Metabolite Translator: Prediction of drug metabolites using a Transformer model

Eleni Litsa[1], Payel Das[2,3], Lydia Kavraki[1]
[1] Department of Computer Science, Rice University, Houston, TX 77005, USA
[2] IBM Research AI, IBM Thomas J. Watson Research Center, NY 10598, USA
[3] Applied Physics and Applied Mathematics, Columbia University, NY 10027, USA

Drugs may be metabolised in the human body, through enzymatic reactions, forming metabolites which can lead to complications such as toxicity. Computational tools that enable the prediction of drug metabolites can greatly benefit safety studies in the drug development process. Existing approaches [1] are rule-based methods with the rules encoding the action of specific enzymes. Such rules are extracted from reaction datasets involving manual work from experts and therefore raising scalability issues. Additionally, these methods lack generalisation since a rule is applied only when there is an exact match between the rule pattern and the query molecule. We present Metabolite Translator: a rule-free, end-to-end learning based method for predicting drug metabolites. We approached the problem of predicting drug metabolites as a sequence translation problem where molecules are represented using a sequence-based representation called SMILES. Metabolite Translator is built upon the Transformer model [2] for Neural Machine Translation. For training the model, we constructed a dataset by collecting human metabolism data from openly accessible databases. Our dataset includes metabolism of drugs but also endogenous compounds in an effort to develop a model that captures human metabolism in general without restricting to specific enzyme families. Due to the limited amount of human metabolism data, we used Transfer Learning: we pretrained a Transformer model on general chemical reactions and subsequently fine-tuned it on human metabolic reactions. Additionally, we created an ensemble model by combining the outputs of multiple models trained with different model hyper-parameters to account for multiple and diverse metabolites. We evaluated our method on a set of 65 drugs and compared it against 3 existing rule-based tools [3,4,5]. Our evaluation showed that our method, which was trained on general human reactions, had comparable performance with existing tools that were specifically developed for drugs. Additionally our method was able to predict metabolites through enzymes which are not commonly involved in drug metabolism and were missed by existing methods. The presented methodology provides a more scalable approach that can leverage larger datasets of chemical reactions and therefore obtain better generalisability and coverage of human metabolism which in turn can facilitate a more comprehensive study of drug metabolism in the drug development process.

References:

[1] 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., 106: 54-64, 2019.

[2] Vaswani A. Shazeer N., Parmar N., Uszkoreit, J., Jones L. , Gomez A. N., Kaiser L., Polosukhin I., Attention is all you need, Advances in Neural Information Processing Systems 30, pages 5998–6008. Curran Associates, Inc., 2017.

[3] de Bruyn Kops C., Šícho M. , Mazzolari A. , Kirchmair J. , GLORYx: Prediction of the Metabolites Resulting from Phase 1 and Phase 2 Biotransformations of Xenobiotics,
Chem. Res. Toxicol., Aug 2020.

[4] Ridder, L., Wagener, M., SyGMa: Combining Expert Knowledge and Empirical Scoring in the Prediction of Metabolites. ChemMedChem, 3: 821-832, 2008.

[5] 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 Cheminform* 11, 2019.