Journal of Chemical Information and Modeling

Article

pubs.acs.org/jcim

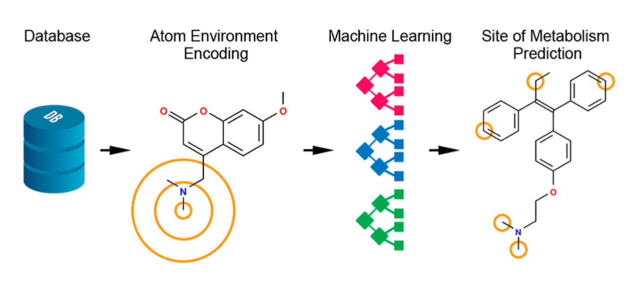
FAME 2: Simple and Effective Machine Learning Model of Cytochrome P450 Regioselectivity

Martin Šícho,[†,‡] Christina de Bruyn Kops,[†] Conrad Stork,[†] Daniel Svozil,[‡] and Johannes Kirchmair*[,†]

[†]Faculty of Mathematics, Informatics and Natural Sciences, Department of Computer Science, Center for Bioinformatics, Universität Hamburg, Hamburg, 20146, Germany

[‡]CZ-OPENSCREEN: National Infrastructure for Chemical Biology, Laboratory of Informatics and Chemistry, Faculty of Chemical Technology, University of Chemistry and Technology Prague, 166 28 Prague 6, Czech Republic

Supporting Information



ABSTRACT: We report on the further development of FAst MEtabolizer (FAME; *J. Chem. Inf. Model.* **2013**, *53*, 2896−2907), a collection of random forest models for the prediction of sites of metabolism (SoMs) of xenobiotics. A broad set of descriptors was explored, from simple 2D descriptors such as those used in FAME, to quantum chemical descriptors employed in some of the most accurate models for SoM prediction currently available. In line with the original FAME approach, our objective was to keep things simple and to come up with accurate and robust models that are based on a small number of 2D descriptors. We found that circular descriptions of atoms and their environments with such descriptors in combination with an extremely randomized trees algorithm can yield models that perform equally well compared to more complex approaches. Thorough evaluation experiments on an independent test set showed that the best of these models obtained a Matthews correlation coefficient, area under the receiver operating characteristic curve, and Top-2 accuracy of 0.57, 0.91 and 94.1%, respectively. Models for the prediction of isoform-specific regioselectivity of CYP 3A4, 2D6, and 2C9 were also developed and showed competitive performance. The best models have been integrated into a newly developed software package (FAME 2), which is available free of charge from the authors.

INTRODUCTION

The metabolic system has evolved as a primary line of defense of organisms against potentially harmful xenobiotics. Biotransformation can yield metabolites that differ substantially from their parent compounds with respect to biological and physicochemical properties.[1] For example, only about 3% of all drug metabolites are confirmed to maintain pharmacological activity, and even though the primary function of the metabolic system is to detoxify substances, at least 7% of metabolites are known as toxic or reactive.[2] Therefore, understanding metabolism is essential for the development of effective and safe compounds, in particular drugs, agrochemicals and cosmetics.

The ability to predict and identify metabolically labile atom positions in a molecule (i.e., sites of metabolism, SoMs) can aid in the derivation of likely metabolites and, consequently, strategies for the optimization of the metabolic properties of small molecules. However, experimental identification of SoMs or metabolites is an expensive and time-consuming task. Therefore, being able to accurately predict SoMs without the need to carry out laboratory experiments would greatly reduce costs and time needed to evaluate each compound.[3]

In the past few years, in silico methods for SoM prediction have shown their ability to carry out such predictions with good accuracy,[4] and multiple approaches have been presented and reviewed in the literature.[3,5−9]

In principle, a SoM prediction methodology should account for three main factors: (i) reactivity of atom positions in a molecule, (ii) pharmacophoric and shape constraints imposed by the binding site of the enzyme, and (iii) the accessibility of potential sites from the catalytic center. In the case of cytochrome P450 enzymes (CYPs), CYP 3A4 in particular, reactivity often plays a more significant role than the two other factors because several CYP isoforms have large and flexible binding sites and can accommodate a plethora of substrates.[10] However, taking accessibility and pharmacophoric constraints into account as well can often increase accuracy and has more impact when modeling metabolism of less flexible and promiscuous isoforms such as CYP 2D6.[11]

In silico methods for SoM prediction can be roughly divided into three classes: (i) rule-based, (ii) structure-based, and (iii) ligand-based. Rule-based methods attempt to derive likely metabolites and SoMs by applying a dictionary of biotransformation rules (usually compiled by human experts) to molecular fragments.[12−21] In relation to rule-based methods, structure-

Received: May 4, 2017
Published: August 7, 2017

ACS Publications © 2017 American Chemical Society

DOI: 10.1021/acs.jcim.7b00250

based approaches are on the opposite side of the spectrum. Their objective is to focus on the substrate−protein interaction in more detail and not to abstract it away by looking at the transformations and their outcomes alone.[8,22] Therefore, very different information can be obtained from structure-based modeling. The most significant advantage is that more insight is given into the orientation of a ligand at the binding site, which can provide additional benefits when optimizing the metabolic properties of compounds. In principle, it is possible to evaluate whether a compound is a plausible ligand or not, and how ligand stereochemistry influences the outcome of the biotransformation. However, this more detailed information comes at a higher computational cost and requires expert input. Moreover, it is often difficult to accurately capture the flexibility of the CYP isoforms.[9] In fact, the observed conformational space of the binding sites of CYPs observed in X-ray structures does not account for a substantial number of known substrates.[3] Structure-based approaches also suffer from typical docking problems. This means that the scoring functions are usually not sufficiently accurate to predict the relative binding free energy.[23] Additionally, water is not always explicitly modeled or modeled at all. Thus, some important hydrogen bonds that stabilize the substrate−enzyme complex or important effects related to entropy can be easily missed. Overall, structure-based approaches can be directly interpreted and provide a lot of information, but accurate modeling requires high quality crystal structures covering the relevant conformational space, sophisticated scoring methods, more computation time and considerable human effort during the preparation and analysis of the data. Therefore, structure-based approaches are more suited for a thorough investigation of a particular substrate−enzyme pair of interest than for a general screen of compounds.

Ligand-based methods do not explicitly take any structural information on substrate-enzyme complexes into account but rely heavily on available experimental data on SoMs and metabolites. They can have higher throughput and still rival structure-based methods in performance. It can also be argued that, with a large amount of high-quality data and a reliable set of descriptors, both flexibility of the enzymes and reactivity of the substrates can be captured with comparable accuracy from empirical data, albeit with the important exception of metabolic stereoselectivity. A lot of methods have been developed over the years,[3] and most of the modern approaches employ machine learning and atom-level descriptors to model CYP regioselectivity. One of the first such methods, published by Sheridan et al.,[24] was a reasonably predictive random forest model which relied on simple 2D topological descriptors that encoded local chemical environments and atom accessibility. This method was more accurate than some mechanistic approaches that explicitly modeled reactivity but did not take accessibility into account. This outcome showed that accessibility is an important factor, even for flexible isoforms such as CYP 3A4, and that a relatively simple empirical approach can be quite successful when enough data are available.

SMARTCyp is a freely available software package[25] and web service[26] for SoM prediction which does not rely on empirical data but on precomputed activation energies from density functional theory (DFT) calculations of model compounds. These energies were used to derive a lookup table which assigns an activation energy for a particular fragment in the query molecule as a measure of a site's reactivity. Accessibility of different sites was also included via a simple descriptor that was used to correct the final reactivity score.[11] This approach was able to predict a correct SoM within the top two highest ranking positions for 76% and 83% of the molecules of two validation sets.[11] Later, an additional accessibility descriptor was added, which resulted in a modest increase in performance.[27]

Two methods that have made heavy use of machine learning algorithms are RS-predictor[28−30] and its successor, XenoSite.[31,32] Both methods used quite a large number of different descriptors to capture the reactivity and accessibility of potential SoMs. RS-predictor was further improved by inclusion of SMARTCyp reactivities to model reactivity more accurately.[29] XenoSite is the final evolutionary step of RS-predictor so far and employs a neural network as a machine learning algorithm.[31,32] Using a leave-one-out cross-validation procedure, XenoSite was found to give at least one correct SoM prediction in the top two ranking positions for 89% of the molecules. A major contribution to SMARTCyp, RS-predictor and XenoSite was a large publicly available ("Zaretzki") data set, which includes SoM data for multiple CYP isoforms and which has been used to develop and test many other approaches as well.

A rather simple but effective machine learning method was implemented by Kirchmair et al. in FAME.[33] It relies on just seven descriptors based on the 2D representation of a molecule and uses a random forest model to facilitate predictions. Despite its simplicity, this approach was able to identify the correct SoM in the top two ranking positions for up to 81% of molecules from an external validation set, and it also allows the prediction of biotransformations catalyzed by non-CYP enzymes.

A progressive method in terms of estimation of activation energies was recently presented by Tyzack et al.[34] This method does not rely on a library of precomputed fragments to estimate the relative activation energy of different sites like SMARTCyp but performs effective semiempirical quantum chemical calculations of the whole substrate. For most CYP isoforms, this method was able to give a correct prediction in the top two ranked positions for more than 85% of the molecules in an external validation set.

Another recent contribution to the world of ligand-based approaches is a method published by He et al.[35] In this case, several models were created for each CYP-mediated reaction. The method relies on 56 quantum chemical descriptors, 29 physicochemical properties and 27 topological descriptors. Several methods from the machine learning toolbox were tested and the built models were shown to perform well on an external test set, with nearly 88% of the reactions predicted correctly.

More recently, Finkelmann et al.[36] reported on a new SoM predictor based on ab initio calculations that were used to describe electronic properties of molecules in the open source data set published with XenoSite. A scheme was introduced to encode partial charge distributions into atomic descriptors capable of not only describing potential SoMs themselves but also the properties of neighboring atoms. In this way, each SoM was described in a larger context rather than being treated as an independent sample. It was demonstrated that using such descriptors had a positive influence on machine learning model performance. The models presented in the study were able to predict a correct SoM within the first two ranked atom positions for 91% of the compounds in leave-one-out cross-validation experiments.

DOI: 10.1021/acs.jcim.7b00250

In this work, we present an accurate SoM prediction methodology further developed from the simple machine learning approach originally implemented in FAME.[33] Rather than using the random forest machine learning algorithm used in the original method, in this work extremely randomized trees[37] are applied. Using a revised version of the Zaretzki data set, we investigate the influence of various molecular descriptors (both topological and quantum chemical) on the accuracy of this ligand-based approach. We also describe a relatively simple 2D method to capture atomic environments of potential SoMs and show that models built using such descriptors have comparable or better performance than other more complex approaches reported in the literature. Finally, a freely available software package is introduced in order to make it possible for researchers to readily use the models presented in this work.

RESULTS

All models were trained and tested on a revised version of the Zaretzki data set (see Methods for detail). The Zaretzki data set has been successfully applied in several other studies[27,36,38−40] and, thus, is well-established in the field. In total, the prepared data contained information on 15,233 atoms, a sufficient number of data points for machine learning. It should, however, be noted that these are not entirely independent samples. Prior to any experiments, this data set was randomly split on molecules into a training set and an independent test set using an 80:20 ratio (Figure 1). This resulted in a training set that contained data on 542 compounds and a test set of 136 molecules (Table 1).

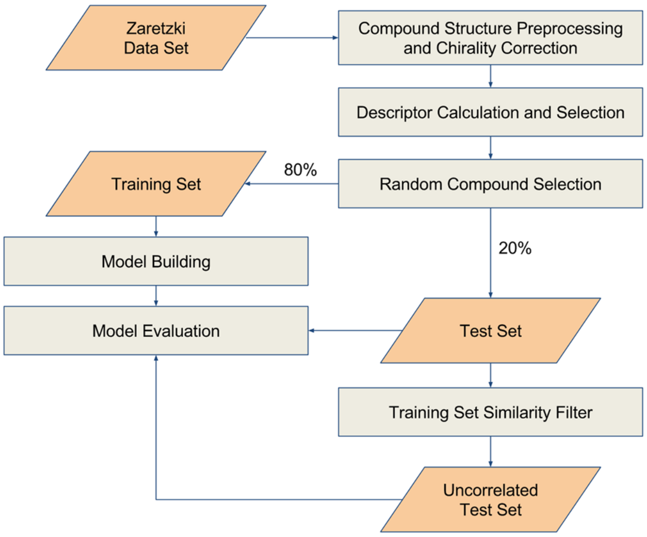

Zaretzki
Data Set
Compound Structure Preprocessing and Chirality Correction
Descriptor Calculation and Selection
Training Set
80%
Random Compound Selection
Model Building
20%
Model Evaluation
Test Set
Training Set Similarity Filter
Uncorrelated
Test Set

Figure 1. Overview of the data preparation, model building, and model evaluation workflow.

Model Evaluation Measures.

In order to evaluate the performance of the developed models, an appropriate set of evaluation measures needed to be selected. We used three different metrics: (i) the Matthews correlation coefficient (MCC), (ii) area under the ROC curve (AUC), and (iii) Top-k. The MCC was used as the main classification quality metric in this study. The MCC is a balanced measure which not only takes into account true and false positives but also negatives. It is calculated according to

Table 1. Number of Compounds in Each Dataset Used in the Current Study

CYP isoform[a]	total[b]	training set[c]	test set[d]	uncorr test set[e]
global	678	542	136	71
3A4	473	378	95	52
2D6	269	215	54	32
2C9	225	180	45	32

[a]The code of the isoform for which a specific model was built. The joined global model which includes all isoforms in the Zaretzki data set is marked as "global". [b]Total number of compounds that had SoMs annotated for the particular isoform in the data set. [c]Compounds randomly selected for training and hyperparameter optimization. [d]Compounds randomly selected for testing. [e]Compounds that remained in the uncorrelated test set after the similarity filter was applied.

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}} \tag{1}$$

where TP is the number of true positive, TN the number of true negative, FP the number of false positive and FN the number of false negative predictions. Because the MCC includes all items of the confusion matrix, it is regarded as a rather robust and balanced measure which is suitable for data sets with imbalanced classes.[41] Therefore, it was used as the main quality measure in this work and as a scoring function during model parameter optimization. In addition to the MCC, we also used the AUC, which provides insight into how well the models are able to rank the putative sites according to the probabilities given by the ensemble of decision trees.

Related to the AUC is the third quality measure reported in this study, the Top-k metric, which is commonly used for model evaluation in metabolism regioselectivity prediction. The Top-k measure provides a means to evaluate performance in terms of molecules rather than individual atoms. When calculating the measure, the atoms in the molecule are sorted according to the ranking obtained from the model and if at least one known SoM is discovered within the top k positions in the sorted list, the molecule is marked as correctly predicted. The percentage of the correctly predicted compounds in the data set is then reported. The most commonly used values of k are 2 and 3.

Model Performance as a Function of Descriptor Types and Model Parameters.

Different types of descriptors were explored for model development: (i) CDK atomic descriptors in analogy to those used in FAME[33] ("CDK"), (ii) atom type fingerprints ("ATF") used in various approaches,[42−44] (iii) quantum chemical ("QC") descriptors, and (iv) combinations thereof (Table 2). Additionally, these descriptors were used to derive circular representations of atom environments with a depth of up to 6 bonds (Figure 2), giving rise to a total of 32 different models.

Optimized Model Parameters. For each model, the *decision_threshold*, *max_features*, *class_weight*, and maximum number of features allowed in the model (Table 4) were optimized using a 10-fold cross-validated grid search on the training data set (a subset of the original data set that contained data on 542 molecules; see Methods for details). For this optimization, the MCC was used as a scoring function and the model with the highest average MCC over all folds was selected as optimal.

Table 2. Various Descriptor Types Investigated in this Study

class	group[a]	description
2D[b]	CDK	This group comprises 15 basic 2D descriptors implemented in CDK. Those are the original 15 descriptors considered in FAME, and they encode various properties of individual atoms in the molecule (Table S1).
	circCDK	This contains circular descriptors derived from the 15 CDK descriptors. In other words, it is equal to the CDK set enriched by the circular versions of the CDK descriptors of bond depth 1−6.
	CDK + ATF	This is a combination of the 15 basic CDK descriptors and the circular atom type fingerprints. The fingerprints represent occurrence counts of various atom types within a certain distance from the considered atom.
	circCDK + ATF	This group includes atom type circular fingerprints in addition to all descriptors from the circCDK set.
QC-enhanced[c]	CDK + QC	This is the simplest quantum chemistry (QC) enhanced set. It represents the complete set of noncircular atomic descriptors used in the current study (Table S1 and Table 3).
	CDK + circQC	This combination comprises the 15 basic CDK descriptors that are combined with circular versions of the 10 quantum chemical descriptors. In other words, this set contains the CDK + QC set and the circular versions of the QC descriptors for bond depths of 1−6.
	circCDK + ATF + circQC	These represent the combination of all descriptors investigated in the current study.

[a] Codes representing the descriptor groups explored in the current study. [b] Simple 2D descriptor sets that only use topological information and their circular counterparts. [c] Descriptor sets enhanced by quantum chemical descriptors.

Table 3. Quantum Chemical Descriptors Calculated by MOPAC[45,46]

name	description
piS(r)	atom self-polarizability[47,48]
De(r)	electrophilic delocalizabilities of an atom[47,49]
Dn(r)	nucleophilic delocalizabilities of an atom[47,49]
s-Pop	population of the s-orbital (i.e., formal electron density of the s-orbital)
p-Pop	population of the p-orbitals (i.e., formal electron density of the p-orbitals)
NumOfElecs	number of valence electrons localized on the atom (i.e., electron density)
NetCharge	net charge (i.e., formal number of valence electrons− NumOfElecs)
valence	sum of bonds for an atom (i.e., molecular orbital valency)[50]
mull_charge	partial charge of an atom determined by Mulliken[51,52]
mull_pop	electron population of an atom determined by Mulliken[51,52]

Decision Threshold and Class Weights. Adjusting the decision threshold provides a means to affect the trade-off between false-positives and false-negatives. This adjustment can remove the bias of the models toward the overrepresented class (non-SoMs) by lowering the percentage of estimators in the ensemble that are needed to make a decision that an atom is a SoM. The same strategy was also employed by Finkelmann et al.[36] to balance their models.

In addition to the decision threshold, the class balancing parameter of the classifier was optimized. This parameter adjusts the weight of the SoM and non-SoM classes during model building to compensate for the imbalance of the two classes in the training set. The weights are set to be inversely proportional to the number of samples available for each particular class in the training data.

A clear preference of decision thresholds lower than the default value of 0.5 was observed almost uniformly across all models (Table S2). The most commonly selected value was 0.4, but even values as low as 0.2 were sometimes selected. Especially low decision threshold values were observed for the models built using atom type fingerprints as the only circular descriptor ("CDK + ATF"). The CDK + ATF model was also the only one that made use of class balancing uniformly across all experiments. Therefore, one could argue that such models were probably more susceptible to class imbalance than others and that it was thus necessary to compensate for the influence of the negative class by lowering the decision threshold and adjusting the weights of the classes during model building.

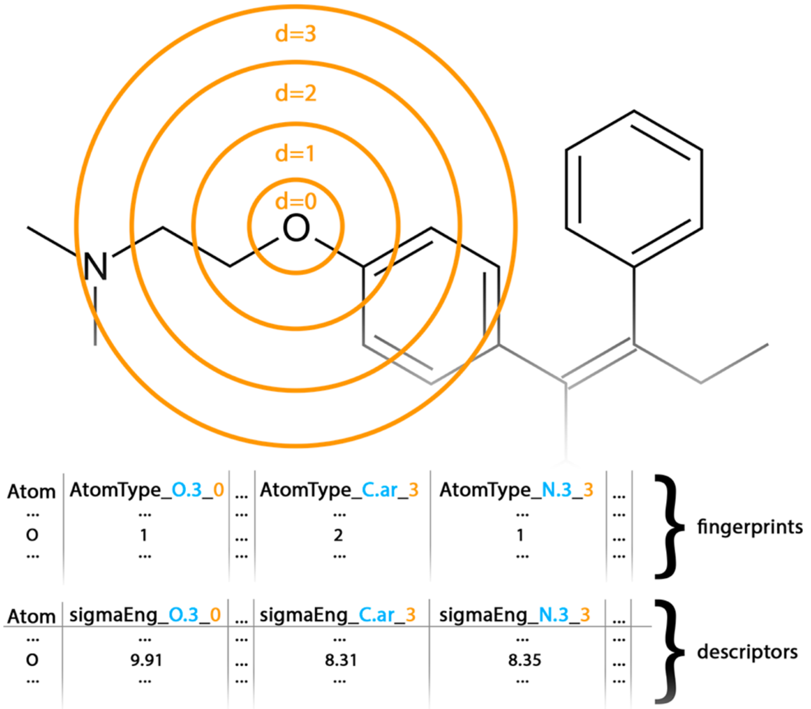
Maximum Number of Considered Features per Split. The optimal maximum percentage of features allowed to be sampled at each tree branch varied, but the most commonly selected value of the *max_features* parameter was 0.9, meaning that at most 90% percent of features were sampled (Table S2). Usually, the more features were available in total, the higher was the percentage of features required to be sampled when looking for a split. This was probably due to the fact that circular descriptors introduced more noise to the data and, therefore, a more extensive sampling of features was needed at each step. However, the CDK + ATF model was again an exception here. It seems that sampling more than 30% of features per split mostly did not provide any benefit in this case. This might have been a result of encoding less noisy information that covers only shallow relationships in the data (just atom type occurrences). There were also much fewer features overall in these models, which might have further contributed to the observed trend.

ANOVA F-Test. Prior to modeling, the number of all features was reduced using a univariate ANOVA F-test on each variable. No more than 400 features per model were allowed in our experiments because a greater number of features increases model complexity and, thus, the variance and risk of overfitting.

The median number of features per model was 275, and it was observed that the maximum number of features selected for modeling using the ANOVA F-test increased with the maximum bond depth of circular descriptors (Table S2). This behavior could be expected since encoding a wider neighborhood of an atom must mean that more detailed information is available about each potential SoM.

Influence of Descriptors on Model Performance During Training. Good models were obtained with each of the seven explored descriptor sets. The MCC values obtained for the models during cross-validation ranged from 0.51 to 0.59, whereas the Top-2 prediction rates ranged from 84.1 to 88.4% (all results provided in Table S3, including AUC values). The best models included the circCDK model at a bond depth of 6 (MCC 0.59; Top-2 86.9%) and the circCDK + ATF + circQC model at a bond depth of 2 (MCC 0.58; Top-2 88.4%).

Models based on 2D descriptors gradually improved in MCC with increasing bond depth of circular descriptors during cross-validation while QC-enhanced models did not show any growth after bond depth 1 (Figure 3a). Other metrics generally correlated well with the MCC. The only exception was the CDK + circQC model, for which there was no apparent correlation between the MCC, Top-2, and Top-3. However,



Atom	AtomType_O.3_0	...	AtomType_C.ar_3	AtomType_N.3_3	...
...	...	...	...	...	...
O	1	...	2	1	...
...	...	...	...	...	...

Atom	sigmaEng_O.3_0	...	sigmaEng_C.ar_3	sigmaEng_N.3_3	...
...	...	...	...	...	...
O	9.91	...	8.31	8.35	...
...	...	...	...	...	...

Figure 2. Example showing how circular atom type fingerprints and descriptors of up to bond depth 3 were calculated for a single atom. In this example, neighboring atoms up to three bonds away from the sp^3 hybridized oxygen were encoded. Two examples of data matrix parts corresponding to neighbors encoded in bond depths 0 and 3 are shown. In both the fingerprints and descriptors, atoms were grouped by their atom type (highlighted in blue in the column names) in each bond depth (highlighted in orange). In the case of the fingerprints, the occurrences of each atom type were noted, while for real-valued descriptors, the average value of the basic descriptor among the grouped atoms was calculated and assigned to the corresponding atom type. Therefore, for the *sigmaElectronegativity* descriptor in this example, the third layer would contain 8.35 (8.35/1 = 8.35) for the sp^3 hybridized nitrogen type and 8.31 ((8.31 + 8.31)/2 = 8.31) for the aromatic carbon type. Consequently, what can be derived from this matrix fragment is that there are two aromatic carbons (which are topologically identical) within three bonds of the oxygen atom that have a *sigmaElectronegativity* equal to 8.31 each.

Table 4. Hyperparameter Grid

model parameter	explored values
decision_threshold	0.2, 0.4, 0.5, 0.6
class_weight	balanced, balanced_subsample, none
max_features	sqrt, 0.3, 0.6, 0.9
max_features_ANOVA	100, 150, 200, 250, 300, 350, 400

there was a good correlation between the MCC and the AUC for all models (Pearson correlation coefficient of 0.92). Therefore, the further discussion will mostly focus on the MCC and only highlight other metrics where necessary.

The faster increase in accuracy and the early performance plateau of the QC-enhanced models could be interpreted as a result of having the 3D conformation of the molecule available and the fact that QC calculations give a much more precise description of the electron density around each atom. If the 3D conformation is available and the electron density is, therefore, more accurately described, more detailed properties of each atom can be calculated and more detailed information about its neighborhood is already implicitly included in the description of the atom itself. Hence, encoding a wider neighborhood of a potential SoM did not result in substantial improvements when quantum chemical descriptors were used, but it was necessary in the case of the 2D descriptors to reach comparable performance. Therefore, the QC approach, albeit computationally more costly, may have an advantage in that good model performance can be obtained with fewer features.

One more important conclusion that can be drawn from Figure 3a is that when sufficient area around a potential SoM is encoded using a purely 2D methodology, the resulting models have comparable performance to those enhanced by quantum chemical calculations.

Model Performance on an Independent Test Set. In order to assess the generalization of the models and their ability to perform in practice, an external validation set of 136 compounds was randomly drawn from the whole data set prior to any modeling and cross-validation experiments. The resulting scores obtained by each model on this independent test set (Figure 3b and Table 5) suggest similar trends to those observed during cross-validation (Figure 3a and Table S3), with the exception of lower performance improvement of the circular QC-enhanced models relative to the basic CDK + QC methodology. Also in contrast to the cross-validation results, the combined circCDK + ATF models seem to show greater performance improvement in comparison to the basic CDK method but with no clear increase in MCC beyond bond depth 1.

For a more robust comparison of the performance of different models, 100 random subsamples comprised of 50 molecules each were selected (with replacement) from the original test set and treated as separate external sets. Therefore, each subsample represented a possible classification problem of fixed size but varying difficulty.

In order to determine which models performed statistically differently from others, the MCC scores obtained on the subsamples were subjected to both parametric (repeated

DOI: 10.1021/acs.jcim.7b00250

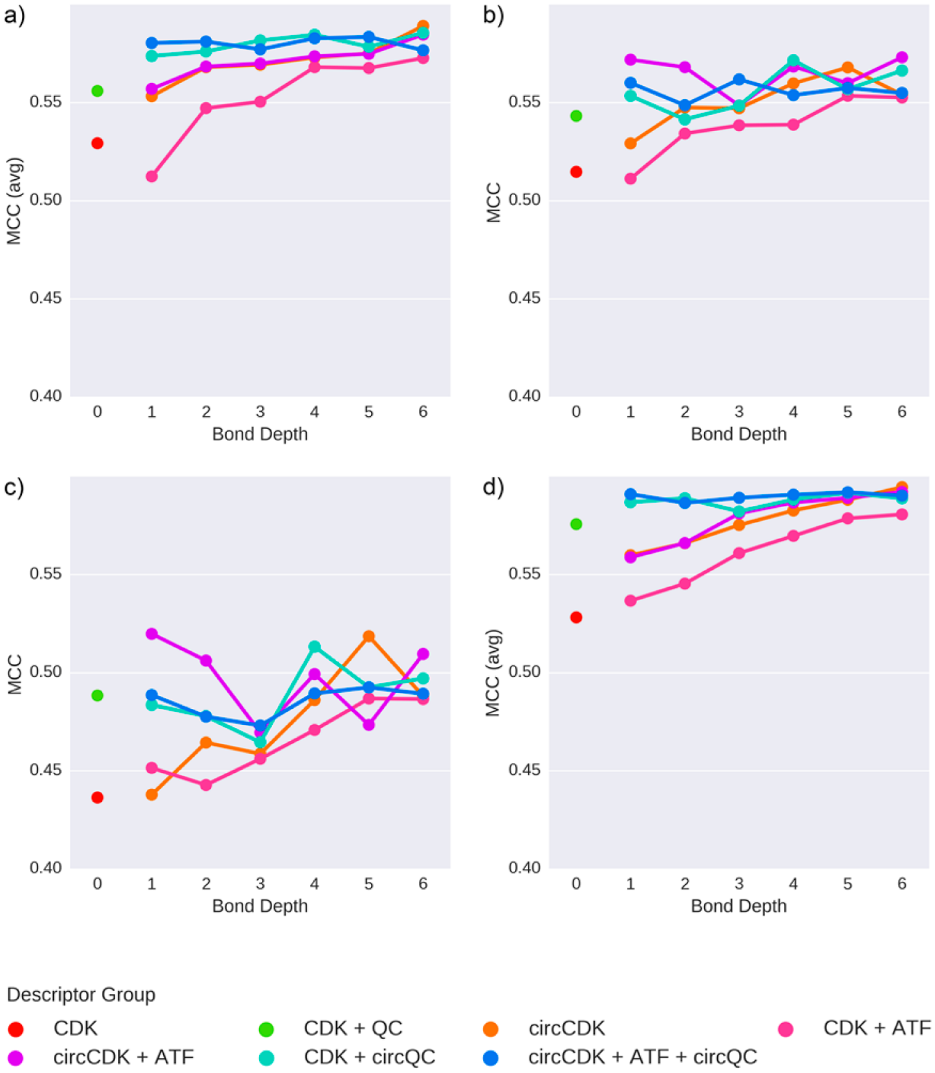


Figure 3. Visualization of the relationship between maximum bond depth of circular descriptors and model performance on the (a) training set (measured as the mean MCC across the 10 folds in cross-validation of the optimized model), (b) independent test set, (c) uncorrelated test set, and (d) full Zaretzki data set (measured as the mean MCC across the 10 folds in cross-validation of the optimized model). Models that do not use circular descriptors are shown to have a bond depth equal to 0.

measures ANOVA as an omnibus test and paired *t* tests with *p*-values corrected according to Holm[53] for posthoc analysis) and nonparametric (Friedman test[54−56] as an omnibus test and the Nemenyi test for posthoc analysis) statistical tests. We concluded that the use of parametric tests was possible in our case since neither visual observation of Q−Q plots nor the Shapiro−Wilk test[57] highlighted any issues with the underlying normality assumption. The data only violated the sphericity condition of the repeated measures ANOVA test (determined by the Mauchly test[58]). Therefore, the correction methods of Greenhouse−Geisser[59] and Huynh−Feldt[60] were employed to obtain the final *p*-values.

A significant statistical difference between at least two models was confirmed by both the repeated measures ANOVA with the aforementioned corrections and the Friedman omnibus tests (all p-values were lower than 2.2×10^{-16}). There was also a good agreement between the parametric and nonparametric posthoc tests regarding the rejection of the underlying null hypotheses (Figure 4). The two approaches differed in 52 out of all 496 pairwise comparisons conducted. This could be attributed to the Holm correction which is more conservative than the Nemenyi test. However, these differences do not influence the main results of our study.

As shown in Figure 5, most models were positively affected by the bond depth of the circular descriptors. This effect is most apparent for the models based on 2D descriptors for which the performance gradually increased with bond depth, and most 2D models with six layers significantly outperformed their simpler counterparts. The only exception was the circCDK + ATF model with bond depth 1 that showed comparable performance to the more complex models.

The QC-enhanced models were only modestly improved by circular descriptors, with only six of them statistically significantly outperforming the simplest CDK + QC model. Additionally, no QC-enhanced model showed significantly better performance than any of the top scoring 2D models (see the circCDK + ATF models with bond depth 1, 5, and 6 and the circCDK models with bond depth 5 and 6, for instance).

Therefore, in terms of performance, there was no model that showed a significantly better success rate than all others. However, it should be noted that these results favor the use of 2D approaches in practice since they are considerably easier

DOI: 10.1021/acs.jcim.7b00250
J. Chem. Inf. Model. 2017, 57, 1832−1846

Table 5. Model Performance on an Independent Test Set[a]

Descriptor Set[b]	Top-2[c]	Top-3[c]	AUC[c]	MCC[c]
CDK [0]	89.7%	92.6%	0.89	0.51
circCDK [1]	94.1%	97.1%	0.91	0.53
circCDK [2]	94.1%	97.1%	0.91	0.55
circCDK [3]	92.6%	97.8%	0.91	0.55
circCDK [4]	91.2%	97.1%	0.91	0.56
circCDK [5]	93.4%	97.8%	0.91	0.57
circCDK [6]	93.4%	96.3%	0.91	0.55
CDK + ATF [1]	91.2%	94.9%	0.90	0.51
CDK + ATF [2]	91.2%	94.9%	0.90	0.53
CDK + ATF [3]	89.7%	97.1%	0.91	0.54
CDK + ATF [4]	91.9%	97.1%	0.91	0.54
CDK + ATF [5]	91.2%	98.5%	0.91	0.55
CDK + ATF [6]	92.6%	97.8%	0.91	0.55
circCDK + ATF [1]	92.6%	95.6%	0.91	0.57
circCDK + ATF [2]	93.4%	98.5%	0.91	0.57
circCDK + ATF [3]	91.9%	97.8%	0.91	0.55
circCDK + ATF [4]	91.9%	98.5%	0.92	0.57
circCDK + ATF [5]	94.1%	98.5%	0.91	0.56
circCDK + ATF [6]	94.1%	97.8%	0.91	0.57
CDK + QC [0]	94.1%	97.1%	0.91	0.54
CDK + circQC [1]	94.9%	98.5%	0.91	0.55
CDK + circQC [2]	94.9%	97.1%	0.91	0.54
CDK + circQC [3]	91.9%	97.1%	0.91	0.55
CDK + circQC [4]	92.6%	96.3%	0.92	0.57
CDK + circQC [5]	92.6%	97.1%	0.91	0.56
CDK + circQC [6]	94.1%	97.8%	0.91	0.57
circCDK + ATF + circQC [1]	94.1%	97.1%	0.91	0.56
circCDK + ATF + circQC [2]	94.9%	97.8%	0.91	0.55
circCDK + ATF + circQC [3]	93.4%	97.8%	0.91	0.56
circCDK + ATF + circQC [4]	91.9%	97.1%	0.92	0.55
circCDK + ATF + circQC [5]	91.9%	97.8%	0.91	0.56
circCDK + ATF + circQC [6]	91.9%	97.8%	0.91	0.56
AVERAGE	92.7%	97.2%	0.91	0.55
MIN	89.7%	92.6%	0.89	0.51
MAX	94.9%	98.5%	0.92	0.57

[a] Model performance is indicated by a color gradient, ranging from dark red (worst performance among all models) via white to dark green (best performance among all models). [b] Model code according to the descriptor set (Table 2) and maximum bond depth (indicated by a number in brackets) used. [c] Value of the given performance metric calculated using predictions of 136 compounds in the independent test set.

and faster to both use and implement. Simpler approaches are also likely to be robust and easier to interpret. Therefore, the circCDK + ATF model with maximum bond depth of 1 seems to offer a good trade-off between accuracy and robustness. Even though the wider atomic environment was not considered in this case, the model was still able to perform well on data not used to develop the model. This is probably due to the fact that only the most important information needed to describe the reactive center, that is the types, counts, and continuous properties of the immediate neighbors, was encoded. Thus, the circCDK + ATF model with maximum bond depth of 1 seems to be a good choice for use in practice.

The subsampling experiments outlined above were also conducted using the random forest algorithm rather than extra trees. In these random forest experiments, the same parameters, data samples and random states were used as already outlined above. The only difference was in the used algorithm and that only 2D models were trained. In vast majority of the cases, using the random forest algorithm did not show any advantage over extra trees and there were a few cases where its average performance over the subsamples was slightly lower (compare Figure S1 and Figure 5).

Considering the Top-2 score, there was a positive difference in performance between the test set and the cross-validation average. The average difference between the cross-validation means and the test set scores was +6.5%, with a standard deviation of 1.3%. Though the performance was generally better on the test set in terms of the Top-2 metric, this was not the case for the MCC, for which the performance mostly remained the same with a difference of only −0.02 (standard deviation 0.01). This difference might have occurred because the whole training set of 542 compounds was used to train the final models rather than just 90% of the training set during cross-validation. Since Top-2 only depends on the first two predictions and is only calculated on a per-molecule basis, it might have been more sensitive to this effect.

Model Performance in a Structurally Uncorrelated Test Set. To investigate the structural correlation between the test set and the training set, an analysis of molecular similarity was conducted using MACCS keys as structural fingerprints and Tanimoto similarity. It was found that the median maximum similarity between the compounds from the testing set and the training set was 0.78, which hints at possible structural correlation of some compounds in the test set with compounds in the training set. This was also confirmed by plotting the histogram of maximum similarities shown in Figure 6a. In total, there were 65 compounds in the test set with a similarity to the closest compound in the training set higher than or equal to 0.8.

Therefore, a second (structurally uncorrelated) test set was constructed as a subset of the original test data. This second test set comprised of 71 compounds that showed maximum similarity to any of the compounds in the training data lower than 0.80. This second test set had median maximum similarity of 0.70 relative to the training set, and the mean maximum similarity within this test set was 0.55 with a standard deviation of 0.17 (Figure 6b). Therefore, the second test set contained molecules that were much less similar to the training data and also represented a rather diverse set of compounds.

In the case of the uncorrelated test set, the MCC changed by −0.09 on average relative to the mean from cross-validation, but the Top-2 remained very high with an average difference of +4% (Table S4). Therefore, even in the uncorrelated test set, the Top-2 performance had generally improved relative to the mean from cross-validation.

The fact that Top-2 did not decrease for the independent and the uncorrelated test sets might also be related to the nature of the metric itself. Top-2 results in a correctly predicted

DOI: 10.1021/acs.jcim.7b00250

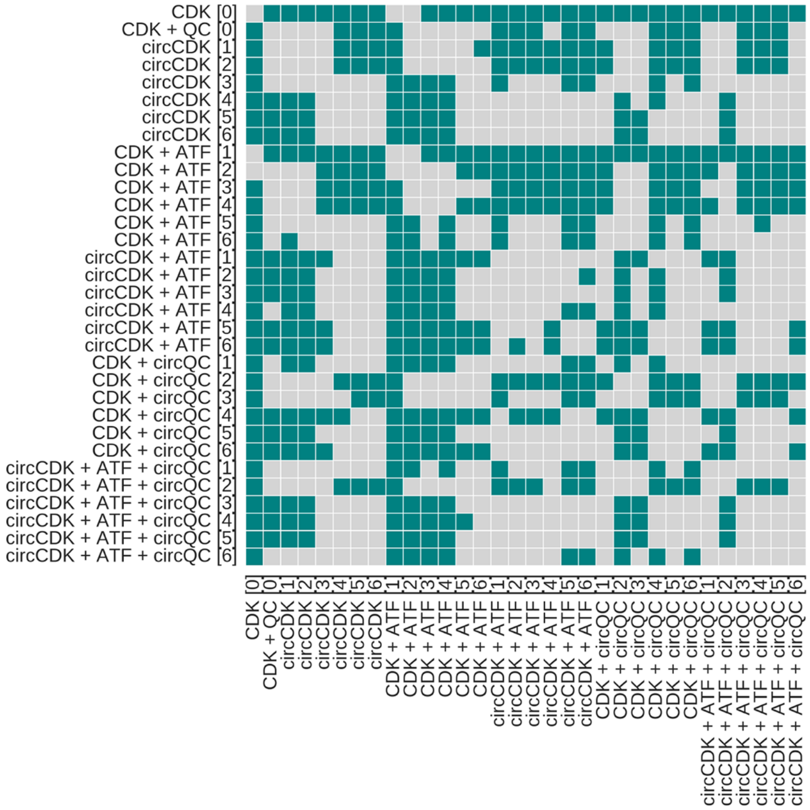


Figure 4. Consensus matrix combining results from parametric and nonparametric posthoc tests. The pairs of models for which the null hypothesis (the equality of means for *t* test or similar performance ranking for Nemenyi test) was rejected at an α level of 0.05 by both the parametric and nonparametric method are highlighted in teal.

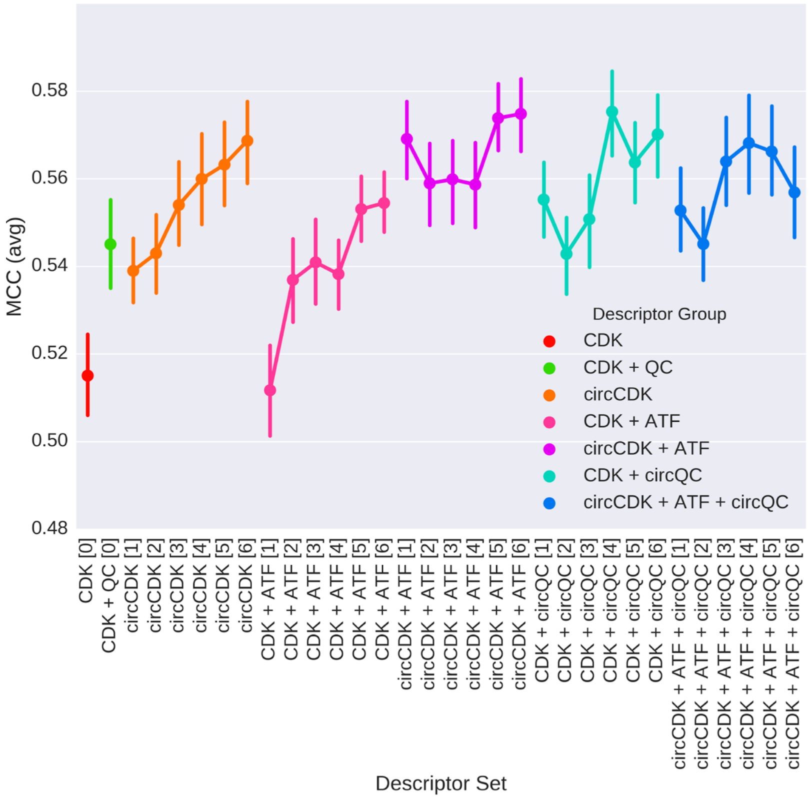


Figure 5. Point plot which represents the expected performance of each model by indicating 95% confidence intervals for mean MCC. The confidence intervals were obtained from a statistical analysis of predictions of 100 random subsamples (50 molecules each, sampled with replacement) of the original test set.

DOI: 10.1021/acs.jcim.7b00250

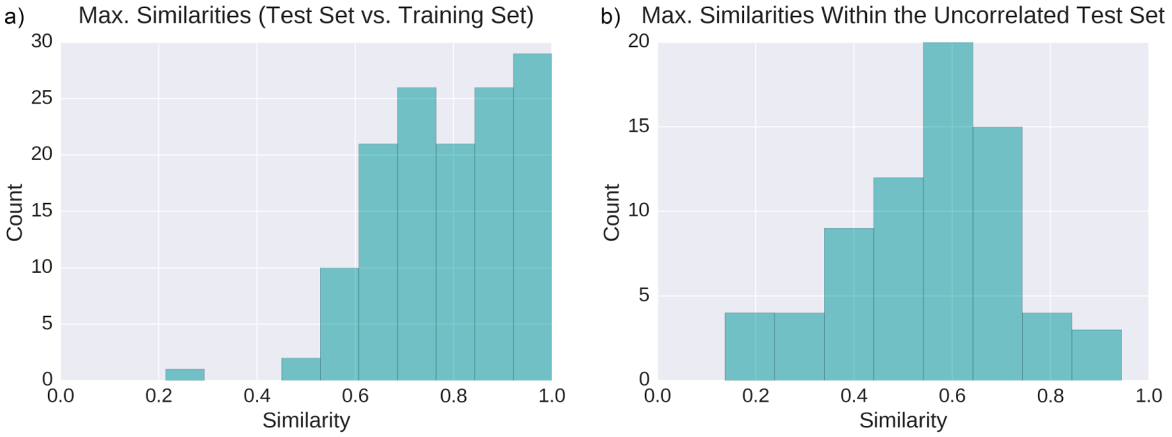


Figure 6. Histogram of maximum Tanimoto similarities (a) between compounds in the training set and the independent test set (136 molecules) and (b) between compounds within the uncorrelated test set (71 molecules).

molecule whenever at least one true SoM is within the top two ranked atoms, which not only is a rather loose criterion but also fails to capture the effect of false positives and, to some extent, false negative predictions (unlike the MCC). Therefore, this result might also hint at the unsuitability of the Top-2 metric to correctly reflect overall classification performance, for which the MCC is much more robust and reliable.

However, it should also be noted that the observed Top-2 values show that the best models were able to correctly predict at least one true SoM within the two top ranked positions in more than 90% of the molecules in both test sets. Since false negative results are usually more prevalent in SoM predictions (because of the imbalance of the classes), this could mean that the most obvious SoMs were still correctly identified by the models regardless of structural similarity of the test compounds to the training set.

Just like in the cross-validation experiments, the MCC obtained for models based on 2D descriptors generally improved with increasing bond depth (Figure 3c). However, this was not the case for the circCDK + ATF model, for which the performance seems to have decreased slightly overall.

For QC-enhanced models, no general improvement in the MCC with increasing bond depth was observed. Most of these models scored in the same range as the basic CDK + QC model. Moreover, the fact that the simple circCDK + ATF model with bond depth 1 also scored highly on the uncorrelated test set suggests that this model is fairly robust and will probably perform well in practical applications.

Comparison with Other Methods. Two of the latest and most accurate models for SoM prediction are Xenosite[31] and the models reported by Finkelmann et al.[36] Both approaches are based on the Zaretzki data set. While in the original publications the performance of XenoSite was evaluated by leave-one-out experiments using the Top-2 measure, the performance of the Finkelmann models was assessed using 10-fold cross-validation across the Zaretzki data set (and, in addition, was tested on a set of 25 undisclosed compounds from Bayer drug discovery projects). We followed both evaluation protocols as closely as possible to allow direct comparison of these models with ours. In order to compare the performance of our models with those presented by Finkelmann et al., we performed 10-fold cross-validation with hyper-parameter optimization on the full Zaretzki data set. For comparison of our models with Xenosite, we conducted experiments with leave-one-out cross-validation.

Performance Comparison with the Finkelmann Models. In general, the performance of our models and the Finkelmann models was comparable. Their final model (i.e., the best out of 10 random reruns) obtained an average MCC of 0.63 across all folds with a standard deviation of 0.06. They also devised a simpler model with lower variance by employing ANOVA F-test feature selection. This model performed with an average MCC of 0.59 and a standard deviation of 0.03. The result is comparable with the majority of the models developed in the current study.

Although our favorite simple model (the circCDK + ATF model with a bond depth of 1) obtained a lower average MCC score of 0.56 with a standard deviation of 0.03, its six layer version reached a performance comparable to the Finkelmann models, with an MCC of 0.59 and a standard deviation of 0.03 (Table S5 and Figure 3d). However, it should be noted that the mean score from the cross-validation was used to optimize the hyper-parameters of the models and, thus, a slight overtraining might have occurred which is more likely to happen for more complex models. Therefore, we still encourage the use of the simplest circCDK + ATF model in practice since it is expected to be less susceptible to such phenomena (see previous sections).

These findings indicate that simple 2D approaches can be just as effective in SoM prediction as more complex models based on descriptors derived from semiempirical or even ab initio quantum chemical calculations. Therefore, at the moment, there might be no need to work with a 3D representation of the molecule and calculations can be done much faster. Thanks to their simplicity, the resulting 2D models are also expected to be more robust and less prone to overfitting. The simplest circCDK + ATF model showed especially good generalization ability since it showed comparable or better performance than more complex QC-enhanced models on both test sets (Figure 3b and c) and in the resampling experiments (Figures 4 and 5). Additionally, the simple 2D models developed herein can be used without relying on a number of (often proprietary) software packages and applications, making usability and distribution to the scientific community easy and straightforward.

Performance Comparison with Xenosite. Most of the models presented herein showed comparable performance to both XenoSite and the models presented by Finkelmann et al.[36] that reached leave-one-out Top-2 accuracy of 89.4% and 90.9%, respectively. The average Top-2 accuracy of our models was

Table 6. Model Performance of Isoform-Specific Models[a]

	CYP 3A4[c]			CYP 2D6[c]			CYP 2C9[c]		
Descriptor Set[b]	MCC (CV)[d]	MCC (TS1)[e]	MCC (TS2)[f]	MCC (CV)[d]	MCC (TS1)[e]	MCC (TS2)[f]	MCC (CV)[d]	MCC (TS1)[e]	MCC (TS2)[f]
CDK [0]	0.46	0.46	0.35	0.51	0.50	0.41	0.41	0.42	0.40
circCDK [1]	0.49	0.54	0.44	0.52	0.44	0.31	0.46	0.51	0.49
circCDK [2]	0.51	0.57	0.49	0.55	0.50	0.46	0.49	0.51	0.47
circCDK [3]	0.53	0.56	0.44	0.54	0.53	0.51	0.48	0.52	0.50
circCDK [4]	0.52	0.52	0.41	0.51	0.54	0.53	0.50	0.57	0.56
circCDK [5]	0.53	0.55	0.45	0.55	0.51	0.45	0.50	0.55	0.53
circCDK [6]	0.53	0.56	0.46	0.57	0.55	0.49	0.49	0.59	0.57
CDK + ATF [1]	0.47	0.51	0.42	0.52	0.52	0.42	0.44	0.46	0.45
CDK + ATF [2]	0.50	0.54	0.45	0.52	0.51	0.45	0.48	0.50	0.47
CDK + ATF [3]	0.50	0.54	0.44	0.53	0.51	0.42	0.48	0.52	0.49
CDK + ATF [4]	0.52	0.54	0.42	0.53	0.51	0.47	0.48	0.56	0.53
CDK + ATF [5]	0.53	0.52	0.41	0.54	0.56	0.50	0.48	0.58	0.59
CDK + ATF [6]	0.54	0.53	0.42	0.54	0.57	0.53	0.50	0.58	0.57
circCDK + ATF [1]	0.50	0.51	0.44	0.54	0.51	0.41	0.49	0.47	0.46
circCDK + ATF [2]	0.51	0.57	0.49	0.55	0.50	0.46	0.49	0.50	0.48
circCDK + ATF [3]	0.53	0.56	0.46	0.53	0.51	0.47	0.49	0.55	0.51
circCDK + ATF [4]	0.53	0.53	0.41	0.54	0.50	0.43	0.50	0.58	0.57
circCDK + ATF [5]	0.53	0.55	0.43	0.56	0.52	0.45	0.48	0.58	0.56
circCDK + ATF [6]	0.54	0.55	0.45	0.57	0.53	0.47	0.50	0.55	0.52
CDK + QC [0]	0.51	0.54	0.42	0.57	0.54	0.50	0.49	0.52	0.49
CDK + circQC [1]	0.52	0.55	0.46	0.58	0.51	0.44	0.52	0.55	0.53
CDK + circQC [2]	0.54	0.54	0.45	0.59	0.54	0.45	0.54	0.53	0.51

DOI: 10.1021/acs.jcim.7b00250

Table 6. continued

Descriptor Set[b]	CYP 3A4[c] MCC (CV)[d]	 MCC (TS1)[e]	 MCC (TS2)[f]	CYP 2D6[c] MCC (CV)[d]	 MCC (TS1)[e]	 MCC (TS2)[f]	CYP 2C9[c] MCC (CV)[d]	 MCC (TS1)[e]	 MCC (TS2)[f]
CDK + circQC [3]	0.55	0.54	0.43	0.55	0.54	0.50	0.53	0.58	0.56
CDK + circQC [4]	0.54	0.54	0.40	0.58	0.52	0.47	0.52	0.57	0.55
CDK + circQC [5]	0.54	0.56	0.46	0.56	0.54	0.52	0.53	0.58	0.55
CDK + circQC [6]	0.54	0.56	0.48	0.58	0.54	0.49	0.54	0.54	0.53
circCDK + ATF + circQC [1]	0.54	0.55	0.45	0.59	0.51	0.44	0.52	0.57	0.54
circCDK + ATF + circQC [2]	0.54	0.55	0.46	0.57	0.52	0.47	0.53	0.55	0.51
circCDK + ATF + circQC [3]	0.54	0.55	0.46	0.57	0.54	0.52	0.52	0.56	0.55
circCDK + ATF + circQC [4]	0.53	0.55	0.46	0.57	0.54	0.48	0.53	0.54	0.49
circCDK + ATF + circQC [5]	0.54	0.56	0.46	0.57	0.52	0.47	0.53	0.53	0.50
circCDK + ATF + circQC [6]	0.54	0.57	0.47	0.57	0.52	0.46	0.52	0.56	0.52
AVERAGE	0.52	0.54	0.44	0.55	0.52	0.46	0.50	0.54	0.52
MIN	0.46	0.46	0.35	0.51	0.44	0.31	0.41	0.42	0.40
MAX	0.55	0.57	0.49	0.59	0.57	0.53	0.54	0.59	0.59

[a] Model performance is indicated by a color gradient, ranging from dark red (worst performance among all models) via white to dark green (best performance among all models). [b] Model code according to the descriptor set (Table 2) and maximum bond depth (indicated by a number in brackets) used. [c] Isoform code. [d] Average value of the given performance metric across ten cross-validation folds on the training set. [e] Value of the given performance metric as calculated using predictions on the independent test set. [f] Value of the given performance metric as calculated using predictions on the uncorrelated test set.

between 86.3% for the simplest CDK model and 90.3% for the CDK + QC model (Table S6). The best 2D models were the circCDK and circCDK + ATF models with maximum Top-2 equal to 89.4%. The favored circCDK + ATF model with bond depth 1 again scored slightly lower with Top-2 of 88.5%, but this difference is, in our opinion, almost negligible.

It should also be noted that we report results from one run with a fixed random seed, while Finkelmann et al.[36] and Zaretzki et al.[31] reported the best model out of multiple random reruns. Additionally, the two other methods used the full set of descriptors while we always limited their number to no more than 400 using an ANOVA F-test prior to modeling, since we used the same set of optimal hyper-parameters as determined by 10-fold cross-validation on the whole data set of 678 compounds.

Isoform-Specific Models. The same workflow outlined in Figure 1 was employed to obtain isoform-specific models for CYP 3A4, 2D6, and 2C9. The original data set did not contain SoM annotation for all isoforms and all compounds. Therefore, an appropriate subset of the original data was identified for each isoform (Table 1 for composition of each isoform-specific set).

The prediction of the SoMs of CYP 3A4 substrates proved to be more challenging than for the other two CYP isoforms. The best-performing model during cross-validation was the CDK + circQC model with circular descriptors of up to bond depth 3 with an MCC of 0.55 (Table 6). The best performance on the uncorrelated test set was obtained with circCDK and circCDK + ATF models with a maximum circular descriptor bond depth of 2. Both models reached MCC values as high as 0.49.

Models built for CYP 2D6 had the best performance among the three different isoforms, with an average MCC of 0.55. The best model in cross-validation was the CDK + circQC model with a maximum circular descriptor bond depth of 2 that obtained an MCC of 0.59. However, the CYP 2D6 models also showed the biggest difference in MCC between the cross-validation and the uncorrelated test set (−0.09 on average). The best performing models on the uncorrelated test set were the circCDK model with circular descriptors up to bond depth

4 and the CDK + ATF model with a maximum bond depth of 6.

The CYP 2C9 models showed the lowest mean model performance during the 10-fold cross-validation on the training set (mean MCC of 0.50 across the averages from folds) but the best performance on the uncorrelated test set. The best-performing models during the cross-validation were the CDK + circQC models with maximum circular bond depths of 2 and 6 and an MCC of 0.54. The best model in the uncorrelated test set was the CDK + ATF model with maximum circular descriptor bond depth of 5, which reached an MCC as high as 0.59. Therefore, even though the models did not perform as well during cross-validation, they were still able to generalize well.

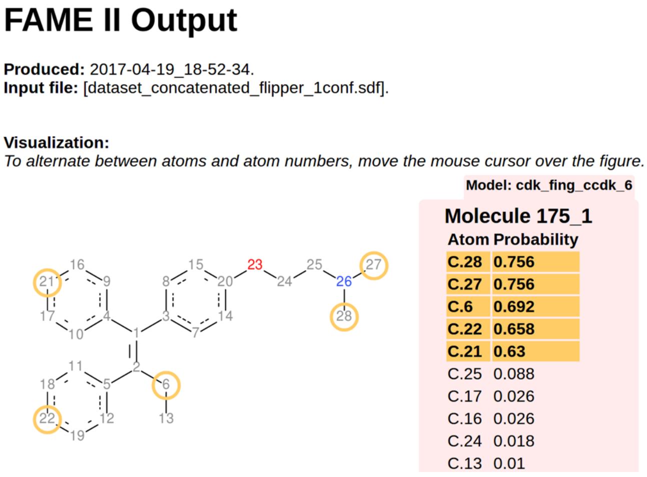
It should be pointed out that also in the case of the isoforms, the best performing models in the uncorrelated test set were based on 2D descriptors. This outcome suggests that using a 2D method to encode molecular properties is indeed a robust approach that results in models with good generalization ability. However, it should also be noted that the isoform-specific models often showed lower overall performance in comparison to the global model, which is most likely related to the lack of training data (particularly in the case of the CYP 2D6 and 2C9 isoforms). Therefore, it is expected that such models could greatly improve if more experimental data were available.

The choice of the best model also becomes more difficult in this case, since with less data in the test sets the variability of the score increases, making a comparison more difficult. Nevertheless, for the CYP 3A4 isoform, the circCDK and circCDK + ATF models with a bond depth of 2 seem to have been most successful on both test sets. The success of simpler models like these might again be attributed to their robustness and simplicity, but also to the fact that for CYP 3A4 reactivity is a more important factor than the shape of the substrate and, thus, the immediate neighbors of the SoM are more influential. This conclusion also seems to be reinforced with the results obtained for CYP 2C9 and 2D6 for which the best performing model on both test sets was the CDK + ATF model with a bond depth of 6. Since these isoforms are expected to be more selective regarding their substrates, encoding a wider area may be necessary to include the influence of the other interacting atoms.

Computation Time. Average execution time for the predictor was measured for the full data set of 678 compounds. A workstation equipped with an Intel Core i5 processor (i5-6200U), 8 GB of RAM, and a Linux operating system was used to make predictions using the circCDK + ATF model with a maximum bond depth of 6, since this was the most complex 2D model. The median execution time per compound was 265 ms, with only 24 compounds taking longer than 5 s to compute. Therefore, very fast predictions were possible for the vast majority of compounds.

Software Package Description. A Java software package named FAME 2 was developed to make the 2D models that are likely to be most successful in practice available for anyone to use. FAME 2 contains the 2D model that showed the best average performance during test set resampling (circCDK + ATF with bond depth 6) and the simplest 2D models that were not found to perform significantly differently than the best model (the circCDK model with bond depth 4 and the circCDK + ATF model with bond depth 1). These final global metabolism models were developed using the whole data set and 10-fold cross-validation to optimize the parameters. The software uses the simplest model (circCDK + ATF with bond depth 1) by default, but it also enables the user to easily select any of the two other models. The software package is available from the authors free of charge.

FAME 2 uses a slightly modified version of the visualization developed by Patrik Rydberg and implemented in SMARTCyp.[25] The predictions are generated as a simple HTML page which displays the structure of the compound with the predicted SoMs highlighted with yellow circles (Figure 7).



FAME II Output

Produced: 2017-04-19_18-52-34.
Input file: [dataset_concatenated_flipper_1conf.sdf].

Visualization:
To alternate between atoms and atom numbers, move the mouse cursor over the figure.

Model: cdk_fing_ccdk_6

Molecule 175_1

Atom	Probability
C.28	0.756
C.27	0.756
C.6	0.692
C.22	0.658
C.21	0.63
C.25	0.088
C.17	0.026
C.16	0.026
C.24	0.018
C.13	0.01

Figure 7. Example of prediction visualization by FAME 2.

The software also displays the probabilities generated by the extra trees model (the percentage of trees that predicted a given atom as a SoM) and highlights the atoms that scored above the decision threshold of the current model. Some metadata and helpful information as well as the model used for the predictions are also displayed.

CONCLUSIONS

Current models for SoM prediction often rely on complex sets of (quantum) chemical descriptors and accurate representation of substrate conformations. In this work we explored the use of extremely randomized trees for modeling the regioselectivity of cytochrome P450 enzymes and showed that simple, circular descriptions of atoms and their neighborhood with 2D descriptors can yield models that are robust and just as accurate as more complex approaches. This conclusion is supported by a thorough analysis of model performance during which we also postulate that the simplest one layer model based on 2D circular CDK descriptors and atom type fingerprints (model "circCDK + ATF [1]") is most suitable to use in practice since it showed good (or comparable) generalization ability in comparison to other models (MCC of 0.57 on an independent test set).

The method presented here is simple and effective, and we distribute the models that we found most likely to perform well in practice in the form of an easy-to-use software package named FAME 2. This will also allow researchers to use these models with confidential data as no transmission of such to external sources is required (as in the case of the many available web services).

It appears that with the data on enzyme regioselectivity available at present we are approaching a plateau with respect to the accuracy and applicability of models. It is expected that with the availability of additional data in the future the performance of SoM predictors can be further improved and their applicability extended.

■ METHODS

Data Set. The study is based on a revised version of the Zaretzki data set[31] published in ref 61. The main difference between the revised and the original version of the data set is a more accurate annotation of stereochemical information for chiral molecules. The general workflow of the experiments in this study is outlined in Figure 1. All stereoisomers of molecules with at least one undefined stereocenter were enumerated with the *flipper* option in OMEGA[62,63] (v2.5.1.4), and the conformation of each isomer optimized with OMEGA (default settings, with *maxconfs* = 1 and *canonOrder* = true in order to keep SoM annotation consistent with the atom sequence). From all possible stereoisomers generated this way, one was selected at random and used throughout all experiments. The preprocessed molecules were then used to calculate various atomic descriptors (see Descriptors).

The final data set contained information about 678 molecules. Two molecules of the original set[31] could not be processed due to missing parameters in MOPAC for the boron atom in bortezomib and a failure of CDK to correctly assign the atom type of the nitrogen atom in the nitro group of imidacloprid.

Topologically symmetrical atoms were considered to be equivalent when purely 2D descriptors were used for modeling. For the calculation of quantum chemical descriptors, symmetrical atoms were all assigned one probability value when making the predictions, which was the maximum probability calculated by the model for potential SoMs in one particular symmetry class.

Finally, the molecules were randomly assigned into a training and testing set. After the split, the test set contained 136 molecules (20% of the original data), while the remaining 542 compounds were designated for training. The initial test set was further reduced to 71 compounds by applying a similarity filter that only retained the molecules with maximum structural similarity to any of the compounds in the training set lower than 0.8. MACCS keys and Tanimoto similarity as implemented in RDKit[64] were used for compound representation and similarity calculation.

Descriptors. Various sets of atomic descriptors were investigated in this work (see Table 2 for an overview of descriptor types and abbreviations). The most basic set of descriptors was comprised of the same set of 15 atomic descriptors investigated in FAME[33] and implemented in CDK[65,66] (Table S1). In addition to these simple 2D descriptors, ten quantum chemical descriptors were extracted from geometry optimization calculations by MOPAC[45,46] with the semiempirical Austin Model 1 (AM1), which is based on the Neglect of Differential Diatomic Overlap (NDDO) with the VSTO-3G basis set (standard basis set in MOPAC). The quantum chemical descriptors in Table 3 were used (generated by the keywords AM1, XYZ, MMOK, VECTORS, BONDS, PI, PRECISE, ENPART, EF, MULLIK, and SUPER).

Furthermore, circular atom type fingerprints were calculated for each potential SoM as described by Tyzack et al.[44] In general, the fingerprints encode how many atoms of a particular type are within a certain topological distance (i.e., number of bonds) from a putative SoM. The AtomType descriptor was used to group the neighboring atoms by their types. Therefore, each layer of the fingerprint represents occurrences of Sybyl atom types that lie a certain number of bonds away from the putative SoM in question (Figure 2). In our experiments, fingerprints of up to bond depth 6 were used.

A more detailed set of circular descriptors was also calculated using both the CDK descriptors and the quantum chemical descriptors described above. These descriptors were generated in a similar fashion as the fingerprints, but the actual descriptor values (instead of simple atom type occurrence counts) were encoded in each layer (Figure 2). To be more precise, the atoms in each layer were grouped by atom types and the values of a particular descriptor calculated for all atoms in a group were averaged and assigned to the corresponding atom type. The rest of the atom types then had an unspecified value attached in this layer. This process was repeated for every atomic descriptor and all atoms up to a distance of six bonds from a possible SoM were considered as neighbors and encoded in this manner.

Descriptor Selection. In order to reduce model complexity, a univariate ANOVA F-test was conducted before each model was built. The maximum number of best features to retain was optimized using a cross-validated grid search on the training data using the MCC (described in the Results section) as a scoring function. Additionally, all variables with less than three unique values (including an unspecified value) were also removed prior to modeling.

Value Imputation. In order to replace the unspecified values in circular descriptors, a simple imputation strategy was implemented according to

$$i_d = \overline{x_d} + 10 \times s_d \qquad (2)$$

where i_d is the imputed value of descriptor d while $\overline{x_d}$is the mean of all numerical values of descriptor d and s_d is their standard deviation. The purpose of the factor 10 in the formula was to ensure (with reasonable confidence) that the imputed values lie outside the expected range of the known numeric values. The imputed values for the test set were equal to those determined from the training set according to eq 2.

Model Building. An extra trees classifier implementation from the machine learning Python package scikit-learn (v0.18)[67] was used to build all models in the present study. In all experiments, bootstrap sampling and the out-of-bag error estimate were used during training. Each model ensemble contained 500 estimators. All other hyperparameters were kept at their default values during training, but a 10-fold cross-validated grid search was conducted to optimize some of the most important ones. The grid shown in Table 4 was used for parameter optimization.

The *decision_threshold* parameter denotes the minimum probability needed to decide whether an atom is categorized as a SoM or non-SoM when the atom positions in a predicted molecule are classified. The *max_features* parameter is the maximum number of available features to consider when looking for an optimum split. The value "sqrt" means that the number was set as the square root of all available features. The *class_weight* parameter adjusts the weight of the two atom classes in the data set (SoM vs non-SoM) during model building to compensate for the imbalance in the training set. The weights are inversely proportional to class frequencies and can be either set using global frequencies in the whole training data ("balanced") or frequencies in the bootstrap sample of each tree ("balanced_subsample"). The value of "none" means that class balancing was turned off (all classes had the same weight). The *max_features_ANOVA* parameter is the maximum number of best features to select for modeling according to the ANOVA F-test. The original number of features was retained if it did not exceed 100.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.7b00250.

> Additional figures and tables with calculated CDK descriptors, hyperparameter optimization results, model validation results, and performance of the random forest algorithm (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: kirchmair@zbh.uni-hamburg.de. Tel.: +49 (0)40 42838 7303 (J.K.).

ORCID

Martin Šícho: 0000-0002-8771-1731
Christina de Bruyn Kops: 0000-0001-8890-2137
Conrad Stork: 0000-0002-5499-742X
Daniel Svozil: 0000-0003-2577-5163
Johannes Kirchmair: 0000-0003-2667-5877

Funding

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—project number KI 2085/1-1—and by the Ministry of Education of the Czech Republic—project numbers MSMT No 20-SVV/2017, NPU I-LO1220, and LM2015063. M.S. was supported by the Erasmus+ Programme of the European Commission.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum provided under the programme "Projects of Large Research, Development, and Innovations Infrastructures" (CESNET LM2015042), is greatly appreciated. OpenEye Scientific Software is acknowledged for providing licenses for OMEGA via the Free Academic Licensing Program.

■ LIST OF ABBREVIATIONS

AM1, Austin Model 1; ATF, atom type fingerprints; AUC, area under the ROC curve; CYP, cytochrome P450 enzyme; DFT, density functional theory; MCC, Matthews correlation coefficient; NDDO, Neglect of Differential Diatomic Overlap; QC, quantum chemistry; SOM, site of metabolism

■ REFERENCES

(1) Kirchmair, J.; Howlett, A.; Peironcely, J. E.; Murrell, D. S.; Williamson, M. J.; Adams, S. E.; Hankemeier, T.; van Buren, L.; Duchateau, G.; Klaffke, W.; Glen, R. C. How Do Metabolites Differ from Their Parent Molecules and How Are They Excreted? *J. Chem. Inf. Model.* **2013**, *53*, 354−367.

(2) Testa, B.; Pedretti, A.; Vistoli, G. Reactions and Enzymes in the Metabolism of Drugs and Other Xenobiotics. *Drug Discovery Today* **2012**, *17*, 549−560.

(3) 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*, 387−404.

(4) Campagna-Slater, V.; Pottel, J.; Therrien, E.; Cantin, L.-D.; Moitessier, N. Development of a Computational Tool to Rival Experts in the Prediction of Sites of Metabolism of Xenobiotics by P450s. *J. Chem. Inf. Model.* **2012**, *52*, 2471−2483.

(5) Crivori, P.; Poggesi, I. Computational Approaches for Predicting CYP-Related Metabolism Properties in the Screening of New Drugs. *Eur. J. Med. Chem.* **2006**, *41*, 795−808.

(6) Tarcsay, Á.; Keserű, G. M. In Silico Site of Metabolism Prediction of Cytochrome P450-Mediated Biotransformations. *Expert Opin. Drug Metab. Toxicol.* **2011**, *7*, 299−312.

(7) 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*, 617−648.

(8) Raunio, H.; Kuusisto, M.; Juvonen, R. O.; Pentikäinen, O. T. Modeling of Interactions between Xenobiotics and Cytochrome P450 (CYP) Enzymes. *Front. Pharmacol.* **2015**, *6*, 123.

(9) Bezhentsev, V. M.; Tarasova, O. A.; Dmitriev, A. V.; Rudik, A. V.; Lagunin, A. A.; Filimonov, D. A.; Poroikov, V. V. Computer-Aided Prediction of Xenobiotic Metabolism in the Human Body. *Russ. Chem. Rev.* **2016**, *85*, 854.

(10) Rydberg, P. Reactivity-Based Approaches and Machine Learning Methods for Predicting the Sites of Cytochrome P450-Mediated Metabolism. In *Drug Metabolism Prediction*; Kirchmair, J., Ed.; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, 2014; pp 265−292.

(11) Rydberg, P.; Olsen, L. Predicting Drug Metabolism by Cytochrome P450 2C9: Comparison with the 2D6 and 3A4 Isoforms. *ChemMedChem* **2012**, *7*, 1202−1209.

(12) Darvas, F. Predicting Metabolic Pathways by Logic Programming. *J. Mol. Graphics* **1988**, *6*, 80−86.

(13) Klopman, G.; Dimayuga, M.; Talafous, J. META. 1. A Program for the Evaluation of Metabolic Transformation of Chemicals. *J. Chem. Inf. Model.* **1994**, *34*, 1320−1325.

(14) Talafous, J.; Sayre, L. M.; Mieyal, J. J.; Klopman, G. META. 2. A Dictionary Model of Mammalian Xenobiotic Metabolism. *J. Chem. Inf. Model.* **1994**, *34*, 1326−1333.

(15) Greene, N.; Judson, P. N.; Langowski, J. J.; Marchant, C. A. Knowledge-Based Expert Systems for Toxicity and Metabolism Prediction: DEREK, StAR and METEOR. *SAR QSAR Environ. Res.* **1999**, *10*, 299−314.

(16) Hou, B. K.; Wackett, L. P.; Ellis, L. B. M. Microbial Pathway Prediction: A Functional Group Approach. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1051−1057.

(17) Hatzimanikatis, V.; Li, C.; Ionita, J. A.; Henry, C. S.; Jankowski, M. D.; Broadbelt, L. J. Exploring the Diversity of Complex Metabolic Networks. *Bioinformatics* **2005**, *21*, 1603−1609.

(18) Ridder, L.; Wagener, M. SyGMa: Combining Expert Knowledge and Empirical Scoring in the Prediction of Metabolites. *ChemMedChem* **2008**, *3*, 821−832.

(19) Gao, J.; Ellis, L. B. M.; Wackett, L. P. The University of Minnesota Pathway Prediction System: Multi-Level Prediction and Visualization. *Nucleic Acids Res.* **2011**, *39*, W406−W411.

(20) Mu, F.; Unkefer, C. J.; Unkefer, P. J.; Hlavacek, W. S. Prediction of Metabolic Reactions Based on Atomic and Molecular Properties of Small-Molecule Compounds. *Bioinformatics* **2011**, *27*, 1537−1545.

(21) Yousofshahi, M.; Manteiga, S.; Wu, C.; Lee, K.; Hassoun, S. PROXIMAL: A Method for Prediction of Xenobiotic Metabolism. *BMC Syst. Biol.* **2015**, *9*, 94.

(22) Sun, H.; Scott, D. O. Structure-Based Drug Metabolism Predictions for Drug Design. *Chem. Biol. Drug Des.* **2010**, *75*, 3−17.

(23) Kingsley, L. J.; Wilson, G. L.; Essex, M. E.; Lill, M. A. Combining Structure- and Ligand-Based Approaches to Improve Site of Metabolism Prediction in CYP2C9 Substrates. *Pharm. Res.* **2015**, *32*, 986−1001.

(24) Sheridan, R. P.; Korzekwa, K. R.; Torres, R. A.; Walker, M. J. Empirical Regioselectivity Models for Human Cytochromes P450 3A4, 2D6, and 2C9. *J. Med. Chem.* **2007**, *50*, 3173−3184.

(25) Rydberg, P.; Gloriam, D. E.; Zaretzki, J.; Breneman, C.; Olsen, L. SMARTCyp: A 2D Method for Prediction of Cytochrome P450-Mediated Drug Metabolism. *ACS Med. Chem. Lett.* **2010**, *1*, 96−100.
(26) Rydberg, P.; Gloriam, D. E.; Olsen, L. The SMARTCyp Cytochrome P450 Metabolism Prediction Server. *Bioinformatics* **2010**, *26*, 2988−2989.
(27) Rydberg, P.; Rostkowski, M.; Gloriam, D. E.; Olsen, L. The Contribution of Atom Accessibility to Site of Metabolism Models for Cytochromes P450. *Mol. Pharmaceutics* **2013**, *10*, 1216−1223.
(28) Zaretzki, J.; Bergeron, C.; Rydberg, P.; Huang, T.-W.; 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*, 1667−1689.
(29) Zaretzki, 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*, 1637−1659.
(30) Zaretzki, J.; Bergeron, C.; Huang, T.-W.; Rydberg, P.; Swamidass, S. J.; Breneman, C. M. RS-WebPredictor: A Server for Predicting CYP-Mediated Sites of Metabolism on Drug-like Molecules. *Bioinformatics* **2013**, *29*, 497−498.
(31) Zaretzki, J.; Matlock, M.; Swamidass, S. J. XenoSite: Accurately Predicting CYP-Mediated Sites of Metabolism with Neural Networks. *J. Chem. Inf. Model.* **2013**, *53*, 3373−3383.
(32) Matlock, M. K.; Hughes, T. B.; Swamidass, S. J. XenoSite Server: A Web-Available Site of Metabolism Prediction Tool. *Bioinformatics* **2015**, *31*, 1136−1137.
(33) Kirchmair, J.; Williamson, M. J.; Afzal, A. M.; Tyzack, J. D.; Choy, A. P. K.; Howlett, A.; Rydberg, P.; Glen, R. C. FAst MEtabolizer (FAME): A Rapid and Accurate Predictor of Sites of Metabolism in Multiple Species by Endogenous Enzymes. *J. Chem. Inf. Model.* **2013**, *53*, 2896−2907.
(34) Tyzack, J. D.; Hunt, P. A.; Segall, M. D. Predicting Regioselectivity and Lability of Cytochrome P450 Metabolism Using Quantum Mechanical Simulations. *J. Chem. Inf. Model.* **2016**, *56*, 2180−2193.
(35) He, S.-B.; Li, M.-M.; Zhang, B.-X.; Ye, X.-T.; Du, R.-F.; Wang, Y.; Qiao, Y.-J. Construction of Metabolism Prediction Models for CYP450 3A4, 2D6, and 2C9 Based on Microsomal Metabolic Reaction System. *Int. J. Mol. Sci.* **2016**, *17*, E1686.
(36) Finkelmann, A. R.; Göller, A. H.; Schneider, G. Site of Metabolism Prediction Based on Ab Initio Derived Atom Representations. *ChemMedChem* **2017**, *12*, 606−612.
(37) Geurts, P.; Ernst, D.; Wehenkel, L. Extremely Randomized Trees. *Mach. Learn.* **2006**, *63*, 3−42.
(38) Tyzack, J. D.; Williamson, M. J.; Torella, R.; Glen, R. C. Prediction of Cytochrome P450 Xenobiotic Metabolism: Tethered Docking and Reactivity Derived from Ligand Molecular Orbital Analysis. *J. Chem. Inf. Model.* **2013**, *53*, 1294−1305.
(39) Huang, T.-W.; Zaretzki, J.; Bergeron, C.; Bennett, K. P.; Breneman, C. M. DR-Predictor: Incorporating Flexible Docking with Specialized Electronic Reactivity and Machine Learning Techniques to Predict CYP-Mediated Sites of Metabolism. *J. Chem. Inf. Model.* **2013**, *53*, 3352−3366.
(40) Zaretzki, J. M.; Browning, M. R.; Hughes, T. B.; Swamidass, S. J. Extending P450 Site-of-Metabolism Models with Region-Resolution Data. *Bioinformatics* **2015**, *31*, 1966−1973.
(41) Powers, D. Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. *J. Mach. Learn. Technol.* **2011**, *2*, 37−63.
(42) Adams, S. E. *Molecular Similarity and Xenobiotic Metabolism*; University of Cambridge, 2010.
(43) Boyer, S.; Arnby, C. H.; Carlsson, L.; Smith, J.; Stein, V.; Glen, R. C. Reaction Site Mapping of Xenobiotic Biotransformations. *J. Chem. Inf. Model.* **2007**, *47*, 583−590.
(44) 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*, 29.
(45) Stewart, J. J. MOPAC: A Semiempirical Molecular Orbital Program. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 1−105.
(46) MOPAC2016. http://openmopac.net/home.html (accessed Apr 7, 2017).
(47) Schüürmann, G. Quantitative Structure-Property Relationships for the Polarizability, Solvatochromic Parameters and Lipophilicity. *Quant. Struct.-Act. Relat.* **1990**, *9*, 326−333.
(48) Coulson, C. A.; Longuet-Higgins, H. C. The Electronic Structure of Conjugated Systems. II. Unsaturated Hydrocarbons and Their Hetero-Derivatives. *Proc. R. Soc. London, Ser. A* **1947**, *192*, 16−32.
(49) Fukui, K.; Kato, H.; Yonezawa, T. A New Quantum-Mechanical Reactivity Index for Saturated Compounds. *Bull. Chem. Soc. Jpn.* **1961**, *34*, 1111−1115.
(50) Gopinathan, M. S.; Siddarth, P.; Ravimohan, C. Valency and Molecular Structure. *Theor. Chim. Acta* **1986**, *70*, 303−322.
(51) Mulliken, R. S. Electronic Population Analysis on LCAO-MO Molecular Wave Functions. I. *J. Chem. Phys.* **1955**, *23*, 1833−1840.
(52) Mulliken, R. S. Criteria for the Construction of Good Self-Consistent-Field Molecular Orbital Wave Functions, and the Significance of LCAO-MO Population Analysis. *J. Chem. Phys.* **1962**, *36*, 3428−3439.
(53) Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. *Scand. Stat. Theory Appl.* **1979**, *6*, 65−70.
(54) Friedman, M. A Comparison of Alternative Tests of Significance for the Problem of m Rankings. *Ann. Math. Stat.* **1940**, *11*, 86−92.
(55) Friedman, M. A Correction: The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *J. Am. Stat. Assoc.* **1939**, *34*, 109−109.
(56) Friedman, M. The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *J. Am. Stat. Assoc.* **1937**, *32*, 675−701.
(57) Shapiro, S. S.; Wilk, M. B. An Analysis of Variance Test for Normality (Complete Samples). *Biometrika* **1965**, *52*, 591−611.
(58) Mauchly, J. W. Significance Test for Sphericity of a Normal N-Variate Distribution. *Ann. Math. Stat.* **1940**, *11*, 204−209.
(59) Greenhouse, S. W.; Geisser, S. On Methods in the Analysis of Profile Data. *Psychometrika* **1959**, *24*, 95−112.
(60) Huynh, H.; Feldt, L. S. Estimation of the Box Correction for Degrees of Freedom from Sample Data in Randomized Block and Split-Plot Designs. *J. Educ. Behav. Stat.* **1976**, *1*, 69−82.
(61) de Bruyn Kops, C.; Friedrich, N.-O.; Kirchmair, J. Alignment-Based Prediction of Sites of Metabolism. *J. Chem. Inf. Model.* **2017**, *57* (6), 1258−1264.
(62) *OMEGA*, version 2.5.1.4; OpenEye Scientific Software: Santa Fe, NM, 2011; https://www.eyesopen.com (accessed Apr 7, 2017).
(63) Hawkins, P. C. D.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database. *J. Chem. Inf. Model.* **2010**, *50*, 572−584.
(64) RDKit 2016.03.4. https://github.com/rdkit/rdkit/releases/tag/Release_2016_03_4 (Accessed April 7, 2017).
(65) Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 493−500.
(66) Chemistry Development Kit 1.4.19. https://github.com/cdk/cdk/releases/tag/cdk-1.4.19 (accessed Apr 7, 2017).
(67) scikit-learn 0.18. http://scikit-learn.org/0.18/documentation.html (accessed Apr 7, 2017).