Predicting Metabolic Reactions with Molecular Transformers

Anonymous Authors

Abstract—**Metabolism prediction is a crucial step of drug development, as the biotransformations a drug candidate undergoes inside the human body can affect the clinical outcome. Computer-aided drug design has been extensively employed to speed up the process and enhance its efficiency and effectiveness, but among the investigated areas, metabolism has received less attention. This project aimed at leveraging machine learning to analyze large metabolic datasets, make predictions and recognize patterns, in order to fill this knowledge gap and enhance our understanding of metabolism and its impact on drug development. To achieve this goal, we developed a Deep Learning model for metabolism prediction using natural language processing techniques trained on molecular string representations, i.e., Simplified Molecular Input Line Entry Systems (SMILES) strings. To this end, we employ a Molecular Transformer, because of its ability to capture sequential and contextual information within strings (in this case, SMILES) enabling the learning of complex relationships. The transformer was trained using a high quality dataset, MetaQSAR, from which we derived approximately 100 000 instances of metabolic reactions. In this work, we investigate whether the Transformer architecture bears the potential to learn a mapping between the input molecular structures and their corresponding metabolites, in order to expedite drug discovery and improve patient safety.**

Index Terms—**drug discovery, metabolism, transformers, artificial intelligence, neural networks**

I. Introduction

Drug development is a time-consuming and resource-demanding process, consisting of multiple steps and relying on an iterative trial-and-error approach. [1] This intricate journey can be divided into three distinct phases – discovery, preclinical, and clinical development – that together can take over a decade and several million dollars to put just one medicinal product on the market, even with significant financial investment, advanced laboratory facilities and skilled researchers merging various disciplines [1].

The drug discovery pipeline starts with target identification and validation, already carrying along two significant challenges. The first is the vastness of the chemical space, the conceptual territory inhabited by all possible drug-like compounds. It is estimated to be in the order of 10^{60} molecules, rendering its exploration and understanding extremely difficult. [2]. The second challenge is represented by the complexity of the biological systems [3]: numerous molecules dynamically interact with each other, engaging in complex signaling pathways and regulatory networks, that need to be understood in order to develop effective therapies able to modulate specific targets without unintended consequences. The multifactorial nature of diseases makes it even more challenging to comprehend the intricate crosstalk between these molecules and promptly grasp the underlying mechanisms for identifying potential therapeutic goals [4].

Once the target is chosen, molecules that can interact with it in a specific and effective manner (the so-called "hits") have to be selected among a multitude of candidates studied employing high-throughput screenings, an extremely time-consuming technique. Additionally, hit compounds should be optimized to enhance their potency, selectivity, and safety profiles, as most drug candidates fail during preclinical and clinical stages due to inadequate efficacy, unexpected toxicity, or other safety concerns. Complexity and time are additionally increased by the need to comply with complex regulations and guidelines.

Metabolism plays a key role in drug development, because it influences a drug pharmacokinetics (PK), pharmacodynamics (PD), and overall its safety and efficacy [5]. The investigation of metabolism is a complex and challenging area of study [6], due to several reasons including (1) the production of several metabolites via multistep biotransformation reactions, (2) the physicochemical and pharmacological properties of these metabolites, significantly different from those of the parent drug, (3) the genetic and environmental influence determining inter-individual variations [7]. Understanding how these dynamic metabolic processes affect potential drug candidates would be of great benefit to medicinal chemistry: it would help prioritize the most promising compounds based on their metabolic profile, greatly reducing both the time and the resources needed for screening molecules, thus increasing the cost-effectiveness ratio and success rate of the process. [8]

Computational models to simulate and predict metabolic behavior already exist [9] [10], e.g., to predict the interaction of molecules with metabolic enzymes [11], predict sites on the molecule where metabolism occurrs, [12], and generate putative metabolite structures [13] However, building comprehensive models encompassing the entirety of metabolic pathways and their regulation is still a challenge, often fed by limited domain knowledge and poor quality of the available data, which are both responsible for undermining the accuracy and reliability of predictions. [14] Unfortunately, acquiring high-quality data is not always straightforward, as experimental techniques are often limited by technical constraints, giving experts only snapshots of the complex dynamics occurring within the cell. In this context, Artificial Intelligence (AI) is exceptionally well-suited to address the challenges of drug discovery and development. [15]

Identify applicable funding agency here. If none, delete this.

Computer-Aided Drug Design (CADD) consists in the application of computational tools, algorithms, and simulations to aid in various aspects of drug development. Within drug discovery, CADD has been applied to various domains – such as target identification, virtual screening, and lead optimization [16] – while drug metabolism has received less attention. In this work, we built a Machine Learning model that uses end-to-end learning-based method, relying on the application of the molecular transformer architecture [17], used to cast metabolism as a sequence-to-sequence translation, between molecules and their metabolites. This is achieved by representing molecules as strings, via the Simplified Molecule Input Entry Line Systems (SMILES) representation [18], describing 2D chemical information. Our model does not rely on human-engineered rules to learn relevant aspects of drug-to-metabolite mapping. This work is inspired by the usage of language models to predict organic reactions [19]–[21]. Moreover, our work expands upon other usages of Molecular Transformers for metabolism [22], by including high-quality data, and incorporating metabolism reaction classes (RC) in the predictions. Figure 1 provides a schematic representation of the proposed approach.

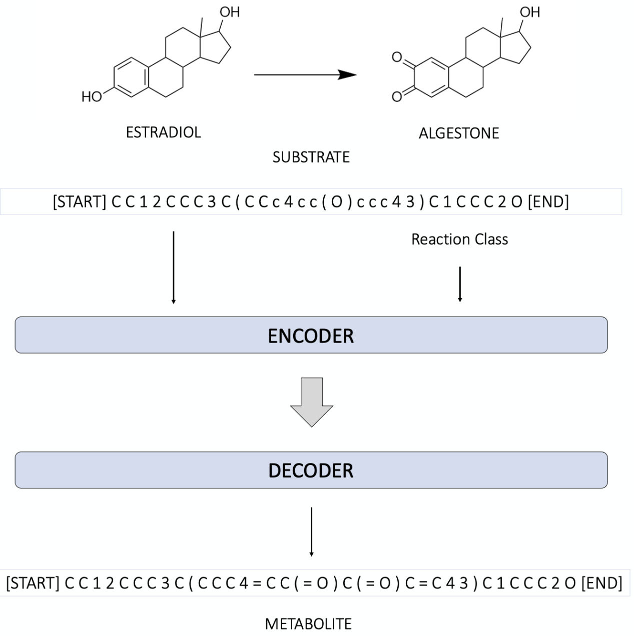
OH
HO
ESTRADIOL
OH
O
O
ALGESTONE
SUBSTRATE
[START] C C 1 2 C C C 3 C (C C c 4 c c (O) c c c 4 3) C 1 C C C 2 O [END]
Reaction Class
ENCODER
DECODER
[START] C C 1 2 C C C 3 C (C C C 4 = C C (= O) C (= O) C = C 4 3) C 1 C C C 2 O [END]
METABOLITE

Fig. 1. Schematic representation of a transformer model for metabolism prediction.

This work aims to leverage the power of machine learning and computational tools to enhance our understanding of how drugs are processed by the body, a knowledge that can help in the design and development of safer and more effective drugs, as well as aid in the identification of potential drug-drug interactions and the optimization of dosing regimens, therefore contributing to the advancement of the drug discovery process, and ultimately benefiting patients in need of new and improved treatments.

After providing a brief description of metabolic processes, we delineate the main methods used in this paper, and discuss the results and opportunities ahead.

II. Metabolism

Metabolism involves a large and diversified set of chemical reactions, mainly mediated by specific enzymes, occurring to endogenous compounds and xenobiotics, serving the dual purpose of providing energy and essential building blocks to cells, as well as removing potentially harmful substances [5].

Xenobiotics are chemical substances not naturally produced or expected to be present within the organism where they are found. They can be introduced either through diet or environmental exposure, and they also include drugs. Since they are recognized as non-self, the body biotransforms them in order to facilitate their removal, and this takes place in almost every tissue of the human body, but the liver is of major significance.

Metabolic reactions occur in two phases: phase I reactions include oxidation, reduction, or hydrolysis reactions, while phase II is dedicated to conjugation reactions mediated by transferases. Inside each of these major classes, metabolic reactions can be classified into several classes and subclasses. The final aim of these biotransformations is to deactivate toxic compounds or increase the polarity to facilitate their elimination, as hydrophilic compounds have higher excretion rates from the body. [23]

All these processes are needed to sustain life, but due to the effects they have on the pharmacokinetics (PK) and pharmacodynamics (PD) of small molecules, they also affect the efficacy and safety of drugs.

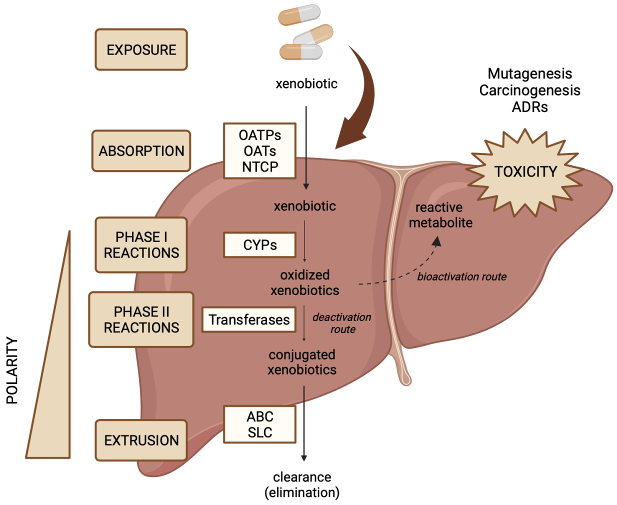

Fig. 2. Scheme of liver metabolism.

The therapeutic effect of a drug is correlated to its bioavailability, which refers to the rate and extent a substance becomes completely available to its intended biological destinations. Metabolic reactions highly influence this property: it can occur that some lead compounds demonstrating strong potency in *in* *vitro* studies exhibit limited effectiveness in *in vivo* models, due to the interference of biotransformation. For example, drug metabolism plays an important role in drug clearance: drugs characterized by a high elimination rate constant have a low half-life, meaning that they are metabolized so fast that systemic concentration does not remain in the therapeutic window long enough to elicit the desired effect. This is due to an intrinsic characteristic of the substance, but also enzymes can be responsible for the alteration of the PK/PD profile of drugs.

The investigation of drug metabolism is an essential aspect of the drug development process: its consequences on the PK/PD profile and safety need to be carefully taken into consideration when assessing whether a compound is suitable as a drug or how it can be modified to have a better metabolic profile.

III. Methods

In this section, we delineate the state-of-the-art techniques and methodologies employed to represent molecules and build a machine learning model able to address the challenges of metabolism prediction. In order to make the paper self-contained, we provide also a detailed explanation of the SMILES encoding.

A. Simplified Molecular Input Line Entry System (SMILES)

The Simplified Molecular Input Line Entry System (SMILES) serves as a linear notation system tailored for chemical information processing. SMILES sequences encapsulate identical information as chemical graphs but adopt a character string format without spaces, occupying 50% to 70% less space compared to an equivalent connection table, without sacrificing interpretability. SMILES employs a unique alphabet with a straightforward vocabulary and few grammar rules. Non-hydrogen atoms are denoted by their atomic symbols within square brackets, except for those in the “organic subset” (B, C, N, O, P, S, F, Cl, Br, I) or with hydrogens inconsistent with the lowest normal valence. Aromatic ring atoms employ lowercase letters, with attached hydrogen or formal charges always specified within brackets.

Single, double, triple, and aromatic bonds are represented by the symbols $-, =, \#$, and : , respectively. Branches, cyclic structures, and disconnected compounds are represented by parentheses, ring-breaking bonds, and periods, respectively. Isomer specification, configuration around double bonds, isotopic specifications, tetrahedral centers, and tautomeric structures are addressed, too. Figure 3 shows an example of SMILES representation and the associated molecule.

To preserve the uniqueness of the representation, algorithms are used to generate “canonical” SMILES, meaning that each string corresponds to one single molecule, and that the information it provides is enough to reconstruct the corresponding 2D structure. Relying on the SMILES representation, molecules can be seen as linguistic constructs, and therefore be treated as sentences to be translated: among the several algorithms and applications in the field of deep learning, for the purpose of this project, we will focus on the sequence-to-sequence (seq2seq) models and, specifically, on Transformer models.

B. Transformers

The transformer model [17] is built upon an encoder-decoder architecture and aims to overcome the limitations of the RNN framework by incorporating several components and mechanisms enabling more efficient and effective processing of sequential data. A Transformer serves as a sequence-to-sequence model commonly employed in Machine Translation. It comprises an Encoder, a Decoder, and an Attention layer typically situated at the Encoder-Decoder interface. The Encoder consists of a series of identical layers, each comprising a multi-head self-attention mechanism and a position-wise fully connected feed-forward network. This network includes two linear transformations separated by a ReLU activation. The decoder layer shares similarities with the encoder but introduces a third sub-layer, which performs multi-head attention on the output of the encoder stack. The output of the decoder is a vector, which is passed through a Softmax function, responsible for converting it into a probability distribution: the predicted token is the most probable choice among the available options. Self-attention dynamically assigns weights to different parts of the input data based on their significance. Specifically, the mechanism here mentioned is:

- Masked: prevents the TransformerDecoder from seeing the whole sequence at once, to avoid the utilization of future information during training, which would render the model unusable at test time.
- Multi-head: allows the model to focus on different aspects or relationships between words in the sequence, as each head learns a different relationship.

We show an example of the architecture used in this work in Figure 4.

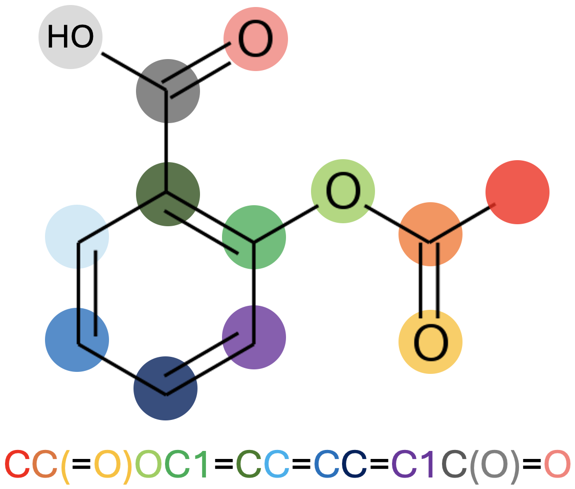

Fig. 3. Example of SMILES representation and corresponding molecule.

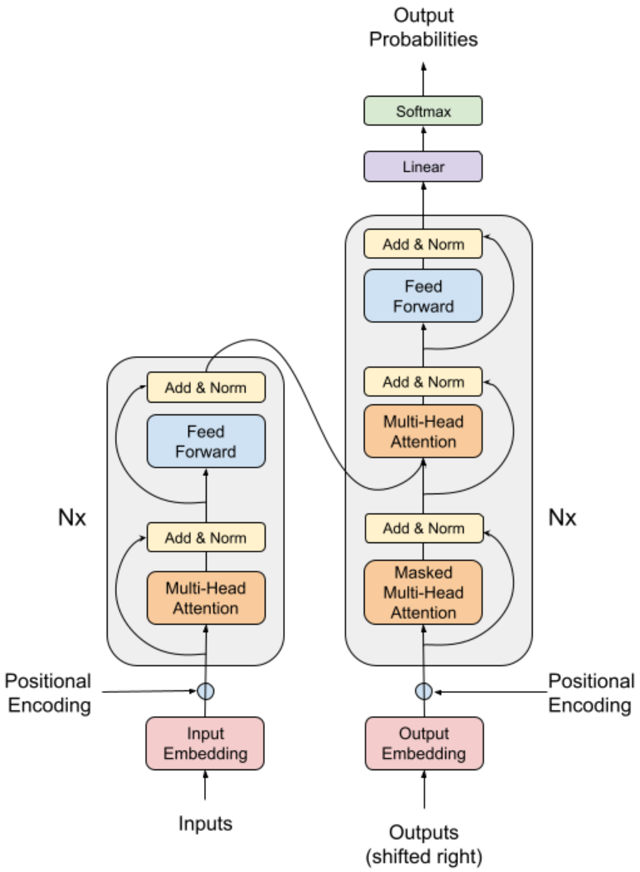

Fig. 4. General architecture of a Transformer model.

C. Tanimoto Similarity

Notably, compounds with similar molecular structures tend to exhibit similar biological and physicochemical characteristics, therefore the similarity between the known metabolites and the ones predicted by our model can be used as a parameter to evaluate the model performance.

Tanimoto similarity [24] is a metric commonly used in molecular structure analysis to quantify the similarity between two sets of data, typically representing chemical compounds: in this case, the reference metabolites and the structures predicted by the model. To calculate the Tanimoto similarity, molecules need to be represented by molecular descriptors, mathematical representations able to capture various aspects of molecular structure. They provide a way to encode the characteristics of molecules in a format that can be analyzed and used for various purposes in computational chemistry, including similarity assessment. The descriptors here used are called Extended Connectivity Fingerprints (ECFPs), circular fingerprints encoding the presence or absence of substructures in a molecule. They typically consist of binary strings, with each bit representing the presence or absence of a specific substructure. It is computed by taking the intersection of the two sets and dividing it by the sum of the sizes of the two sets, as described by the formula:

$$S = \frac{c}{a + b - c}, \tag{1}$$

where a and b represent the bits equal to 1 (”on bits”) in molecules A and B, respectively, while c represents the number of on bits in both molecules. The similarity score S is hence a value between 0 and 1, where 0 indicates no similarity and 1 indicates perfect similarity.

D. The MetaQSAR Dataset

The model was trained using the MetaQSAR database [25], a manually curated resource of published measured data on xenobiotic metabolism, which includes expert curated sites of metabolism (SoMs) and reaction annotations for discovery compounds and drugs. The reactions covered in MetaQSAR are divided into three main reaction classes: redox reactions (3858 reactions); hydrolysis and other nonredox reactions (697 reactions); conjugation reactions (1765 reactions). A notable feature of MetaQSAR lies in the rigorous expert validation applied to each annotated reaction, preventing mistakes and inaccuracies typically associated with automated compilation, as well as avoiding mistakes found in the primary literature. Despite its strengths, it’s worth noting that MetaQSAR is a relatively small dataset. To address this limitation, we employed data augmentation techniques [26]: we leveraged the fact that each molecule can be represented by multiple SMILES strings depending on the atom chosen as a starting point, allowing us to utilize each substrate up to 10 times. For the analysis, these representations will be mapped back to the canonical SMILES, ensuring a 1:1 match with the corresponding 2D structure.

Using the information from MetaQSAR, we created a dataset in which each combination “substrate (augmented 10x) + reaction class” corresponds to the translated metabolite (canonical) deriving from the specified reaction class.

This first dataset was then integrated with information coming from MetaTree [27], another meticulously curated repository of metabolic data. Specifically, it is a collection of complete metabolic trees, meaning that the substrates mentioned are not known to undergo any reaction that’s not annotated in the dataset itself. We employed this insight to make the model aware that there are instances where the substrate remains unchanged, aiming to expand our model’s comprehension of metabolism.

Then, we tokenised substrates and metabolites atom-wise: each atom corresponds to a token, and each token is separated from the adjacent ones by a space. We also added two special tokens, namely [START] and [END], to the metabolites’ strings.

Differently from chemical reaction prediction, where each input gives a single output, in metabolite prediction each parent molecule may yield multiple metabolites. [13] That is, our dataset includes cases that share the same parent molecule but can differ with respect to the resulting metabolite, due to the reaction class involved. We expected that incorporating this information as input of the model could overcome limits in generalization and applicability of the method.

The final dataset was divided in training and test sets, making sure that each combination “molecule + reaction” of the test set was not present in the training one.

E. The Model

To map the substrate SMILES sequences to the SMILES sequences of the corresponding metabolites, we started with the existing implementation of a Transformer [17] commonly employed in Natural Language Processing, making some adjustments to tailor it to the specific task of metabolism prediction. Our sequence-to-sequence model consists of three key components: a TransformerEncoder, a TransformerDecoder, and a PositionalEmbedding layer that ensures that the model is aware of the sequential order of the atoms. We use three instances of the TextVectorization layer to vectorize the SMILES sequences, one for the substrates, one for the metabolites, and one for the reaction classes. The vectorized source sequences are two: the vectorized SMILES for substrate and the vectorized RC. Both are passed through the TransformerEncoder, which will generate two new representations referred to as ”smiles_encoder_inputs” and “rc_encoder_inputs”. The latter passes through a reshape operation and a dense layer to allow its concatenation with the “smiles_encoder_inputs”. Next, the TransformerDecoder receives this concatenation, together with the “decoder_inputs”, which is the vector representing the current state of the target sequence (“target sequence so-far”) and tries to predict the next token in the target sentence (“decoder_outputs”). In order to allow our model to predict multiple metabolites for each couple of ”substrate + reaction”, we also implemented temperature sampling. It is a technique introduced to control the randomness of the sampling process, where low temperature means stagnation, while high temperature makes the model more creative.

F. Software and Code

Experiments were performed in Python, using Tensorflow version 2.14.0 and Pandas version 2.1.2. For the similarity evaluation, we employed the RDKit toolkit version 2023.09.1. The source code of this work is available upon request.

IV. Results

Given the relatively unexplored nature of metabolism prediction with chemical language models in the scientific literature, our primary goal was to determine the optimal hyperparameters, listed in Table I: we wanted to see which combination would lead to the highest performance of our model, established by low loss and high accuracy.

During the initial tuning phase, an extensive examination of various values for each hyperparameter was conducted to determine the optimal direction to follow. Batch Size is reported in Figure 5 as an example: the observed loss exhibits no significant disparity among diverse batch size values, indicating that it likely has minimal impact on model performance. On the contrary, a learning rate of 0.0001 is a more suitable choice compared to LR = 0.01, while LR = 0.005 remains a value of interest for further exploration, as shown in Figure 6.

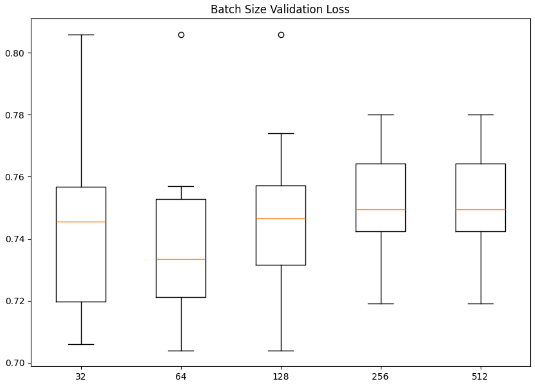

Fig. 5. Validation loss for different values of batch size, first tuning step

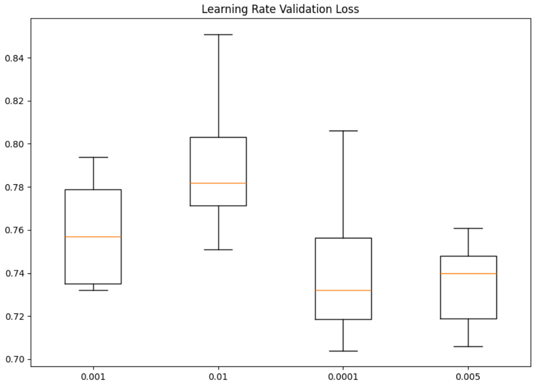

Fig. 6. Validation loss for different values of initial learning rate, first tuning step

TABLE I
Tuned hyperparameters and relative values - second tuning step.

Hyperparameter	**Value 1**	**Value 2**
Vocabulary size (VS)	**50**	-
Batch size (BS)	**64**	-
Embeddings dimension (ED)	64	**256**
Latent dimension (LS)	256	**1024**
Number of heads (H)	6	**10**
Initial learning rate (LR)	**0.0001**	0.005
Reaction class embeddings dimension (RCED)	**32**	-
Reaction class dense dimension (RCDD)	**32**	-

Following the insights gained from the analysis in tuning step 1, we narrowed down the range of values for each hyperparameter and executed a comprehensive grid search encompassing all possible combinations of values listed in Table I. The values corresponding to the best configuration are highlighted in bold.

The model predicted 10 metabolites for each couple “substrate + reaction” of the test set, and to determine its performance, we analyzed validity, which refers to the number of valid designs on the total number of generated metabolites, and recall. For this latter, we used fingerprint similarity based on the Tanimoto coefficient computed on ECFPs: we captured how close the model output resembled the correct metabolite, in terms of substructures and overall connectivity. We chose a threshold value for the Tanimoto coefficient equal to 0.8, meaning that all the predicted metabolites exhibiting a fingerprint similarity with the reference metabolite higher than 80% were considered true positives. We obtained 69.60% of validity and 62.64% of recall.

In Figure 7, we provide a visual example of the predictions considered as true positives based on fingerprint similarity. Related SMILES and involved reaction are listed in Table II.

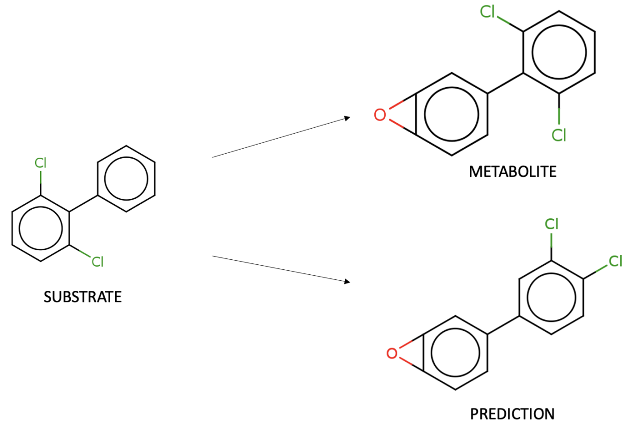

Fig. 7. Visual example of the predictions considered as true positives based on fingerprint similarity - oxidation.

As we aimed, the information pertaining to the reaction involved in the metabolism of molecules proves to be valuable. As illustrated in Figure 7, the predicted metabolite aligns with the description provided by the input reaction. Moreover, this is not limited solely to phase I reactions, which, being frequently predominant, tend to be more accurately recognized. It extends to phase II reactions as well, such as conjugation with glutathione, as depicted in Figure 8. Glutathione conjugation is a crucial phase II detoxification process, essential for protecting cells from the harmful effects of xenobiotics. Therefore, these results underscore the clinical relevance that this model can possess in providing insights into these processes for more effective and tailored medical treatments. Table III summarizes related SMILES and involved reaction.

To deeper investigate the significance and contribution provided by the inclusion of the reaction class, further analysis were conducted in two steps. We tested:

1) Our model on the test set after removing the reaction class information.
2) MetaTrans, another computational tool used for predicting metabolic reactions within biological systems, that, unlike our model, does not incorporate information about reaction classes during its training phase.

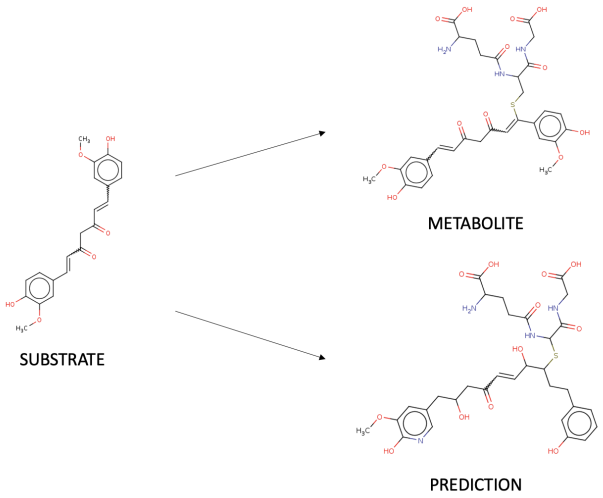

Fig. 8. Visual example of the predictions considered as true positives based on fingerprint similarity - conjugation.

The results, summarized in Table IV, suggest a significant enhancement in model performance with the inclusion of reaction class information. Our model, when trained with reaction class data, achieved higher overall accuracy (67%) compared to the version trained without it (59%). In contrast, MetaTrans, which does not utilize reaction class information, exhibited lower performance (38%), indicating the superiority of our approach in leveraging reaction class data for better predictive outcomes.

V. Conclusion and Future Developments

The objective of this project was the development of a deep learning model able to solve the metabolism prediction task as a machine translation problem, relying on the SMILES representation of molecules. This proof-of-concept illustrates how intricate is to capture metabolic reactions using chemical language models, yet it also highlights the potential promise of such an undertaking: even being a first draft, our model was already able to both learn chemical grammar rules and recognize the reactions molecules undergo. Notably, the inclusion of reaction class information led to a substantial improvement in correctly identifying instances, clarifying that that our model can effectively discern between different metabolic pathways and predict the outcomes of chemical transformations with greater accuracy. This will offer valuable insights into potential metabolites and their impacts on human health: it will aid in the early identification of potential toxicities or adverse effects during drug development, thereby enhancing patient safety. Furthermore, the model will contribute to the identification of drug-drug interactions and exploration of underlying biological mechanisms, ultimately advancing scientific knowledge in the field.

While our model exhibits its strengths, it is important to acknowledge its limitations, too. Addressing these challenges is crucial to enhance its flexibility, accuracy, and comprehensiveness: broadening the validation of predictions to en-

TABLE II
SMILES STRINGS AND REACTION CLASS ID RELATED TO FIGURE 7

Substrate	c1(c(-c2ccccc2)c(ccc1)Cl)Cl
Reaction class	11 - Oxidation of aryl compounds to epoxides, phenols or other metabolites
Known metabolite	Clc1cccc(Cl)c1-c1ccc2c(c1)O2
Prediction	Clc1ccc(-c2ccc3c(c2)O3)cc1Cl

TABLE III
SMILES STRINGS AND REACTION CLASS ID RELATED TO FIGURE 8

Substrate	Oc1ccc(cc1OC)C=CC(CC(C=Cc1cc(c(cc1)O)
Reaction class	103: Nucleophilic additions of glutathione (to a,ß-unsaturated carbonyls, quinones and analogues, isocyanates and isothiocyanates, epoxides, etc)
Known metabolite	COc1cc(C=CC(=O)CC(=O)C=C(SCC(NC(=O)CCC(N)C(=O)O)C(=O)NCC(=O)O)c2ccc(O)c(OC)c2)ccc1O
Prediction	COc1cc(CC(O)CC(=O)C=CC(O)C(CCc2cccc(O)c2)SC(NC(=O)CCC(N)C(=O)O)C(=O)NCC(=O)O)cnc1O

TABLE IV
RECALL COMPARED AMONG THE THREE MODELS

	Recall
Our model with RC	67%
Our model without RC	59%
MetaTrans	38%

compass a wider and more varied range of molecules, as well as conducting experimental validation and integrating supplementary options for ranking predictions is likely to prove beneficial in further enhancing the model's performance, in turn resulting in an increasingly impactful contribution to the field of drug discovery.

ACKNOWLEDGMENT

This work was partially supported by DAIS - Ca' Foscari University of Venice, within the ADIR program. FG and RÖ acknowledge support from the Irene Curie Fellowship and the Centre for Living Technologies.

REFERENCES

[1] D. Sun, W. Gao, H. Hu, and S. Zhou, "Why 90% of clinical drug development fails and how to improve it?" *Acta Pharmaceutica Sinica B*, vol. 12, no. 7, pp. 3049–3062, 2022.

[2] T. Cernak, "Synthesis in the chemical space age," *Chem*, vol. 1, no. 1, pp. 6–9, 2016.

[3] J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah, M. Spitzer *et al.*, "Applications of machine learning in drug discovery and development," *Nature reviews Drug discovery*, vol. 18, no. 6, pp. 463–477, 2019.

[4] R. R. Ramsay, M. R. Popovic-Nikolic, K. Nikolic, E. Uliassi, and M. L. Bolognesi, "A perspective on multi-target drug discovery and design for complex diseases," *Clinical and Translational Medicine*, 2018.

[5] Z. Zhang and W. Tang, "Drug metabolism in drug discovery and development," *Acta Pharmaceutica Sinica B*, vol. 8, no. 5, pp. 721–732, 2018.

[6] R. Caspi, K. Dreher, and P. D. Karp, "The challenge of constructing, classifying, and representing metabolic pathways," *FEMS microbiology letters*, vol. 345, no. 2, pp. 85–93, 2013.

[7] J. Kirchmair, A. H. Göller, D. Lang, J. Kunze, B. Testa, I. D. Wilson, R. C. Glen, and G. Schneider, "Predicting drug metabolism: experiment and/or computation?" *Nature reviews Drug discovery*, vol. 14, no. 6, pp. 387–404, 2015.

[8] N. Issa, H. Wathieu, A. Ojo, S. Byers, and S. Dakshanamurthy, "Drug metabolism in preclinical drug development: A survey of the discovery process, toxicology, and computational tools," *Current Drug Metabolism*, 2017.

[9] J. Kirchmair and et al., "Computational prediction of metabolism: Sites, products, sar, p450 enzyme dynamics, and mechanisms," *J. Chem. Inf. Model.*, 2012.

[10] J. Zhai and et al., "Comparison and summary of in silico prediction tools for cyp450-mediated drug metabolism," *Drug discovery today*, 2023.

[11] J. D. Tyzack and J. Kirchmair, "Computational methods and tools to predict cytochrome P450 metabolism for drug discovery," *Chemical biology and drug design*, 2019.

[12] A. Mazzolari, P. Perazzoni, E. Sabato, F. Lunghini, A. Beccari, G. Vistoli, and A. Pedretti, "Metaspot: A general approach for recognizing the reactive atoms undergoing metabolic reactions based on the metaqsar database," *International Journal of Molecular Sciences*, 2023.

[13] C. de Bruyn Kops, M. Šícho, A. Mazzolari, and J. Kirchmair, "Gloryx: Prediction of the metabolites resulting from phase 1 and phase 2 biotransformations of xenobiotics," *Chemical Research in Toxicology*, 2021.

[14] C. Li, M. Liakata, and D. Rebholz-Schuhmann, "Biological network extraction from scientific literature: state of the art and challenges," *Briefings in Bioinformatics*, 2014.

[15] C. Sarkar, B. Das, V. S. Rawat, J. B. Wahlang, A. Nongpiur, I. Tiewsoh, N. M. Lyngdoh, D. Das, M. Bidarolli, and H. T. Sony, "Artificial intelligence and machine learning technology driven modern drug discovery and development," *International Journal of Molecular Sciences*, 2023.

[16] J. Jiménez-Luna, F. Grisoni, N. Weskamp, and G. Schneider, "Artificial intelligence in drug discovery: recent advances and future perspectives," *Expert Opinion on Drug Discovery*, 2021.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[18] D. Weininger, "Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules," *Journal of Chemical Information and Computer Sciences*, 1988.

[19] J. Nam and J. Kim, "Linking the neural machine translation and the prediction of organic chemistry reactions," *ArXiv*, vol. abs/1612.09529, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:18208656

[20] P. Schwaller, T. Gaudin, D. Lanyi, C. Bekas, and T. Laino, "'found in translation': predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models," *Chemical Science*, 2018.

[21] P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas, and A. A. Lee, "Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction," *American Chemical Society*, 2019.

[22] 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.

[23] E. Croom, "Metabolism of xenobiotics of human environments," *Progress in Molecular Biology and Translational Science*, 2012.

[24] R. Todeschini, V. Consonni, H. Xiang, J. Holliday, M. Buscema, and P. Willett, "Similarity coefficients for binary chemoinformatics data: overview and extended comparison using simulated and real data sets," *Journal of chemical information and modeling*, vol. 52, no. 11, pp. 2884–2901, 2012.

[25] A. Pedretti, A. Mazzolari, G. Vistoli, and B. Testa, "MetaQSAR: an integrated database engine to manage and analyze metabolic data," *Journal of medicinal chemistry*, vol. 61, no. 3, pp. 1019–1030, 2018.

[26] A.-P. Josep, V. J. Simon, P. Oleksii, J. B. Esben, T. Christian, R. Jean-Louis, C. Hongming, and E. Ola, "Randomized SMILES strings improve the quality of molecular generative models," *Journal of Cheminformatics*, 2019.

[27] A. Mazzolari, L. Sommaruga, A. Pedretti, and G. Vistoli, "Metatree, a novel database focused on metabolic trees, predicts an important detoxification mechanism: The glutathione conjugation," *Molecules*, 2021.