



MetaPredictor: *in silico* prediction of drug metabolites based on deep language models with prompt engineering

Keyun Zhu, Mengting Huang, Yimeng Wang, Yaxin Gu, Weihua Li, Guixia Liu , Yun Tang *

Shanghai Frontiers Science Center of Optogenetic Techniques for Cell Metabolism, Shanghai Key Laboratory of New Drug Design, School of Pharmacy, East China University of Science and Technology, 130 Meilong Road, Shanghai 200237, China

*Corresponding author. Shanghai Frontiers Science Center of Optogenetic Techniques for Cell Metabolism, Shanghai Key Laboratory of New Drug Design, School of Pharmacy, East China University of Science and Technology, 130 Meilong Road, Shanghai 200237, China. E-mail: ytang234@ecust.edu.cn

Abstract

Metabolic processes can transform a drug into metabolites with different properties that may affect its efficacy and safety. Therefore, investigation of the metabolic fate of a drug candidate is of great significance for drug discovery. Computational methods have been developed to predict drug metabolites, but most of them suffer from two main obstacles: the lack of model generalization due to restrictions on metabolic transformation rules or specific enzyme families, and high rate of false-positive predictions. Here, we presented MetaPredictor, a rule-free, end-to-end and prompt-based method to predict possible human metabolites of small molecules including drugs as a sequence translation problem. We innovatively introduced prompt engineering into deep language models to enrich domain knowledge and guide decision-making. The results showed that using prompts that specify the sites of metabolism (SoMs) can steer the model to propose more accurate metabolite predictions, achieving a 30.4% increase in recall and a 16.8% reduction in false positives over the baseline model. The transfer learning strategy was also utilized to tackle the limited availability of metabolic data. For the adaptation to automatic or non-expert prediction, MetaPredictor was designed as a two-stage schema consisting of automatic identification of SoMs followed by metabolite prediction. Compared to four available drug metabolite prediction tools, our method showed comparable performance on the major enzyme families and better generalization that could additionally identify metabolites catalyzed by less common enzymes. The results indicated that MetaPredictor could provide a more comprehensive and accurate prediction of drug metabolism through the effective combination of transfer learning and prompt-based learning strategies.

Keywords: drug metabolism; metabolite prediction; deep language model; prompt learning; transfer learning

Introduction

Drug metabolism alters drug molecules through chemical modification catalyzed by various drug-metabolizing enzymes. The resulting metabolites may have physicochemical, pharmacological and even toxicological properties that are distinct from those of the original parent molecule [1]. Drug metabolism is normally divided into two phases, which commonly take place in the liver [2]. Phase I metabolism often involves the introduction or revelation of polar groups of the drug and is catalyzed primarily by cytochromes P450 (CYP450) enzyme family, other oxidoreductases and hydrolases. Phase II metabolism, which is mediated by transferases, serves as binding molecules to some endogenous small molecules to make their easy excretion from the body. Despite that xenobiotics generally become detoxified and deactivated via metabolism, metabolism of drugs may reduce efficacy and cause safety concerns. Toxicity can be triggered by the reactive metabolites that are formed through phase I and, less frequently, phase II reactions [3, 4]. Beyond that, drug metabolism may lead to drug–drug interactions and influence bioavailability [5]. Hence, it is significantly instructive to analyze the metabolism

process for effectively developing drugs. Traditionally, the study of drug metabolism requires the use of sophisticated analytical techniques, which are both resource intensive and labor intensive.

Computational methods for drug metabolism prediction have been developed to assist experimental assessment [6]. Some of these methodologies are very effective in the identification of the atoms within the molecule modified by metabolic transformation (known as sites of metabolism), such as SMARTCyP [7], FAME2 [8], SOMP [9] and Xenosite [10], but they are almost exclusively specific to CYP isoforms. The correct identification of SoMs could help infer the metabolite structures or suggest where a molecule might be rationally designed. In contrast to *in silico* SoM prediction, the computational task that infers metabolite structures from parent compounds is more difficult. Current methods for this task are dominated by rule-based approaches, such as SyGMA [11], Biotransformer [12] and GLORYx [13]. The challenges for rule-based approaches are as follows. Firstly, their application domain is bound by the coverage of transformation rules, which are manually compiled by experts. They may fail to generate predictions when there is no match between the substrate and the rule pattern. Secondly, growth in the number of rules may

Received: April 30, 2024. Revised: July 2, 2024. Accepted: July 16, 2024

© The Author(s) 2024. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

lead to low precision performance, as the number of false positives is increased. To reduce the number of false positives, some approaches rank the predicted metabolites based on statistical analysis [11]. Another attempt is to build machine learning models to identify substrate specificity [14] as a preliminary step to the application of rules. However, most machine learning models are designed just for phase I metabolism.

Artificial intelligence (AI) approaches for chemical reaction prediction have made significant progress in the last few years [15]. A representative work is molecular transformer [16], which tackled the reaction prediction problem as a machine translation task. Inspired by this, Litsa et al. proposed MetaTrans [17], a transformer-based deep-learning approach to directly convert parent molecules to metabolites and bypass the process of extracting metabolic transformation rules. However, MetaTrans was still performed with relatively low precision, which is a common challenge for metabolite prediction.

Prompt-based learning [18] is an up-and-coming paradigm in natural language processing (NLP) that provides opportunities to enhance interactions with AI systems. In NLP, prompts are text-based additional instructions or context provided to a model to help it understand tasks better and generate expected outputs. Thakkar et al. [19] adapted this prompt to fit the context of the retrosynthesis prediction by describing the disconnection site in SMILES strings of a molecule and achieved an improvement in prediction accuracy up to 39%. Inspired by this, we presented a prompt-based method named MetaPredictor to predict human metabolites for small molecules. We introduced prompts by embedding specific annotations within the SMILES string of parent molecules. The prompts specify the SoMs and are used to steer the transformer-based language model to translate a parent molecule into corresponding metabolites. Also, we validated this novel scheme on human metabolic transformations. The results demonstrated 30.4% performance improvement and 16.8% reduction of false positives over the baseline model, which confirmed that the SoM prompt could help the deep language model focus on critical regions and make more accurate predictions about the metabolites for small molecules.

Materials and methods

Overview of MetaPredictor

MetaPredictor is a rule-free, end-to-end and prompt-based tool to predict human metabolites for small molecules. We characterized molecules with SMILES sequence, so the metabolite prediction task could be tackled as a sequence translation problem. The sequence-to-sequence Transformer architecture [20] was used for molecular language modeling. (More details about Transformer are provided in ESI: S2.1†.)

MetaPredictor consists of two modules: SoM identifier and prompt-based metabolite predictor. It was designed as a two-stage schema; the SoM identifier was trained to automatically label the SoMs of small molecules, making the prompt-based metabolite predictor compatible with automatic or non-expert global predictions. More specifically, we auto-tagged the potential SoMs of the interesting parent molecule using SoM identifier followed by metabolite inference with prompt-based metabolite predictor. The interactive nature of prompt-based learning enables the effective integration of external knowledge including human-prompted input, resulting in local metabolite predictions for specific enzymes or specific SoMs.

Our novel methodology incorporates both transfer learning and prompt-based learning. Both transformer models were

firstly pre-trained on general chemical reaction dataset and subsequently fine-tuned on the metabolic reaction dataset. A description of the problem can be found in ESI: S2.2†. In addition, we used ensemble strategy to generate diverse predictions to accommodate metabolism prediction tasks. The full workflow of MetaPredictor is illustrated in Fig. 1.

Data collection and preparation

Chemical reaction data

The dataset for pre-training the models was derived from Lowe's work on mining chemical reaction data [21], which had been widely utilized in forward prediction [16] and retrosynthetic analysis of chemical reactions [22]. By removing duplicates and filtering, there were ~1.2 million training instances with a single product. The components in each chemical reaction were represented using the canonical SMILES and the reagents were removed.

Metabolic reaction data

Pairs of parent molecules and human metabolites that were represented by canonical SMILES constituted the metabolic reaction dataset. To derive a broad-coverage human metabolism dataset, we collected experimentally validated and structurally available human metabolites of both xenobiotic and endogenous compounds from open-access databases, literatures and Lee's Handbook of Metabolic Pathways of Xenobiotics [23]. The open-access databases included Human Metabolome Database (version 5.0) [24], Recon3D (version 3.01) [25], HumanCyc from MetaCyc (version 23.0) [26], DrugBank (version 5.1.10) [27] and the reaction database of BioTransformer (MetXBioDB) [14]. The metabolites were produced by single-step enzymatic reactions. The parent molecule could be a drug or a drug metabolite in the case of drugs with multi-step metabolic transformations. We split the decomposition reaction into two distinct training instances and kept the reaction in both directions when it was indicated as reversible. Concerning the metabolic reactions of endogenous compounds, we retained the pairs where the atom number of maximum common substructure exceed 40% of the atoms of the parent molecule. This process could keep metabolite to maintain a significant degree of structural similarity to the parent molecule and filter out less relevant metabolites. Finally, we merged the metabolic reactions from various sources and removed duplicates. RDKit toolkit [28] was employed for data processing. More details about data collection and source distribution are shown in ESI: S1.1† and Figure S1.

Validation and test sets

Since our work is focused on metabolic prediction for small molecules and specifically drug-like molecules, both the validation set and the test set consisted of drugs and drug metabolites. The validation set was handpicked by Litsa et al. [17] for selecting the hyperparameters for fine-tuned model and supervising model training, while the test set was used for evaluating the prediction performance of the model. In detail, the validation set consists of metabolic reactions derived from 96 parent compounds of DrugBank, catalyzed by not only CYP450 enzymes, but also non-CYP450 enzymes. The test set was sourced from the dataset manually curated by the developers of the GLORY method and DrugBank, and 135 drugs with 283 identified human metabolites were handpicked to create a more diverse test set in terms of metabolizing enzymes. Since distinct metabolic processes may exist for parent molecules, we ensured that instances sharing the same parent molecules were not partitioned into different data partitions (training, validation, test).

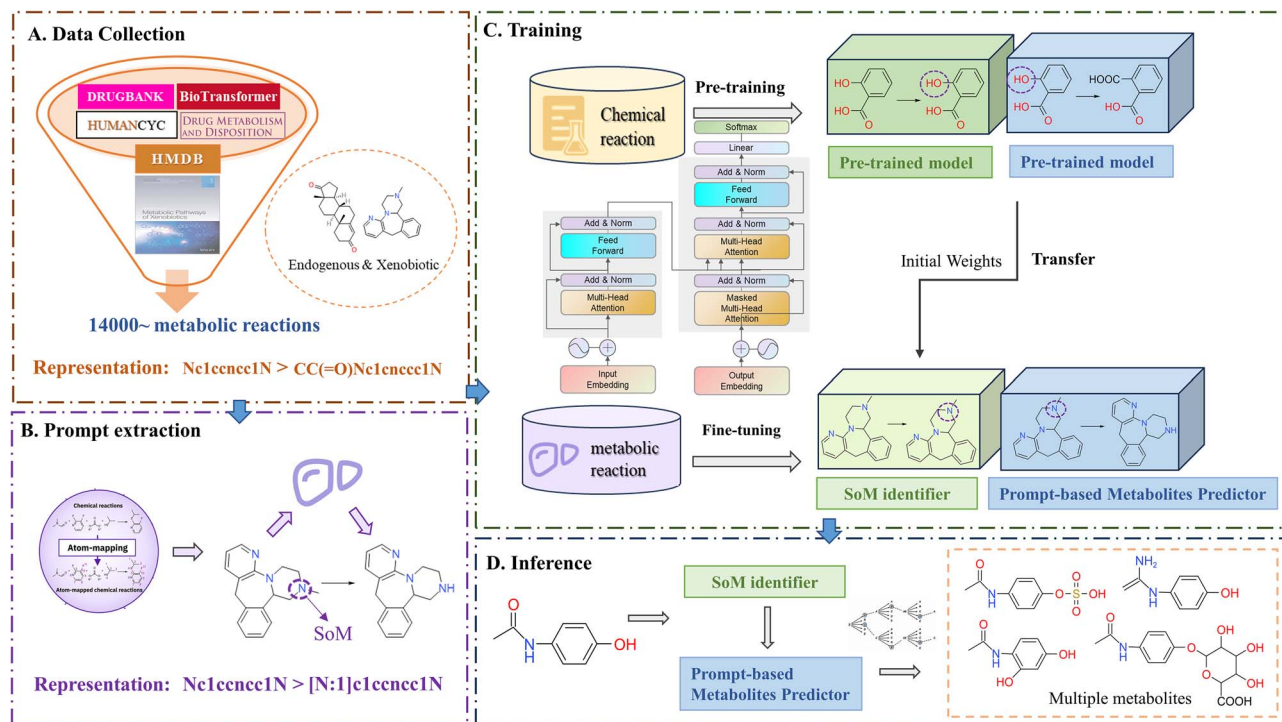


Figure 1. The workflow of MetaPredictor. (A) Collection of metabolic reaction dataset. (B) Extraction of prompt that specified SoM of the parent compound. (C) Training for SoM identifier and prompt-based metabolite predictor. (D) Model inference for potential metabolites.

Extracting prompt

In this study, the prompts were automatically extracted to substitute manually labeling for model building. Firstly, all reactions in both chemical reaction dataset and metabolic reaction dataset were atom-mapped using RXNMapper [29] to find out which atoms had altered the atomic environment during the reaction. The atomic environment includes the atom and all the bonds connected to the atom. For a given atom-mapped metabolic reaction SMILES, the atoms for which environment differed between parent molecule and metabolite were labeled in the parent molecule's SMILES and served as prompt corresponding to the site of metabolism. Similarly, the prompt-represented reactive atom was tagged in the reactant SMILES for a given atom-mapped chemical reaction SMILES. It should be stated that the prompts were introduced by using the SMARTS notation $[*:1]$ [30], where ‘*’ resembles any atom. The atom-mapping information was removed after extracting prompts. The pseudo-code is provided in ESI: S1.4†.

Model building

We processed different datasets for pre-training and fine-tuning to ensure that they aligned with the source and target sequences of the SoM identifier and prompt-based metabolite predictor. The distribution about the instances of the training, validation and test sets for different tasks are shown in Table S1 and Table S2. For the prompt-based metabolite predictor, we pre-trained it on the general chemical reaction dataset where the SMILES of reactants were labeled with reactive atom and then fine-tuned it on the dataset of metabolic reactions where the SMILES of parent molecules were prompted with SoMs. As for the SoM identifier, we also used transfer learning strategy to obtain generalized chemical knowledge about atomic reactivity. To enable the SoM identifier to automatically tag the SoM, the source sequences were SMILES of parent molecules and the target sequences were the SMILE of parent molecules with markers of SoMs. Before

training the transformer model, the input sequence and the output sequence were tokenized by using a regex pattern as described by molecular transformer.

All models used supervised learning and a seq-to-seq Transformer architecture as deployed in the OpenNMT-py library version 2.3.0 [31]. The transfer learning strategy employed in our study was to take the pre-trained model as a starting point for fine-tuning the model. For the parameters of the pre-trained model, minor changes were performed based on the molecular transformer. Regarding the fine-tuned model, diverse parameters including the SMILES augmentation strategy [32] and batch size were experimented and we selected models based on the accuracy of the validation set. More information about training parameter is provided in Table S3 and Table S4.

Model inference

Considering the fact that different metabolites may form through diverse enzymes for drugs, we constructed ensemble models based on the standard beam search algorithm to infer multiple possible sequences from integrated perspectives.

Beam search

The beam search algorithm is a popular search algorithm based on heuristics, which explores all likely characters and maintains the k most probable sequences [33]. Through the application of beam search algorithms, SoM identifier can generate several potential site-of-metabolism predictions for a given parent compound and prompt-based metabolite predictor can predict multiple metabolites for a SoM-prompt drug. By manipulating the beam size, the number of generated predictions can be varied. To strike a balance between enlarging the search space and obtaining the best prediction results, we tried to calculate the prediction performance on the validation metabolic set of all fine-tuned models with different beam sizes. In practice, beam sizes between 3 and 15 have been shown to provide an appropriate compromise

between precision and recall. Specific analysis is shown in Figure S2 and Figure S3.

Processing and ranking of predicted results

Since the model expands the SMILES sequence by selecting only the characters with higher probability when generating predictions, we need to filter the model outcomes to drop invalid SMILES and unreasonable metabolites. Invalid SMILES points to sequences that cannot be recognized by the RDKit toolkit. The unreasonable metabolites include metabolites that have far fewer atoms than the parent compound (<25%) and metabolites that contain different types of atoms from the parent compounds' atoms or the organic compounds' general atoms (H, C, N, O, S, P). After filtering, the models' predictions could be ranked in accordance with the cumulated log likelihood. The greater the log likelihood accumulated across the generated sequence, the stronger the model's confidence that the sequence is the correct result.

Ensemble strategy

To mitigate the potential biases of a single model's performance, ensemble strategy is often utilized in research to obtain better generalization performance [34]. In our study, we adopted this approach by averaging the prediction distributions from several decodes that were trained under different strategies. In order to reach a balance between maximizing the correct prediction rate and keeping the false-positive rate low, we constructed the ensemble models by combining different individual models and evaluated their model performance on the validation set. The top-5 ensemble models are shown in Table S5 and Table S6. Finally, we chose four individual models for SoM identifier and five individual models for prompt-based metabolite predictor. As a complementary note, the output size for individual models and ensemble models with a beam size of k were restricted to a maximum of k due to the filtering of some predictions.

Model evaluation

The performance evaluation of MetaPredictor primarily depends on its accuracy in predicting SoMs and metabolites, as well as its false-positive rate of predictions. The accuracy was assessed mainly by calculating recall, which is the rate of the number of reference metabolites correctly identified by the model to the total number of reference metabolites. The precision that indicates the percentage of false positives was calculated by the proportion of correctly identified reference metabolites to the output size of the model. We also calculated the percentage of input molecules for which at least one, at least half and all SoMs or metabolites were correctly identified by each model to observe the scope of drug retrieval.

To assess the efficacy of the algorithm to rank the decoded sequences, we calculated the top- N metrics when the model generates N predictions. For the evaluation of the prompt-based metabolite predictor, we compared the fingerprint similarity between predicted and reference metabolites. The prediction is considered correct if the fingerprint-based Tanimoto coefficient between the predicted metabolite and the reference is equal to 1.

Results and discussion

Analysis of metabolic reaction datasets

The human metabolic reaction dataset used to train the model consisted of 14 782 unique pairs of parent molecules and metabolites. To better understand the metabolic reaction dataset, we

analyzed metabolic reactions in terms of the distribution of EC classifications (EC-levels 1) for the metabolizing enzymes. Despite that metabolizing enzyme information was not provided for a substantial portion of the dataset, it can be observed that all enzyme classes are covered in the labeled pairs as illustrated in Fig. 2A. Among the distribution of enzymes, oxidoreductases (EC1), hydrolases (EC3) and transferases (EC2) are the most prominent categories. The former two are primarily responsible for catalyzing phase I drug metabolism, while the latter is dominant in phase II drug metabolism.

We visualized the metabolic reaction dataset using the dimensionality reduction algorithm TMAP. As shown in Fig. 2B, each point represents a metabolic reaction based on similarities calculated by the reaction fingerprint RXNFP [35]. Color coding the TMAP by the EC classification number, these enzyme families formed relatively discrete clusters of reactions, and an observation could be made that the majority of unlabeled metabolic reactions were likely catalyzed by oxidoreductases, transferases and hydrolases. The aforementioned analysis demonstrated that the metabolic reaction dataset covered the full spectrum of enzymes, with the expected biases toward the most frequent catalytic enzymes in the field of drug metabolism, thus providing a basis for training our model about the scope and specificity of metabolizing enzymes.

In addition, we analyzed the types of metabolic reactions in the test set and presented the results in ESI: S1.3†. It indicated that the test set covers a diverse range of metabolizing enzymes, allowing a more comprehensive assessment of the prediction performance of the model.

Model evaluation

Comparisons with baseline models

As a start, we evaluated the efficacy of transfer learning and ensemble strategy by presenting in juxtaposition the performance of four models on the metabolite test set, as shown in Fig. 3. Since the two modules of the MetaPredictor handled different prediction tasks, we made separate comparisons based on the ability of each model to identify reference SoMs or metabolites when generating five predictions. The results in Fig. 3 emphasized the significance of transfer learning and ensemble strategy.

To be more specific, the average performance of the individual fine-tuned models that compose the ensemble models showed obvious improvement to the pre-trained models and the models that trained only on metabolic reaction data. Taking prompt-based metabolite predictor as an example, the adoption of transfer learning led to an average 37.65% increase in recall and an average 6.2% improvement in precision. Also, the predictions from the pre-trained models had a low proportion of invalid SMILES but the highest proportion of unreasonable SMILES, suggesting that while the pre-trained models understand the rules of general chemical reactions and the syntax of SMILES language, they lack the necessary expertise in metabolic transformations. In addition, the models trained only on metabolic data performed poorly in prediction accuracy, predicting the highest number of invalid SMILES but a lower number of unreasonable SMILES than pre-trained models. This reflected the inadequacy of general chemistry training due to the relatively small amount of data, despite some knowledge of metabolic transformations being acquired. Furthermore, it can be seen that the ensemble models achieved not only a wider range of drug retrieval but also a higher recall rate and a lower false positive compared to the average individual models. Specifically, the ensemble model for the SoM identifier increased the recall rate by 8% and the precision by 2.6%

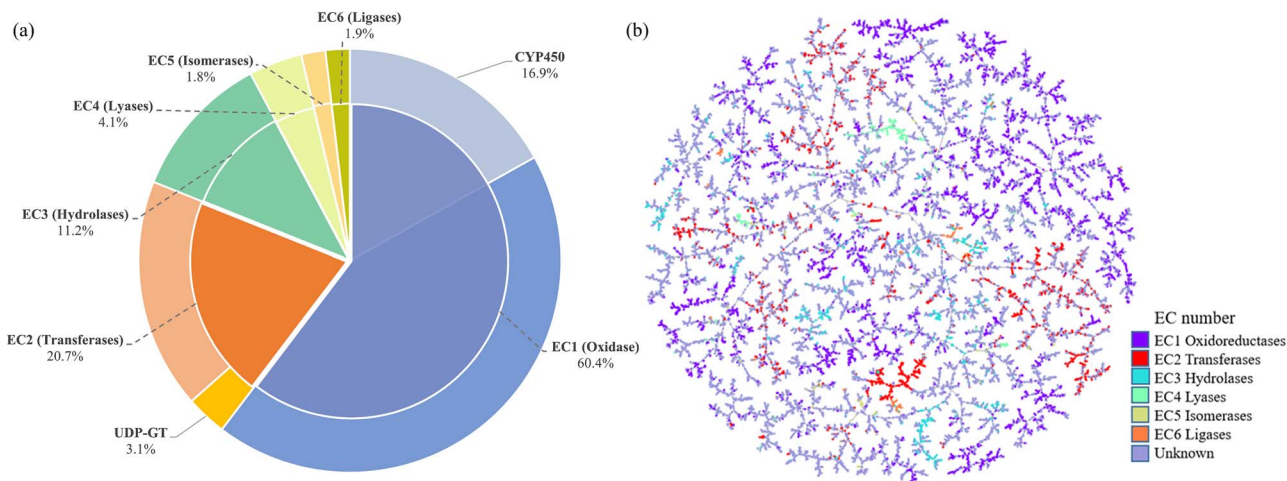


Figure 2. Analysis of the metabolic reaction dataset. (A) The composition of the dataset regarding the metabolizing enzymes based on the EC classification (exclusion of cases with no specified enzymes). (B) TMAP visualising the reaction similarity between metabolic reactions based on EC classification.

compared to the individual fine-tuned models. This demonstrated that the ensemble strategy is an effective approach for enhancing the output diversity and reducing the occurrence of false positives, while not increasing the output size of the models.

Subsequently, we respectively evaluated the prediction performance of the two ensemble models that comprised MetaPredictor on the metabolic test set with different top-N metrics. As shown in Table S7, the results indicated that both modules of MetaPredictor performed good prediction accuracy for their respective metabolic site or metabolite prediction tasks and had a relatively wide range of drug retrieval scopes.

To assess the efficacy of introducing prompts that specify the SoMs in improving the task of metabolite prediction, we used the same dataset and methodology without introducing prompt information to train the baseline model. Specifically, the source sequences of the baseline model are SMILES of the parent compound without SMARTS notation [*:1] that is used to tag the atoms that undergo metabolic transformation. The prompt-based metabolite predictor can use an additional input prompt to guide the metabolite translation, as opposed to the baseline model, which only generated metabolite predictions based on the underlying probability distribution of metabolism transformations in the training dataset. Then, we compared the performance of the baseline model and the prompt-based metabolite predictor on the metabolic test set, and the results are displayed in Fig. 4. The prompt-based metabolite predictor achieves superior performance metrics than the baseline model in all cases where the output size of both models is comparable. Specifically, the implementation of guided prompts resulted in a remarkable average increase of 30.4% in recall and 16.8% in precision. This demonstrated that the introduction of prompt learning could indeed guide the model to generate more accurate metabolite predictions while simultaneously reducing the occurrence of false positives.

Comparison with other prediction tools

In this study, we further evaluated the performance of MetaPredictor by comparison with four existing drug metabolism prediction tools: GLORYx, SyGMa, BioTransformer and MetaTrans. We compared the ability to identify and rank reference metabolites of these five methods on the metabolic test set. All methods generated metabolites through a single-step reaction and were evaluated using fingerprint similarity. For the ranking capability analysis, we compared the top-N ($N = 5, 10, 15$) prediction results

generated by MetaPredictor, MetaTrans, GLORYx and SyGMa. The top-12 performance was also chosen to ensure a fair comparison with BioTransformer, which had an average output size of ~ 12 on the metabolic test set. As for MetaPredictor, we used its automatic prediction pattern that generates SoM prompts without any human intervention. The introduction and implementation of other four methods are displayed in ESI: S3.3†.

The results, as shown in Table 1, proved that MetaPredictor presented great prediction performance and ranking capabilities. Even though MetaPredictor was not trained on a drug-specific dataset, its performance was comparable to models that had been specifically developed for drug metabolism. We could observe that the identified metabolites by MetaPredictor had a relatively larger coverage of the dataset compared to the rule-based methods, which means generating at least one correct metabolite prediction for a larger proportion of the dataset. This might be explained by the fact that rule-based methods relied on the exact matching of compounds and metabolic rules for predictions, whereas MetaPredictor did not. The specific example was that BioTransformer could not generate metabolite predictions for four compounds in the test set resulting in relatively low coverage of the dataset. Moreover, MetaPredictor exhibited better prediction performance and ranking capability when compared with MetaTrans, which was also based on an end-to-end learning method. This improvement could be attributed to the introduction of prompt learning and the expansion of the training dataset, which provided another illustration of the effectiveness of prompt learning in guiding the model to generate more accurate predictions. However, as the number of predictions increases, we cannot be sure whether other approaches would show better performance. Table 1 shows that the model performance of the rule-based approaches improved more substantially with an increase in the model output size, albeit at the expense of precision.

We further analyzed the top-12 performance of each method by concerning the various enzyme families, as shown in Fig. 5A. This evaluation focused on three enzyme families that play important roles in the metabolic processes: oxidation enzymes, the most prevalent of which are CYP450s, transferases, primarily including the UDP-glucuronosyltransferases (UGT) and sulfotransferases, and hydrolases. As we could see from Fig. 5A, MetaPredictor, BioTransformer and SyGMa showed some advantages in terms of identification of metabolites related to oxidation reactions. When considering phase II

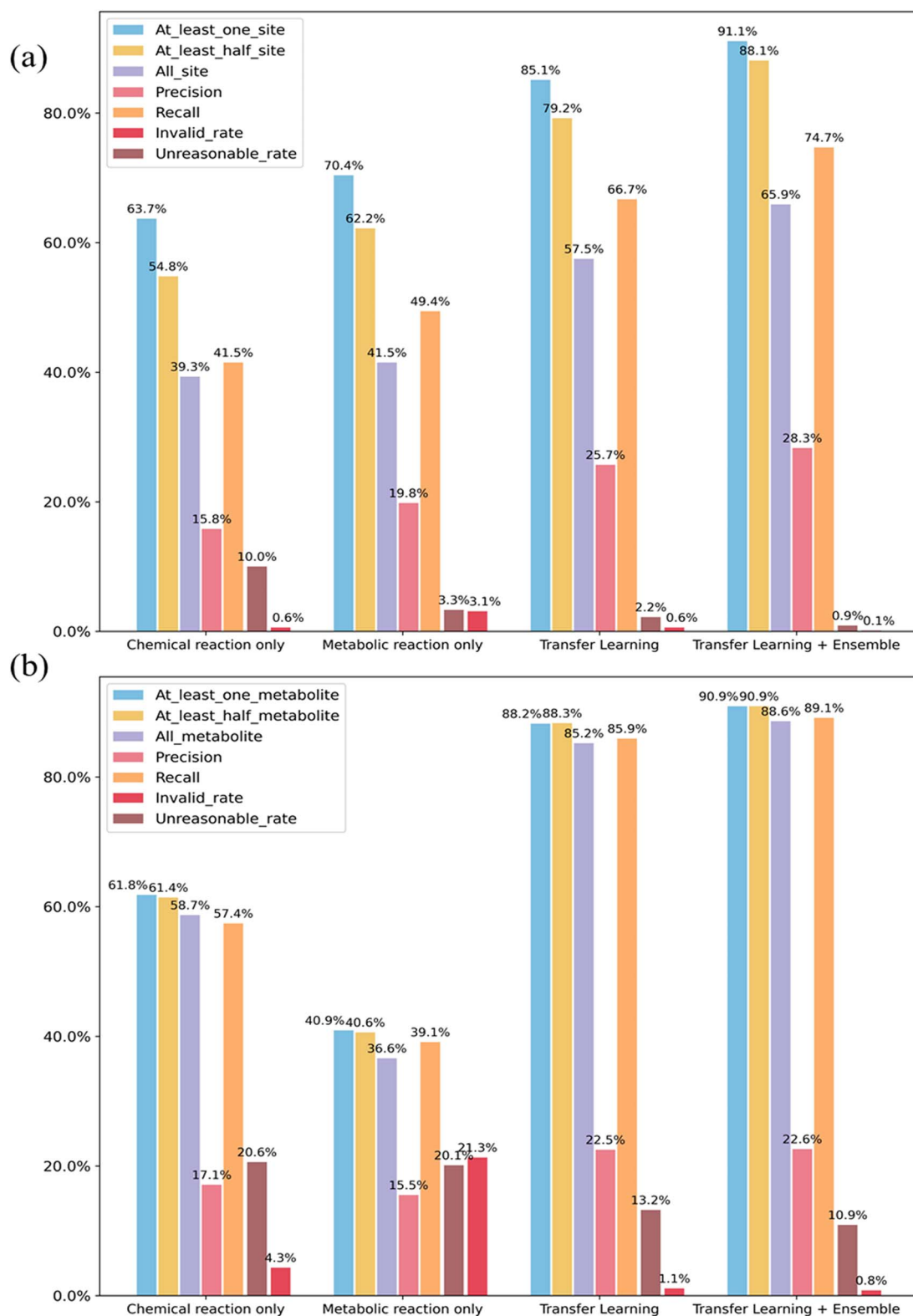


Figure 3. Top-5 prediction performance of the pre-trained models (chemical reaction only), the models trained only on metabolic data (metabolic reaction only), the average performance of the individual fine-tuned models (transfer learning) and the ensemble models (transfer learning + ensemble) for (A) SoM identifier and (B) prompt-based metabolite predictor on the metabolic test set.

metabolism-related metabolites, MetaPredictor, GLORYx and SyGMA performed slightly better. MetaPredictor and SyGMA could correctly predict more hydrolase metabolites. Overall, all methods seem to have the capability to cover these enzyme families that are important for metabolism.

Regarding MetaPredictor, the great diversity of the training set allowed the model to achieve metabolite prediction without being restricted to any specific classes of enzymes. More importantly, it not only showed better prediction performance in terms of the primary enzyme families of phase I and phase

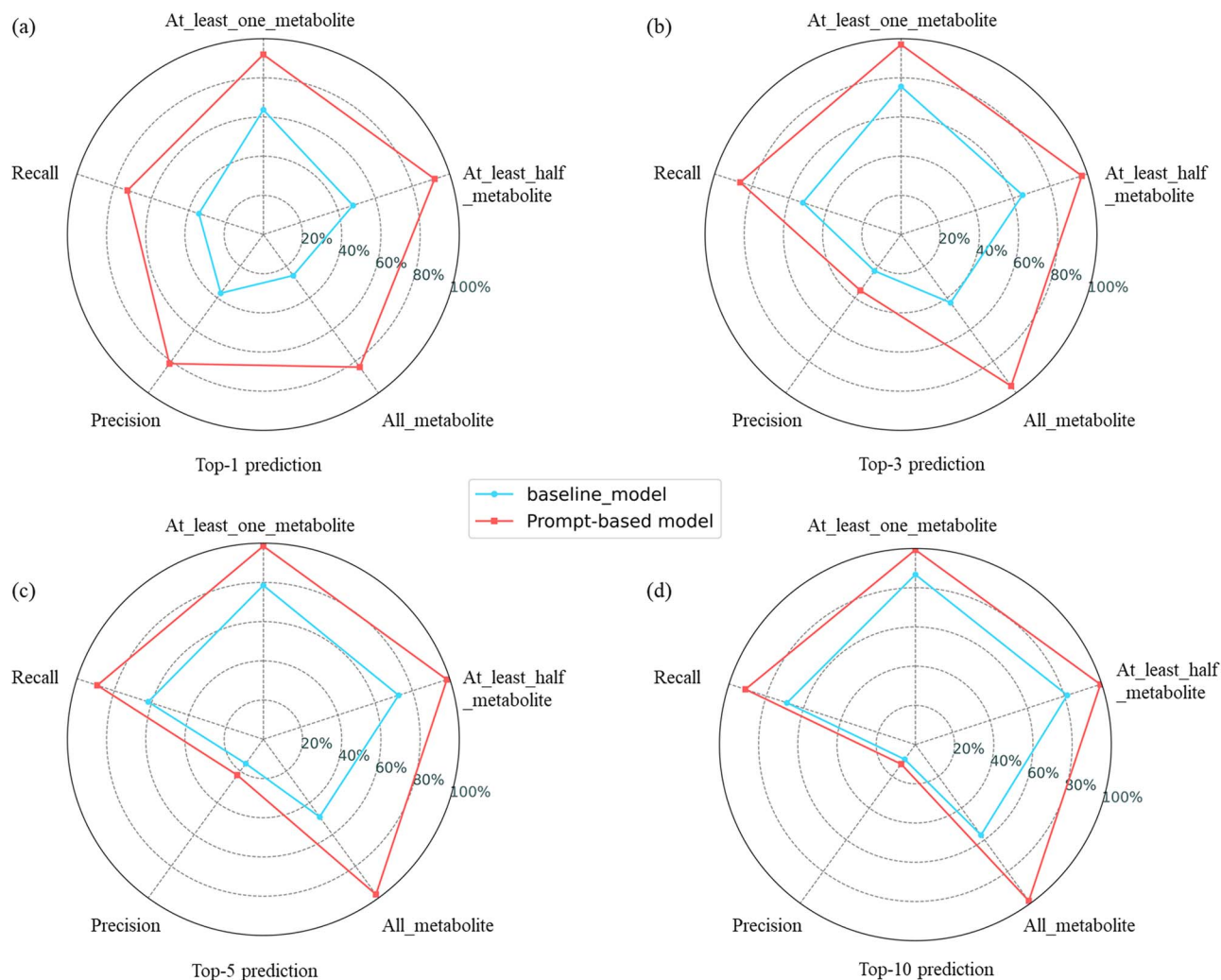


Figure 4. Comparison of prediction performance between the prompt-based metabolite predictor aware model (prompt-based model) and the baseline model that trained and evaluated on the same dataset when output size of models is the same.

Table 1. Comparison of prediction performance between MetaPredictor, MetaTrans, GLORYx, SyGMa and BioTransformer on the metabolic test set

	Method	At least one metabolite (%)	At least half metabolite (%)	All metabolites (%)	Total identified metabolites	Precision (%)	Recall (%)	Output size
Top5	MetaPredictor	77	67.4	45.2	154	20.6	54.4	748
	MetaTrans	72.6	63	31.1	129	17.2	45.6	748
	GLORYx	60.7	44.4	28.1	110	16.3	38.9	675
	SyGMa	68.1	62.2	35.6	135	20	47.4	675
Top10	MetaPredictor	88.9	81.5	57	192	15.7	67.8	1221
	MetaTrans	81.5	73.3	44.4	161	11.4	56.9	1411
	GLORYx	71.9	63	40.7	159	11.8	56.2	1343
	SyGMa	81.5	74.1	51.1	181	13.5	64	1338
Top12	MetaPredictor	91.1	85.9	61.5	205	12.2	72.4	1686
	MetaTrans	83.7	75.6	48.9	172	9.9	60.8	1732
	GLORYx	77.8	71.9	48.1	175	10.9	61.8	1606
	SyGMa	83.7	77.8	54.8	191	11.9	67.5	1600
Top15	BioTransformer	69.6	64.4	42.2	173	10.8	61.1	1596
	MetaPredictor	91.9	86.7	64.4	209	10.9	73.9	1915
	MetaTrans	83	76.3	50.4	178	8.6	62.9	2068
	GLORYx	83.7	77.8	57	195	9.8	68.9	1985
	SyGMa	84.4	79.3	57	198	9.9	70	1992

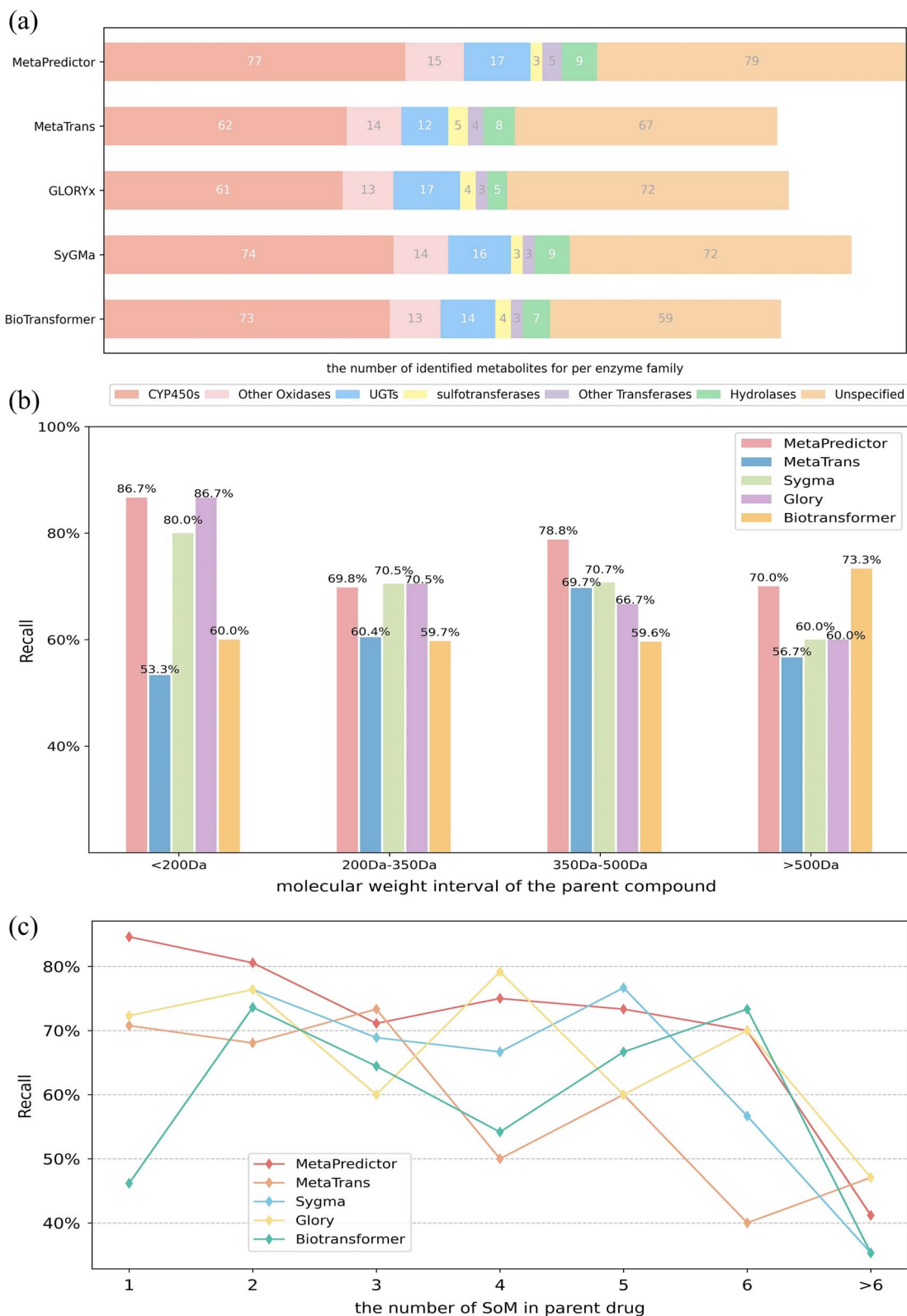


Figure 5. The comparison of model performance between MetaPredictor, MetaTrans, GLORYx, SyGMa and BioTransformer considering (A) number of identified metabolites for each enzyme family on the metabolic test set; (B) recall for different molecular weights of parent drug molecules; (C) recall for different numbers of SoM in parent drug molecules.

II metabolism, but also could find metabolites catalyzed by enzymes that are less frequent in drug metabolism, which may be missed by rule-based methods. One of these specific cases is that the drug fingolimod is transformed into an active

compound fingolimod phosphate (Fig. 6) through the metabolism process catalyzed by the enzyme sphingosine kinase (EC 2.7.1.91) [36]. This metabolite was also identified by MetaTrans, but not by the other three tools. The second specific case is the

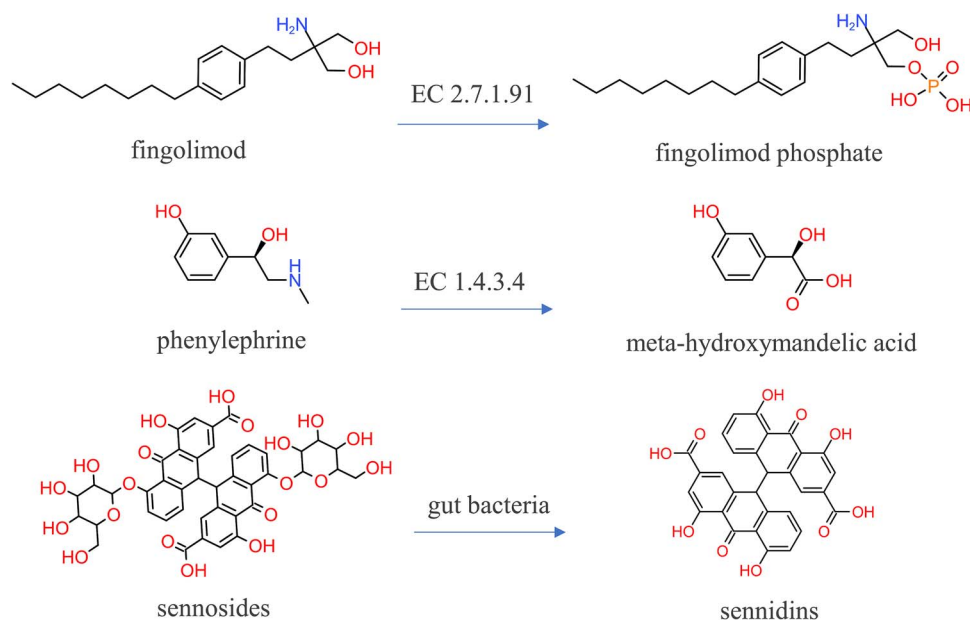


Figure 6. MetaPredictor correctly identified metabolites that transformed through uncommon enzymes.

alpha-1 adrenergic agonist, phenylephrine, which is metabolized through a monoamine oxidase (EC 1.4.3.4) into the meta-hydroxymandelic acid (Fig. 6) which is an inactive metabolite [37]. Another noteworthy case is sennidins (Fig. 6), the metabolite of anthraquinone derivatives sennosides, which is derived through the hydrolysis of gut bacteria [38]. MetaPredictor correctly predicted this metabolic process, demonstrating its applicability beyond the range of hepatic metabolizing enzymes.

To further explore the applicability and benefits of our models, we further evaluated and compared the performance between MetaPredictor and other tools considering different molecular weights of parent compounds and numbers of SoM in parent compounds. As shown in Fig. 5B, MetaPredictor showed competitive recall in all molecular weight ranges, especially in the <200 Da and 350–500 Da ranges, where the recall reached 86.7% and 78.8%, respectively, much higher than those of other methods. This result indicates that MetaPredictor has stable and great prediction abilities for parent compounds of different molecular weights, and especially performs well for parent compounds of smaller and medium molecular weights. It can be seen from Fig. 5c that MetaPredictor showed favorable recall on parent compounds containing from one to five SoMs, especially on parent compounds with only one SoM, where the recall was close to 85%. While the recall of MetaPredictor gradually decreased as the number of SoM increased, probably due to the more complex metabolic pathways increasing the difficulty of model prediction, MetaPredictor still showed competitive prediction performance compared with other methods. These results highlight the robustness and versatility of MetaPredictor in prediction of drug metabolites.

Challenging cases

To better understand the potential capabilities and limitations of the proposed approach, we scrutinized the predicted metabolites for drugs in the metabolic test set. The maximum average similarity based on molecular fingerprints between the model mispredictions and the reference metabolites was calculated to be 0.73, and Fig. 7 illustrates several representative cases where the predicted metabolite deviated from the reference metabolite.

Occasionally, the difference between the reference metabolite and the nearest prediction could arise from just one non-reactive atom. A prime example is DBMET00112 (Case 1 in Fig. 7). Despite the model predicting that DBMET00112 would undergo hydroxylation and correctly identifying the hydroxylation reaction site, it erroneously replaced the non-reactive hydroxymethyl of the original structure with a methyl group, a transformation that is unlikely observed in human metabolic pathways. For certain cases, the error may be attributed to the reference metabolite, as shown in Cases 2 and 3 in Fig. 7. In fact, we found evidence in the literature that ticlopidine in Case 2 could undergo two distinct oxidation reactions at the nitrogen atom, one involving the oxidation of the nitrogen atom and the other resulting in the N–C bond breakage [39]. Our model successfully predicted the latter reaction, whereas DrugBank only recorded the former as a reference metabolite. Similarly, the predicted metabolite of metoclopramide in Case 3 was derived from hydroxylation at the arylamino group, while the reference metabolite collected by Glory was generated from the oxidation of the same group. However, the N–O-glucuronide of metoclopramide was experimentally detected in its human metabolites [40]. It was formed by sequential metabolism via P450 followed by UGT, and the intermediates in this process matched our model predictions.

We also observed a problem in the model predictions regarding successive oxidation reactions. Although the model correctly identified the position and type of oxidation reaction, the predicted structure sometimes does not align exactly with the reference metabolite. For example, in the case of the drug TAK-438 (Case 4 in Fig. 7), the reference metabolite is a carboxylic acid, but the model predicted the metabolites to be the corresponding aldehyde. According to the literature, aldehydes are often identified as intermediates that are subsequently metabolized by CYP450 enzymes into carboxylic acids [41]. It is particularly challenging for the model to address the case of metabolites that are formed through multiple transformations at diverse sites, as our model was trained on a single-step metabolic reaction dataset. Such instances include the drugs molsidomine and bupropion (Cases 5 and 6 in Fig. 7). In the case of molsidomine, the reference metabolite was possibly derived through a multi-step process

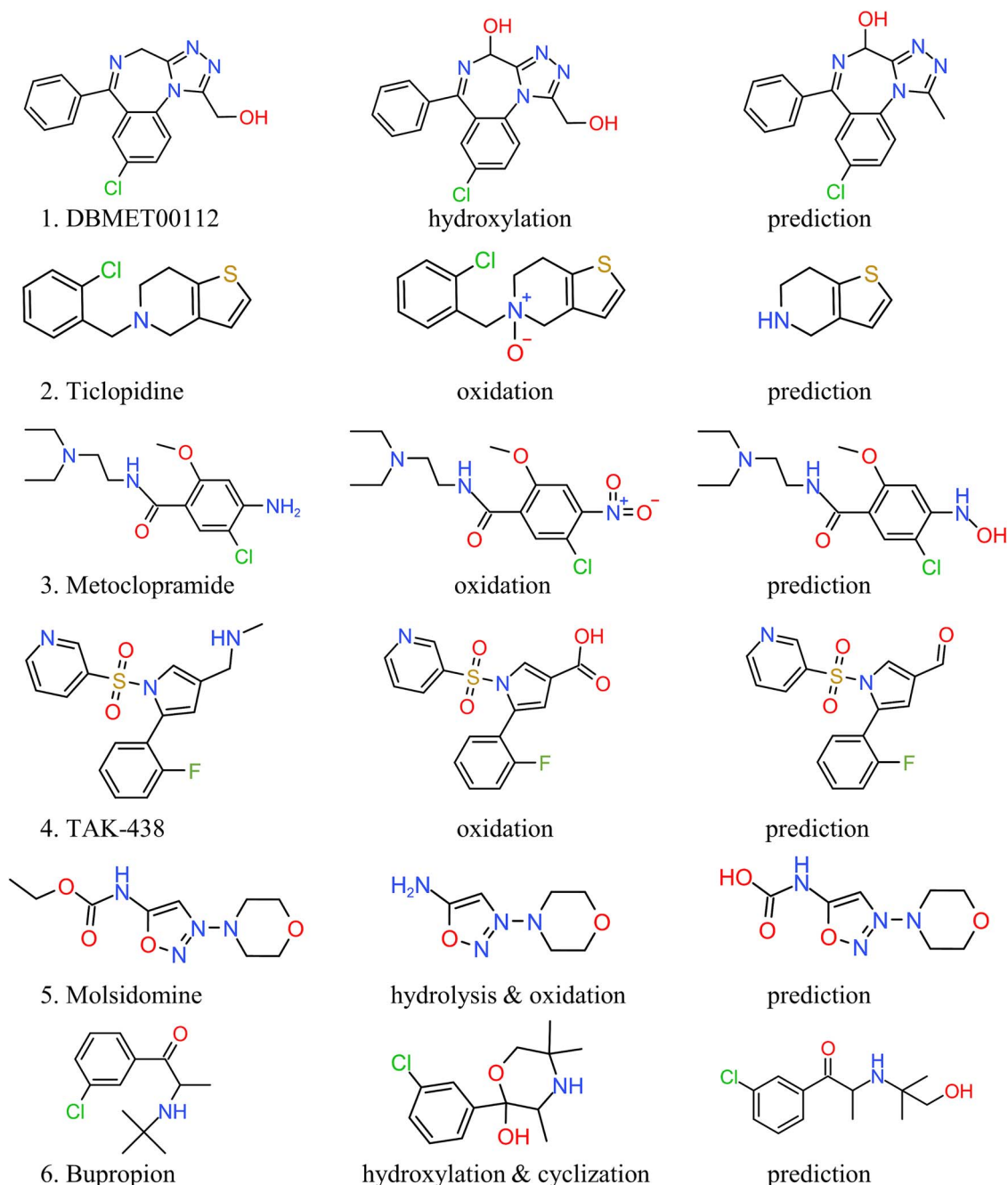


Figure 7. Drug structure, corresponding reference metabolite and nearest prediction for several representative cases that MetaPredictor mispredicted.

that involved esterase hydrolysis and oxidative decarboxylation [42]. Despite correctly identifying the product of the esterase hydrolysis, the model failed to simultaneously predict the oxidation reaction. Regarding bupropion, the model identified the hydroxylation reaction type and hydroxylation site, but missed one cyclization reaction required for converting this prediction to the reference metabolite [43].

Overall, despite certain instances where the predicted metabolites did not perfectly match with the reference ones, our inspection concluded that the model predictions still provided valuable insights for drug metabolism research. More specifically, many cases showed that the predictions correctly identified the reaction type, site of metabolism in the parent compound and even the intermediate of metabolic reactions.

Model attention analysis

With the help of the visualization of attention weights, we can see to some extent how the model learned metabolic transformations. Figure 8 shows the attention weights assigned by the model to tokens within SMILES of a parent drug molecule during the prediction of its metabolites. Higher attention weights indicate that the model considers those specific tokens more important for making predictions. It can be seen from Fig. 8A that the prompt-based metabolite predictor assigns higher attention weights to the [CH2:1] token (with a darker color). This indicates that the model focuses on the crucial token that corresponds to the SoM of the parent drug molecule and consequently predicts the sequence of correct reference metabolite. Conversely, the heatmap in Fig. 8B shows a more scattered attention pattern, with higher weights

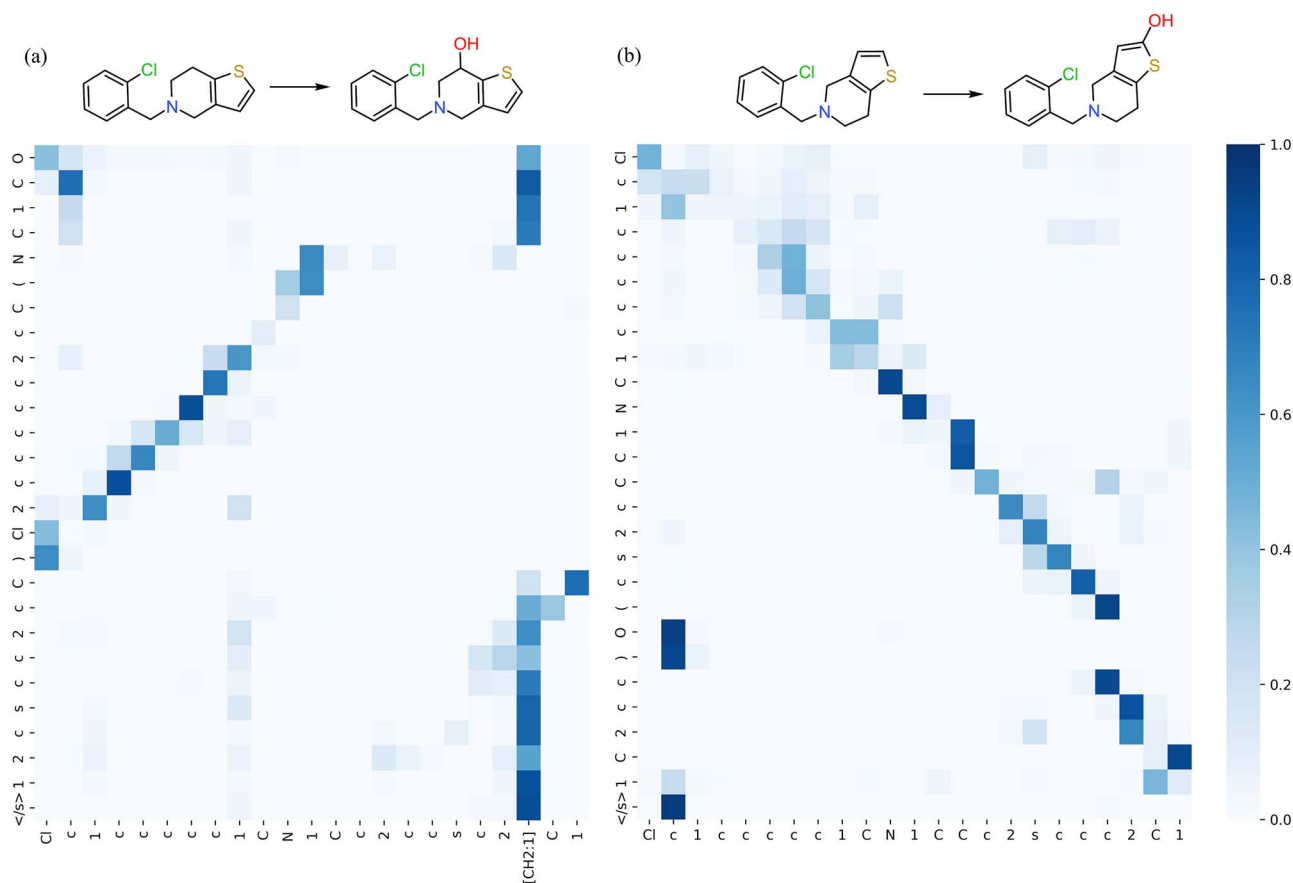


Figure 8. Visualization of attention weights assigned by the model to tokens within SMILES of a parent drug molecule during the prediction of its metabolite (A) based on the prompt-based metabolite predictor that correctly identified reference metabolite and (B) based on the baseline model that mispredicted metabolite.

assigned to atoms that are not involved in the metabolic transformation. This may reveal why the baseline model generated incorrect prediction due to the lack of ability to effectively identify critical regions necessary for accurate metabolite prediction. This comparison further highlights the advantage of introducing prompt-based learning. By providing SoM prompts, the model can be better guided to focus on tokens that were critical for metabolic transformation and generate more accurate predictions.

Conclusions

In this study, we proposed a prompt-based learning approach named MetaPredictor to predict metabolites of small molecules in the human body. MetaPredictor consists of two transformer models: SoM identifier and prompt-based metabolite predictor. It was designed to automate workflows for a wider application. For the first time in metabolite prediction task, we integrated prompts that specified the SoMs with deep language model to enrich domain knowledge and navigate the translation of a parent molecule into correct metabolites. The prompt-based metabolite predictor achieved up to 89% recall and 22.6% precision for top-5 predictions, which were ~30% improvement of performance and 16% reduction of false positives over the baseline model that was trained on the same dataset. This demonstrated the validity of prompt-based learning that partly mitigated the challenges related to metabolite prediction, i.e. decreased precision due to increased output size. Furthermore, transfer learning strategy was utilized to acquire generalized knowledge of chemical reactions to tackle the limited availability of human metabolic reactions.

Although MetaPredictor was not specifically trained on drug metabolic data, it showed improved or comparable performance with other drug-specific methods. The variety of the dataset enabled the MetaPredictor to predict metabolites catalyzed by uncommon enzymes, expanding the predictability of metabolic reactions and model generalization when the existing rule-based methods were focused on the major enzyme families. In addition, MetaPredictor performs reliably effectively across a range of molecular weights and can handle compounds with different numbers of SoMs effectively. MetaPredictor infers metabolites in a similar way to human experts, and the use of a prompt-based language steers the inference of metabolite prediction models toward chemical transformations taking place around the SoMs, which also makes this inference more interpretable. When compared to the rule-free method, MetaPredictor not only provides a more comprehensive analysis of drug metabolism, but also allows for the integration of external knowledge and experience. This human-in-the-loop approach makes prediction of metabolites more accurate or more expected and provides a new paradigm for knowledge introduction for deep-learning models in the field of drug discovery. We expect that MetaPredictor could contribute to accelerating and enhancing safety and efficacy assessment in the early stage of drug discovery.

Key Points

- We presented a rule-free, end-to-end and prompt-based method named MetaPredictor to predict possible human

metabolites of small molecules and offered a solution to address the challenges associated with drug metabolite prediction.

- The introduction of prompt engineering can steer deep language model to generate more accurate metabolite prediction, which provides a new paradigm for knowledge introduction for deep-learning models in the field of drug discovery.
- MetaPredictor was designed as a two-stage schema for a wider application and showed improved performance when compared to four available drug metabolite prediction tools. It could provide a more comprehensive and accurate prediction of drug metabolism.

Supplementary data

Supplementary data are available at *Briefings in Bioinformatics* online.

Conflict of interest: None declared.

Funding

National Key Research and Development Program of China (Grant 2023YFF1204904), the National Natural Science Foundation of China (Grants U23A20530 and 82173746), the 111 Project (Grant BP0719034) and the Shanghai Frontiers Science Center of Optogenetic Techniques for Cell Metabolism (Shanghai Municipal Education Commission).

Data availability

The running code and trained models are available at our GitHub repository: <https://github.com/zhukeyun/Meta-Predictor>.

References

1. Testa B, Pedretti A, Vistoli G. Reactions and enzymes in the metabolism of drugs and other xenobiotics. *Drug Discov Today* 2012;**17**:549–60. <https://doi.org/10.1016/j.drudis.2012.01.017>.
2. Croom E. Metabolism of xenobiotics of human environments. *Prog Mol Biol Transl Sci* 2012;**112**:31–88. <https://doi.org/10.1016/B978-0-12-415813-9.00003-9>.
3. Holt M, Ju C. Drug-induced liver injury. *Handb Exp Pharmacol* 2010;**196**:3–27.
4. Antoine DJ, Williams DP, Park BK. Understanding the role of reactive metabolites in drug-induced hepatotoxicity: state of the science. *Expert Opin Drug Metab Toxicol* 2008;**4**:1415–27. <https://doi.org/10.1517/17425255.4.11.1415>.
5. Tang W, Lu AY. Metabolic bioactivation and drug-related adverse effects: current status and future directions from a pharmaceutical research perspective. *Drug Metab Rev* 2010;**42**:225–49. <https://doi.org/10.3109/03602530903401658>.
6. Kirchmair J, Göller AH, Lang D, et al. Predicting drug metabolism: experiment and/or computation. *Nat Rev Drug Discov* 2015;**14**:387–404. <https://doi.org/10.1038/nrd4581>.
7. Rydberg P, Gloriam DE, Zaretski J, et al. SMARTCyp: a 2D method for prediction of cytochrome P450-mediated drug metabolism. *ACS Med Chem Lett* 2010;**1**:96–100. <https://doi.org/10.1021/ml100016x>.
8. Šicho M, de Bruyn KC, Stork C, et al. FAME 2: simple and effective machine learning model of cytochrome P450 regioselectivity. *J Chem Inf Model* 2017;**57**:1832–46. <https://doi.org/10.1021/acs.jcim.7b00250>.
9. Rudik A, Dmitriev A, Lagunin A, et al. SOMP: web server for in silico prediction of sites of metabolism for drug-like compounds. *Bioinformatics* 2015;**31**:2046–8. <https://doi.org/10.1093/bioinformatics/btv087>.
10. Zaretski J, Matlock M, Swamidass SJ. XenoSite: accurately predicting CYP-mediated sites of metabolism with neural networks. *J Chem Inf Model* 2013;**53**:3373–83. <https://doi.org/10.1021/ci400518g>.
11. Ridder L, Wagener M. SyGMA: combining expert knowledge and empirical scoring in the prediction of metabolites. *ChemMedChem* 2008;**3**:821–32. <https://doi.org/10.1002/cmdc.200700312>.
12. Wishart DS, Tian S, Allen D, et al. BioTransformer 3.0—a web server for accurately predicting metabolic transformation products. *Nucleic Acids Res* 2022;**50**:W115–23. <https://doi.org/10.1093/nar/gkac313>.
13. de Bruyn KC, Šicho M, Mazzolari A, et al. GLORYx: prediction of the metabolites resulting from phase 1 and phase 2 biotransformations of xenobiotics. *Chem Res Toxicol* 2020;**34**:286–99. <https://doi.org/10.1021/acs.chemrestox.0c00224>.
14. Djoumbou-Feunang Y, Fiamoncini J, Gil-de-la-Fuente A, et al. BioTransformer: a comprehensive computational tool for small molecule metabolism prediction and metabolite identification. *J Chem* 2019;**11**:1–25. <https://doi.org/10.1186/s13321-018-0324-5>.
15. Nair VH, Schwaller P, Laino T. Data-driven chemical reaction prediction and retrosynthesis. *CHIMIA Int J Chem* 2019;**73**:997–1000. <https://doi.org/10.2533/chimia.2019.997>.
16. Schwaller P, Laino T, Gaudin T, et al. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS Cent Sci* 2019;**5**:1572–83. <https://doi.org/10.1021/acscentsci.9b00576>.
17. Litsa EE, Das P, Kaviraki LE. Prediction of drug metabolites using neural machine translation. *Chem Sci* 2020;**11**:12777–88. <https://doi.org/10.1039/D0SC02639E>.
18. Liu P, Yuan W, Fu J, et al. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. *ACM Comput Surv* 2023;**55**:1–35. <https://doi.org/10.1145/356081>.
19. Thakkar A, Vaucher AC, Byekwaso A, et al. Unbiasing retrosynthesis language models with disconnection prompts. *ACS Cent Sci* 2023;**9**:1488–98. <https://doi.org/10.1021/acscentsci.3c00372>.
20. Vaswani A, Shazeer N, Parmar N et al. Attention is all you need. *Adv Neural Inf Process Syst* 2017.
21. Lowe DM. *Extraction of Chemical Structures and Reactions from the Literature*. Ph.D. Thesis, University of Cambridge, 2012.
22. Schwaller P, Petraglia R, Zullo V, et al. Predicting retrosynthetic pathways using transformer-based models and a hypergraph exploration strategy. *Chem Sci* 2020;**11**:3316–25. <https://doi.org/10.1039/C9SC05704H>.
23. Lee PW, Aizawa H, Gan LL, et al. *Handbook of Metabolic Pathways of Xenobiotics* (Vol. 1 - Vol. 5). John Wiley & Son Ltd., 2014.
24. Wishart DS, Feunang YD, Marcu A, et al. HMDB 4.0: the human metabolite database for 2018. *Nucleic Acids Res* 2018;**46**:D608–17. <https://doi.org/10.1093/nar/gkx1089>.
25. Brunk E, Sahoo S, Zielinski DC, et al. Recon3D enables a three-dimensional view of gene variation in human metabolism. *Nat Biotechnol* 2018;**36**:272–81. <https://doi.org/10.1038/nbt.4072>.
26. Caspi R, Billington R, Fulcher CA, et al. The MetaCyc database of metabolic pathways and enzymes. *Nucleic Acids Res* 2018;**46**:D633–9. <https://doi.org/10.1093/nar/gkx935>.
27. Wishart DS, Feunang YD, Guo AC, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic*

- Acids Res* 2018;**46**:D1074–82. <https://doi.org/10.1093/nar/gkx1037>.
28. RDKit. Open-Source Cheminformatics Software. <https://www.rdkit.org/>.
 29. Schwaller P, Hoover B, Reymond J-L, et al. Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. *Sci Adv* 2021;**7**:eabe4166. <https://doi.org/10.1126/sciadv.abe4166>.
 30. Theory D. SMARTS - A Language for Describing Molecular Patterns. <https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html>.
 31. Klein G, Kim Y, Deng Y et al. OpenNMT: Open-Source Toolkit for Neural Machine Translation *Proceedings of ACL 2017, System Demonstrations* 2017.
 32. Tetko IV, Karpov P, Van Deursen R, et al. State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis. *Nat Commun* 2020;**11**:1–11. <https://doi.org/10.1038/s41467-020-19266-y>.
 33. Freitag M, Al-Onaizan Y. Beam Search Strategies for Neural Machine Translation. *Proceedings of the First Workshop on Neural Machine Translation* 2017.
 34. Dong X, Yu Z, Cao W, et al. A survey on ensemble learning. *Front Comp Sci* 2020;**14**:241–58. <https://doi.org/10.1007/s11704-019-8208-z>.
 35. Schwaller P, Probst D, Vaucher AC, et al. Mapping the space of chemical reactions using attention-based neural networks. *Nat Mach Intell* 2021;**3**:144–52. <https://doi.org/10.1038/s42256-020-00284-w>.
 36. David OJ, Kovarik JM, Schmouder RL. Clinical pharmacokinetics of fingolimod. *Clin Pharmacokinet* 2012;**51**:15–28. <https://doi.org/10.2165/11596550-000000000-00000>.
 37. Gelotte CK, Zimmerman BA. Pharmacokinetics, safety, and cardiovascular tolerability of phenylephrine HCl 10, 20, and 30 mg after a single oral administration in healthy volunteers. *Clin Drug Investig* 2015;**35**:547–58. <https://doi.org/10.1007/s40261-015-0311-9>.
 38. Hardcastle J, Wilkins J. The action of sennosides and related compounds on human colon and rectum. *Gut* 1970;**11**:1038–42. <https://doi.org/10.1136/gut.11.12.1038>.
 39. Farid NA, Kurihara A, Wrighton SA. Metabolism and disposition of the thienopyridine antiplatelet drugs ticlopidine, clopidogrel, and prasugrel in humans. *J Clin Pharmacol* 2010;**50**:126–42. <https://doi.org/10.1177/0091270009343005>.
 40. Argikar UA, Gomez J, Ung D, et al. Identification of novel metoclopramide metabolites in humans: in vitro and in vivo studies. *Drug Metab Dispos* 2010;**38**:1295–307. <https://doi.org/10.1124/dmd.110.033357>.
 41. Guengerich FP, Sohl CD, Chowdhury G. Multi-step oxidations catalyzed by cytochrome P450 enzymes: processive vs. distributive kinetics and the issue of carbonyl oxidation in chemical mechanisms. *Arch Biochem Biophys* 2011;**507**:126–34. <https://doi.org/10.1016/j.abb.2010.08.017>.
 42. Lorenc-Koci E, Czarnecka A, Lenda T, et al. Molsidomine, a nitric oxide donor, modulates rotational behavior and monoamine metabolism in 6-OHDA lesioned rats treated chronically with L-DOPA. *Neurochem Int* 2013;**63**:790–804. <https://doi.org/10.1016/j.neuint.2013.09.021>.
 43. Sager JE, Choiniere JR, Chang J, et al. Identification and structural characterization of three new metabolites of bupropion in humans. *ACS Med Chem Lett* 2016;**7**:791–6. <https://doi.org/10.1021/acsmedchemlett.6b00189>.