicting molecules that are almost correct may be insufficient. Nevertheless, these results suggest that the model has great potential.

Using Chemformer for metabolite prediction has proven to be less successful than the task of retrosynthesis, i.e., predicting the building blocks of a compound. Despite these tasks being similar, there are differences that could explain why metabolite prediction has shown to be more difficult. To begin with, there is more reliable data on synthetic routes. Additionally, retrosynthetic training data is based on synthetic pathways devised by humans, which are significantly simpler than the complex biosynthetic processes shaped by over four billion years of evolutionary development in metabolic systems.

Despite the challenges, the results of this project suggest that there is great potential of using transformer-based models for metabolite prediction in the near future. For further research and potential optimisation, it could be valuable to explore ensemble models with differently split data, providing the model with other chemical information than what has been attempted, incorporating other data sources, and combining both new and attempted methods.

Bibliography

[1] M. I. Hutchings, A. W. Truman, and B. Wilkinson, "Antibiotics: Past, present and future," *Current Opinion in Microbiology*, vol. 51, pp. 72–80, 2019, Antimicrobials, ISSN: 1369-5274. DOI: https://doi.org/10.1016/j.mib.2019.10.008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1369527419300190.

[2] A. S. Pina, A. Hussain, and A. C. A. Roque, "An historical overview of drug discovery," *Ligand-Macromolecular Interactions in Drug Discovery: Methods and Protocols*, pp. 3–12, 2010.

[3] L. Di, "The role of drug metabolizing enzymes in clearance," *Expert opinion on drug metabolism & toxicology*, vol. 10, no. 3, pp. 379–393, 2014.

[4] N. Parker, M. Schneegurt, A.-H. T. Tu, P. Lister, and B. M. Forster, *Microbiology*. OpenStax, 2016.

[5] E. E. Litsa, P. Das, and L. E. Kavraki, "Prediction of drug metabolites using neural machine translation," *Chemical science*, vol. 11, no. 47, pp. 12 777–12 788, 2020.

[6] J. Kirchmair, A. H. Göller, D. Lang, *et al.*, "Predicting drug metabolism: Experiment and/or computation?" *Nature reviews Drug discovery*, vol. 14, no. 6, pp. 387–404, 2015.

[7] C. de Bruyn Kops, M. Sícho, A. Mazzolari, and J. Kirchmair, "Gloryx: Prediction of the metabolites resulting from phase 1 and phase 2 biotransformations of xenobiotics," *Chemical research in toxicology*, vol. 34, no. 2, pp. 286–299, 2020.

[8] Y. Djoumbou-Feunang, J. Fiamoncini, A. Gil-de-la-Fuente, R. Greiner, C. Manach, and D. S. Wishart, "Biotransformer: A comprehensive computational tool for small molecule metabolism prediction and metabolite identification," *Journal of cheminformatics*, vol. 11, pp. 1–25, 2019. DOI: 10.5281/zenodo.4056560.

[9] L. Ridder and M. Wagener, "Sygma: Combining expert knowledge and empirical scoring in the prediction of metabolites," *ChemMedChem: Chemistry Enabling Drug Discovery*, vol. 3, no. 5, pp. 821–832, 2008.

[10] J. Langowski and A. Long, "Computer systems for the prediction of xenobiotic metabolism," *Advanced Drug Delivery Reviews*, vol. 54, no. 3, pp. 407–415, 2002, Computational Methods for the Prediction of ADME and Toxicity, ISSN: 0169-409X. DOI: https://doi.org/10.1016/S0169-409X(02)00011-X. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169409X0200011X.

[11] R. Irwin, S. Dimitriadis, J. He, and E. J. Bjerrum, "Chemformer: A pre-trained transformer for computational chemistry," *Machine Learning: Science and Technology*, vol. 3, no. 1, p. 015 022, 2022.

[12] A. Dehlén and P. Aronsson, "Prediction of drug metabolites using a deep learning language model," 2024.

[13] S. Grogan and C. V. Preuss, *Pharmacokinetics*. U.S. National Library of Medicine, 2023, Accessed: 2025-11-03. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK557744/.

[14] S. K. Bhal, "Logp making sense of the value," *Adv. Chem. Dev*, pp. 1–4, 2007.

[15] T. A. Berger, B. K. Berger, and K. Kogelman, "10.18 - chromatographic separations and analysis: Supercritical fluid chromatography for chiral analysis and semi-preparative purification," in *Comprehensive Chirality (Second Edition)*, J. Cossy, Ed., Second Edition, Oxford: Academic Press, 2024, pp. 355–393, ISBN: 978-0-323-90645-6. DOI: https://doi.org/10.1016/B978-0-32-390644-9.00013-5. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780323906449000135.

[16] J. C. Lewis, P. S. Coelho, and F. H. Arnold, "Enzymatic functionalization of carbon–hydrogen bonds," *Chemical Society Reviews*, vol. 40, no. 4, pp. 2003–2021, 2011.

[17] E. F. Barreto, T. R. Larson, and E. J. Koubek, "Drug excretion," in *Reference Module in Biomedical Sciences*, Elsevier, 2021, ISBN: 978-0-12-801238-3. DOI: https://doi.org/10.1016/B978-0-12-820472-6.99999-7. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128204726999997.

[18] D. T. Manallack, "The pka distribution of drugs: Application to drug discovery," *Perspectives in medicinal chemistry*, vol. 1, 2007.

[19] rdkit.org, *Getting started with the rdkit in python*, Accessed on 2025-01-30. [Online]. Available: https://www.rdkit.org/docs/GettingStartedInPython.html#list-of-available-descriptors.

[20] D. Weininger, "Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules," *Journal of chemical information and computer sciences*, vol. 28, no. 1, pp. 31–36, 1988.

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

[22] D. Weininger, A. Weininger, and J. L. Weininger, "Smiles. 2. algorithm for generation of unique smiles notation," *Journal of chemical information and computer sciences*, vol. 29, no. 2, pp. 97–101, 1989.

[23] J. Arús-Pous, S. V. Johansson, O. Prykhodko, *et al.*, "Randomized smiles strings improve the quality of molecular generative models," *Journal of cheminformatics*, vol. 11, pp. 1–13, 2019.

[24] S. Riniker and G. A. Landrum, "Similarity maps - a visualization strategy for molecular fingerprints and machine-learning methods," *Journal of cheminformatics*, vol. 5, pp. 1–7, 2013.

[25] H. L. Morgan, "The generation of a unique machine description for chemical structures - a technique developed at chemical abstracts service," *Journal of chemical documentation*, vol. 5, no. 2, pp. 107–113, 1965.

[26] P. Willett, "Similarity-based virtual screening using 2d fingerprints," *Drug discovery today*, vol. 11, no. 23-24, pp. 1046–1053, 2006.

[27] D. Butina, "Unsupervised data base clustering based on daylight's fingerprint and tanimoto similarity: A fast and automated way to cluster small and large data sets," *Journal of Chemical Information and Computer Sciences*, vol. 39, no. 4, pp. 747–750, 1999.

[28] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[29] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

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

[31] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," in *International conference on machine learning*, PMLR, 2017, pp. 1243–1252.

[32] H. Naveed, A. U. Khan, S. Qiu, *et al.*, "A comprehensive overview of large language models," *arXiv preprint arXiv:2307.06435*, 2023.

[33] H. Zhao, H. Chen, F. Yang, *et al.*, "Explainability for large language models: A survey," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 2, pp. 1–38, 2024.

[34] E. D. Liddy, "Natural language processing," 2001.

[35] S. Chithrananda, G. Grand, and B. Ramsundar, "Chemberta: Large-scale self-supervised pretraining for molecular property prediction," *arXiv preprint arXiv:2010.09885*, 2020.

[36] J. Li and X. Jiang, "Mol-bert: An effective molecular representation with bert for molecular property prediction," *Wireless Communications and Mobile Computing*, vol. 2021, no. 1, p. 7 181 815, 2021.

[37] Y. Wang, H. Zhao, S. Sciabola, and W. Wang, "Cmolgpt: A conditional generative pre-trained transformer for target-specific de novo molecular generation," *Molecules*, vol. 28, no. 11, p. 4430, 2023.

[38] Q. Ai, F. Meng, J. Shi, B. Pelkie, and C. W. Coley, "Extracting structured data from organic synthesis procedures using a fine-tuned large language model," *Digital Discovery*, vol. 3, no. 9, pp. 1822–1831, 2024.

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

[40] A. Mao, M. Mohri, and Y. Zhong, "Cross-entropy loss functions: Theoretical analysis and applications," in *International conference on Machine learning*, PMLR, 2023, pp. 23 803–23 828.

[41] A. Sarah, G. Nencioni, and M. M. I. Khan, "Resource allocation in multi-access edge computing for 5g-and-beyond networks," *Computer Networks*, vol. 227, p. 109 720, 2023, ISSN: 1389-1286. DOI: `https://doi.org/10.1016/j.comnet.2023.109720`. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S1389128623001652`.

[42] C. Meister, T. Vieira, and R. Cotterell, "Best-first beam search," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 795–809, 2020.

[43] 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*, vol. 11, no. 1, p. 5575, 2020.

[44] P. Villalobos, A. Ho, J. Sevilla, T. Besiroglu, L. Heim, and M. Hobbhahn, "Position: Will we run out of data? limits of llm scaling based on human-generated data," in *Forty-first International Conference on Machine Learning*, 2024.

[45] S. Gunasekar, Y. Zhang, J. Aneja, *et al.*, "Textbooks are all you need," *arXiv preprint arXiv:2306.11644*, 2023.

[46] M. C. Ramos, C. J. Collison, and A. D. White, "A review of large language models and autonomous agents in chemistry," *Chemical Science*, 2025.

[47] T. Sterling and J. J. Irwin, "Zinc 15–ligand discovery for everyone," *Journal of chemical information and modeling*, vol. 55, no. 11, pp. 2324–2337, 2015.

[48] W. J. Thompson, "Poisson distributions," *Computing in Science & Engineering*, vol. 3, no. 3, pp. 78–82, 2001.

[49] D. S. Wishart, Y. D. Feunang, A. C. Guo, *et al.*, "Drugbank 5.0: A major update to the drugbank database for 2018," *Nucleic acids research*, vol. 46, no. D1, pp. D1074–D1082, 2018.

[50] S. Preissner, K. Kroll, M. Dunkel, *et al.*, "Supercyp: A comprehensive database on cytochrome p450 enzymes including a tool for analysis of cyp-drug interactions," *Nucleic acids research*, vol. 38, no. suppl_1, pp. D237–D243, 2010.

[51] S. Heller, A. McNaught, S. Stein, D. Tchekhovskoi, and I. Pletnev, "Inchi-the worldwide chemical structure identifier standard," *Journal of cheminformatics*, vol. 5, pp. 1–9, 2013.

[52] C. Knox, M. Wilson, C. M. Klinger, *et al.*, "Drugbank 6.0: The drugbank knowledgebase for 2024," *Nucleic acids research*, vol. 52, no. D1, pp. D1265–D1275, 2024.

[53] M. Sícho, C. Stork, A. Mazzolari, *et al.*, "Fame 3: Predicting the sites of metabolism in synthetic compounds and natural products for phase 1 and phase 2 metabolic enzymes," *Journal of chemical information and modeling*, vol. 59, no. 8, pp. 3400–3412, 2019.

[54] J. O. Miners and P. I. Mackenzie, "Drug glucuronidation in humans," *Pharmacology & therapeutics*, vol. 51, no. 3, pp. 347–369, 1991.

[55] D. Dimova and J. Bajorath, "Systematic design of analogs of active compounds covering more than 1000 targets," *Med. Chem. Commun.*, vol. 7, pp. 859–863, 5 2016. DOI: 10.1039/C5MD00585J. [Online]. Available: http://dx.doi.org/10.1039/C5MD00585J.

[56] B. Zdrazil, E. Felix, F. Hunter, *et al.*, "The chembl database in 2023: A drug discovery platform spanning multiple bioactivity data types and time periods," *Nucleic Acids Research*, vol. 52, no. D1, pp. D1180–D1192, Nov. 2023, ISSN: 0305-1048. DOI: 10.1093/nar/gkad1004. eprint: https://academic.oup.com/nar/article-pdf/52/D1/D1180/55040046/gkad1004.pdf. [Online]. Available: https://doi.org/10.1093/nar/gkad1004.

[57] D. Dimova and J. Bajorath, *Systematic design of analogs of active compounds covering more than 1000 targets*, Zenodo, Feb. 2016. DOI: 10.5281/zenodo.45807. [Online]. Available: https://doi.org/10.5281/zenodo.45807.

[58] *Chembl ftp directory*, Dec. 2024. DOI: 10.6019/CHEMBL.database.35.

[59] S. Multari, R. Özçelik, A. Mazzolari, M. S. Nobile, and F. Grisoni, "Predicting metabolic reactions with a molecular transformer for drug design optimization," in *2024 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, IEEE, 2024, pp. 1–8.

A

Appendix 1

Similarly as for the ChemVA model, the different variants of the VA model were evaluated based on SMILES string validity. These results are displayed in Figure A.1, which shows that the Comb model outperforms the other models. This aligns with the conclusion made for the ChemVA variants.

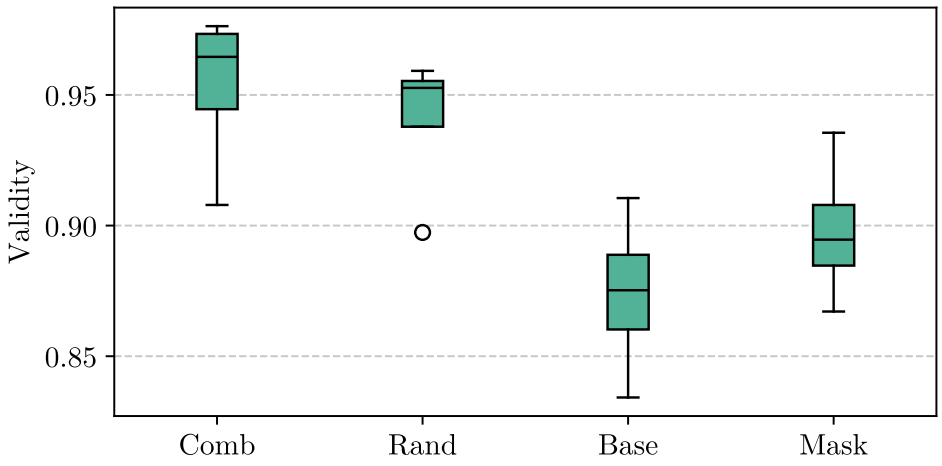


Figure A.1: SMILES validity of the predictions of the variants of VA. The box represents the middle 50% of the data points, with lines extending to the data's full range. Outliers are shown beyond. The line inside the box indicates the median.