Chemical Information

FAME 3: Predicting the Sites of Metabolism in Synthetic Compounds and Natural Products for Phase 1 and Phase 2 Metabolic Enzymes

Martin Šícho, Conrad Stork, Angelica Mazzolari, Christina de Bruyn Kops, Alessandro Pedretti, Bernard Testa, Giulio Vistoli, Daniel Svozil, and Johannes Kirchmair

J. Chem. Inf. Model., **Just Accepted Manuscript** • DOI: 10.1021/acs.jcim.9b00376 • Publication Date (Web): 30 Jul 2019

FAME 3: Predicting the Sites of Metabolism in Synthetic Compounds and Natural Products for Phase 1 and Phase 2 Metabolic Enzymes

Martin Šícho,[1,2] Conrad Stork,[1] Angelica Mazzolari,[3] Christina de Bruyn Kops,[1] Alessandro Pedretti,[3] Bernard Testa,[4] Giulio Vistoli,[3] Daniel Svozil,[2] and Johannes Kirchmair[1,5,6]*

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

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

[3] Università degli Studi di Milano, Facoltà di Scienze del Farmaco, Dipartimento di Scienze Farmaceutiche "Pietro Pratesi", I- 20133 Milan, Italy

[4] Emeritus Professor, University of Lausanne, 1015 Lausanne, Switzerland

[5] University of Bergen, Department of Chemistry, N-5020 Bergen, Norway

[6] University of Bergen, Computational Biology Unit (CBU), N-5020 Bergen, Norway

[*]J. Kirchmair. E-mail: kirchmair@zbh.uni-hamburg.de. Tel.: +49 40 42838 7303.

ABSTRACT

In this work we present the third generation of FAst MEtabolizer (FAME 3), a collection of extra trees classifiers for the prediction of sites of metabolism (SoMs) in small molecules such as drugs, drug-like compounds, natural products, agrochemicals and cosmetics. FAME 3 was derived from the MetaQSAR database (Pedretti, A.; Mazzolari, A.; Vistoli, G.; Testa, B. MetaQSAR: An Integrated Database Engine to Manage and Analyze Metabolic Data. *J. Med. Chem.* **2018**, *61*, 1019–1030), a recently published data resource on xenobiotic metabolism that contains more than 2,100 substrates annotated with more than 6,300 experimentally confirmed SoMs related to redox reactions, hydrolysis and other non-redox reactions, and conjugation reactions. In tests with holdout data, FAME 3 models reached competitive performance, with Matthews correlation coefficients (MCCs) ranging from 0.50 for a global model covering phase 1 and phase 2 metabolism, to 0.75 for a focused model for phase 2 metabolism. A model focused on cytochrome P450 metabolism yielded an MCC of 0.57. Results from case studies with several synthetic compounds, natural products and natural product derivatives demonstrate the agreement between model predictions and literature data even for molecules with structural patterns clearly distinct from those present in the training data. The applicability domains of the individual models were estimated by a new, atom-based distance measure ("FAMEscore") that is based on a nearest neighbor search in the space of atom environments. FAME 3 is available via a public web service at https://nerdd.zbh.uni-hamburg.de/ and as a self-contained Java software package, free for academic and non-commercial research**.**

INTRODUCTION

Detailed understanding of the metabolic fate of small molecules is essential to the development of safe and efficacious drugs, cosmetics and agrochemicals. A wide range of advanced in vitro and in vivo methods are at our disposal today. Paired with powerful analytical methods, they allow the determination of small-molecule metabolism at an unprecedented level of detail but remain resource-demanding.[1] At the same time, increasingly mature in silico methods for the prediction of (i) the interaction of xenobiotics with metabolic enzymes, (ii) atom positions in small molecules liable to metabolism (i.e. sites of metabolism; SoMs) and (iii) the molecular structures of likely metabolites are becoming available.[1–4]

Machine learning approaches have shown high potential in modeling the increasingly large and complex sets of measured data on xenobiotic metabolism.[5] One of the best-explored categories of models in this context is predictors of SoMs, several of which are accessible as free web services or software packages. The most prominent examples of free tools for SoM prediction include SMARTCyp,[6,7] XenoSite,[8] SOMP[9] and FAME.[10,11]

Most SoM predictors are limited to cytochrome P450 (CYP) mediated metabolism, which is related to the important role of CYPs in xenobiotic metabolism but also to the fact that substantially more measured data are available for this family of metabolizing enzymes than to any other.[12] In recent years, specialized predictors of SoMs related to biotransformations catalyzed by uridine 5'-diphospho-glucuronosyltransferases (UGTs) have been reported. These have been integrated, for example, into SOMP[9] and XenoSite.[13] A machine learning model for the prediction of SoMs related to redox, conjugation and further types of metabolic reactions is MetScore.[14] MetScore has undergone thorough validation and yields a high prediction accuracy.

However, the model has not been released for use by the scientific community. A second SoM predictor covering a broad range of phase 1 and phase 2 enzymes is FAME. The predictor is trained on the Metabolite Database,[15] which has recently been discontinued. FAME is based on a random forest approach that relies on just seven simple descriptors to encode essential atom properties such as electronegativity, atom type or its steric accessibility from the perspective of the enzyme's reaction center. The successor of FAME, FAME 2,[11] utilizes circular descriptors that include atom type information as well as numerical values for partial charges, hybridization states, atom accessibility and other features of the encoded atom and its neighbors. Additionally, in FAME 2 the random forest algorithm has been replaced by an extremely randomized trees algorithm. Together, these enhancements produced more descriptive, accurate and robust models, which allowed for a substantial reduction of required training instances and, hence, for the use of smaller-sized, non-commercial data sets. Specifically, FAME was trained on more than 20,000 molecules with computationally annotated SoMs whereas FAME 2 was trained on a revised version[16] of the manually curated Zaretzki data set consisting of 678 compounds.[8] However, as a result of the scarcity of the publicly available data, FAME 2 is limited to CYP-mediated metabolism only. Besides the quantity of the data and the limitation to CYPs, two further important differences must be observed. First, for the large training set of FAME, SoMs were assigned by an automated approach based on the structural differences observed between the parent compound and its known metabolites, whereas in the case of FAME 2, the SoMs were manually assigned by experts. This is a qualitative difference, because mechanistic SoMs can only be predicted with a few methods, including FAME 2.

Here, we present FAME 3, which aims to address three major constraints shared by most SoM predictors: (i) the limited coverage of metabolizing enzymes and reactions, (ii) the absence of

means to estimate prediction accuracy for individual atoms and (iii) the limited accessibility for the use of the models by the scientific community. FAME 3 is trained on a new, comprehensive data set of drug-like molecules annotated with expert-curated SoMs, originating from the MetaQSAR database.[17] Compared to FAME 2, FAME 3 enables the prediction of SoMs not only for CYP-mediated metabolism but also for other types of phase 1 metabolism and for phase 2 metabolism. In addition, FAME 3 features a new method for the estimation of the reliability of predictions for individual atom positions of query molecules. FAME 3 is available via a public web service at https://nerdd.zbh.uni-hamburg.de/ and as a stand-alone software package free of charge for academic and non-commercial research.

METHODS

Data Sources and Preprocessing

MetaQSAR served as the data source for model development. It is a manually compiled resource of published measured data on xenobiotic metabolism, including expert-curated SoMs and reaction annotations for discovery compounds and drugs. The version of the database used in this work[18] contains 2,314 compounds with annotated SoMs that were compiled from articles published in *Chemical Research in Toxicology* (2004-2012), *Xenobiotica* (2004-2012) and *Drug Metabolism and Disposition* (2004-2015). The reactions covered in MetaQSAR are divided into three main reaction classes: redox reactions (3,458 reactions), hydrolysis and other non-redox reactions (640 reactions), and conjugation reactions (1,302 reactions). While the first two main reaction classes are associated with enzymes taking part in phase 1 metabolism, the third group consists of reactions specific to phase 2 enzymes. This allows the separation of phase 1 and phase 2 reactions during modeling. The individual reaction classes are further divided into

reaction subclasses (Tables S1, S2 and S3), which enables training of models focused on individual well-represented biotransformations. Only compounds meeting all of the following criteria were considered for model building and testing (the numbers in brackets report the numbers of molecules not meeting the respective criterion):

- Has at least one experimentally confirmed SoM annotated (32)
- Has a molecular weight between 100 and 1000 Da (79)
- Does not consist of element types other than C, N, S, O, H, F, Cl, Br, I, P, B, Si (20)
- Can be successfully parsed and descriptors successfully calculated by the Chemistry Development Kit (CDK)[19,20] framework (16)

Therefore, the preprocessed data set ("FAME 3 data set") consisted of 2,167 compounds in total.

Regarding the annotation of SoMs, the following special cases are considered: If a metabolic transformation relates to a bond rather than a single atom, the participating atoms are considered individually, and both are labeled as SoMs in the training data. This is to provide the model with more information on the participating atoms, which can each be subjected to different effects of the corresponding atomic environment. Overall this should render a more complete picture about the effects determining the reactivity of the bond, but also, and perhaps more importantly, indicate which atoms react together when they are connected by a particular bond type. It is also possible that more than one reaction subclass or enzyme is assigned to a single atom. In such cases, the atom is considered as a valid SoM if at least one reaction class or enzyme known for that atom is relevant for the model being built.

Descriptors

Calculation of Descriptors and Atom Type Fingerprints

Circular atom descriptors combined with atom type fingerprints ("circCDK+ATF") were calculated with CDK as reported in ref 11. However, rather than exploring the "circCDK + ATF" descriptors of up to only six layers, as we did for FAME 2,[11] in this study, we explored descriptors with up to ten layers. This allowed to gain a better understanding of the impact of higher descriptor complexity on model performance and come up with a precise approach for the definition of the applicability domain.

FAME fingerprint

A new circular, atom-based binary fingerprint ("FAME fingerprint") was generated by assigning a 32-bit string to every combination of a Sybyl atom type and topological distance (bond depth) from a given atom (Figure 1). The first *n* bits in this 32-bit string are switched to "1" if the atom has *n* neighbors of a particular atom type in the given topological distance. The bit string length was chosen taking into consideration the maximum number of atoms of the same type and topological distance observed in molecules of potential interest to small-molecule drug discovery. This means that for the given topological distance, a particular atom type can be present in the molecule up to 32 times (which should cover even rare, large and complex structures). The final atom fingerprint is created by concatenating the list of all 32-bit strings sorted by atom type and distance. Overall, there were 23 distinct atom types in the training set and, thus, the largest fingerprint generated for environments up to the bond depth of 10 (i.e. 11 layers) had a total of 8096 bits (resulting from 23 atom types × 32 bits describing the atom neighborhood × 11 layers).

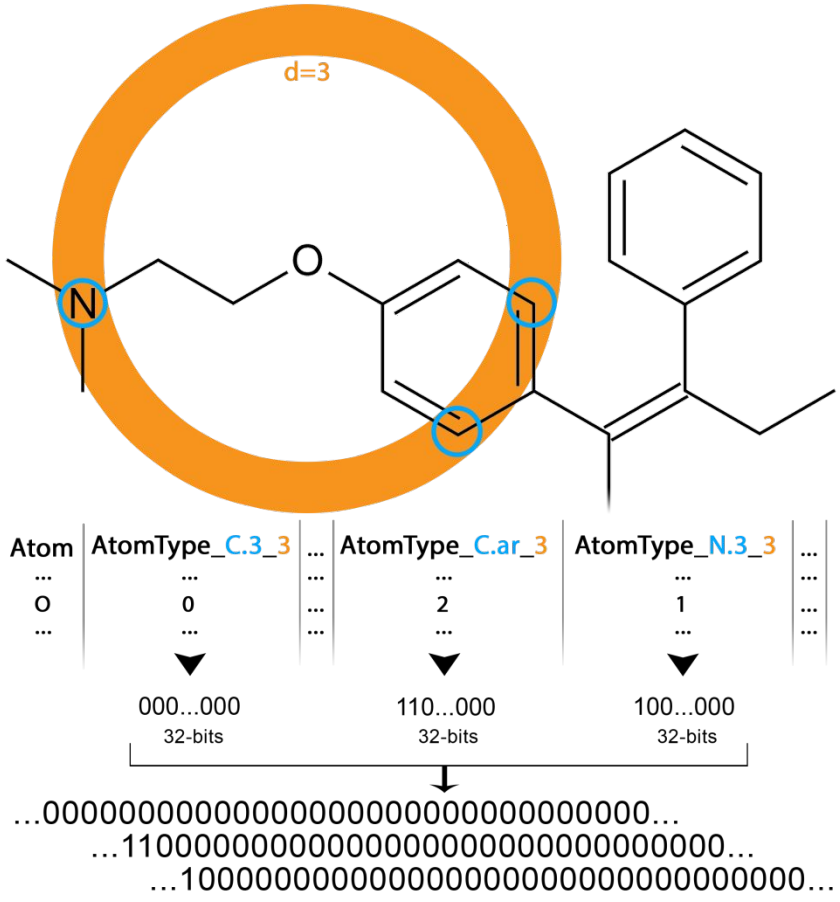

Figure 1: Illustration of the binarization scheme underlying the FAME fingerprint. An example is given for the oxygen atom in tamoxifen and the third layer (bond depth, d) of the binary fingerprint. The oxygen atom has two aromatic carbon atoms (Sybyl atom type C.ar) in a distance of three bonds. Therefore, the first two bits are turned on in the 32-bit string that represents the C.ar type residing in a distance of three bonds from the encoded atom. Similarly, only the first bit is turned on for the 32-bit string associated with the sp^3 hybridized nitrogen (Sybyl atom type N.3) in the same distance. No bits are turned on in the 32-bit string associated with atom types that are not present 3 bonds away from the oxygen atom (such as the sp^3 hybridized carbon shown in this example). All of the generated 32-bit strings are then concatenated to form the FAME fingerprint.

Model Building

Model construction and related data analysis tasks were implemented using scikit-learn.[21] Prior to any model development, 80% of the molecules of the individual preprocessed data sets were dedicated for model training (training sets) and 20% for testing (test sets; holdout data) by a random split of the original data set. All molecules were checked for topologically symmetric atoms as described in ref 11.

Extremely randomized trees were already successfully utilized in FAME 2 development and were also used in the present study. However, in this study the number of trees was reduced from 500 to 250. Also, the decision threshold was set to a fixed value of 0.4 (this threshold value was found to work best for almost all models in FAME 2[11]). Because the value of the *class_weight* parameter (which enables different class balancing strategies within the extra trees model) was found to have little effect on model performance, the value of this parameter was also kept fixed in FAME 3 rather than optimized as in FAME 2.[11] During the optimization of the "circCDK+ATF" models of FAME 2, the most commonly chosen value for the *class_weight* parameter was “balanced_subsample”. Thus, we decided to use this setting for all FAME 3 models. As in FAME 2,[11] the *max_features* and *max_features_ANOVA* parameters were optimized during cross-validation, although the size of the parameter grid was slightly reduced (Table 1). While *max_features* is a parameter of the extra trees classifier (it is the maximum number of available features to consider when searching for the optimum split), *max_features_ANOVA* affects a data preprocessing step which is useful in removing irrelevant features and, thus, reducing computation complexity and removing potential sources of noise from the data set (see ref 11 for details of the feature selection step).

Table 1: Overview of Model Hyperparameters and Their Values Optimized During Grid Search.

Parameter	Explored values
max_features	0.3, 0.6, 0.9
max_features_ANOVA	200, 400

Measures for the Evaluation of Model Performance

Model performance was assessed by the Matthews correlation coefficient (MCC), the area under the receiver operating characteristic curve (AUC), and the Top-*k* metric. The MCC is a balanced measure that takes into account the proportion of all classes in the confusion matrix. It is generally considered one of the best measures of performance of binary classifiers and hence has been used in this study as the primary metric for performance assessment. The AUC quantifies the ability of a model to correctly rank SoMs and non-SoMs based on the probabilities given by the ensemble approach. Related to the AUC measure, the Top-*k* metric denotes the percentage of molecules for which at least one known SoM is listed among the *k* highest-ranked atom positions in a molecule (again, the ranking is based on probabilities given by the ensemble approach). In the context of SoM prediction, the most commonly applied value for *k* is 2 (which was also used in this study).

FAMEscore Atom-based Distance Measure

FAMEscore is an atom-based distance measure. It is calculated with a *k*-nearest neighbor approach that determines the distance of a query atom (defined as Tanimoto coefficient calculated on the FAME fingerprint) to a defined number of nearest atoms in the training set (Eq 1):

$$FAMEscore = 1 - \frac{\sum_{i=1}^{k} d_i}{k} \qquad \text{(Eq 1)}$$

Here, k corresponds to the number of nearest neighbors and was set to 3 for our experiments. $d_i$is the Tanimoto distance between the i-th nearest neighbor and the query atom. Therefore, the closer this metric is to 1, the more examples of similar atoms are present in the training data.

RESULTS & DISCUSSION

Data Analysis

The preprocessed data set derived from MetaQSAR ("FAME 3 data set"; see Methods for details) consists of a total of 2,167 substrates annotated with 6,307 experimentally confirmed SoMs (Table 2, "FAME 3 P1+P2 data set"). Among these, 1,106 compounds are CYP substrates annotated with 3,517 SoMs (Table 2, "FAME 3 CYPs subset"). Compared to the Zaretzki data set (678 substrates annotated with 1,672 SoMs), which was used to develop FAME 2, the MetaQSAR database contains nearly twice as many CYP substrates and covers more CYP-related SoMs, suggesting that the CYP-related SoM data utilized by FAME 3 are more complete. In addition, the FAME 3 data set contains 622 substrates annotated with 1,551 SoM records related to phase 1 metabolism mediated by enzymes other than CYPs, and 784 substrates annotated with 1,239 SoMs related to phase 2 metabolism (Table 2, "FAME 3 P2 subset"). Natural products remain the most productive resource of inspiration for the development of new small-molecule drugs.[22–24] In order to understand the extent to which MetaQSAR covers synthetic compounds and natural products, we employed NP-Scout, a random forest-based classifier developed by some of us.[25] According to NP-Scout, 58% of the substrates included in

the data set are predicted to be of synthetic origin (i.e. they are assigned a probability of less than 0.5 of belonging to the class of natural products; Figure 2). For 22% of all compounds, a minimum probability of 0.9 for a compound to belong to the class of natural products was calculated. Overall, this indicates a solid representation of natural products by the FAME 3 data set.

Table 2: Comparison of the FAME 3 and the Zaretzki Data Sets.[a]

	No. of substrates	No. of atoms	No. of SoMs	SoMs per molecule	SoM %[b]
FAME 3 P1+P2 data set	2,167	49,045	6,307	2.91	12.9%
FAME 3 CYPs subset	1,106	25,581	3,517	3.18	13.8%
FAME 3 P1 subset	1,728	40,192	5,068	2.93	12.6%
FAME 3 P2 subset	784	16,462	1,239	1.58	7.5%
Zaretzki et al.	678	15,233	1,672	2.47	11.0%

[a] All values refer to the preprocessed FAME 3 data set and a revised version[16] of the Zaretzki data set.[8]

[b] Percentage of heavy atoms annotated as SoMs.

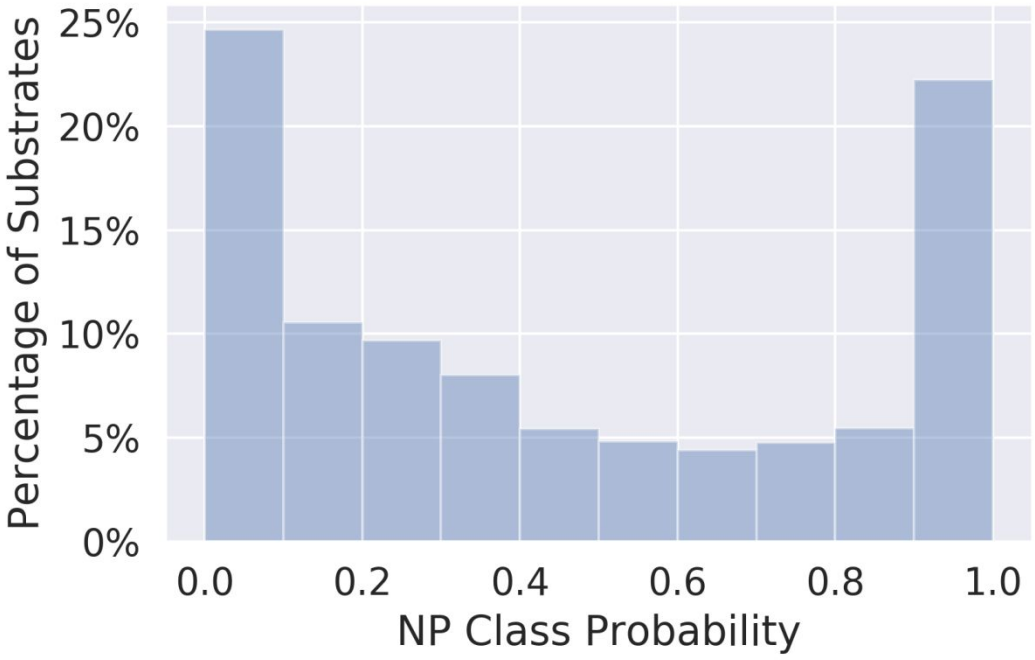

Figure 2: Natural product (NP) class probability distribution in the FAME 3 P1+P2 data set, calculated with NP-Scout.

A further important indicator of the relevance of a data set for modeling is its diversity with respect to the covered atom environments. In order to describe and compare the diversity of atom environments in the FAME 3 and Zaretzki data sets, we developed a new circular, atom-based binary fingerprint ("FAME fingerprint"; see Methods for details) that we used as input for principal component analysis (PCA). In the score plot in Figure 3 it can be observed that the atom environments covered by the Zaretzki et al. data set are essentially a subset of those covered by the FAME 3 data set. In many areas the density of information is higher for the FAME 3 data set, providing better statistical support in particular of reactions that are less frequently observed or that occur in atom environments that are less common. It is therefore

expected that models derived from the FAME 3 data set will benefit from the more complete and fine-graded picture of metabolism with respect to accuracy and domain of applicability.

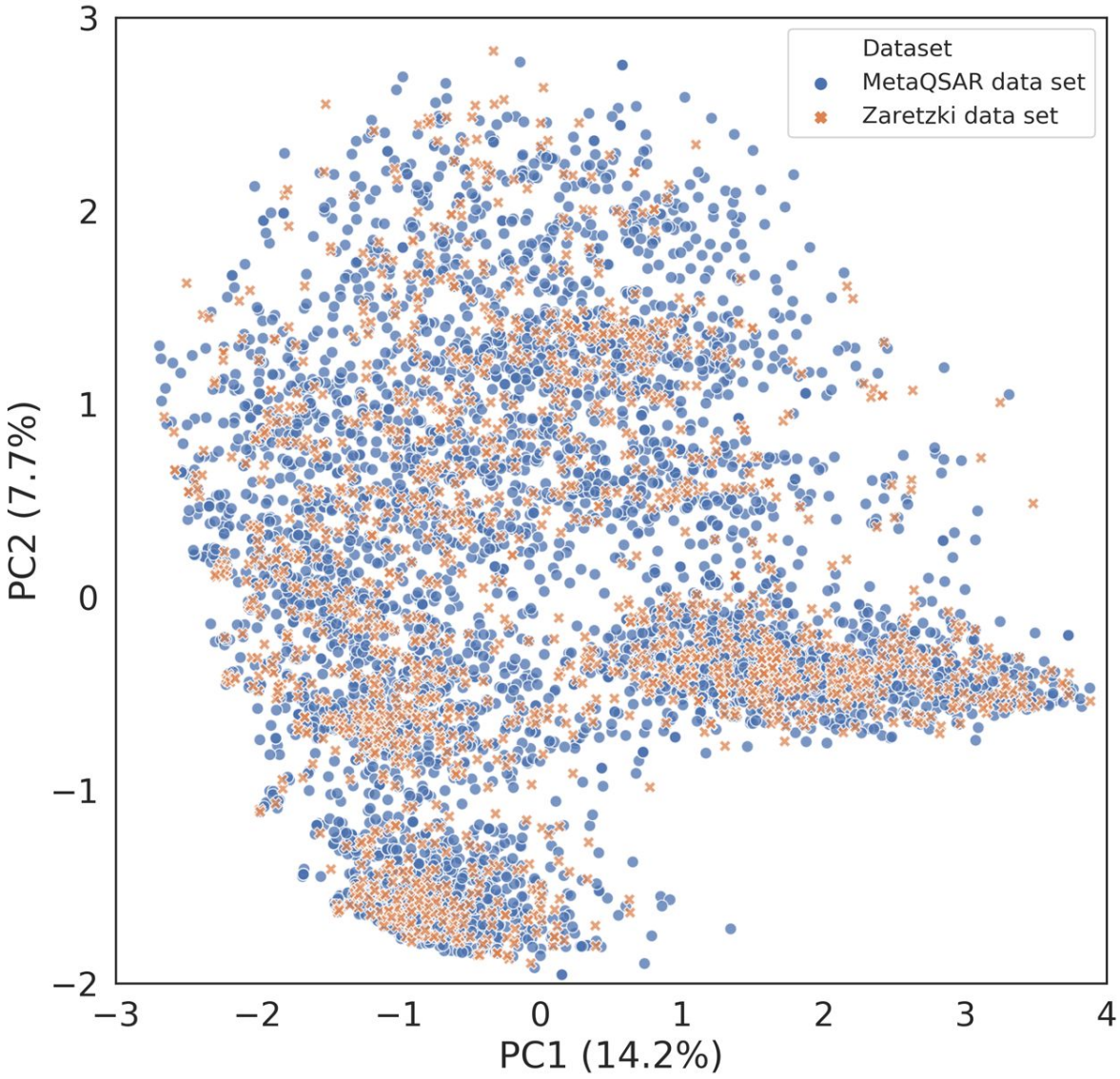

Figure 3: PCA score plot depicting atom neighborhoods in the MetaQSAR database and the revised version of the Zaretzki data set. The plot was generated by projection of the FAME fingerprints (maximum bond depth 6; length 5152 bits) generated for all atoms in both data sets onto a plane defined by the first two principal components (PC1 and PC2). For the sake of clarity only a random 10% sample of projected points from each data set is depicted. Note that the proportion of variance explained by the two principal components (reported in parentheses with

the axis labels) is low. For this reason, the plot should be considered only as a coarse representation of the diversity of the two data sets.

Model Building

Models were built on 80% of the molecules of the individual preprocessed data sets; 20% of the molecules (selected by random split) were held back for testing (Table 3). The machine learning approach was adopted from that of FAME 2.[11] It is based on the extra trees classifier algorithm using a combination of the circular representation of 15 basic 2D CDK descriptors with circular atom-type fingerprints as inputs ("circCDK + ATF"). Previously, this combination of machine learning algorithms and descriptors resulted in the overall best-performing models.[11]

Table 3: Training and Test Set Sizes for All FAME 3 Models.

Model	Molecules (training set)	Molecules (test set)	Atoms (training set)	Atoms (test set)
P1+P2	1733	434	39131	9914
CYP	884	222	20520	5061
P1	1382	346	32313	7879
P2	627	157	12986	3476
P1+P2 100+	1104	277	25786	6491
CYP 100+	763	191	17807	4487
P1 100+	872	219	20636	5335
P2 100+	460	116	9891	2427

Four different types of extremely randomized trees models were developed:

- “P1+P2 model”: Model covering both metabolic phases
- "CYP model": Model covering CYP-mediated metabolic reactions
- "P1 model": Model covering phase 1 metabolic reactions (both CYP and non-CYP)
- “P2 model”: Model covering phase 2 metabolic reactions

In addition, we also constructed models that cover only reaction subclasses represented by at least 100 SoM annotations in the training set. This will allow to determine the impact of the quantity of data available for model building on model accuracy. Typical examples of well-represented reaction subclasses are oxidation reactions of aryl compounds to epoxides, phenols and other metabolites, or O-glucuronidation reactions of alcohols. We refer to these reaction subclass-restricted models as the “P1+P2 100+ model”, “CYP 100+ model”, “P1 100+ model” and “P2 100+ model”. The “P1+P2 100+ model” covers 18 out of 93 reaction subclasses (Tables S1 and S3), the “CYP 100+ model” 5 out of 44 (Table S2), the “P1 100+ model” 13 out of 62 (Table S1), and the “P2 100+ model” 5 out of 31 (Table S3). The low number of reaction subclasses covered by at least 100 SoM annotations shows that data on xenobiotic metabolism are still sparse and, thus, represent a bottleneck in the development of in silico models.

Parameter Optimization

A cross-validated grid search was conducted to identify optimum parameters (see Methods for details). The differences in performance observed across the searched parameter grid were minor, but some trends emerged, nonetheless. For example, the optimal value of the *max_features_ANOVA* parameter, affecting a data preprocessing step useful in removing irrelevant features, was usually 200 for models relying on descriptors with lower bond depth (mostly for bond depth 1) and 400 for models based on descriptors with higher bond depth

(Table S4). This behavior is expected since by branching out further from an atom the model may find more useful patterns that utilize a much wider variety of descriptors. The most commonly selected values for the *max_features* parameter were 0.6 and 0.9, but no clear relationship was observed between the value of this parameter and the bond depth.

Internal Evaluation of the Models by Cross-Validation

Model performance was assessed by the MCC, AUC and Top-*k* metric (see Methods for details). Depending on the fingerprint bond depth, the 10-fold cross-validation MCCs of the "P1+P2 model" were between 0.49 and 0.51, whereas AUC values were around 0.89 and the Top-2 success rate around 82% (Figure 4A and Table S5). The "CYP model" showed similar performance, with MCCs ranging from 0.47 to 0.52 and AUC and Top-2 success rate values around 0.89 and 82%, respectively (Figure 4C and Table S6). The MCCs for the "P1 model" ranged from 0.50 to 0.53, while AUC values were around 0.90 and Top-2 success rates around 82% (Figure 4E and Table S7). In contrast, the dedicated "P2 model" yielded even higher predictive power, reflected by MCCs between 0.70 and 0.72, AUC values around 0.97, and Top-2 success rates around 90% (Figure 4G and Table S8). The higher predictive performance of phase 2 models as compared to phase 1 models is consistent with previous reports[10,14] and can be attributed to the characteristics of phase 2 reactions, which are in general more specific with respect to the atom environments at which they occur.

In our previous study,[11] we showed that the explicit encoding of atom neighborhoods improves model performance. However, improvements beyond the bond depth of 2 were minor in most cases. Similar behavior was also observed in this study (see the curve progression in Figure 4). Interestingly, no substantial increase in performance was observed for models trained and tested only on well-represented reaction subclasses (i.e. reaction subclasses represented by at least 100

SoMs in the training data). The MCCs for these models were higher by a maximum of only 0.05 than for the models trained and tested on all reaction subclasses represented by the training data (Figure 4B, D, F, H and Tables S9 to S12). From this we conclude that the number of annotations per reaction class has only a minor effect on model performance. There likely are other factors at play such as individual enzymes' substrate selectivity and catalytic mechanisms and the structural diversity of the training data available per reaction class. A slightly stronger uptrend in model performance with increasing bond depth was observed for models with reaction subclass restriction as compared to those models without restriction. Models of neither type showed any substantial improvements in performance when going beyond a bond depth of 5. For this reason, we chose this bond depth as the optimum descriptor complexity, and the "circCDK+ATF" models of bond depth 5 were used for performance experiments on the holdout data (see Evaluation of the Final Models on Test Sets).

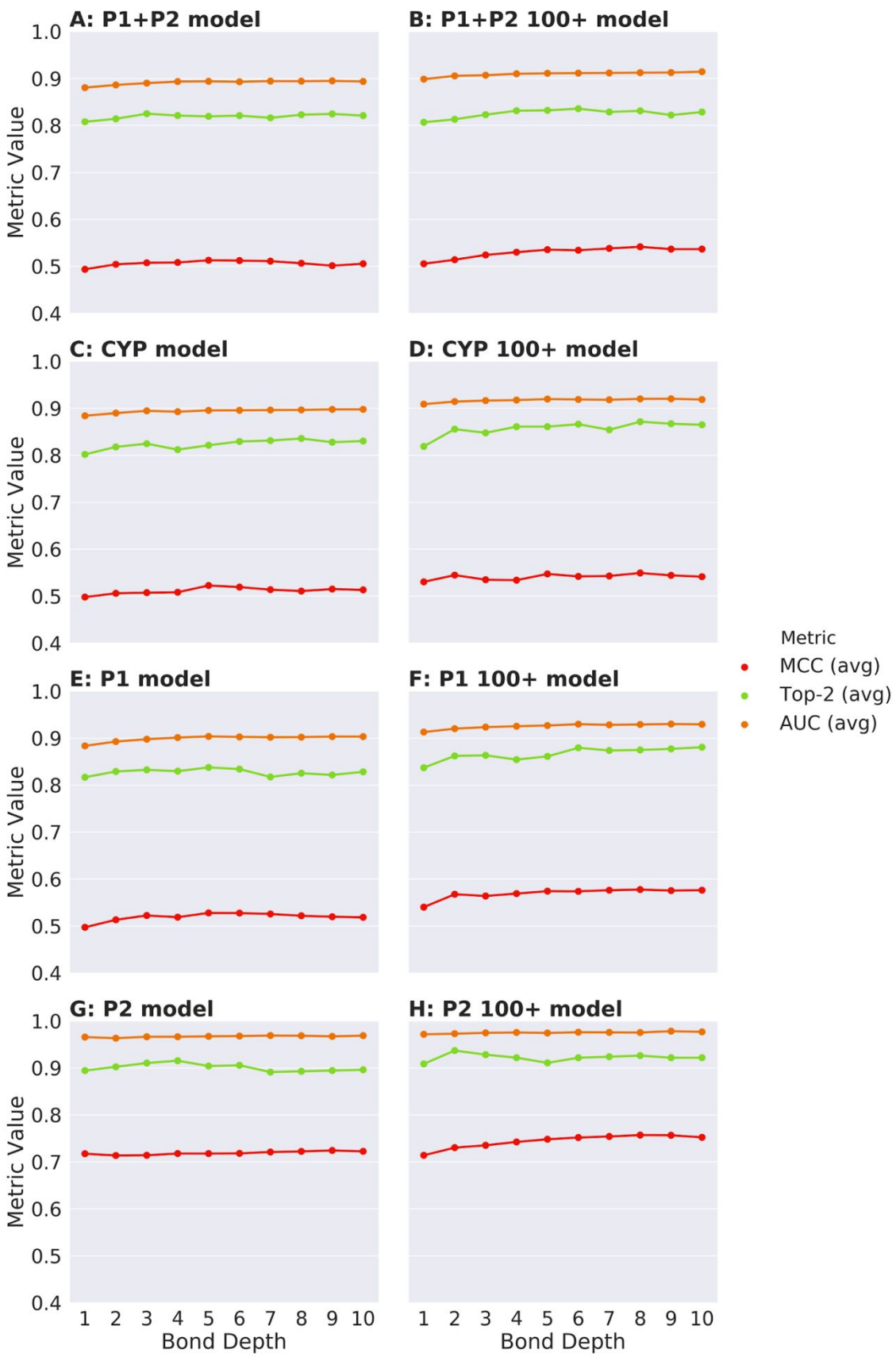
A: P1+P2 model
B: P1+P2 100+ model
C: CYP model
D: CYP 100+ model
E: P1 model
F: P1 100+ model
G: P2 model
H: P2 100+ model
Metric Value
Bond Depth
Metric
MCC (avg)
Top-2 (avg)
AUC (avg)

Figure 4: Dependence of the internal model performance of FAME 3 on the bond depth. Internal performance of FAME 3 models was evaluated by 10-fold cross-validation and all FAME 3 models were built using the "circCDK+ATF" descriptor set. The performance is given as MCC, AUC and Top-2 success rates (in this figure reported as fractions rather than percentages) averaged over 10 cross-validation runs. (A) “P1+P2 model”, (B) "P1+P2 100+ model", (C) "CYP model", (D) "CYP 100+ model", (E) "P1 model", (F) "P1 100+ model", (G) "P2 model" and (H) "P2 100+ model".

Comparison of Model Performance to Established Models

Direct comparison of the performance of FAME 3 with that of FAME 2 is only possible to a limited extent because of the different scopes of the two methods (in particular, FAME 2 is limited to CYP metabolism). Therefore, we compared the selected FAME 3 “CYP model” (based on the "circCDK+ATF" descriptor set with a bond depth of 5) to the equivalent FAME 2 model (identical descriptor set and bond depth). During cross-validation, the FAME 3 model obtained an average MCC of 0.52 whereas the FAME 2 reached an average MCC of 0.57. We attribute the slightly lower MCC obtained by FAME 3 to the higher diversity of the FAME 3 data set (with respect to the diversity of atom environments; see the PCA score plot Figure 3). Direct comparison of the performance of FAME 3 with that of MetScore, a leading SoM predictor capable of handling both phase 1 and phase 2 reactions, must also be considered with caution. The training sets of both models differ substantially with respect to coverage and the annotation approach. The cross-validation MCC average of the FAME 3 “P1+P2 model” with a bond depth of 5 was 0.51, which is comparable to the MCC of 0.53 obtained by the MetScore phase 1 and phase 2 composite model for the MetScore calibration data set.[14] Cross-validation MCC values for the phase 1-specific models suggest that FAME 3 performs slightly weaker than

the corresponding MetScore model (MCC 0.53 versus 0.61). We believe that this could be attributed to the phase 1 training set of MetScore, which roughly contains double as many annotated SoMs as that of FAME 3 (note that for MetScore SoMs have been annotated with a (semi-) automated procedure whereas for the training of FAME 3, SoMs have been manually assigned by experts; also note that the MetScore training data is restricted to well-defined one-step transformations represented by more than 100 instances[14]). Similar factors are believed to be involved in the also slightly lower performance of the phase 2-specific FAME 3 model in comparison to the respective MetScore model (MCCs 0.72 versus 0.76).

Evaluation of the Final Models on Test Sets

In addition to internal validation, FAME 3 models were also validated on holdout data sampled randomly prior to modeling (Tables 3 and 4; see Methods section). For the four models without a reaction subclass restriction (i.e. the "P1+P2 model", "CYP model", "P1 model" and "P2 model"), the performance on the test sets was not worse than that observed during the cross-validation experiments. More specifically, between the cross-validation experiments and testing on unseen data, the MCC, AUC and Top-2 success rate values dropped by a maximum of only 0.01, 0.02 and 1 percentage points, respectively (Table 4). For the "CYP model", a slight increase in performance was noted even (+0.05 in MCC, +0.02 in AUC and +8 percentage points in Top-2 success rate). Overall, these results demonstrate the robustness and good generalization capability of the FAME 3 models.

Table 4: Cross-Validation and Test Set Performance of Selected Models.

Model[a]	MCC (cross-validation)	MCC (test set)	AUC (cross-validation)	AUC (test set)	Top-2 (cross-validation)	Top-2 (test set)
P1+P2	0.51	0.50	0.89	0.90	82%	82%
CYP	0.52	0.57	0.90	0.92	82%	90%
P1	0.53	0.53	0.90	0.88	84%	83%
P2	0.72	0.71	0.96	0.97	90%	92%
P1+P2 100+	0.54	0.55	0.91	0.92	83%	87%
CYP 100+	0.55	0.63	0.92	0.94	86%	86%
P1 100+	0.57	0.52	0.93	0.92	86%	80%
P2 100+	0.75	0.75	0.97	0.97	91%	91%

[a] All models based on the "circCDK+ATF" descriptor set with a maximum bond depth of 5.

The increase in performance of the “CYP model” on the test set may be related to the good coverage of CYP-catalyzed reactions in the MetaQSAR database. Roughly 64% of the phase 1 data are on CYPs. Therefore, it is likely that the CYP data are more complete. Also, the diversity of CYP-catalyzed reactions is lower, which likely boosts the generalization ability of the model. In comparison to the “P1 model” and “P1+P2 model” the “CYP model” also shows good performance for atoms with lower FAMEscore values, which are more abundant in the “CYP model” test set (Table 5 and Figure 5C).

The results obtained for the models with the reaction subclass restriction were, in general, comparable to those obtained during cross-validation. In fact, the MCC, AUC and Top-2 success rate values never dropped below the cross-validation average for all but one model (Table 4). In line with the observations made for the "CYP model", the "CYP 100+" model for well-represented CYP reactions performed slightly better on the test set than during cross-validation (+0.08 in MCC, +0.02 in AUC and identical values for the Top-2 success rates). On the other hand, a minor decrease in performance was observed for the "P1 100+ model" (-0.05 in MCC, -0.01 in AUC and -6 percentage points in Top-2 success rate).

In order to understand the reason as to why the "P1 model" reached comparable performance during cross-validation and testing on holdout data whereas the "P1 100+ model" had a slightly better internal performance, a detailed analysis of the composition of training and test sets is necessary. This is where the developed distance measure, FAMEscore, can be useful (see Methods for details on FAMEscore). Table 5 and the histograms in Figure 5 demonstrate that the test set for the "P1 model" contains more atom environments closely related to the training data (higher FAMEscore) and less dissimilar atom environments (lower FAMEscore). In the case of the P1 100+ test set the opposite is true. The test set has a higher proportion of atom environments that are dissimilar from those in the training data and fewer atom environments that are closer to the training set (Figure 5F and Table 5). This likely makes the P1 100+ test set more challenging, which is reflected by the lower performance of the model.

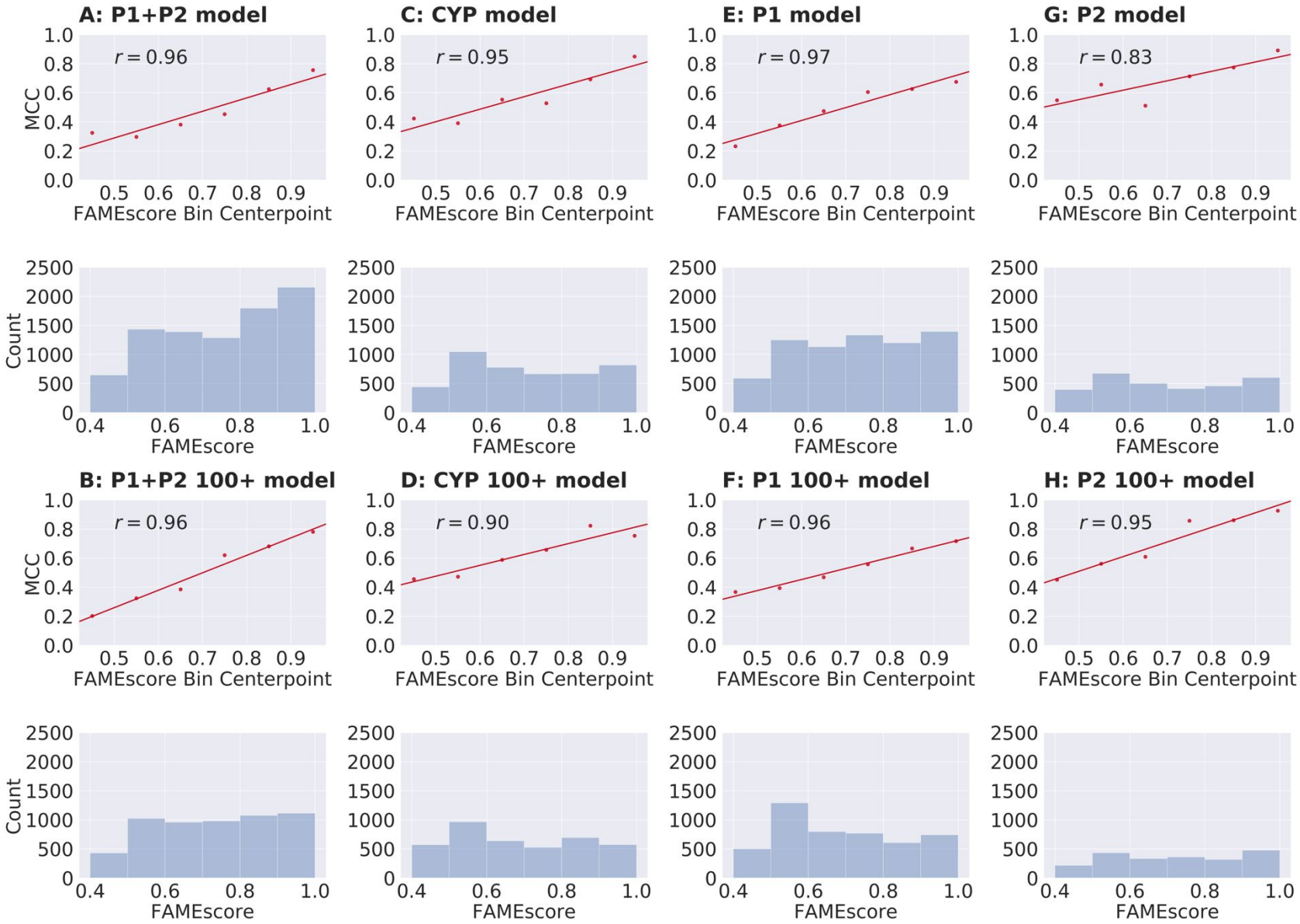
A: P1+P2 model
C: CYP model
E: P1 model
G: P2 model
B: P1+P2 100+ model
D: CYP 100+ model
F: P1 100+ model
H: P2 100+ model
r = 0.96
r = 0.95
r = 0.97
r = 0.83
r = 0.96
r = 0.90
r = 0.96
r = 0.95
MCC
FAMEscore Bin Centerpoint
Count
FAMEscore

Figure 5: Test set composition and its impact on performance indicators (for models based on the “circCDK+ATF” descriptor set of a maximum bond depth of 5). The graph reports the regression line based on the MCCs obtained by the individual models as a function of the coverage of atom environments of the test set by the training data. For this purpose, the individual test sets were binned according to FAMEscore values and the MCC was calculated for each bin. In the graph, each bin is represented by its FAMEscore center point on the x-axis while the MCC value calculated for each bin is indicated on the y-axis. Pearson’s correlation coefficient (r) between FAMEscore center points and the calculated MCC is shown in the top left corner of each graph. The histograms show the FAMEscore distribution among the atoms in each test set. The total number of atoms in each bin is indicated on the y-axis and the endpoints of each bin are indicated on the x-axis. (A) “P1+P2 model”, (B) "P1+P2 100+ model", (C) "CYP model", (D) "CYP 100+ model", (E) "P1 model", (F) "P1 100+ model", (G) "P2 model" and (H) "P2 100+ model".

Evaluation of the Applicability Domain Score on Test Sets

It is important to understand the coverage of the specific molecule or atom of interest by the training data but few SoM predictors provide such information. A notable exception is SMARTCyp, which, in its third version, offers a fingerprint-based method to calculate the similarity between the matched substructure in the input structure and the exact molecule fragment used for the SoM prediction.[7]

We explored the applicability of FAMEscore as an estimator of prediction accuracy. FAMEscore is a measure of how far a sample atom is from the training data and, thus, how easy it should be for the model to make a correct prediction for it. Therefore, the FAMEscore values obtained for atoms in the test sets should correlate with model performance.

The regression lines in Figure 5 estimate the MCC of models as a function of binned FAMEscore values for each test set. From this graphical representation and also from the Pearson's correlation coefficients calculated for the MCC and the FAMEscore center points, a clear linear relationship between FAMEscore and model performance is apparent across all test sets.

All types of models built in this study obtained high MCCs for atom environments with a high FAMEscore. This was to be expected since high FAMEscore values should be associated with atom environments that are well-represented by the training data and, thus, the model performance should be considerably better for such atoms. Conversely, atom environments with a low FAMEscore should be less often correctly predicted, which was also confirmed in our experiments. For all models in this study, MCCs between 0.63 and 0.93 were achieved if the atoms belonged to a bin with a FAMEscore higher than 0.8 (Table 5). On average, 36% of atoms in all test sets satisfied this condition. On the other hand, MCCs between 0.20 and 0.66 were

recorded for atoms belonging to bins with FAMEscore lower than 0.6. On average, this affected 32% of atoms in our test sets.

Table 5: MCCs Obtained for Subsets of the Test Set that are Represented by the Training Data to Varying Degrees.

Model[a]	FAMEscore $\leq$ 0.6 %[b]	FAMEscore $\geq$ 0.8 %[c]	Min. MCC (FAMEscore $\leq$ 0.6)[d]	Max. MCC (FAMEscore $\leq$ 0.6)[e]	Min. MCC (FAMEscore $\geq$ 0.8)[f]	Max. MCC (FAMEscore $\geq$ 0.8)[g]
P1+P2	24%	45%	0.30	0.32	0.63	0.76
CYP	34%	33%	0.39	0.42	0.69	0.85
P1	27%	37%	0.23	0.38	0.63	0.67
P2	36%	35%	0.55	0.66	0.77	0.89
P1+P2 100+	27%	39%	0.20	0.32	0.68	0.78
CYP 100+	39%	32%	0.46	0.47	0.75	0.82
P1 100+	38%	28%	0.37	0.39	0.67	0.72
P2 100+	31%	37%	0.45	0.56	0.86	0.93
Average	32%	36%	0.37	0.44	0.71	0.80
Min	24%	28%	0.20	0.32	0.63	0.67
Max	39%	45%	0.55	0.66	0.86	0.93

[a] All models based on the "circCDK+ATF" descriptor set with a maximum bond depth of 5.

[b] Percentage of atoms in the test set with FAMEscore lower than or equal to 0.6.

[c] Percentage of atoms in the test set with FAMEscore higher than or equal to 0.8.

[d] Minimum MCC calculated for a bin of the test set where the FAMEscore for all atoms in this bin is lower than or equal to 0.6.

[e] Maximum MCC calculated for a bin of the test set where the FAMEscore for all atoms in this bin is lower than or equal to 0.6.

[f] Minimum MCC calculated for a bin of the test set where the FAMEscore for all atoms in this bin is higher than or equal to 0.8.

[g] Maximum MCC calculated for a bin of the test set where the FAMEscore for all atoms in this bin is higher than or equal to 0.8.

Case Studies

The utility of FAME 3 in real world applications was assessed in two sets of case studies. In both of these sets, the FAME 3 "P1+P2 model" with bond depth 5 and trained on the complete (preprocessed) MetaQSAR database (rather than the FAME 3 training set presented above) was used to predict SoMs for various molecules. In the first set of case studies, high-quality data set consisting of drug-like molecules and their metabolites compiled and published by Finkelmann et al.[14] ("MetScore Validation Set") was utilized. Data used for the second set contained interesting cases of pharmaceutically relevant natural products and their derivatives. We refrained from defining thresholds for class assignment and FAMEscore (applicability domain) in order to avoid introducing a bias to these case studies. Instead, we focused on analyzing the ability of the models to correctly rank sites of metabolism early in the ordered list of atoms of the individual query molecules.

MetScore Validation Set

It was determined that seventeen out of the 24 compounds of the original MetScore validation set are part of the FAME 3 training set (the complete preprocessed MetaQSAR database in this case) and were therefore not included in the case studies. In addition, paritaprevir was excluded because of incomplete data on its metabolism.[26] Thus, in the first set of case studies predictions of metabolic liability of atoms in six molecules were investigated (Figure 6A-F). It should be noted that the MetScore validation set has been deemed as very challenging by its authors.[14] In addition, most of the atoms in this set are characterized by generally lower FAMEscore values. Therefore, we can expect the data set to be challenging for FAME 3 as well.

AZD1 (Figure 6A) is a selective glucokinase activator with seven annotated SoMs related to phase 1 and one SoM related to phase 2 metabolism. FAMEscore values lower than 0.65 indicate that the atom environments of this compound differ somewhat from those included in the training data. The FAME 3 model placed seven out of the nine true SoMs at the top of the generated rank-ordered list, including C.14, C.35, C.33, C.24, C.11, C.30 and C.15. However, interestingly, C.34 is ranked at the very top of the list, although not labeled as a SoM in the MetScore validation set. MetScore was also quite successful in this case and correctly labelled five out of the nine SoMs. Unfortunately, neither FAME 3 nor MetScore predicted well the phase 1 SoM at C.12 and the phase 2 SoM at O.13. Nevertheless, both methods perform reasonably well in this case and we see some agreement between them.

AZD7 (Figure 6B) is a chemokine receptor 2 antagonist with three annotated phase 1 SoMs and one phase II SoM. The FAMEscore values of AZD7 atoms were again rather low (mostly below 0.5). Nevertheless, the FAME 3 model fared quite well and correctly placed O.11 and C.10 at the second and third position in the ranked-ordered list, respectively. MetScore was also successful

in this case and labeled O.11 and C.10 correctly. MetScore also predicted S.12 as a SoM, which is at the top of the ranked-ordered list of FAME 3 as well. This is intriguing since S.12 is not labeled as a true SoM in the MetScore validation set. However, the fact that both tools predicted this atom as a possible SoM could be an important indication that an unreported metabolite could form here. In addition, MetScore also labeled C.8 as a SoM, which is also quite high on the atom list generated by FAME 3, so further research of this potential site could also be warranted by these results. Unfortunately, both the FAME 3 model and MetScore failed to identify C.24 and N.22 as SoMs.

Dasabuvir (Figure 6C) is a non-nucleoside inhibitor of the hepatitis C virus RNA-dependent RNA polymerase. The MetScore Validation Set lists four phase 1 SoMs for dasabuvir and one phase 2 SoM for one of its metabolites. The FAMEscore values are not particularly high for this compound either (around 0.55 or lower). However, FAME 3 was still able to rank all experimental SoMs at the very top of the sorted list. MetScore was less successful in this example and only predicted one true SoM (located at either C.12, C.13 or C.14 due to symmetry). This SoM is involved in the formation of the hydroxylated metabolite of dasabuvir which then undergoes a phase 2 transformation. The oxygen atom involved in this phase 2 reaction was also correctly predicted by both FAME 3 and MetScore (data not shown).

Epacadostat (Figure 6D) has two phase 1 SoMs and one phase 2 SoM annotated in the MetScore Validation set. It is clear from the FAMEscore values (often lower than 0.4) that many atom environments in this molecule are quite far from those represented by the training data. The accuracy of the predictions also reflects this fact. There is no clear relationship between the order of atoms in the list and the recorded experimental results. The phase 2 SoM O.8 and the phase 1 SoM N.7 are both ranked at high positions in the list but are still preceded by three non-SoM

atoms. C.19, a phase 1 SoM, is also ranked poorly. MetScore correctly labeled N.7 as a SoM but also failed to recognize the remaining SoMs. It should also be noted that some of the transformations recorded for this molecule are rare, as the authors of MetScore point out in their paper.[14]

Levobunolol (Figure 6E) is a nonselective beta-adrenoceptor antagonist with two annotated phase 1 SoMs (note that in the case of atoms C.5 and O.4 the SoM is not unambiguously defined in the available literature, for which reason both atoms are highlighted in this case[27]) and two annotated phase 2 SoMs. The FAMEscore values for this molecule are not high but never drop below 0.6. FAME 3 correctly ranks O.21, O.4 and C.5 at the top of the rank-ordered list. Both O.21 and C.5 are also correctly marked by MetScore. MetScore in addition labels C.1, C.3 and C.6 as potential SoMs but they are not recorded as such in the validation set. FAME 3 ranks those three atoms higher than others, but the calculated probabilities are still rather low. Unfortunately, both algorithms failed to highlight nitrogen N.18 as a phase 2 SoM and carbons C.14, C.15 and C.16 as phase 1 SoMs.

A compound named “mGluR5 compound 2” (Figure 6F), an allosteric modulator of metabotropic glutamate receptor subtype 5, has two phase 2 SoMs assigned in the MetScore Validation Set. The FAMEscore values are higher for the atoms in this molecule (mostly between 0.7 and 0.8), likely prompted by the presence of several examples of compounds with similar scaffolds in the MetaQSAR database. Therefore, we can expect the FAME 3 model to give more reliable results, which it does. Both C.7 and C.8 are correctly placed at the top of the rank-ordered list. MetScore only marks O.28 in this case, but this atom is not labeled as a SoM in its validation set.

Concluding on this case study, the MetScore Validation Set proves challenging to SoM prediction methods. However, it was shown that despite the low similarity between the atom environments of molecules of the MetScore Validation Set and those of the training data, the FAME 3 model was in many cases able to rank atoms according to their metabolic lability quite reliably.

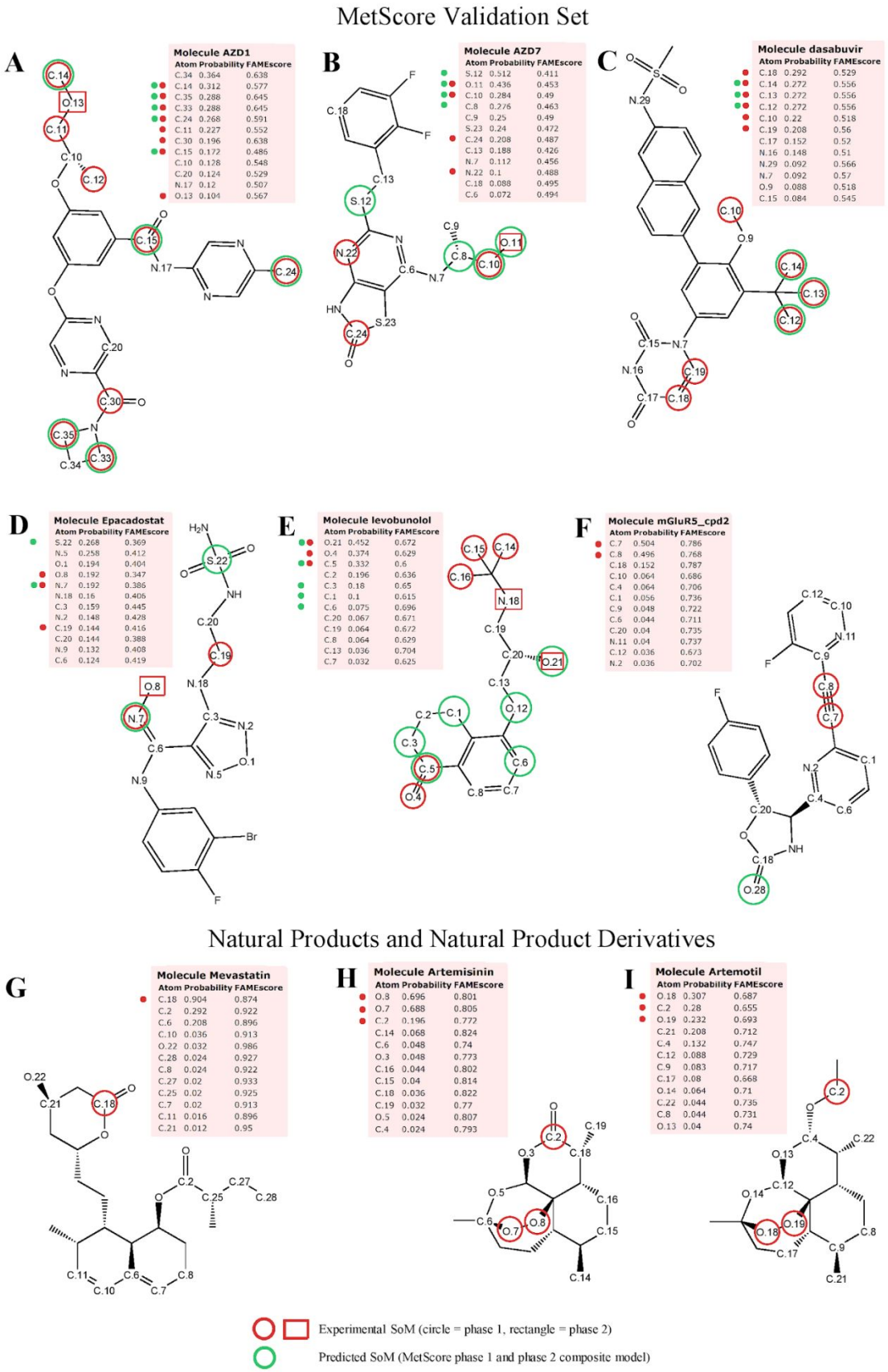
MetScore Validation Set
A
Molecule AZD1
Atom Probability FAMEscore
C.34 0.364 0.638
C.14 0.312 0.577
C.35 0.288 0.645
C.33 0.288 0.645
C.24 0.268 0.591
C.11 0.227 0.552
C.30 0.196 0.638
C.15 0.172 0.486
C.10 0.128 0.548
C.20 0.124 0.529
N.17 0.12 0.507
O.13 0.104 0.567
B
Molecule AZD7
Atom Probability FAMEscore
S.12 0.512 0.411
O.11 0.436 0.453
C.10 0.284 0.49
C.8 0.276 0.463
C.9 0.25 0.49
S.23 0.24 0.472
C.24 0.208 0.487
C.13 0.188 0.426
N.7 0.112 0.456
N.22 0.1 0.488
C.18 0.088 0.495
C.6 0.072 0.494
C
Molecule dasabuvir
Atom Probability FAMEscore
C.18 0.292 0.529
C.14 0.272 0.556
C.13 0.272 0.556
C.12 0.272 0.556
C.10 0.22 0.518
C.19 0.208 0.56
C.17 0.152 0.52
N.16 0.148 0.51
N.29 0.092 0.566
N.7 0.092 0.57
O.9 0.088 0.518
C.15 0.084 0.545
D
Molecule Epacadostat
Atom Probability FAMEscore
S.22 0.268 0.369
N.5 0.258 0.412
O.1 0.194 0.404
O.8 0.192 0.347
N.7 0.192 0.386
N.18 0.16 0.406
C.3 0.159 0.445
N.2 0.148 0.428
C.19 0.144 0.416
C.20 0.144 0.388
N.9 0.132 0.408
C.6 0.124 0.419
E
Molecule levobunolol
Atom Probability FAMEscore
O.21 0.452 0.672
O.4 0.374 0.629
C.5 0.332 0.6
C.2 0.196 0.636
C.3 0.18 0.65
C.1 0.1 0.615
C.6 0.075 0.696
C.20 0.067 0.671
C.19 0.064 0.672
C.8 0.064 0.629
C.13 0.036 0.704
C.7 0.032 0.625
F
Molecule mGluR5_cpd2
Atom Probability FAMEscore
C.7 0.504 0.786
C.8 0.496 0.768
C.18 0.152 0.787
C.10 0.064 0.686
C.4 0.064 0.706
C.1 0.056 0.736
C.9 0.048 0.722
C.6 0.044 0.711
C.20 0.04 0.735
N.11 0.04 0.737
C.12 0.036 0.673
N.2 0.036 0.702
Natural Products and Natural Product Derivatives
G
Molecule Mevastatin
Atom Probability FAMEscore
C.18 0.904 0.874
C.2 0.292 0.922
C.6 0.208 0.896
C.10 0.036 0.913
O.22 0.032 0.986
C.28 0.024 0.927
C.8 0.024 0.922
C.27 0.02 0.933
C.25 0.02 0.925
C.7 0.02 0.913
C.11 0.016 0.896
C.21 0.012 0.95
H
Molecule Artemisinin
Atom Probability FAMEscore
O.8 0.696 0.801
O.7 0.688 0.805
C.2 0.196 0.772
C.14 0.068 0.824
C.6 0.048 0.74
O.3 0.048 0.773
C.16 0.044 0.802
C.15 0.04 0.814
C.18 0.036 0.822
C.19 0.032 0.77
O.5 0.024 0.807
C.4 0.024 0.793
I
Molecule Artemotil
Atom Probability FAMEscore
O.18 0.307 0.687
C.2 0.28 0.655
O.19 0.232 0.693
C.21 0.208 0.712
C.4 0.132 0.747
C.12 0.088 0.729
C.9 0.083 0.717
C.17 0.08 0.668
O.14 0.064 0.71
C.22 0.044 0.735
C.8 0.044 0.731
O.13 0.04 0.74
Experimental SoM (circle = phase 1, rectangle = phase 2)
Predicted SoM (MetScore phase 1 and phase 2 composite model)

Figure 6: FAME 3 and MetScore predictions for six compounds selected from the MetScore validation set (A-F) and FAME 3 predictions for three examples of pharmaceutically relevant natural products and their derivatives (G-I). All FAME 3 predictions were made with the "P1+P2 model" with bond depth 5, trained on the complete, preprocessed MetaQSAR database. The SoMs predicted by MetScore are taken from the publication of Finkelmann et al.[14] For each compound, the first twelve rows of the FAME 3 output are listed in the pink box. Each atom in the box is identified by its ID (see the "Atom" column). The atoms are sorted by their SoM likelihood as assigned by FAME 3 (see the "Probability" column). For each atom, the calculated FAMEscore is reported in the "FAMEscore" column. MetScore predictions and experimental SoMs are indicated next to the atom list by green and red dots, respectively. If one of multiple atoms can be a SoM due to symmetry, all plausible atom positions are annotated in both the structure and the ordered atom list.

Natural Products and Natural Product Derivatives

In addition to the six compounds selected from the MetScore validation set, we also investigated the performance of FAME 3 on three cases of interesting natural products and natural product derivatives for which metabolism plays a major role in their biological effect. MetScore predictions for these three compounds are not shown because MetScore was not accessible to the authors.

Mevastatin (also known as compactin) is an inhibitor of the HMG-coenzyme A (HMG-CoA) reductase. The prodrug takes its active form by hydrolysis of the lactone ring (Figure 6G). FAME 3 performs well in this case because the FAMEscore values are quite high (mostly between 0.8 and 0.95). This is due to the fact that similar statins are present in the MetaQSAR database. Carbon C.18 is involved in the conversion of mevastatin to its active metabolite, and

the FAME 3 "P1+P2 model" correctly ranks this atom at the top of the list with very high probability.

The final two case studies involve the natural product artemisinin (Figure 6H) and its semi-synthetic derivative, artemotil (arteether; Figure 6I). These compounds are potent antimalarial agents.[28] Artemisinin is known to undergo at least the following two metabolic reactions: (i) deactivation by the formation of deoxydihydroartemisinin through the reduction of the endoperoxide moiety (oxygens O.7 and O.8 in Figure 6H), which is also held responsible for its antimalarial activity,[29,30] and (ii) activation by reduction of the lactone moiety (carbon C.2 in Figure 6H) yielding dihydroartemisinin (artenimol), which is the well-known active metabolite of all artemisinin-type compounds.[31]

The structure of artemisinin is present in the MetaQSAR database as well as the SoMs involved in its deactivation (oxygen O.7 and O.8). On the other hand, the SoM involved in the activation reaction (carbon C.2) is not annotated in the database. FAMEscore values are slightly lower than in the case of mevastatin, but quite high nonetheless (around 0.8 and not lower than 0.7; Figure 6H). Unsurprisingly, the two oxygen atoms involved in the deactivation reaction (O.8 and O.7) are placed at the top of the rank-ordered list generated by FAME 3. However, despite the fact that the MetaQSAR database is lacking an annotation for the formation of the active metabolite, the C.2 atom involved in this transformation is ranked just below the two oxygens in this example. This suggests that the FAME 3 model is able to generalize and balance out the incompleteness of the training data to some extent. In other words, the model was able to rank C.2 higher and, thus, hint at the possibility that this atom could be implicated in a metabolic transformation despite the lack of direct evidence in the training data.

Artemotil has a very similar structure to artemisinin. The key difference between the structure of these two molecules is that the carbonyl group of artemisinin is replaced by an ethyl-ether substituent in the structure of artemotil (Figure 6I). Therefore, the active metabolite dihydroartemisinin is created not by the reduction of the lactone, but through a dealkylation reaction on carbon C.2.[32,33] The endoperoxide moiety of artemisinin is preserved in artemotil and it is also eliminated during its metabolism.[34] The metabolic transformations and the structure of artemotil are not annotated in the MetaQSAR database. This is reflected by slightly lower FAMEscore values than obtained for artemisinin. However, since the two structures are related, the values still remain quite high (around 0.7 and not lower than 0.65). The order of atoms in the rank-ordered list generated by FAME 3 reflects the true sites of metabolism quite well (Figure 6I). Both of the endoperoxide oxygens (O.18 and O.19) responsible for the deactivation are on top of the list, as well as C.2, at which the dealkylation reaction occurs during the formation of the active metabolite. In particular, the prediction for the C.2 atom is an interesting result since this environment and SoM annotation is missing from the database for the training case of artemisinin. This suggests that the FAME 3 model is able to extract knowledge from the information it has on other reactions involving different structures and successfully apply the learned rules to unknown atomic environments.

FAME 3 Public Web Server and Software Package

FAME 3 is available via a public web service at https://nerdd.zbh.uni-hamburg.de/ and as a self-contained Java software package for local execution. Both web service and software package provide the “P1+P2”, “P1” and “P2” models, trained and optimized as described in this study but featuring the complete MetaQSAR data set. This includes the “P1+P2” model with bond depth 5 used in our case studies.

The web service accepts various types of inputs, including the upload of larger sets of compounds in SD file format. Upon submission of a job, users are provided a web link, allowing them to collect their predictions at a later point in time. Usually, for individual molecules, predictions will only take a few seconds. Upon completion of the calculations, interactive HTML depictions of the model's predictions for each compound are generated (similar to those shown in Figure 6) using components of the open source SMARTCyp. Users are offered options for downloading and for deleting all results from the server.

The command line interface of the FAME 3 software package accepts input structures in either SMILES format or as an SDF file and then proceeds to generate the interactive HTML depictions described above. In addition to the HTML page, results of FAME 3 are also reported as a CSV file.

CONCLUSIONS

The third generation of FAME models for SoM prediction is based on a new, comprehensive data set of expert-derived SoMs. FAME 3 includes a collection of models for both phase 1 and phase 2 metabolism and is as such, to our knowledge, the most broadly applicable SoM predictor that is freely available for academic and non-commercial research. As we show in comprehensive tests, the FAME 3 models reach competitive performance, with MCCs ranging from 0.50 for the combined phase 1 and phase 2 model ("P1+P2 model") to 0.75 for a focused phase 2 model ("P2 100+ model"). A key feature of FAME 3 is the newly developed FAMEscore, an atom-based distance measure allowing the estimation of the applicability domain. FAME 3 thus enables researchers to understand the quality of the representation of any atom in their molecules of interest by the training data. Our benchmarking results suggest that the applicability domain of the FAME 3 models should be defined by a minimum FAMEscore of

0.6. However, in some cases accurate predictions (in particular with respect to atom ranking) can still be obtained even below this threshold. In addition to statistical analysis, this was also shown by several case studies with synthetic compounds and natural products.

A general conclusion that can be drawn from this and others' works is that models for SoM prediction are approaching a performance plateau defined primarily by the data available for model development. We therefore recognize that the generation and publication of additional data on xenobiotic metabolism is of utmost importance to the further progress of the field. Here we would hope that industry in particular will continue to strengthen their efforts in sharing data with the scientific community.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: Additional figures and tables: Number of reaction subclasses reported in the MetaQSAR database, model hyperparameter optimization results and 10-fold cross-validation performance of each model reported in the study.

AUTHOR INFORMATION

Corresponding Author

*J. Kirchmair. E-mail: kirchmair@zbh.uni-hamburg.de. Tel.: +49 40 42838 7303.

ORCID

Martin Šícho: 0000-0002-8771-1731

Conrad Stork: 0000-0002-5499-742X

Angelica Mazzolari: 0000-0003-1352-1094

Christina de Bruyn Kops: 0000-0001-8890-2137

Alessandro Pedretti: 0000-0001-5916-2029

Giulio Vistoli: 0000-0002-3939-5172

Bernard Testa: 0000-0002-3218-4612

Daniel Svozil: 0000-0002-9398-4357

Johannes Kirchmair: 0000-0003-2667-5877

NOTES

The authors declare a potential financial interest in the event that FAME 3 is licensed for a fee to non-academic institutions in the future.

ACKNOWLEDGEMENTS

The authors thank Ya Chen from the Center for Bioinformatics of the University of Hamburg for analyzing the natural product-likeness of compounds contained in the MetaQSAR database. Computational resources were supplied by the Ministry of Education, Youth and Sports of the Czech Republic under the Projects CESNET (Project No. LM2015042) and CERIT-Scientific Cloud (Project No. LM2015085) provided within the program Projects of Large Research, Development and Innovations Infrastructures.

FUNDING

M.S. and D.S. are supported by the Ministry of Education, Youth and Sports of the Czech Republic – project number LM2015063 and by RVO 68378050-KAV-NPUI. M.S. also received financial support from specific university research (MSMT No 21-SVV/2018). C.S. and J.K. are supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number KI 2085/1-1. J.K. is also supported by the Bergen Research Foundation (BFS) – grant no. BFS2017TMT01.

DECLARATION OF CONFLICT OF INTEREST

The authors declare a potential financial interest in the event that FAME 3 is licensed for a fee to for-profit institutions in the future.

ABBREVIATIONS

AUC, area under the ROC curve

CDK, Chemistry Development Kit

CoA, coenzyme A

CSV, comma-separated values

CYP, cytochrome P450

HMG-CoA, β-Hydroxy β-methylglutaryl-CoA

HTML, hypertext markup language

MCC, Matthews correlation coefficient

PCA, principal component analysis

RNA, ribonucleic acid

ROC, receiver operating characteristic curve

SDF, structure data file

SoM, site of metabolism

REFERENCES

(1) 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 Discov.* **2015**, *14*, 387–404.

(2) 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.

(3) Dixit, V. A.; Lal, L. A.; Agrawal, S. R. Recent Advances in the Prediction of Non-CYP450-Mediated Drug Metabolism. *WIREs Comput. Mol. Sci.* **2017**, *7*, e1323.

(4) de Bruyn Kops, C.; Stork, C.; Šícho, M.; Kochev, N.; Svozil, D.; Jeliazkova, N.; Kirchmair, J. GLORY: Generator of the Structures of Likely Cytochrome P450 Metabolites Based on Predicted Sites of Metabolism. *Front. Chem.* **2019**, *7*, 402.

(5) Tyzack, J. D.; Kirchmair, J. Computational Methods and Tools to Predict Cytochrome P450 Metabolism for Drug Discovery. *Chem. Biol. Drug Des.* **2019**, *93*, 377–386.

(6) 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.

(7) Olsen, L.; Montefiori, M.; Tran, K. P.; Jørgensen, F. S. SMARTCyp 3.0: Enhanced Cytochrome P450 Site-of-Metabolism Prediction Server. *Bioinformatics* **2019**.

https://doi.org/10.1093/bioinformatics/btz037.

(8) 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.

(9) 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*, 2046–2048.

(10) 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.

(11) Šícho, 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*, 1832–1846.

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

(13) Dang, N. L.; Hughes, T. B.; Krishnamurthy, V.; Swamidass, S. J. A Simple Model Predicts UGT-Mediated Metabolism. *Bioinformatics* **2016**, *32*, 3183–3189.

(14) Finkelmann, A. R.; Goldmann, D.; Schneider, G.; Göller, A. H. MetScore: Site of Metabolism Prediction Beyond Cytochrome P450 Enzymes. *ChemMedChem* **2018**, *13*, 2281–2289.

(15) Accelrys, Inc.: San Diego, CA, 2011. Accelrys Metabolite Database.

(16) de Bruyn Kops, C.; Friedrich, N.-O.; Kirchmair, J. Alignment-Based Prediction of Sites of Metabolism. *J. Chem. Inf. Model.* **2017**, *57*, 1258–1264.

(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*, 1019–1030.

(18) Pedretti, A.; Mazzolari, A.; Vistoli, G.; Testa, B. MetaQSAR Database (snapshot from March 15, 2018). *MetaQSAR: An Integrated Database Engine to Manage and Analyze Metabolic Data*, 2018.

(19) Chemistry Development Kit 1.4.19 https://github.com/cdk/cdk/releases/tag/cdk-1.4.19.

(20) 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.

(21) scikit-learn 0.18. Documentation of scikit-learn 0.18 http://scikit-learn.org/0.18/documentation.html (accessed Jun 21, 2019).

(22) Rodrigues, T.; Reker, D.; Schneider, P.; Schneider, G. Counting on Natural Products for Drug Design. *Nat. Chem.* **2016**, *8*, 531–541.

(23) Harvey, A. L.; Edrada-Ebel, R.; Quinn, R. J. The Re-Emergence of Natural Products for Drug Discovery in the Genomics Era. *Nat. Rev. Drug Discov.* **2015**, *14*, 111–129.

(24) Newman, D. J.; Cragg, G. M. Natural Products as Sources of New Drugs from 1981 to 2014. *J. Nat. Prod.* **2016**, *79*, 629–661.

(25) Chen, Y.; Stork, C.; Hirte, S.; Kirchmair, J. NP-Scout: Machine Learning Approach for the Quantification and Visualization of the Natural Product-Likeness of Small Molecules. *Biomolecules* **2019**, *9*, 43.

(26) Shen, J.; Serby, M.; Reed, A.; Lee, A. J.; Zhang, X.; Marsh, K.; Khatri, A.; Menon, R.; Kavetskaia, O.; Fischer, V. Metabolism and Disposition of the Hepatitis C Protease Inhibitor Paritaprevir in Humans. *Drug Metab. Dispos.* **2016**, *44*, 1164–1173.

(27) Argikar, U. A.; Dumouchel, J. L.; Dunne, C. E.; Saran, C.; Cirello, A. L.; Gunduz, M.

Ocular Metabolism of Levobunolol: Historic and Emerging Metabolic Pathways. *Drug Metab. Dispos.* **2016**, *44* (8), 1304–1312.

(28) Guo, Z. Artemisinin Anti-Malarial Drugs in China. *Acta Pharm Sin B* **2016**, *6*, 115–124.

(29) Lee, I. S.; Hufford, C. D. Metabolism of Antimalarial Sesquiterpene Lactones. *Pharmacol. Ther.* **1990**, *48*, 345–355.

(30) Sharma, S.; Anand, N. Chapter 14 - Natural Products. In *Pharmacochemistry Library*; Sharma, S., Anand, N., Eds.; Elsevier, Amsterdam, 1997; Vol. 25, pp 347–383.

(31) Navaratnam, V.; Mansor, S. M.; Sit, N. W.; Grace, J.; Li, Q.; Olliaro, P. Pharmacokinetics of Artemisinin-Type Compounds. *Clin. Pharmacokinet.* **2000**, *39*, 255–270.

(32) Melendez, V.; Peggins, J. O.; Brewer, T. G.; Theoharides, A. D. Determination of the Antimalarial Arteether and Its Deethylated Metabolite Dihydroartemisinin in Plasma by High-Performance Liquid Chromatography with Reductive Electrochemical Detection. *J. Pharm. Sci.* **1991**, *80*, 132–138.

(33) Grace, J. M.; Aguilar, A. J.; Trotman, K. M.; Brewer, T. G. Metabolism of β-Arteether to Dihydroqinghaosu by Human Liver Microsomes and Recombinant Cytochrome P450. *Drug Metab. Dispos.* **1998**, *26*, 313–317.

(34) Chi, H. T.; Ramu, K.; Baker, J. K.; Hufford, C. D.; Lee, I. S.; Zeng, Y. L.; McChesney, J. D. Identification of the in Vivo Metabolites of the Antimalarial Arteether by Thermospray High-Performance Liquid Chromatography/Mass Spectrometry. *Biol. Mass Spectrom.* **1991**, *20*, 609–628.

For Table of Contents use Only

FAME 3: Predicting the Sites of Metabolism in Synthetic Compounds and Natural Products for Phase 1 and Phase 2 Metabolic Enzymes

Martin Šícho, Conrad Stork, Angelica Mazzolari, Christina de Bruyn Kops, Alessandro Pedretti, Bernard Testa, Giulio Vistoli, Daniel Svozil, and Johannes Kirchmair