Prediction of UGT-mediated Metabolism Using the Manually Curated MetaQSAR Database

Angelica Mazzolari,^{*,†,∇} Avid M. Afzal,^{‡,§,∇} Alessandro Pedretti,^{†,‡} Bernard Testa,^{||,‡} Giulio Vistoli,^{†,‡} and Andreas Bender^{‡,‡}

[†]Dipartimento di Scienze Farmaceutiche, Facoltà di Scienze del Farmaco, Università degli Studi di Milano, Via Mangiagalli, I-20133 Milano, Italy

[‡]Centre for Molecular Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, CB2 1EW Cambridge, U.K.

^{||}Lausanne University, CH-1015 Lausanne, Switzerland

S Supporting Information

ABSTRACT: Even though glucuronidations are the most frequent metabolic reactions of conjugation, both in quantitative and qualitative terms, they have rather seldom been investigated using computational approaches. To fill this gap, we have used the manually collected MetaQSAR metabolic reaction database to generate two models for the prediction of UGT-mediated metabolism, both based on molecular descriptors and implementing the Random Forest algorithm. The first model predicts the occurrence of the reaction and was internally validated with a Matthew correlation coefficient (MCC) of 0.76 and an area under the ROC curve (AUC) of 0.94, and further externally validated using a test set composed of 120 additional xenobiotics (MCC of 0.70 and AUC of 0.90). The second model distinguishes between O- and N-glucuronidations and was optimized by the random undersampling procedure to improve the predictive accuracy during the internal validation, with the recall measure of the minority class increasing from 0.55 to 0.78.

KEYWORDS: Metabolism, predictive modeling, glucuronidation, UGT-mediated metabolism, machine learning, Random Forest

During the last decades, the field of metabolism prediction has gained a major importance in the context of drug discovery.¹ Reaching a satisfactory half-life as well as avoiding the onset of toxic effects are two crucial concerns during the lead generation step, which depend on both dose and metabolic fate of the parent compounds *in vivo*. Many issues related to metabolic liability can potentially lead to failures during drug development. For example, enhanced or reduced clearance levels can produce a rapid loss of pharmacological efficacy or a toxic accumulation of drugs, respectively.² Moreover, the metabolic profile of a given compound can influence the activity of other drugs due to drug–drug interactions,³ which may result in a reduction or exacerbation of therapeutic side effects.⁴

In this context, we approve the development of a wide variety of predictive methods, using both experimental and computational approaches,^{5–8} with the potential to combine measurements and predictions.⁹ While *in vitro* systems and *in vivo* approaches for the determination of metabolites have recently enjoyed significant improvements,^{10–12} computational tools offer the highest throughput at the lowest cost but provide predictions that are usually susceptible to experimental validation.¹³

One approach of metabolism prediction is the data-driven approach, where models are based on experimental data, and

metabolism prediction hence requires learning data sets that meet both quantitative and qualitative criteria. While the former are fulfilled by collecting large-enough learning sets from online freely or commercially available resources,¹⁴ the latter are more demanding. Indeed, the available resources, which are mostly automatically compiled, tend to suffer from sparseness, inaccuracies, or redundancies, and often they are not even specifically focused on xenobiotic metabolism.¹ However, as is well-known, cheminformatics studies base their reliability, above all, on the accuracy of the learning data sets,¹⁵ and moreover, only with accurate learning sets in hand, the attention paid to structural data, as well as to the choice of descriptors, may lead to substantial increases in the predictive power of models.¹⁶

This was one of the main concerns of the current study, which is based on data entirely collected by manual and critical analysis of the literature. Indeed, the models presented here are based on the MetaQSAR database, a metabolic reaction database developed by some of us and periodically updated.¹⁷ In the current version, the database contains 7962 molecules

Special Issue: Highlighting Medicinal Chemistry in Italy

Received: November 30, 2018

Accepted: February 12, 2019

Published: February 12, 2019

with 10 965 annotated metabolic reactions, classified into 101 subclasses. Each reaction underwent a set of expert checks, thereby preventing mistakes and inaccuracies typically associated with automatic compilation as well as avoiding mistakes found in the primary literature.

Metabolism prediction involves a wide array of computational approaches among which the so-called “local methods” apply to simple biological systems (single enzyme, single reaction).¹⁸ These approaches have been mainly applied to study the redox reaction focusing on the catalytic activity of cytochrome P450 enzymes (CYP450).^{19–24} In contrast, compared to phase I reactions, phase II metabolism has received much less attention, despite its important impact on the modulation of the pharmacological effects.²⁵

With the aim to fill this gap, we decided to focus our studies on glucuronidation reactions, which are the most important reactions in conjugation (phase II) metabolism, both qualitatively (considering the variety of functional groups that are targets in the reaction), as well as in terms of frequency.^{26,27} In quantitative terms, according to the meta-analysis by Testa et al., glucuronosyltransferases (UGTs) account for 14% of all collected metabolites, second in importance only to the occurrence of reactions catalyzed by cytochromes P450.²⁸ The conjugation with glucuronic acid increases the hydrophilic properties of the metabolite compared to the substrate, leading to important changes in the fate of the original drug. Hence, in terms of outcome, glucuronidations have an essential role in detoxification and can reduce the duration of drug action or markedly decrease its pharmacological effect.²⁹ For these reasons, modifying the propensity of a compound to be glucuronated is an important part of the lead optimization process.

In practical terms, the choice to study glucuronic acid metabolism is in line with another important requirement for model building, namely, the use of a data set which guarantees the best possible coverage of the specific metabolic process.³⁰ This need is satisfied considering that UGT metabolism is so relevant that one may suppose that almost all experimental collected studies (when not specifically focused on a single reaction) involve the detection of this type of metabolites. Moreover, the number of general inaccuracies is further reduced by using the MetaQSAR database, as already discussed. In addition, glucuronidations are conceivably so frequent that the number of reactive substrates is large enough to allow the development of statistically robust predictive models.

The goal of this work is hence to develop predictive models for UGT-mediated metabolism based on molecular properties of xenobiotics. Our primary aim is to predict whether a molecule is prone to transformation into a glucuronide (Model 1). The secondary objective is to distinguish between the two main classes of glucuronidation by predicting whether the conjugation occurs on an oxygen or a nitrogen atom (Model 2). This last achievement can help in identifying the site of metabolism on the substrate molecule when the latter features both oxygen (hydroxyl, phenol, carboxyl, hydroxylamine-hydroxylamide) and nitrogen (amino, amido, aza-heterocyclic) functions.

Model 1 was generated on a data set composed of 2192 molecules. Of these, 1400 are reported in MetaQSAR as able to undergo metabolic reactions other than glucuronidation, so these were classified as “UGT non-substrates” (class 0, negative), while 792 molecules are annotated as “UGT-

substrates” (class 1, positive). The database is not perfectly balanced with the positive class being less populated (Figure 1a). Inside class 1, it is possible to distinguish two groups of

Figure 1. Model data sets: relative proportions between the two classes undergoing binary classification modeling. Panel e shows the ratio of first-generation substrates (I-GEN SUBs) and second-or more generation substrates (II-GEN SUBs) of the Model 1 data set.

UGT-substrates. The first group is composed of xenobiotics that carry the electron-rich nucleophile groups undergoing the reaction and are classified as “first-generation substrates” (I-GEN SUBs), while the second group is composed of metabolites deriving from functionalization reactions that provide the required functional groups and are classified as “second-or more generation substrates” (II-GEN SUBs) (Figure 1e).

In order to characterize the chemical space of the collected data, we carried out two analyses for data exploration. The first was principal component analysis (PCA) based on 19 physicochemical and stereoelectronic properties. The second was molecular similarity analysis (MSA) based on Skelspheres descriptors as implemented in DataWarrior,³¹ placing similar molecules onto a 2D plane next to each other. The results of both analysis are shown in Figure 2. It can be seen that there is a sort of data clustering along the second component (as discussed in more detail in the Supporting Information, section 2), but neither in the PCA nor in the MSA plot do the resulting distributions highlight any pattern to separate negatives from I-GEN SUBs and II-GEN SUBs. These results emphasize the intrinsic difficulty of predicting the propensity to glucuronation of a given molecule by linear combinations of molecular properties as well as by structural similarity analyses.

For the above reasons, we next employed machine learning techniques for developing predictive models based on binary classification. Model 1 was validated by two methods for internal validation, namely, Monte Carlo cross validation (MCCV) and leave one out validation (LOO). The first method consists in a set of repeated cycles of training and testing, each time randomly splitting the data set to avoid biases deriving from the selection of a particular test set, while the second method, also known as “rotation estimation”, involves the whole available data set, since it iteratively trains the model on the $n - 1$ samples and predicts the sample excluded from training. Moreover, the model was further validated on an external test set. All of the validation methods were implemented optimizing hyperparameters by a prior nested 5-fold cross validation.

Random Forest³² (RF) was found to be the best performing algorithm, as was observed in the model assessment methods. The Model 1 MCCV box plot shows the predictive power measured on the 30% test set randomly split for 100 runs

Figure 2. Principal component analysis (a,b) and molecular similarity analysis based on Skelspheres descriptors (c) of the Model 1 data set. The data set is composed of 1400 UGT non-substrates and 792 UGT-substrates comprising 400 first-generation substrates (I-GEN SUBs) and 392 second-or more generation substrates (II-GEN SUBs). See the [Supporting Information](#) for a more detailed description of the images.

Figure 3. MCCV performance for Model 1, Model 2, and random undersampled Model 2 based on 100 iterations of model generation and evaluation. The performance for the single class is measured in terms of Precision, Recall, and F1 score, while the overall performance is measured in terms of Matthews Correlation Coefficient (MCC) and Area Under ROC Curve (AUC).

Table 1. Model Performance for Model 1, Model 2, and Undersampled Model 2 Measured in Terms of Precision and Recall for the Single Class, and in Terms of Matthews Correlation Coefficient (MCC) and Area Under the ROC Curve (AUC) for the Two Classes Together

	MCCV Model 1		LOO validation Model 1		external test set Model 1		MCCV Model 2		MCCV undersampled Model 2	
	class 0	class 1	class 0	class 1	class 0	class 1	class 0	class 1	class 0	class 1
precision	0.88	0.89	0.89	0.89	0.83	0.87	0.93	0.85	0.80	0.84
recall	0.95	0.76	0.94	0.79	0.86	0.84	0.98	0.55	0.85	0.78
MCC	0.76		0.75		0.70		0.65		0.64	
AUC	0.93		0.94		0.90		0.92		0.90	

(Figure 3a). The size of the interquartile range parallels the variability of the performances, due to the high diversity inside the data set. The average values of each predictive measure are reported in Table 1. The high value for specificity, of 0.94,

shows a low number of false positives, which are usually the weakness of other predictive models.⁸ The sensitivity shows a still generally acceptable value of 0.76, and this is expected because the minority class of an unbalanced data set (here the

UGT-substrates) is typically affected by a higher rate of wrong predictions. The global predictive power, as measured by a Matthews correlation coefficient (MCC) of 0.76 and an area under the ROC curve of 0.93, also indicates a good quality model.

The Model 1 LOO validation results are very similar to the above numbers (Figure 3b and Table 1). This shows, on the one hand, that the model is robust enough to maintain the same performances even when tested on the whole available data, and on the other hand, this emphasizes that the LOO validation gave rise to an accurate prediction error and did not lead to overfitting.³³ The I-GEN SUBs are responsible for the large majority of the false negatives, affecting the sensitivity measure (0.55), while the II-GEN SUBs are almost perfectly predicted, reaching a true positive rate of 0.97. This remarkable result can be explained by considering that the II-GEN SUBs are in fact molecules that are purposely optimized to become UGT-substrates by specific functionalization reactions. This renders this set of molecules more homogeneous, helping in their correct classification.

A further confirmation of the different ability in predicting the two subclasses of UGT-substrates was obtained by analyzing the probability measure that the RF algorithm assigns to each prediction. The higher this probability, the more confident the prediction. The results are shown in Figure 4, and it can be seen that II-GEN SUBs display probability values higher than those of I-GEN SUBs, meaning that the model predicts II-GEN SUBs with higher certainty.

Figure 4. Distribution of the probability measures assigned by the RF algorithm during Model 1 LOO validation. The analysis involves the probabilities assigned only to the retrieved true positives and true negatives.

Finally, as the last step of the internal validation of the model, we performed an applicability domain (AD) study to define the space inside which Model 1 yields reliable predictions.³⁴ To this aim, the predictive power of the model was correlated to the 3D-structural similarity among molecules, obtaining a clear trend of inverse proportionality between these two measures (see Supporting Information, section 4, Figure S1).

We next validated Model 1 on an external test set. The collection of the external test set was done by searching suitable publications published during the year 2016, which were not included in the training set, within the same journals used as for the compilation of MetaQSAR (i.e., *Chemical Research in Toxicology*, *Drug Metabolism and Disposition*, and

Xenobiotica). The resulting test set consists of 120 molecules with a set of properties inside the space defined by the previous AD analysis. The test set is almost balanced (Figure 1b), and as the consequence, specificity and sensitivity are close to each other, with values of 0.86 and 0.84, respectively. The global predictive performances are only slightly lower than those shown by the internal validation (Table 1).

Overall, despite the relatively small size of the data sets generally, model performance has been found to be consistent between all different model validation methods.

We next performed an analysis of model feature importance when training Model 1 on its data set and taking into consideration the descriptors resulting by the feature selection phase of the model. The modeled descriptors were classified into four main categories, namely, “constitutive”, “geometric”, “physicochemical”, and “electronic”. The distribution of the descriptors in those categories exhibits a marked prevalence of constitutive descriptors (67%), while the physicochemical properties comprise just the 3% of the total number (Figure 5a). The

Figure 5. Analysis of the feature importance in Model 1. Distribution of the number of features (a), distribution of the category importance (b), and distribution of the category importance weighted on the total number of features in each category (c).

distribution of the category importance, which was obtained by summing the importance of the features in each category, reveals a decrease of relevance for the constitutive descriptors (28%) and an increase for the geometric (26%) and even more for the electronic descriptors (40%), meaning that the last ones are the descriptors that give the greatest contribution to correct predictions (Figure 5b). Putting then those numbers into context of the total number of descriptors in each category, we observe a further decrease in importance for the constitutive descriptor class (to 6% of the total number), and an important gain for the physicochemical descriptors (to 28%), pointing out the crucial role played also by this category of features (Figure 5c).

We also trained a second model, Model 2, which is meant to be used after Model 1 in the case this one predicts the query molecule as belonging to the “UGT-substrates” class to further predict whether the reaction occurs on an oxygen or on a nitrogen atom. This model was generated on a selection of the “positive” samples of Model 1 that was obtained by discarding those molecules that undergo the glucuronidation on both an oxygen and a nitrogen atom. This data set is composed of 775

molecules, of which 661 are substrates of O-glucuronidation (class 0) and 114 are substrates of N-glucuronidation (class 1). The glucuronidations occurring on a nitrogen atom are less frequent in nature than those occurring on oxygen by a 6:1 ratio (Figure 1c). The model was generated similarly to Model 1 by applying the RF algorithm and by internally evaluating the predictive power by MCCV and LOO validations.

Both methods for internal validation exhibited comparable predictive performances and revealed modest capability to correctly predict the substrates of N-glucuronidation, as the consequence of the unbalanced starting data set. The MCCV box plot in Figure 3b shows predictive performance for the minority class that is close to random prediction, as indicated by the recall value of 0.55 (Table 1). We assumed that class imbalance represents a problem for this data set, and hence, we employed undersampling of the majority class in the next step. The balanced data set composed of 228 molecules was obtained by randomly undersampling the class 0 to 114 molecules (Figure 1d). The relative MCCV box plot shows that class 1 recall indeed increased to 0.78. However, the global predictive performance did not improve (MCC = 0.64), indicating the commonly observed trade-off of optimizing model parameters (Figure 3c).

A closer analysis of the obtained results reveals only negligible physicochemical differences between the correctly and the incorrectly predicted substrates, a gratifying finding that suggests the absence of biasing effects in the reported predictive analyses. The only significant difference is seen in the ionization state of the incorrectly predicted substrates, which appears particularly enriched in negatively charged compounds. Indeed, we observed that when running Model 1 LOO validation for three times, the 97 constantly incorrectly predicted substrates include 21 negatively charged molecules (i.e., 22%), which is more than double the abundance in the entire database (10%). A complete list of these substrates is provided in the Supporting Information, Figure S4. In contrast, the presence of a positive charge does not affect the predictive performance since the abundance of basic protonated molecules in the incorrectly predicted substrates is in line with that occurring in the entire data set (~20%). Further analysis will clarify if this unsatisfactory performance for anionic molecules is ascribable to a poor physicochemical parametrization of the negatively charged groups in some used descriptors (e.g., log P) or if it is related to the biochemical features of the simulated metabolic reaction that introduces an anionic moiety.

In summary, the present work points out the possibility to predict the occurrence of glucuronidation reactions based on molecular descriptors, also distinguishing between metabolic reactions occurring on oxygen or nitrogen atoms. We are aware that the metabolic learning sets contain both information and noise, in particular the class of “negatives” since the “non-substrates” are mainly defined by exclusion from the class of “substrates”, without specific confirming evidence.³⁵ In light of this, the performance shown by our models underline the relevance of using suitably curated data sets when developing novel approaches for Phase II metabolism prediction and show a practically useful application of the MetaQSAR database.

Nevertheless, two main evidence arise from this study. First, the analysis of feature importance emphasizes the key role of both physicochemical and electronic descriptors for Phase II metabolic transformation, which encode for the two main steps into which the UGT-mediated reaction can be subdivided.

These are a first noncovalent recognition, which is governed by physicochemical features, and a second catalytic process, which is influenced by the reactivity of the substrate, as expressed by electronic descriptors. Second, the better performance obtained when predicting II-GEN SUBs suggests that the structural heterogeneity encoded by these molecules is lower than that defined by the I-GEN SUBs, and this helps in their correct prediction. This means that the functionalization (Phase I) reactions, besides increasing substrate polarity, have evolved to increase the propensity of a metabolite to undergo specific conjugation reactions by reducing the chemical space covered by metabolized xenobiotics.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acsmchemlett.8b00603.

Details with regard to the computational workflows and inherent Random Forest models, principal component analysis, molecular similarity analysis, and applicability domain study; instructions to use the Model 1 pickle file (PDF)

Model_1 (pickle) (ZIP)

■ AUTHOR INFORMATION

Corresponding Author

*Phone: +390250319330. E-mail: angelica.mazzolari@unimi.it.

ORCID

Angelica Mazzolari: 0000-0003-1352-1094

Alessandro Pedretti: 0000-0001-5916-2029

Giulio Vistoli: 0000-0002-3939-5172

Andreas Bender: 0000-0002-6683-7546

Present Address

[§]Quantitative Biology, Discovery Sciences, IMED Biotech Unit, AstraZeneca, Cambridge, United Kingdom.

Author Contributions

[▽]These authors contributed equally. The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

[#]B.T.: Emeritus Professor.

■ ABBREVIATIONS

AUC, area under the ROC curve; CYP, cytochrome P450; LOO, leave one out; MCC, Matthew Correlation Coefficient; MCCV, Monte Carlo cross validation; RF, Random Forest; UGT, UDP-glucuronosyltransferase; I-GEN SUBs, first-generation substrates; II-GEN SUBs, second or more generation substrates

■ REFERENCES

- (1) Kirchmair, J.; Williamson, M. J.; Tyzack, J. D.; Tan, L.; Bond, P. J.; Bender, A.; Glen, R. C. Computational Prediction of Metabolism: Sites, Products, SAR, P450 Enzyme Dynamics, and Mechanisms. *J. Chem. Inf. Model.* **2012**, 52 (3), 617–648.
- (2) Sun, H.; Scott, D. O. Structure-Based Drug Metabolism Predictions for Drug Design. *Chem. Biol. Drug Des.* **2010**, 75 (1), 3–17.

- (3) Almond, L.; Yang, J.; Jamei, M.; Tucker, G.; Rostami-Hodjegan, A. Towards a Quantitative Framework for the Prediction of DDIs Arising from Cytochrome P450 Induction. *Curr. Drug Metab.* **2009**, *10* (4), 420–432.
- (4) Cascorbi, I. Drug Interactions—Principles, Examples and Clinical Consequences. *Dtsch. Arztebl. Int.* **2012**, *109* (33–34), 546–555.
- (5) Caron, G.; Ermondi, G.; Testa, B. Predicting the Oxidative Metabolism of Statins: An Application of the MetaSite®. *Pharm. Res.* **2007**, *24* (3), 480–501.
- (6) Carlsson, L.; Spjuth, O.; Adams, S.; Glen, R. C.; Boyer, S. Use of Historic Metabolic Biotransformation Data as a Means of Anticipating Metabolic Sites Using MetaPrint2D and Bioclipse. *BMC Bioinf.* **2010**, *11* (1), 362.
- (7) Lapins, M.; Worachartcheewan, A.; Spjuth, O.; Georgiev, V.; Prachayasittikul, V.; Nantasenamat, C.; Wikberg, J. E. S. A Unified Protochemometric Model for Prediction of Inhibition of Cytochrome P450 Isoforms. *PLoS One* **2013**, *8* (6), No. e66566.
- (8) Tyzack, J. D.; Mussa, H. Y.; Williamson, M. J.; Kirchmair, J.; Glen, R. C. Cytochrome P450 Site of Metabolism Prediction from 2D Topological Fingerprints Using GPU Accelerated Probabilistic Classifiers. *J. Cheminf.* **2014**, *6*, 1–10.
- (9) Kirchmair, J.; Göller, A. H.; Lang, D.; Kunze, J.; Testa, B.; Wilson, I. D.; Glen, R. C.; Schneider, G. Predicting Drug Metabolism: Experiment and/or Computation? *Nat. Rev. Drug Discovery* **2015**, *14* (6), 387–404.
- (10) Scheer, N.; Wolf, C. R. Genetically Humanized Mouse Models of Drug Metabolizing Enzymes and Transporters and Their Applications. *Xenobiotica* **2014**, *44* (2), 96–108.
- (11) Kitamura, S.; Sugihara, K. Current Status of Prediction of Drug Disposition and Toxicity in Humans Using Chimeric Mice with Humanized Liver. *Xenobiotica* **2014**, *44* (2), 123–134.
- (12) Zhu, M.; Zhang, H.; Humphreys, W. G. Drug Metabolite Profiling and Identification by High-Resolution Mass Spectrometry. *J. Biol. Chem.* **2011**, *286* (29), 25419–25425.
- (13) Testa, B.; Balmat, A. L.; Long, A.; Judson, P. Predicting Drug Metabolism - An Evaluation of the Expert System METEOR. *Chem. Biodiversity* **2005**, *2* (7), 872–885.
- (14) Peach, M. L.; Zakharov, A. V.; Liu, R.; Pugliese, A.; Tawa, G.; Wallqvist, A.; Nicklaus, M. C. Computational Tools and Resources for Metabolism-Related Property Predictions. 1. Overview of Publicly Available (Free and Commercial) Databases and Software. *Future Med. Chem.* **2012**, *4* (15), 1907–1932.
- (15) Fourches, D.; Muratov, E.; Tropsha, A. Trust but Verify: On the Importance of Chemical Structure Curation in Chemoinformatics and QSAR Modeling Research. *J. Chem. Inf. Model.* **2010**, *50* (7), 1189–1204.
- (16) Zhu, H.; Tropsha, A.; Fourches, D.; Varnek, A.; Papa, E.; Gramatica, P.; Öberg, T.; Dao, P.; Cherkasov, A.; Tetko, I. V. Combinatorial QSAR Modeling of Chemical Toxicants Tested against *Tetrahymena Pyriformis*. *J. Chem. Inf. Model.* **2008**, *48* (4), 766–784.
- (17) Pedretti, A.; Mazzolari, A.; Vistoli, G.; Testa, B. MetaQSAR: An Integrated Database Engine to Manage and Analyze Metabolic Data. *J. Med. Chem.* **2018**, *61* (3), 1019–1030.
- (18) Testa, B.; Balmat, A.-L.; Long, A. Predicting Drug Metabolism: Concepts and Challenges. *Pure Appl. Chem.* **2004**, *76* (5), 907–914.
- (19) Cruciani, G.; Carosati, E.; De Boeck, B.; Ethirajulu, K.; Mackie, C.; Howe, T.; Vianello, R. MetaSite: Understanding Metabolism in Human Cytochromes from the Perspective of the Chemist. *J. Med. Chem.* **2005**, *48* (22), 6970–6979.
- (20) Zaretski, J.; Bergeron, C.; Rydberg, P.; Huang, T.; Bennett, K. P.; Breneman, C. M. RS-Predictor: A New Tool for Predicting Sites of Cytochrome P450-Mediated Metabolism Applied to CYP 3A4. *J. Chem. Inf. Model.* **2011**, *51* (7), 1667–1689.
- (21) Zaretski, J.; Rydberg, P.; Bergeron, C.; Bennett, K. P.; Olsen, L.; Breneman, C. M. RS-Predictor Models Augmented with SMARTCyp Reactivities: Robust Metabolic Regioselectivity Predictions for Nine CYP Isozymes. *J. Chem. Inf. Model.* **2012**, *52* (6), 1637–1659.
- (22) Šicho, M.; De Bruyn Kops, C.; Stork, C.; Svozil, D.; Kirchmair, J. FAME 2: Simple and Effective Machine Learning Model of Cytochrome P450 Regioselectivity. *J. Chem. Inf. Model.* **2017**, *57* (8), 1832–1846.
- (23) Zaretski, J.; Matlock, M.; Swamidass, S. J. XenoSite: Accurately Predicting CYP-Mediated Sites of Metabolism with Neural Networks. *J. Chem. Inf. Model.* **2013**, *53* (12), 3373–3383.
- (24) Rudik, A.; Dmitriev, A.; Lagunin, A.; Filimonov, D.; Poroikov, V. SOMP: Web Server for in Silico Prediction of Sites of Metabolism for Drug-like Compounds. *Bioinformatics* **2015**, *31* (12), 2046–2048.
- (25) Smith, P.; Sorich, M.; Low, L. S.; McKinnon, R.; Miners, J. Towards Integrated ADME Prediction: Past, Present and Future Directions for Modelling Metabolism by UDP-Glucuronosyltransferases. *J. Mol. Graphics Modell.* **2004**, *22* (6), 507–517.
- (26) Testa, B.; Kramer, S. *The Biochemistry of Drug Metabolism: Vol. 2: Conjugations, Consequences of Metabolism, Influencing Factors* (v. 2); Wiley-VCH, 2010.
- (27) Breton, C.; Šnajdrová, L.; Jeanneau, C.; Koča, J.; Imberty, A. Structures and Mechanisms of Glycosyltransferases. *Glycobiology* **2006**, *16* (2), 29–37.
- (28) Testa, B.; Pedretti, A.; Vistoli, G. Reactions and Enzymes in the Metabolism of Drugs and Other Xenobiotics. *Drug Discovery Today* **2012**, *17* (11–12), 549–560.
- (29) Miners, J. O.; Smith, P. A.; Sorich, M. J.; McKinnon, R. A.; Mackenzie, P. I. PREDICTING HUMAN DRUG GLUCURONIDATION PARAMETERS: Application of In Vitro and In Silico Modeling Approaches. *Annu. Rev. Pharmacol. Toxicol.* **2004**, *44*, 1–25.
- (30) Kirchmair, J.; Williamson, M. J.; Afzal, A. M.; Tyzack, J. D.; Choy, A. P. K.; Howlett, A.; Rydberg, P.; Glen, R. C. FAME Metabolizer (FAME): A Rapid and Accurate Predictor of Sites of Metabolism in Multiple Species by Endogenous Enzymes. *J. Chem. Inf. Model.* **2013**, *53* (11), 2896–2907.
- (31) Sander, T.; Freyss, J.; von Korff, M.; Rufener, C. DataWarrior: An Open-Source Program For Chemistry Aware Data Visualization And Analysis. *J. Chem. Inf. Model.* **2015**, *55* (2), 460–473.
- (32) Liaw, A.; Wiener, M. Classification and Regression by RandomForest. *R News* **2002**, *2* (3), 18–22.
- (33) Xu, Q. S.; Liang, Y. Z.; Du, Y. P. Monte Carlo Cross-Validation for Selecting a Model and Estimating the Prediction Error in Multivariate Calibration. *J. Chemom.* **2004**, *18* (2), 112–120.
- (34) Sahigara, F. Defining the Applicability Domain of QSAR Models: An Overview. http://www.molecularDescriptors.eu/tutorials/T7_molecularDescriptors_ad.pdf.
- (35) Altman, D. G.; Bland, J. M. Absence of Evidence Is Not Evidence of Absence. *BMJ.* **1995**, *311* (7003), 485.