Published in: *Expert Opinion on Drug Metabolism & Toxicology*

Machine learning models in the prediction of drug metabolism: Challenges and future perspectives

Keywords: drug metabolism, drug metabolites, sites of metabolism, enzymes, machine learning, deep learning

1 Introduction

Metabolism can be the underlying cause of drug adverse effects and diminished efficacy. Metabolic reactions in the human body, mediated mainly by enzymes, may transform the administered drug into metabolites that exhibit different biological activity [1,2]. As a general rule, metabolic reactions deactivate a drug, however, off-target effects or toxicity, resulting from the formed metabolites, cannot be excluded. On the flip side, metabolism is necessary for the formation of the active substance in the case of prodrugs. In scenarios where multiple drugs are co-administered, the presence of a drug may inhibit or further induce the clearance of another setting metabolism as one of the underlying causes of drug-drug interactions. As a result, the metabolic fate of a candidate drug needs to be thoroughly investigated during the drug development process.

Aiming at an accelerated and less resource demanding evaluation than what in vitro studies offer, multiple computational tools for drug metabolism prediction have been developed. Such tools are covering various aspects of drug metabolism, with a focus on the prediction of enzyme specificity, sites of metabolism and metabolites [3,4]. Tools for enzyme specificity attempt to identify the enzymes for which a candidate drug may act as a substrate, inhibitor or inducer. This information is important for assessing metabolic stability and, also, identifying potential drug-drug interactions. Sites of metabolism (SoMs) are atoms in the molecule where the metabolic transformation takes place. The identification of the SoMs for a molecule gives insights on its metabolic stability and also helps medicinal chemists determine structural modifications to manipulate its metabolism. Finally, the identification of potential metabolites is important for predicting possible adverse effects due to off-target activity or toxicity. Early computational approaches rely on expert knowledge. Molecular docking, QSAR modeling, molecular interaction fields and quantum mechanical simulations have been used to model interactions between enzymes and molecules and predict regioselectivity [4]. Regarding the prediction of potential metabolites, rule-based methods have been the main approach where a set of transformation rules, that encode the action of enzymes in general reaction patterns, is used in order to infer the molecular structures of potential metabolites.

As more data on drug metabolism are becoming available, there has been a shift to machine learning (ML) models which offer much faster inference [3,4]. The standard approach for predicting enzyme specificity is to apply a classifier on a vector representation of the molecule for distinguishing between binders and non-binders. Molecular fingerprints as well as other molecular descriptors, indicating physicochemical and structural properties, are widely used for obtaining vector representations for chemical molecules. ML classification models such as Support Vector Machines and Random Forests, as well as shallow Neural Networks have been the main choices. The classifiers are enzyme-specific, that is, each classifier is trained to predict interactions for a specific enzyme. A similar approach is followed for predicting SoMs where the descriptors contain atom-level attributes, and the classifier predicts the probability for a given atom to be a SoM for a specific enzyme. Regarding the prediction of metabolites, ML models have been used alongside the rule-based methods in order to reduce false positives filtering out unlikely predictions. These models are used to either predict substrate specificity or SoMs prior to the application of the transformation rules.

2 Expert Opinion

In the following, we are discussing the challenges and current trends in the development of ML models for the prediction of drug metabolism and we are giving directions for future developments.

2.1 Extending enzyme coverage: Existing tools are focused on the CYP450 enzyme family which is known to metabolize the big majority of existing drugs in phase I metabolism [4]. There is increasing interest though, for extending coverage to phase II enzymes and, also, other metabolic reactions, including gut microbiome metabolism, as complications may arise from various enzymes [2]. Especially as the repertoire of therapeutic agents is expanding from small molecules to biologics an extended enzyme coverage is deemed crucial [5]. However, the current approach of developing enzyme-specific models cannot be applied to enzymes with limited experimental data, that is, for the majority of human enzymes. Therefore, extending enzyme coverage calls for a steer from enzyme-specific models. Training a single model on data from multiple enzymes could not only facilitate predictions for enzymes with limited experimental data but also allow the model to identify shared reactivity patterns between enzymes or enzyme promisquity patterns. The most recent ML approaches for metabolite prediction are indeed trained on data covering human metabolism in its entirety including metabolism of endogenous compounds [6,7].

2.2 Leveraging Deep Learning: Existing ML approaches for drug metabolism prediction are mainly based on shallow ML models and mostly classification models for distinguishing enzyme binders from non-binders, and SoMs from non-SoMs. In these approaches, the selection of the molecule or atom level descriptors seems to be especially critical for the performance of the models and possibly even more critical than the model architecture. On the other hand, deep learning (DL) models can directly operate on structured and semi-structured data, including molecules represented either as molecular graphs or as SMILES sequences, and they are intended to learn task-specific representations. More importantly though, DL models have greater expressiveness as they can facilitate prediction tasks that call for methods that can output molecular structures on top of the standard classification and regression tasks. Graph

Convolutional Neural Networks have attracted a lot of attention, within the field of chemoinformatics, for learning molecule or atom level descriptors while language models have been a suitable choice for outputting molecules or even generating de-novo molecules using line notations such as SMILES [8, 9,10,15]. Although there is controversy on whether the learnt representations can actually offer an advantage [8], the ability to obtain descriptors that are optimized for a given prediction task is certainly appealing while the generation of molecules through language models opens the horizon for more complicated prediction tasks. Indeed, DL models have been applied on general chemical reactions with great success on various prediction tasks such as reaction outcome, reaction conditions, reaction center, reaction atom mapping [10,9,11,12]. The impetus behind these advancements has been the availability of massive datasets on chemical reactions which are challenging to obtain in the metabolism realm. DL though could be the response to the deviation from enzyme-specific models and facilitate greater enzyme coverage [6,7].

2.3 Dealing with small datasets: It turns out that the lack of large datasets is not prohibitive to the application of DL architectures on drug metabolism prediction. Various techniques are being developed from the ML community to tackle the exact same problem including transfer learning, few-shot learning as well as more traditional techniques such as ensembling. In particular, transfer learning seems an intuitive choice since metabolic reactions are a subset of all possible chemical reactions for which massive datasets are available and various applications have been already explored with promising results. Indeed, recent studies have demonstrated that a DL model, pre-trained on general chemical reactions, can be further tuned on metabolic data for predicting the outcome of metabolic reactions or even drug metabolites [13, 6]. Similarly, the tasks of SoMs and enzyme substrate prediction could also benefit from transfer learning starting from tasks such as reaction center or atom mapping prediction, and reaction condition prediction, respectively.

2.4 Introducing explainability: Existing ML-based approaches fail to give insights on how or why a prediction is made as opposed to the traditional knowledge-based computational tools which offer more transparency. This often averts medicinal chemists from using black-box models. Besides boosting confidence on the model predictions, explainability can additionally provide the user with actionable information. For example, Graph Convolutional Neural Networks may offer insights on how a molecule interacts with an enzyme as opposed to simply classifying a molecule as a substrate or non-substrate [14]. Similarly, the internal mechanisms of language models may reveal valuable information such as the reaction mechanism on top of predicting the reaction outcome [12].

2.5 Other data-related challenges: An additional challenge, besides the size of the available datasets, is incomplete information or inconsistencies across different sources. Regarding the first case, a major problem is the lack of negative training instances with regards to enzyme interactions. Lack of information for specific enzymes may mean that there is no interaction or that the interaction has not been studied. For the task of predicting SoMs, there is the additional problem of inconsistent labeling of SoMs across different sources due to different definitions. Inconsistencies regard the metabolites structures as well with some sources having more extensive lists of metabolites than others while often the primary metabolites are not distinguished from the secondary. All these issues hamper not only the development of ML models for drug metabolism prediction but also the evaluation of the methods and comparative assessment. Besides improving the data collection processes, ML models as well as model evaluation should also account for such uncertainties.

2.6 Integrating drug metabolism prediction models in the drug design pipeline: ML models offer fast inference and therefore open up the possibility of integrating metabolic studies in the early stages of drug development. Combined with ML models that perform complementary evaluations, such as prediction of biological activity or toxicity of the predicted metabolites, they can reveal complications that may arise in the later stages enabling a more efficient pipeline. Taking one step further, ML models for drug metabolism prediction can serve a fully automated drug development pipeline where target-specific molecules are generated using generative models and ranked using a set of ML models that evaluate different criteria, including metabolic stability and safety of metabolites [15].

References:

1 Croom E. Metabolism of xenobiotics of human environments. Progr Mol Biol Transl Sci., 2012; 112:31–88.

2 Testa B., Pedretti A., and Vistoli G. Reactions and enzymes in the metabolism of drugs and other xenobiotics. Drug Discov Today, 2012;17(11):549–560.

3 Kazmi S.R., Jun R., Yu M.S., Jung C., Na D. In silico approaches and tools for the prediction of drug metabolism and fate: A review. Comput. Biol. Med., 2019;106:54–64.

4 Tyzack J.D. Kirchmair J. Computational methods and tools to predict cytochrome P450 metabolism for drug discovery. Chem Biol Drug Des., 2019;4:377–386.

5 Hamuro L.L. Kishnani N.S. Metabolism of biologics: biotherapeutic proteins. Bioanalysis, 2012;4(2):189–195.

6 Litsa E.E., Das P., Kavraki L.E. Prediction of drug metabolites using neural machine translation. Chem. Sci., 2020;11:12777–12788.

7 Wang D., Liu W., Shen Z., Jiang L., Wang J., Li S., Li H. Deep learning based drug metabolites prediction. Front. Pharmacol., 2020;10:1586.

8 Jiang D., Wu Z., Hsieh C.-Y., Chen G., Liao B., Wang Z., Shen C., Cao D., Wu J., Hou T. Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. J. Cheminform., 2021;13.

9 Schwaller P., Laino T., Gaudin T., Bolgar P., Hunter C.A., Bekas C., Lee A.A.. Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. ACS Cent. Sci., 2019;5(9):1572–1583.

10 Coley C.W., Jin W., Rogers L., Jamison T.F., Jaakkola T.S., Green W.H., Barzilay R., Jensen K.F. A graph-convolutional neural network model for the prediction of chemical reactivity. Chem. Sci., 2019;10:370–377.

11 Gao H., Struble T.J., Coley C.W., Wang Y., Green W.H., Jensen K.F. Using machine learning to predict suitable conditions for organic reactions. ACS Cent. Sci., 2018; 4(11):1465–1476.

12 Schwaller P., Hoover B., Reymond J-L., Strobelt H., Laino T. Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. Sci. Adv., 2021; 7(15).

13 Kreutter D., Schwaller P., Reymond J.L. Predicting enzymatic reactions with a molecular transformer. Chem. Sci., 2021;12:8648–8659.

14 Jiménez-Luna J., Grisoni F., Schneider, G. Drug discovery with explainable artificial intelligence. Nature Machine Intelligence, 2020;2:573–584.

15 Chenthamarakshan V., Das P., Hoffman S., Strobelt H., Padhi I., Wai Lim K., Hoover B., Manica M., Born J., Laino T., Mojsilovic A. Cogmol: Target-specific and selective drug design for covid-19 using deep generative models. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, Advances in NeurIPS, 2020;33:4320–4332.