Predicting Reactivity to Drug Metabolism: Beyond CYPs

Mario Öeren, Peter A. Hunt, Matthew D. Segall
Optibrium Ltd., Cambridge, UK, info@optibrium.com

Introduction

Predicting sites of metabolism (SoM) enables chemists to be more efficient in optimising the structure of new chemical entities and helps them to identify potentially toxic metabolites early in a project. Historically, predictive models have focused on human isoforms of the Cytochrome P450 (CYP) family of enzymes due to their primary importance in the metabolism of drug-like compounds. However, predictive models for other enzymes, *e.g.,* **Aldehyde Oxidases** (AO), **Flavin-containing Monooxygenases** (FMO), and **Uridine 5'-diphospho-glucuronosyl-transferases** (UGT), are increasing in prevalence.[1,2] Here, we present models that predict the regioselectivity of metabolism for isoforms relevant to the metabolism of drug-like compounds in humans: AO1, FMO1, FMO3, UGT1A1, UGT1A4, UGT1A9, and UGT2B7.

Reactivity-Accessibility Models

Our approach combines a mechanistic element to estimate the **reactivity** of potential sites of metabolism with a machine learning model to capture steric and orientation effects (**accessibility**) within the active site. The reactivity of a potential SoM is described using quantum mechanical calculations that estimate the activation energy (E_a) of product formation. The accessibility descriptors capture distances from the potential SoM to specified functional groups (*e.g.*, acidic and basic groups) as counts of bonds. The reactivity and accessibility descriptors for each potential SoM are then associated with the **data from the experiments**, to build quantitative structure-activity relationship models.

Experimental Data

The data is curated from public sources that provide detailed information on the experimentally observed SoM. Since the models are intended to distinguish the observed SoM from all potential SoM, the molecules included in the datasets have two or more potential SoM, out of which at least one is experimentally observed to be metabolised. Each potential SoM on a molecule was labelled as either experimentally observed or not by the corresponding isoform.

Enzyme	Isoform	No. of Substrates	No. of Potential SoM	No. of Potential SoM Metabolised
AO	AO1	157	865	160
FMO	FMO1	56	172	56
	FMO3	67	209	69
UGT	UGT1A1	98	297	146
	UGT1A4	54	146	66
	UGT1A9	137	390	187
	UGT2B7	90	223	115

Studies Using Density Functional Theory (DFT)

For DFT calculations, the simulated systems must be small, but retain their chemical characteristics. We tested a series of simplifications for AO, FMO[2], and UGT[2] enzymes to ensure that the reactivity of the reaction centres was not significantly modified. The substrate structures were not simplified, to ensure that long-range effects within the compounds were considered. The results were validated using experimental data on site-specific rates of metabolism.

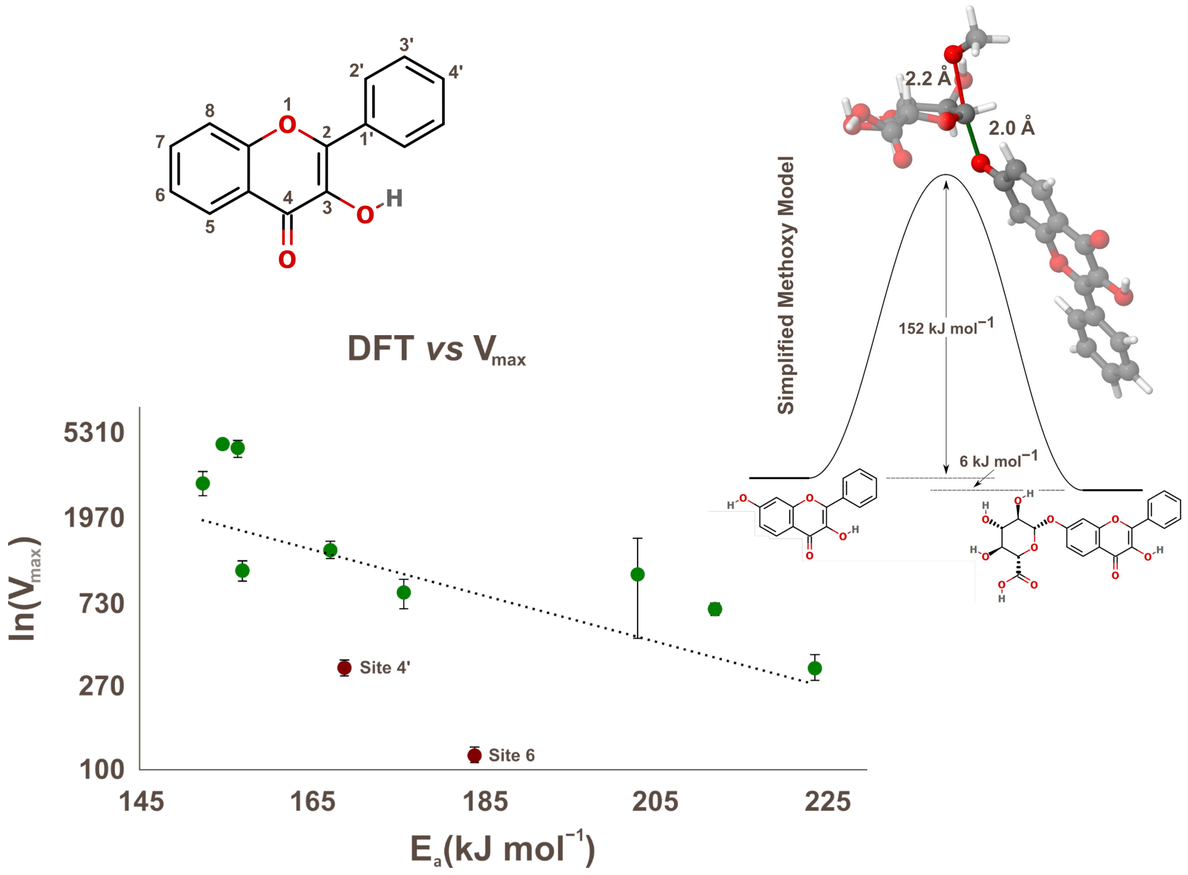

Figure 1. The simplified transition state for UGT and the correlation between calculated activation energy and the reaction rate (B3LYP/SVP).

Studies Using Semi-empirical Methods

DFT calculations can take hours to days. Therefore, to calculate reactivity within a reasonable timeframe, we use semi-empirical methods, reducing the calculation time to minutes. However, semi-empirical methods are known to introduce systematic errors depending on the environment of the SoM, which must be corrected to achieve accurate predictions.

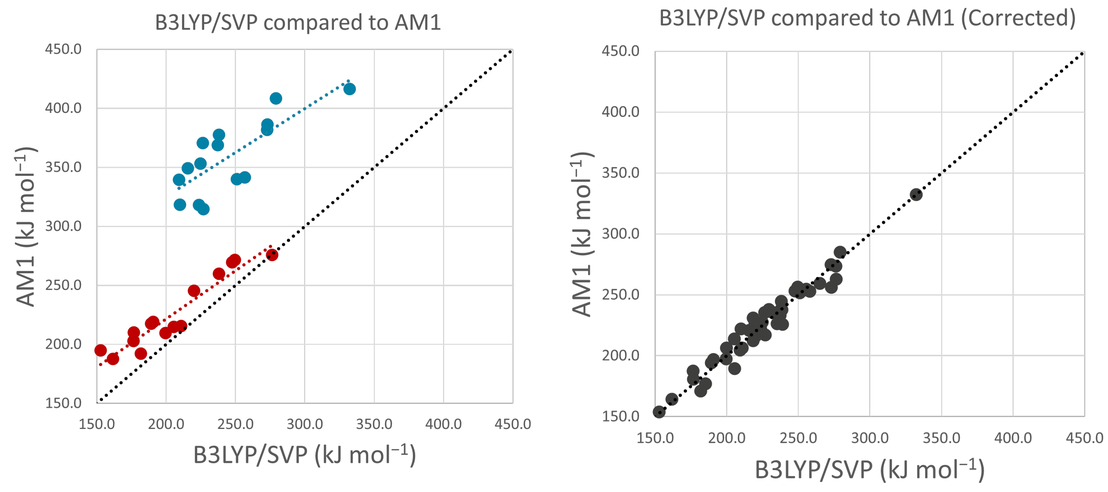

Figure 2. The correlation of E_a between B3LYP/SVP and AM1. The blue points refer to *N*-glucuronidation (primary, secondary, and tertiary amines), the red points refer to *O*-glucuronidation (alcohols, enols, and phenols), and the black points refer to corrected E_a for all *N*- and *O*-glucuronidation types.

QSAR Models

For small data sets, the data for each isoform was split into training and test sets (80:20). For larger data sets, the data was split into training, validation and test sets (70:15:15). The split was made randomly by compound; thus, all potential SoM of one substrate were in the same subset set. The models were trained using Gaussian Processes (GP) method in StarDrop™.

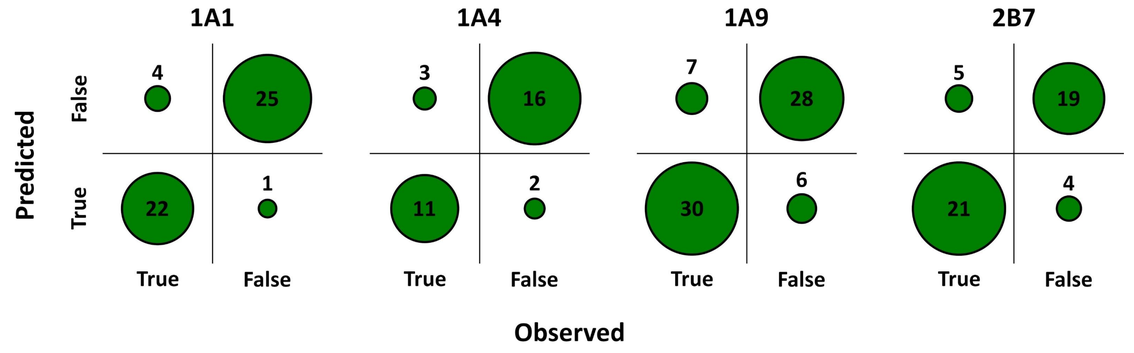

Figure 2. The confusion matrices for the independent test sets of reactivity-accessibility models for four UGT isoforms.

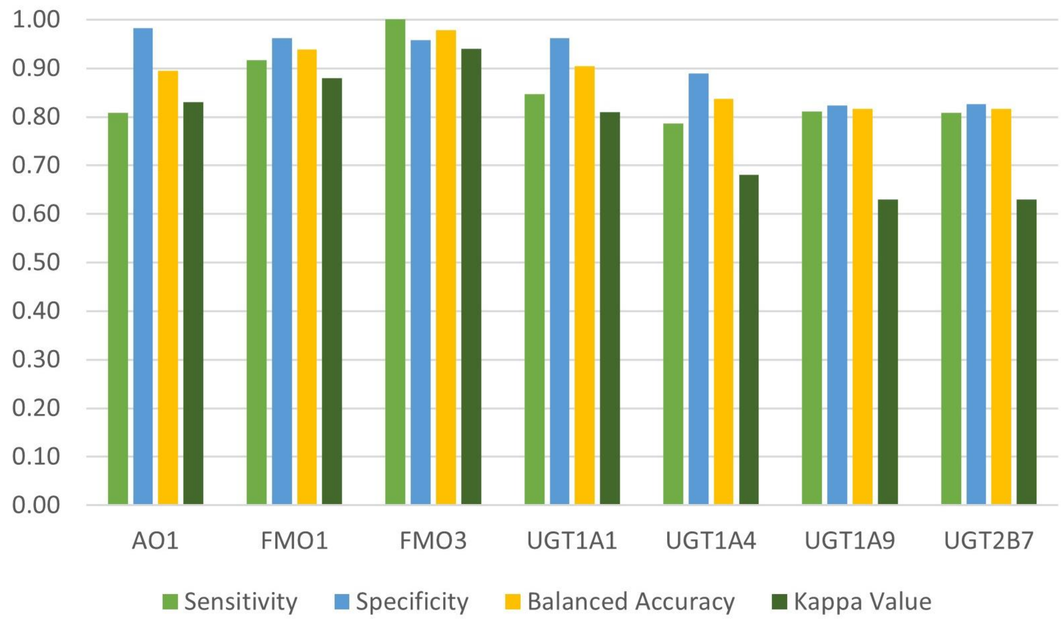

Figure 3. The sensitivity, specificity, balanced accuracy, and kappa values of the trained models on independent test sets.

Conclusions

The presented work adds **seven novel models**, which predict the regioselectivity of metabolism for relevant enzyme families and isoforms for metabolism of drug-like compounds. The models show **excellent performance** for the prediction of the primary SoM. In combination with the existing CYP models we can cover the majority of observed metabolic pathways.

References

[1] J. D. Tyzack, P. A. Hunt and M. D. Segall, "Predicting Regioselectivity and Lability of Cytochrome P450 Metabolism Using Quantum Mechanical Simulations," Journal of Chemical Information and Modeling, vol. 56, no. 11, pp. 2180-2193, 2016.

[2] M. Öeren, P. J. Walton, P. A. Hunt, D. J. Ponting and M. D. Segall, "Predicting Reactivity to Drug Metabolism: Beyond P450s—Modelling FMOs and UGTs," Journal of Computer-Aided Molecular Design, vol. 35, pp. 541-555, 2020.