Journal of
CHEMICAL INFORMATION
AND MODELING

ARTICLE

pubs.acs.org/jcim

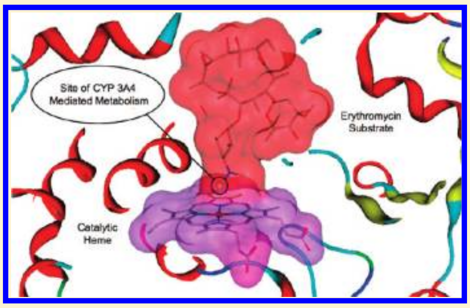
RS-Predictor: A New Tool for Predicting Sites of Cytochrome P450-Mediated Metabolism Applied to CYP 3A4

Jed Zaretzki,[†] Charles Bergeron,[‡] Patrik Rydberg,[¶] Tao-wei Huang,[†] Kristin P. Bennett,[‡] and Curt M. Breneman*[,†]

[†]Department of Chemistry and Chemical Biology, Rensselaer Polytechnic Institute, Troy, New York 12180, United States
[‡]Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, New York 12180, United States
[¶]University of Copenhagen, Universitetsparken 2, DK-2100 Copenhagen, Denmark

S *Supporting Information*

ABSTRACT: This article describes RegioSelectivity-Predictor (RS-Predictor), a new *in silico* method for generating predictive models of P450-mediated metabolism for drug-like compounds. Within this method, potential sites of metabolism (SOMs) are represented as "metabolophores": A concept that describes the hierarchical combination of topological and quantum chemical descriptors needed to represent the reactivity of potential metabolic reaction sites. RS-Predictor modeling involves the use of metabolophore descriptors together with multiple-instance ranking (MIRank) to generate an optimized descriptor weight vector that encodes regioselectivity trends across all cases in a training set. The resulting pathway-independent (O-dealkylation vs N-oxidation vs Csp^3 hydroxylation, etc.), isozyme-specific regioselectivity model may be used to predict potential metabolic liabilities. In the present work, cross-validated RS-Predictor models were generated for a set of 394 substrates of CYP 3A4 as a proof-of-principle for the method. Rank aggregation was then employed to merge independently generated predictions for each substrate into a single consensus prediction. The resulting consensus RS-Predictor models were shown to reliably identify at least one observed site of metabolism in the top two rank-positions on 78% of the substrates. Comparisons between RS-Predictor and previously described regioselectivity prediction methods reveal new insights into how *in silico* metabolite prediction methods should be compared.



INTRODUCTION

Cytochrome P450s are responsible for the observed phase I metabolism of over 90% of all marketed drugs.[1−3] The most promiscuous P450 isoform is CYP 3A4, which metabolizes 50% of the top 200 drugs prescribed in 2002.[4] The P450 isozymes catalyze a variety of biotransformations, including aromatic and aliphatic oxidation, N- and O-dealkylation, S- and N-oxidation, sulfoxide/sulfone formation, oxidative deamination, desulfuration, and dehalogenation.[5] The dominant oxidation mechanism common to all CYPs involves a two-electron reduction of molecular oxygen to form a reactive oxygen species and water, a conversion catalyzed by the heme group situated at the bottom of the CYP active site.[6]

The metabolic fates of potential therapeutic lead compounds are often unknown at the time of their initial discovery, potentially resulting in lost time and wasted resources if systemic metabolic liabilities are later discovered within that class of compounds. Clearly, having *a priori* knowledge of the potential metabolic liabilities of potential lead compounds could have important ramifications in the cost and speed of the drug discovery process. Armed with this information, medicinal chemists are given the opportunity to modify susceptible regions of lead compounds to reduce their susceptibility to unwanted modifications and to optimize their *in vivo* viability. Metabolic regioselectivity models can also assist in the design of pro-drugs, where the metabolized form of an administered compound becomes the active species through one or more predictable metabolic modifications. When combined with toxicity models, the possibility of lead candidate metabolites with undesirable pharmacokinetic profiles or harmful off-target interactions can be identified earlier in the discovery process, creating low cost intervention opportunities.

One classic example of this is the P450-mediated metabolism of acetaminophen, commonly known as Tylenol, illustrated in Figure 1. While only a small proportion of the drug is metabolized by P450s (primarily CYP 2E1), significant amounts of the toxic metabolite NAPQI can build up when acetaminophen is taken in large quantities or in conjunction with alcohol; when either of these events occur the sulfation and glucuronidation pathways that normally metabolize 80% to 90% of acetaminophen become saturated. The resulting systemic excess of acetaminophen induces P450 expression and increases levels of NAPQI metabolites, leading to potentially fatal hepatocyte damage.[7] Ideally,

Received: January 31, 2011
Published: May 02, 2011

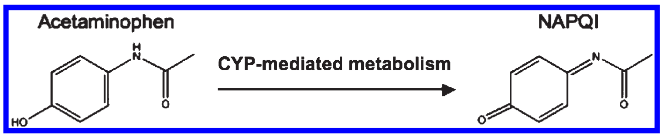


Figure 1. Metabolism of Tylenol by P450s.

all potential lead compounds (and their scaffolds) should be screened for metabolic liabilities before the expenditure of significant developmental resources. Unfortunately, experimental determination of metabolic outcomes for each potential lead compound is both time-consuming and prohibitively expensive. Consequently, there is a clear need for effective *in silico* methods capable of providing reliable guidance during all phases of therapeutic compound development.

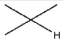
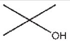
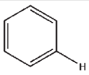
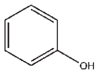
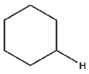
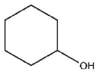
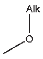
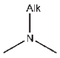
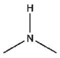
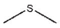
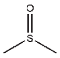
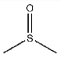
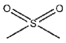
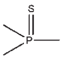
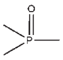
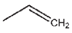
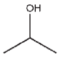
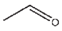
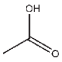
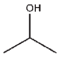
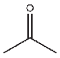
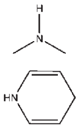
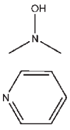
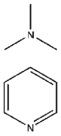
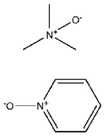
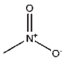
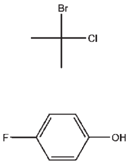
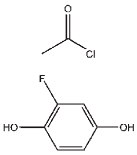
In recent years, a number of approaches have been developed for the purpose of predicting the outcome of small-molecule CYP-mediated metabolism with various degrees of accuracy; some are classification models designed to predict whether target ligands are inhibitors, inducers, or substrates of specific CYP isozymes,[8−16] while others are designed to predict CYP oxidative regioselectivity. Regioselectivity models fall into two broad categories: ligand/reactivity based[17−24] and enzyme-structure based,[25,26] while others are hybrids that utilize substrate QSAR models together with enzyme docking and/or scoring approaches to arrive at a prediction.[27,28] Traditionally, regiospecificity model performance is evaluated by ranking metabolic site predictions on a set of test molecules against their known site(s) of metabolism (SOM(s)): A molecule is considered correctly predicted if any of its experimentally observed sites of metabolism are ranked first, first or second, or first, second or third (depending on the metric: Top-1, Top-2, or Top-3, respectively) among all potential metabolic sites on that molecule. The Top-2 metric is considered to be the most standard method of performance evaluation and is the primary one used during the Discussion section of this work.[18−24,27,28]

An early 3A4 regioselectivity model from Merck[18] utilized a "trend vector" approach that was developed to associate site-specific topological descriptors with hydrogen abstraction energies calculated using the AMPAC AM1 method. The Top-2 prediction rate of 44% reported for 50 substrates by the Merck group was later surpassed by MetaSite, a hybrid model that utilizes precomputed fragment-based reactivity values in conjunction with a recognition component derived from CYP isozyme crystal structures. While the MetaSite group reported a Top-2 prediction rate of 78% for a set of 340 substrates of 3A4, information about structures, observed metabolites, and predicted metabolic sites was not publicly released.[27] Nevertheless, MetaSite was applied to a public data set of 324 substrates of 3A4 by *Sheridan et al.*, comparing results with Merck empirical regioselectivity models expanded with additional descriptors and random forest machine learning.[19] In that study, the Top-2 prediction accuracy was reported to be 77% for the Merck method and 62% for MetaSite (as implemented at Merck), though again the molecule-specific SOM predictions were not released. Two other modern CYP regioselectivity prediction methods are StarDrop and SMARTCyp.[24,29] StarDrop is a commercial application that performs on-the-fly AM1 calculations using a modified version MOPAC97 that are combined with orientation and steric accessibility descriptors to evaluate potential reaction pathways. Consequently, StarDrop predictions can require several minutes of computing per compound on a modern workstation. The SMARTCyp method exploits a set of precomputed transition state energies for potential CYP-mediated reactions on representative molecular fragments to create a reactivity lookup table that covers common SOM environments. In practice, the tabulated reactivity values are used together with an accessibility descriptor to provide a final SMARTCyp ranking of putative SOMs on each molecule.

Random-forest machine learning models such as those used by the Merck group are derived across all potential oxidative metabolism pathways (Csp^3 hydroxylation, N-dealkylation, S-oxidation, etc.) for an entire substrate set. *Sheridan et al.* argued that properly encoded topological information with rational modeling could equal or surpass first-principle methods that utilize local electronic information. First-principle methods potentially provide better descriptions of reactivity, but they also must overcome the difficulties associated with making comparisons between SOMs that oxidatively degrade to metabolites through different pathways. On the other hand, MetaSite simply assigns each potential SOM a weight based on its likely potential reaction mechanism, which is influenced by how often that oxidative pathway is observed in the isozyme training set.[30] In other work, individual regression models by *Henneman et al.* and classification models by *Zheng et al.* have been created for separate pathways using AM1 semiempirical descriptors.[21,22] In those scenarious, potential oxidation sites of test molecules are ranked through the normalized application of separate models. Models generated in this fashion rely upon first-principle signal captured by an AM1 Hamiltonian to differentiate between pathways of different types without explicitly representing or modeling topological information. SMARTCyp models rely on first-order principles as well, using DFT calculations of molecular fragments with a isoform-nonspecific heme group. Reactivity-based models such as these work reasonably well because they contain a high amount of regioselectivity information, but they do not incorporate any isozyme-specific information necessary to create CYP-specific models. To accomplish that, regioselectivity models must be empirically derived for all oxidation pathways using descriptors derived from both molecular topology and quantum chemical reactivities.

The main contribution of this work is a description of RegioSelectivity-Predictor (RS-Predictor), a method for generating isozyme-specific, pathway-independent regioselectivity QSARs using any sufficiently diverse calibration set of substrates for training. Potential SOMs are represented through a combination of 148 topological and 392 quantum chemical atom-specific descriptors that are grouped together as metabolophores. Models are generated using MIRank — a support vector machine (SVM)-like ranking and multiple instance learning method specifically designed to correctly rank metabolophores associated with oxidized SOM(s) over metabolophores of nonoxidized SOMs on the same substrate.[31] In this work, cross-validated RS-Predictor models were generated from a set of 394 substrates of 3A4 culled from the primary literature.[19,32,33] In one of the first applications of rank aggregation within the chemoinformatics community, independently generated regioselectivity rankings for each compound were merged into single consensus predictions — with 78% of the compounds having an observed SOM predicted in one of the top two rank-positions. RS-Predictor models calibrated using an updated 322 compound Merck data set were shown to be robust, with similar performance rates when applied to an additional 72 substrate external test set. Additionally, both calibration and test set prediction rates were shown to equal or surpass those of previously published

Table 1. Analysis of the SOMs from 394 Substrates of 3A4 by Common P450-Mediated Pathways

Pathway	Initial Fragment	Resultant Fragment	Potential SOM	Oxidized SOM	Oxidized SOM/Potential SOM
Aliphatic Hydroxylation			929	92	9.90%
Aromatic Hydroxylation			1563	115	7.46%
Ring Hydroxylation			1342	134	9.99%
O-dealkylation			206	54	26.21%
N-dealkylation			380	172	45.26%
Sulfur(II) Oxidation			56	33	58.93%
Sulfur(IV) Oxidation			10	10	100%
Desulfuration			9	3	33.43%
Csp^2 Hydroxylation			108	8	7.41%
Aldehyde Oxidation			5	3	60.00%
Alcohol Oxidation			109	8	7.44%
Nitrogen Hydroxylation			243	11	4.53%
N-oxide Formation			527	23	4.46%
Nitro-group Reduction			15	0	0.00%
Dehalogenation			183	2	1.09%
Uncommon			2533	17	0.67%

Merck and MetaSite methods. RS-Predictor is shown to outperform both SMARTCyp and StarDrop using either Top-2 or Top-3 metrics. A new Lift metric for assessing prediction quality is introduced here as well, where each substrate is assigned a lift weight that expresses the statistical likelihood of randomly picking the CYP-oxidized SOM(s) out of all putative SOMs of the substrate.

■ METHOD

Data Sets. The 3A4 substrate and metabolite data used in this paper was collected and curated through an extensive analysis of public sources. Prior to March of 2010, the most extensive public collection of 3A4 substrates and metabolites was published by Merck in the form of two data sets consisting of 305 and 19 molecules.[19] Further analysis of the primary literature, relying especially upon review articles by Brown and Rendic[32,33] produced an additional 72 substrates of 3A4 not contained in the Merck database. During the process of curating these new compounds, the structures, responses, and citations from original Merck data set were also reviewed. It was discovered that one molecule, H_259_31 appeared twice in the Merck data, while the ferrocene derivative SSR97193[34] was removed due to low reported oxidation rates. Additionally the experimental responses

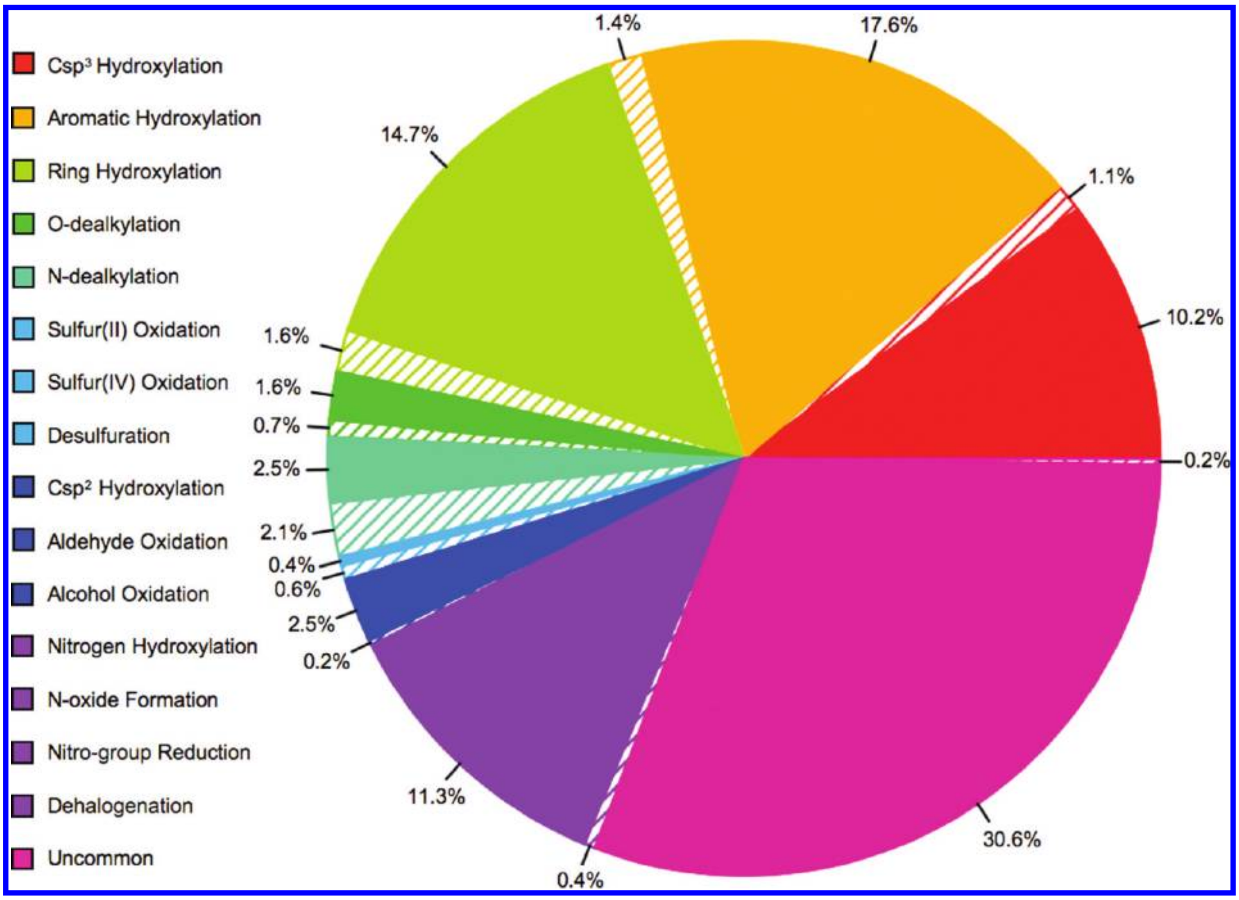


Figure 2. Propensities of 3A4 to catalyze the SOMs of 394 substrates according to established P450-mediated biotransformations. Hashed sections represents SOMs which are oxidized during 3A4-mediated metabolism. Solid sections represent potential SOMs which do not undergo oxidation. To simplify result presentation, similar pathways (ex. Sulfur Oxidations) or those with low population, are placed into the same analysis group.

of 85 compounds were updated based upon new published results, while errors in the structure of aflatoxin_b1 and promazine were identified and fixed. This data set of 394 compounds was originally released as a validation set for SMARTCyp[24] and is available in the Supporting Information.

RS-Predictor models generated using the curated Merck data set were also used to make predictions on a second external validation set consisting of a proprietary database of 20 compounds provided by a collaborating major pharmaceutical company. The prediction results are presented in this manuscript, but for obvious reasons the structures and experimental responses are not included.

Metabolophores. To be useful as a virtual screening tool, a metabolite prediction algorithm must be able to correctly rank-order the relative susceptibility of each potential site on a substrate to metabolism across all possible reaction pathways known to be catalyzed by a given CYP. The curated 3A4 substrate/product database of this work is analyzed in Table 1 according to an augmented set of the general rules proposed by *Korolev et al.* as determining the outcome of CYP-catalyzed biotransformations of specific molecular substructures.[5] Illustrated here are structural representations of common CYP-mediated biotransformations, with a breakdown of the 394 substrates of 3A4 according to the number of potential reactions for each pathway, the number that undergo oxidization during CYP 3A4-mediated metabolism, and the corresponding propensity for 3A4 to catalyze each of these oxidative biotransformations. A single P450 substrate often contains several potential sites of oxidation, each of which may be associated with a specific type of potential biotransformation. The structural visualization of P450-mediated pathways within Table 1 illustrates how multiple distinct heavy atoms and their attached hydrogens as well as topologically equivalent heavy atoms may both individually and collectively represent the same potential SOM. Experimentally, only a subset of these sites will undergo oxidative transformation during P450-mediated metabolism, with rates that differ from substrate to substrate and isozyme to isozyme. For the data set presented in this paper, substrate molecules were found to have, on average, 2.05 observed SOMs and an average of 20.70 potential SOMs. There is significant variance in SOM sites within the substrates in the data set, for example, BPR0L075 has (8 observed)/(20 potential) SOMs, while SDZ IMM 125 has an observed to potential SOM ratio of (3)/(70). A visualization of overall pathway propensities for the 394 substrates is given in Figure 2.

While knowledge of isozyme propensities to follow certain oxidation pathways over others is useful, these propensities are only known for a relatively small number of substrates, each of which has a unique electronic and steric environment and resulting distribution of SOMs. For example, Fluvastatin in Figure 4 does not undergo N-dealkylation — one of the more likely 3A4-mediated pathways — but instead undergoes aromatic ring hydroxylation. The problem of regioselectivity prediction is further compounded by the paucity of reliable kinetic data for known sites of metabolism and the lack of comprehensive metabolite data that would enable the relative ranking of inactive sites on substrate molecules to be performed.[19] Often, only one or two sites with the greatest oxidation rate are known for any given substrate molecule. Since experimental data come from a variety of different laboratories and literature sources and has been obtained using different techniques with different possible motivations, reactivities are typically treated as binary: Either a potential SOM is oxidized during a 3A4-mediated metabolism experiment or it is not.

These data constraints highlight the potential difficulties in creating broadly applicable metabolite prediction models. Binary response values mean that information about relevant differences in oxidation rates between sites are not available within the data. An example of this is illustrated in Figure 3 through the different mediated metabolic pathways of R and S enantiomers of warfarin by CYP 3A4. The standard Top-1, Top-2, and Top-3 metrics do not differentiate between the accurate identification of a primary versus secondary observed SOMs, while regioselectivity methods that do not encode conformation specific information will be unable to differentiate between enantiomer-specific metabolism. As a consequence of these data constraints, regioselectivity models must be able to predict likely metabolites, or sites with high oxidation rates, without knowing the relative differences in reactivity between sites on the same molecule undergoing different reactions, or the observed reaction sites of different substrates.

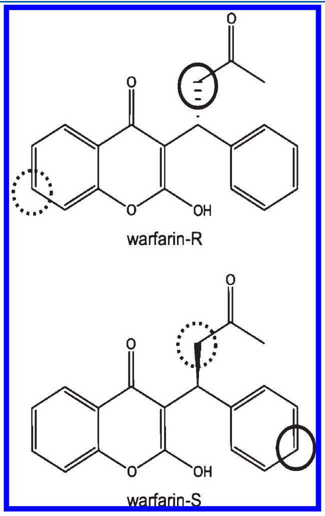


Figure 3. 3A4-mediated metabolism of R and S enantiomers of warfarin. Primary observed SOMs are designated by a solid circle, secondary observed SOMs are designated by a dashed circle.

RS-Predictor determines reaction rate trends by treating substrate molecules as individual competitions between potentially oxidizable sites and extracting ensemble trend information. Within this paragidm, ensuring that known sites of oxidation are ranked above other putative sites becomes part of a multiple-instance ranking (MIRank) optimization problem, where potential SOMs are represented as collections of atoms, and each collection represents a potential P450-mediated pathway illustrated in Table 1. These collections include the base heavy atom for the candidate transformation as well as the bonded hydrogen atom(s). Topologically equivalent atoms that represent the same potential biotransformation site are grouped together; Figure 4 shows an example of this grouping. Each individual atom is then represented by 148 topological and 392 quantum chemical descriptors. For convenience, these collections of atomic and group descriptors are called "metabolophores". A metabolophore is therefore a mathematical representation of an oxidizable region of a molecule that is defined by its local electronic environment and its potential oxidative reaction mechanism. The use of metabolophores as basic modeling units allows potential SOMs to be directly compared with others within the MIRank framework, regardless of which oxidative pathway is operative.

Descriptors. A number of descriptors have been developed to characterize the properties of metabolophores, and they fall into one of three categories: topological, quantum chemical (atom based), or quantum chemical (atom-pair based). In the RS-Predictor workflow, metabolophore descriptors are first computed

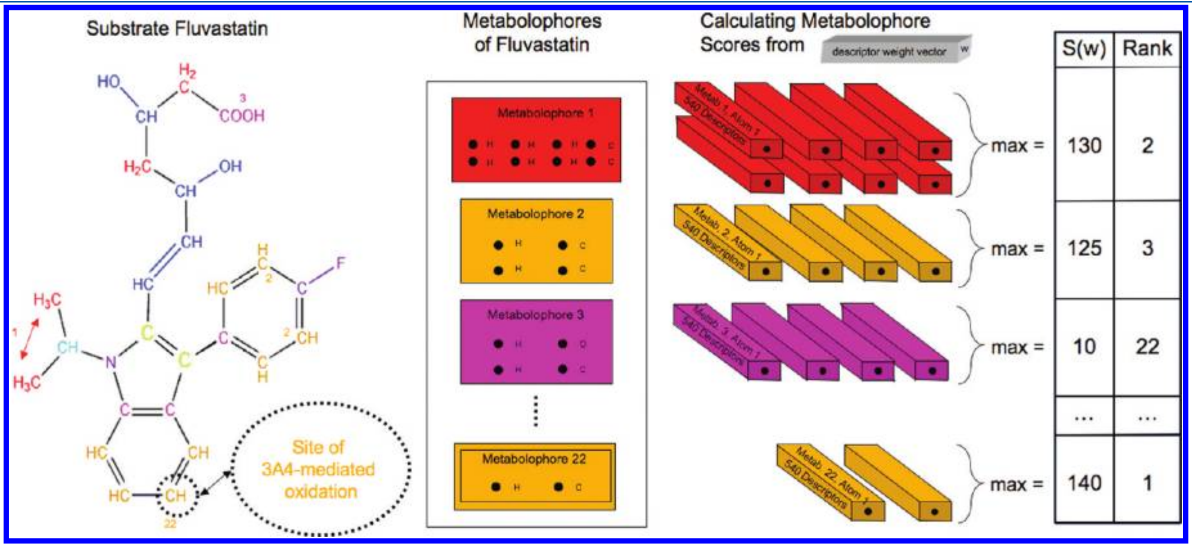


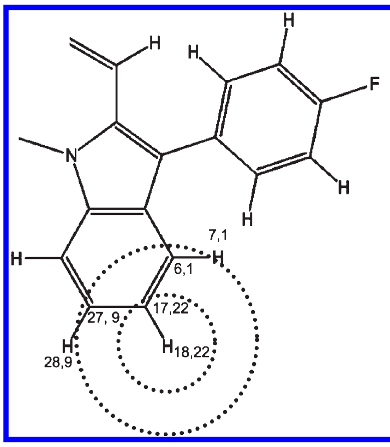
Figure 4. Metabolophores of Fluvastatin with hypothetical $S(w)$ and Rank values. Fluvastatin has a total of 22 distinct metabolophores (pathway-based color scheme shown in Figure 2). Metabolophore 22, representing the 3A4-mediated aromatic ring oxidation, is denoted with a double box. A metabolophore score, $S(w)$, is the maximum dot product between a descriptor weight vector w and the descriptors values of each individual atom contained within the metabolophore. Actual metabolophore labeling and compositions are used, while $S(w)$ and ranking values are artificial.

Table 2. Descriptor Definitions

label	definition	depth range	total number
	Topological Descriptors		148
NA_*d*_*e*	number of atoms depth bonds away of type element	0−4	35
PA_*d*_*e*	percentage of atoms depth bonds away of type element	1−4	28
NA_*d*_*p*	number of atoms depth bonds away of PATTY atom type	0−4	35
PA_*d*_*p*	percentage of atoms depth bonds away of PATTY atom type	1−4	28
span	(maximum path length from current atom)/(maximum path length from all atoms within the molecule)	0	1
$\mathrm{Lip}_{\alpha,\beta}$	lipophilicity according to 5 different (α,β) dependent metrics	0	5
HBonded	number of hydrogen atoms bonded to base atom of metabolophore	0	1
NHBonded	number of non-hydrogen atoms bonded to base atom of metabolophore	0	1
RS	base atom of metabolophore is within a ring of size (0, 3−8)	0	7
AR	base atom of metabolophore is within an aromatic ring	0	1
MR	base atom of metabolophore is within multiple rings	0	1
RB	number of rotatable bonds for base atom of metabolophore (0−3)	0	5
	Quantum Chemical Atom Based		112
HM	the projection of the given atom onto the 3-dimensional vector defined by the hydrophobic atoms of the molecule	0	1
BL	the average distance between the atom and all atoms it is bound too	0	1
SC	the charge kept by the atoms involved in the bond	0−2	11
AC	the charge not involved in bonding	0−2	11
ED	aromatic orbital electron density	0−2	11
F	Fukui reactivity index	0−2	11
N	nucleophilicity index	0−2	11
E	electrophilicity index	0−2	11
EERE	one-center electron−electron repulsion energy	0−2	11
ENAE	one-center electron−nuclear attraction energy	0−2	11
EE	total one-center electronic energy	0−2	11
area	solvent accessible surface area	0−2	11
	Quantum Chemical Atom-Pair Based		280
$\delta_{\sigma-\sigma}$	$\sigma-\sigma$ component of atom−atom interactions from Mulliken Population analysis	0−1	20
$\delta_{\sigma-\pi}$	$\sigma-\pi$ component of atom−atom interactions from Mulliken Population analysis	0−1	20
$\delta_{\pi-\pi}$	$\pi-\pi$ component of atom−atom interactions from Mulliken Population analysis	0−1	20
$P_{\sigma-\sigma}$	$\sigma-\sigma$ bond order	0−1	20
$P_{\sigma-\pi}$	$\sigma-\pi$ bond order	0−1	20
$P_{\pi-\pi}$	$\pi - \pi$ bond order	0−1	20
P	bond degree	0−1	20
ERE	electronic resonance energy	0−1	20
EEE	electronic exchange energy	0−1	20
ERPE	electronic repulsion energy	0−1	20
NEAE	nuclear-electron attraction energy	0−1	20
NNRE	nuclear−nuclear repulsion energy	0−1	20
C	Coulomb interaction energy	0−1	20
TENE	total of electronic and nuclear energy	0−1	20

based on the properties of each individual atom within a metabolophore, but those descriptors are then modified to incorporate properties the atoms surrounding it. Consequently, each descriptor label has a depth range, where each value of the range represents the set of atoms a specific number of bonds away from the individual atom in question. In practice, descriptor values are calculated from each of these sets and then mapped back onto the base atom. The exact number and nature of these descriptors can vary between label and class and are briefly explained below. A complete listing of each descriptor calculated within RS-Predictor is provided in Table 2. Specific examples of the numerical values of descriptors for selected atoms of Fluvastatin are given in Figure 5. Calculation of substrate metabolophores and descriptors is accomplished through a combination of Scientific Vector Language scripts applied within the molecular operating environment (MOE).[35]

Topological. The topological descriptors used in this work are an expanded version of the substructure descriptors described in *Sheridan et al.*[19] Extensions include hydrogen atom representation as well as additional bond and lipophilicity information. Topological descriptors contain bond and atom-type information calculable from 2D structures and therefore are conformationally independent. Some topological descriptors used here are atom specific, while others are specific to whole metabolophores, meaning that all atoms within a particular metabolophore are assigned the same value. The *span* descriptor is a topological



Metabolophore ID	Atom ID	Descriptor	Value
22	18	NA_0_H	1
22	18	PA_3_H	0.5
22	18	NA_0_Hp	0
22	18	AR	1
22	18	RS	6.0
22	18	RS_6	1
22	18	SC_D0	0.1334
22	18	$Lip_{0,1.2A}$	0.1581
22	18	$Lip_{1.2A,2.4A}$	0.3162
6	6	SC_D0	-0.0801
9	27	SC_D0	-0.1093
22	17	SC_D1_mean	-0.0187
22	17	SC_D1_max	0.1334
22	17	RS	6.0
22	17	NA_0_Hp	1

Figure 5. Numerical values of select descriptors for specific atoms within a substructure of Fluvastatin, illustrated earlier in Figure 4. The first number after each atom designates the unique Atom ID, while the second number designates the Metabolophore ID. All atoms in this substrate have the same Molecule ID, which is unique to Fluvastatin, as well as a metabolophore-specific binary response value (not shown). For Fluvastatin, only atoms 17 and 18 have positive response values. Atoms that contribute to two of the five lipophilicity descriptors of atom 18 fall either within the illustrated 1.2 Å concentric circle or the 1.2 Å and 2.4 Å concentric torus. The H after NA and PA descriptor labels denote atoms of element type Hydrogen, while the Hp label denotes atoms that are hydrophobic. The D0 and D1 after the self-charge (SC) descriptor labels denote the set of atoms 0 (self) and 1 bond-lengths away from the given atom.

measure of the position of an atom relative to the middle or an end of a molecule. This descriptor was originally developed by *Sheridan et al.* and is the accessibility descriptor used within SMARTCyp.

Lipophilicity values for each atom *i* from the substrate atom set A are calculated as follows

$$Lip_{\alpha,\beta}(i) = \sum_{j \in A, j \neq i} \rho_{\alpha,\beta}(d_{ij})P(i) \qquad (1)$$

where $P(i)$ is the contribution of atom i to $\log P^{36}$ and ρ determines concentric atom groupings based upon d_{ij}, the distance between atom i and j. Here ρ is an indicator function defined to be 1 if $\alpha \leq d_{ij} \leq \beta$ and 0 otherwise. Different value combinations of α and β are used to calculate five different gauges of lipophilicity. The specific values of (α,β) used are (0,1.2 Å), (1.2 Å,2.4 Å), $(2.4\ \text{Å},(D_i)/(3))$, $((D_i)/(3),(2{*}D_i)/(3))$, and $((2{*}D_i)/(3),D_i)$, where D_i is the longest distance from atom i to all other atoms of the molecule. Examples of the first two gauges are illustrated in Figure 5.

Element types for NA (number of atoms) and PA (percentage of atoms) descriptors are H, C, O, N, S, P, and Other. PATTY atom types are cation, anion, neutral H-bond donor, neutral H-bond acceptor, donor/acceptor, hydrophobe, and none of the above.[37] Rotatable Bond and Ring Size descriptors each have a range of potential values. Each value within the range is represented by a binary descriptor as well as an additional floating point descriptor, providing distribution information that machine learning methods are designed to exploit.

Quantum Chemical Atom Based Descriptors. These features describe atom-specific reactivity through a combination of computed values of polar, electrostatic, and donor−acceptor features,[21] while intramolecular energy distributions are used to represent atomic electron−electron repulsion and electron−nuclear attraction capabilities. These semiempirical features are extracted from the output of a MOPAC 2007[38] calculation using the AM1 Hamiltonian with keywords AM1, XYZ, MMOK, VECTORS, BONDS, PI, PRECISE, ENPART, EF, MULLIK, GNORM = $((\#atoms)^{1/2})/(2)$. Atom-specific descriptor values were calculated for up to 25 unique conformations (generated using the stochastic conformation search function within MOE) and then averaged according to Boltzmann's distribution law

$$\begin{aligned}\frac{N_j}{N} &= \frac{e^{-\varepsilon_j/kT}}{\sum_i e^{-\varepsilon_i/kT}} \\ &= \frac{\text{energetic likelihood of conformation j}}{\text{sum of likelihoods over all conformations}} \\ &= \text{weight for conformation j} \qquad (2)\end{aligned}$$

Most descriptors within this category have a depth range of 0−2. Depth values of 1 and 2 then represent the sets of all atoms 1 and 2 bonds away from the individual atom, with each atom having a discrete descriptor value for depth range 0. The *mean, max, min, norm,* and *sum* of depth 1 and 2 sets are used as individual atom descriptor values, giving a total of 11 discrete values per label. Descriptor values are extracted directly from MOPAC output, except for Fukui reactivity, Nucleophilicity, and Electrophilicity, whose values are calculated from wave function properties using the following functions

$$\text{Fukui reactivity} = \sum_{i \in A} \sum_{j \in A} \frac{C_{i,HOMO}C_{j,LUMO}}{\varepsilon_{HOMO} - \varepsilon_{LUMO}} \qquad (3)$$

$$\text{Nucleophilicity} = \sum_{i \in A} \frac{C^2_{i,HOMO}}{1 - \varepsilon_{HOMO}} \qquad (4)$$

$$\text{Electrophilicity} = \sum_{i \in A} \frac{C^2_{i,LUMO}}{\varepsilon_{LUMO} + 10} \qquad (5)$$

where $C_{i,HOMO}$ = highest occupied molecular orbital MO coefficient i for atom A, $C_{j,LUMO}$ = lowest unoccupied molecular orbital MO coefficient j for atom A, ε_{HOMO} = energy of highest occupied molecular orbital MO, and ε_{LUMO} = energy of lowest unoccupied molecular orbital MO.

Quantum Chemical Atom-Pair Based Descriptors. The interactions between different atoms within the same molecule

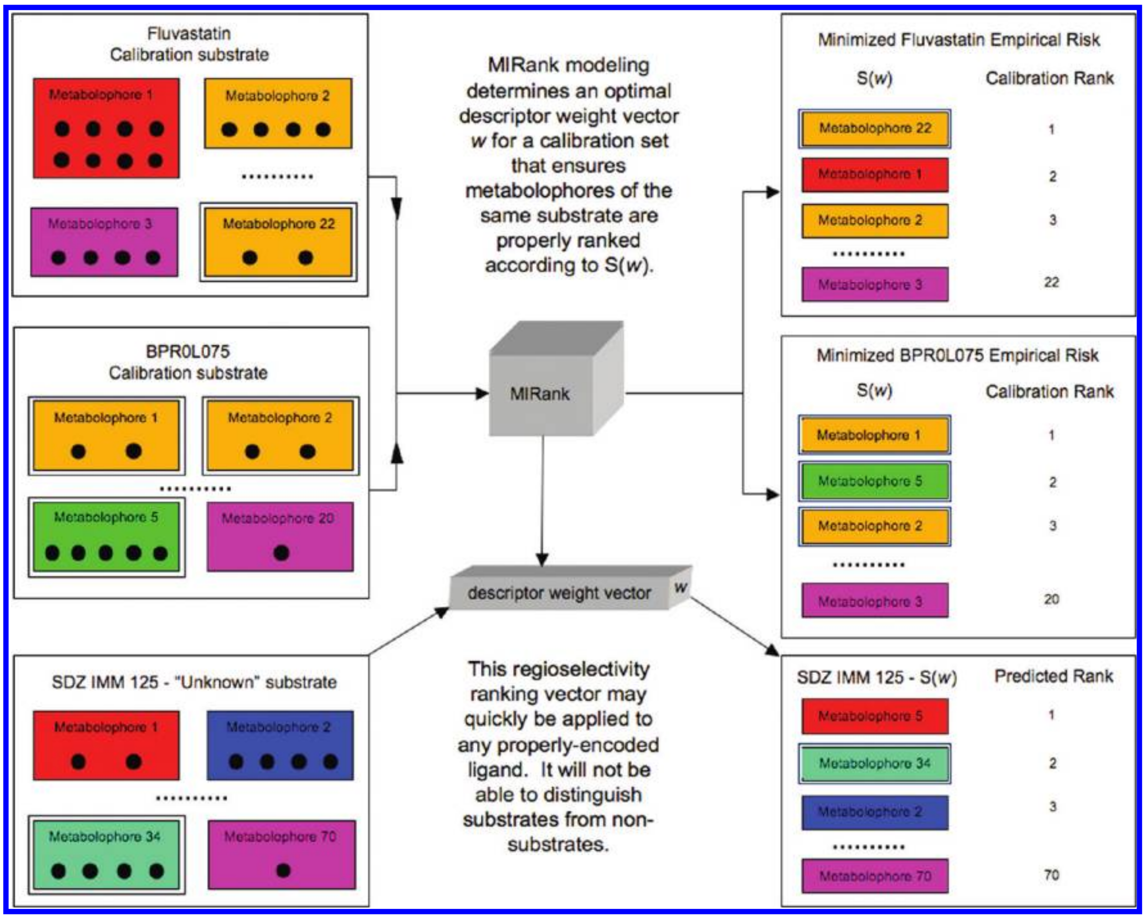


Figure 6. An illustration of MIRank application using actual substrates with actual metabolophore compositions and pathway designations (color scheme in Figure 2). Artificial metabolophores predictions are sorted by decreasing $S(w)$ (eq 9) and Rank. Rankings were chosen to illustrate minimized empirical risk (eq 7) for the calibration set, whereby oxidized metabolophores, denoted by double boxes, are ranked above nonoxidized metabolophores of the same substrate through pairwise comparisons of $S(w)$ (eq 8). The "unknown" external substrate SDZ IMM 125 is correctly predicted using a Top-2 metric. While determining the optimal w, no pairwise comparisons are made between metabolophores of different substrates or metabolophores of the same substrate that are either both oxidized or both not oxidized.

are gauged by orbital overlaps and nuclear attraction/repulsion forces. The bond strengths, electronic energies, and conformational flexibilities of atom pairs potentially correlate to metabolic site lability and are used as indirect measures of site reactivity.[21] MOPAC semiempirical calculations are used to compute the $\sigma - \sigma, \sigma - \pi, \pi - \pi$ orbital overlaps between an individual atom and all other atoms in the molecule. To calculate atom specific pair-based descriptors, all other atoms are divided into two sets: a) topologically bonded atoms and b) all other atoms. For each atom set and orbital type, the overlap between the specific atom and each atom from the set is calculated. The *mean, max, min, norm*, and *sum* of each set are then used as descriptors of atom lability. In this case, two sets are used to prevent the stronger orbital overlap from bound atoms from drowning out additional signal from the other atoms in the molecule. A similar calculation is made for each set derived from atom—atom nuclear attraction/repulsion forces. The same MOPAC calculations and Boltzmann averaging techniques used for atom based quantum mechanical descriptors were used for calculating atom-pair descriptors. Each descriptor label had a depth range of 0—1, with each value from the range being representing by two atom sets, each having five discrete values per set.

MIRank. MIRank is a fusion of support vector machines (SVM) ranking and multiple instance learning specifically developed to address the diffuse response signal and consequential partial ranking challenges inherent to accurate prediction of metabolic regioselectivity.[31] As the primary modeling unit, metabolophores are composed of all atoms potentially involved in a specific metabolic reaction, and the descriptors used to quantify each atom represent diffuse information. Since the rate constants of each observed metabolic reaction are unknown, putative SOMs are either considered oxidized or not oxidized. Consequently, no ranking information is known between metabolophores of different substrates or metabolophores of the same substrate which are either both oxidized or both not oxidized. Regioselectivity modeling then becomes a partial-ranking problem, as the only known response signal is that, for a given substrate, the metabolophores of the CYP-oxidized SOM(s) should be ranked above the metabolophores of the nonoxidized SOMs. The optimal descriptor weight vector w that accomplishes this for a given calibration set of N substrates is determined by MIRank through the minimization of the following loss function

$$MIRank(w) = ||w||^2 + C \sum_{i=1}^{N} ER^i(w) \qquad (6)$$

This loss resembles that of SVM in that the first term penalizes structural risk (using 2-norm regularization to avoid overfitting), the second term penalizes empirical risk or error, and C is the trade-off parameter between both terms. The difference between MIRank and traditional SVM is that empirical risk is measured for each individual calibration substrate i through

$$ER^i(w) = \sum_{A}^{OM} \sum_{B}^{NOM} ER^i_{A,B}(w) \qquad (7)$$

Here pairwise comparisons between all oxidized metabolophores (OM) of the substrate are made with all nonoxidized metabolophores (NOM) of the substrate

$$ER^i_{A,B}(w) = \max[0, 1 - S_A(w) + S_B(w)]^2 \qquad (8)$$

using the descriptor weight vector w to determine the score, $S(w)$, of each metabolophore

$$S_A(w) = \max(w \cdot x_a), \ \forall \ atoms \ a \in metabolophore \ A \qquad (9)$$

with x_a being the descriptor values for the atom a. The results of this change to SVM is an optimization problem that is nonconvex. In previous work, a fast subgradient algorithm was developed to solve problems such as these.[39]

As illustrated in Figure 4 and Figure 6, representing candidate substrates using metabolophores followed by MIRank model formulation has two main benefits:

1 Combining metabolophore scores (eq 9) with a pairwise empirical risk term (eq 8) lets direct comparisons be made between SOMs with different potential reaction pathways (ex. N-oxidation vs Csp^3 hydroxylation) and different electronic environments.

2 Gauging (eq 7) and subsequent minimization (eq 6) of empirical risk over all ordered pairs OM × NOM on a substrate by substrate basis lets MIRank utilize available partial-ranking information to optimize the regioselectivity prediction for an entire substrate calibration set.

Once determined, the optimal descriptor weight vector w is applied to predict the CYP-mediated regioselectivity of any lead candidate. In an effort to create robust models and to determine how well RS-Predictor should work on unknown substrates, MIRank models were generated using 5-fold cross-validation (CV) for model selection. In 5-fold CV, a calibration set of substrates with known responses is randomly divided into 5 equal partitions. One partition is assigned to be the testing set, while the remaining four sets are used for training and validation. Each of the four partitions are used once as the validation set and three times as part of the training set. Validation performance is gauged each time, with the optimal w for the three training partitions being determined for different values of the trade-off parameter C (eq 6). The optimal C value across all validation runs is then used to find the optimal w for all four training/validation partitions. The resulting w is then used to rank-order the putative SOMs of all substrates within the test partition. Optimization of C and w is performed five times, with each partition being used as the test set once. To overcome potential biases inherent to random partitions, the entire process of 5-fold CV is repeated 10 times. As a result, each molecule in the calibration set has 10 independent rank-orderings of its putative SOMs, each of which comes from a model generated without including the predicted molecule. When making predictions on an external test set, the descriptor weight vectors generated from each training/validation model are used, giving 50 predictions per test compound.

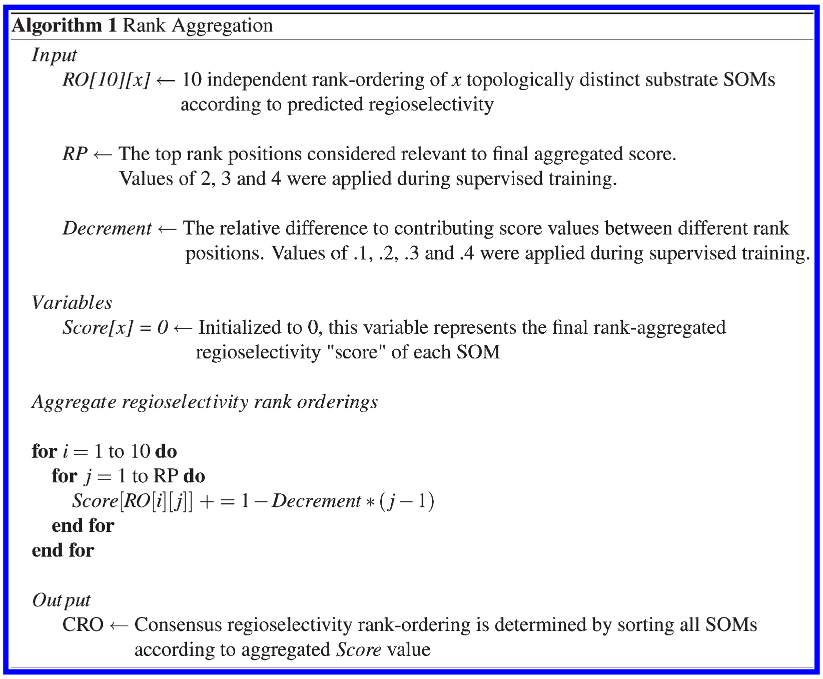
Rank Aggregation. While CV is an accepted method for estimating unbiased prediction error, it would not be effective to present end-users with ten or more slightly different rankings of the potential SOMs for each molecule. As with any statistical learning method, the results of an individual MIRank model may not provide an optimal prediction. Variance in overall accuracy between different CV models of the 394 compound set was found to be ±2%, a maximal difference of 16 correct/incorrect substrate predictions between individual models. Prior work has shown aggregate models can significantly improve QSPR and QSAR models;[40] however, consensus regioselectivity models found by applying the weight vectors from individual CV iterations, and subsequently ranking putative metabolophores according to their average scores, were found to have poor performances.

An explanation for poor consensus results was that SVM returns a floating point number. Within SVM ranking that number is used as a score, which in turn produces a rank. Unlike in regression analysis, the magnitude of this score is largely meaningless since the model loss function is designed to produce a ranking versus a regression value. As a result the magnitude of the difference in the score between different ranks is largely irrelevant as long it is sufficiently large. Consequently, averaging the metabolophore "score" across different models to create a single prediction per substrate led to decreased overall performance relative to averaging the accuracy of ten individual SOM rankings per substrate. Since bootstrap aggregation is known to increase the robustness of a model,[41,42] several methods of rank aggregation were investigated to identify the best method for producing a single "consensus" rank-ording of sites from a set of bootstrapped models. While at first glance this may appear to be a simple extension of the bootstrap aggregation techniques often utilized in regression models, the combination of ranked lists is quite another problem.

Rank aggregation is a classical problem stemming from social choice and voting theory; methods for determining victorious candidates or parties from individual voter rankings vary between different implementations of representative government.[43] In recent years, the computer science community has investigated consensus ranking in a number of settings, including the merging of query results from multiple databases or optimization of Web based queries by combing results from multiple search engines.[44] The "best" aggregation method for a given problem is one which successfully combines the signal from each set or rank-orderings into a single "consensus" ranking that optimizes the problems' objective criteria.

Previous investigations indicate that no single aggregation function is optimal for all problems, especially for real-world situations in which the rankings are noisy, incomplete, or even disjoint.[45] *Klementiev et al.* proposes a general unsupervised framework for learning rank aggregation models but left the inclusion of a dispersion parameter as a future direction.[46] A dispersion parameter links the quality of a vote to the rank-position, thereby representing when different preference information available is expressed differently by different judges (set of rank-orderings). Values for this parameter would indicate whether a specific set of rank-ordering has more "signal" within the top-k ranks, or is the "signal" dispersed equally across all rank-positions. In a different study, supervised rank aggregation was employed; manual hyper-parameters selections for different Borda Count and Markov Chain aggregation functions were made to minimize distances between the aggregated rankings

Chart 1



```
Algorithm 1 Rank Aggregation
  Input
      RO[10][x] ← 10 independent rank-ordering of x topologically distinct substrate SOMs
                  according to predicted regioselectivity

      RP ← The top rank positions considered relevant to final aggregated score.
           Values of 2, 3 and 4 were applied during supervised training.

      Decrement ← The relative difference to contributing score values between different rank
                  positions. Values of .1, .2, .3 and .4 were applied during supervised training.

  Variables
      Score[x] = 0 ← Initialized to 0, this variable represents the final rank-aggregated
                     regioselectivity "score" of each SOM

  Aggregate regioselectivity rank orderings

  for $i = 1$ to 10 do
     for $j = 1$ to RP do
        $Score[RO[i][j]] += 1 - Decrement * (j - 1)$
     end for
  end for

  Output
      CRO ← Consensus regioselectivity rank-ordering is determined by sorting all SOMs
            according to aggregated Score value
```

and experimentally labeled rankings.[47] These works[46,47] were drawn upon with domain knowledge of the ranking constraints of regioselectivity prediction to propose the *regioselectivity rank aggregation* function of RS-Predictor.

Existing rank aggregation methods utilize different weighting schemes in order to define the relative importance between different rank-positions, with higher positions being considered more important than lower ones. This paradigm shifts when the ranking "signal" is not equally dispersed between all items to be ranked but is instead concentrated within the top *k* predictions. In this case the rank aggregation method employed should only consider the top *k* rank-positions from each set of rank-orderings. A single cross-validated prediction from MIRank on a candidate substrate gives a complete rank-ordering of putative sites according to predicted regioselectivity; however, that model was trained on substrate molecules whose average signal is the correct ranking of 2.05 sites of CYP-mediated oxidation out of 20.70 potential sites. When training the *regioselectivity rank aggregation* function of RS-Predictor it was hypothesized that only the top four rank-positions from each MIRank generated rank-ordering of sites would contain a relevant prediction signal.

The *regioselectivity rank aggregation* function implemented within RS-Predictor was trained on the cross-validated MIRank predictions of 100 substrates, each quantified by descriptors obtained from a single minimum energy conformation, with 10 predicted rank-orderings of putative sites per compound. The models used to create these predictions are not the final models presented within this paper. Implementation and supervised hyperparameter optimization is shown in Algorithm 1 (see Chart 1).

The "best" *regioselectivity rank aggregation* function is defined as the one that optimizes the consensus-based regioselectivity prediction rate of a given set of candidate substrates. However, as defined previously, there are three Standard metrics (Top-1, Top-2, and Top-3) by which prediction quality is gauged: A molecule is considered correctly predicted if any experimental site of CYP 3A4-mediated oxidation is ranked first, first or second, or first or second or third within the consensus SOM regioselectivity ranking. During training, consensus ranking from the top 4 rank-positions with a 0.3 decrement in weight for each lesser rank gave optimal results using a Top-1 or Top-2 metric, but using the top 3 rank-positions with a 0.4 weight decrement was found to produce optimal results for the Top-3 metric. Similar differences in predictive performance were observed when models were generated using Boltzmann-averaged descriptors instead of those from single low-energy conformations. When an exponential decay aggregation function was applied that utilized all rank-positions over a broad range of decay rates, the results were found to be consistent with the conclusions of *Sculley*: For real-world situations with noisy and incomplete ranking information, no single aggregation function is optimal across all problems or all metrics.[45] Fortunately, our findings indicate that rank aggregation is robust across different weighting schemes; the maximum difference in rank aggregation performance across a wide variety of rank weighting parameters on the 100 molecule training set was found to be only four predictions.

Lift Metric. One drawback of the Top-*k* metrics reported in previous publications[18−24,27,28] is that overall prediction quality is evaluated on a per-substrate basis, without recognizing that the successful prediction of one substrate may be much more difficult to achieve than the successful prediction of another. Consider the two 3A4 substrates ethanol and valspodar (see Figure 13a), each having only a single observed site of 3A4-mediated oxidation. Even though ethanol has only two potential SOMs and valspodar has 66 potential SOMs, the contribution of a correct prediction on either one of these molecules would be exactly the same when using traditional metrics. The statistical likelihoods of randomly predicting the experimental SOM within two guesses for ethanol and valspodar are 1 and 0.03, respectively, illustrating that a new approach is needed to fairly assess the practical performance of different prediction methods on real-world data sets. To address this problem, we developed a new metric based upon lift — a measure of how predictions from a real model compare with those from a random model.

The first step in creating a random Top-k model involves determining the likelihood of randomly picking an observed SOM from all possible sites on a given substrate within k guesses. Conceptually, this is no different from the classic probability problem of having a bag filled with marbles that are one of two colors, and one must pick a specific color combination of marbles from the bag within k guesses with no replacements. For a given substrate i (bag), composed of M metabolophores (marbles), of which OM are CYP-oxidized (colored blue), and NOM (NOM == M − OM) are nonoxidized (colored red), the likelihood of an accurate random Top-k prediction can be determined as follows

$$R_k(i) = P(\textit{at least one blue marble in k picks from M with no replacements})$$

The converse of this event

$$P(\textit{all red marbles in k picks from M with no replacements}) = \frac{\binom{NOM}{k}}{\binom{M}{k}}$$

may be calculated as shown above with the hypergeometric distribution, resulting in the expression

$$R_k(i) = 1 - \frac{\binom{NOM}{k}}{\binom{M}{k}} \qquad (10)$$

The $R_k(i)$ value of substrate i represents the statistical difficulty of randomly predicting an observed oxidation site on that substrate within k rank-positions, which is then used to define the substrate lift weight $(1)/(R_k)$. Substrates with higher lift weights have higher proportional contributions to the overall prediction accuracy when using a Lift metric (eq 11). The $(1)/(R_k)$ lift weights of ethanol and valspodar are 1 and 33.3, respectively. The average Top-2 lift weight for the full set of 394 substrates is 7.58, where each molecule had an average of 20.70 potential sites of oxidation and 2.05 observed sites of metabolism.

Gauging of Prediction Accuracy. To determine the overall accuracy of the predicted rank-orderings (RO) of method X on a data set of N substrates using a Top-k Lift metric, the following equation is employed

$$Top - k - Lift\ Accuracy(N, RO^X, k) = \frac{\sum_{i=1}^{N} P_k(i, RO_i^X)\frac{1}{R_k(i)}}{\sum_{i=1}^{N} \frac{1}{R_k(i)}} \qquad (11)$$

where P_k, descriped below in eq 13, is the rank-order accuracy of the metabolophores on substrate i using method X and a standard Top-k metric. This is the more complicated version of the Standard Top-k metric

$$Top - k - Standard\ Accuracy(N, RO^X, k) = \frac{\sum_{i=1}^{N} P_k(i, RO_i^X)}{size\ N} \qquad (12)$$

which simply represents the percentage of molecules from the set which had a site of oxidation predicted within the top k rank-positions by method X. The Lift metric uses the random likelihood of a making an accurate prediction for a given substrate as a means of gauging the relative contribution of that substrate to the overall prediction rate. Indeed, if each compound from the data set being predicted has the exact same random likelihood of being correctly predicted, and therefore the same $(1)/(R_k)$ values, the two functions become identical.

The Standard Top-k metric P_k (eq 13) is defined as follows: A substrate molecule is considered to be correctly predicted if any observed experimental site of CYP-mediated metabolism is ranked first, first or second, or first or second or third for respective k values of one, two, and three. This approach works well when each potential site is assigned a unique rank, but methods such as StarDrop and SMARTCyp occasionally assign equal ranks to more than one site. For example, in the current work, SMARTCyp and StarDrop were found to assign equal ranks to topologically unique sites in 62 and 30 cases, respectively. This creates an additional level of ambiguity when gauging relative prediction quality, as illustrated by one real and two hypothetical regioselectivity predictions on three example substrates in Figure 7. On the other hand, by grouping all topologically equivalent SOMs within a particular metabolophore and then ranking each metabolophore according to a 540 dimension descriptor weight vector, RS-Predictor ensures that each SOM is given a unique rank-position. When each putative SOM has a unique rank-position, the calculation of P_k is trivial and binary — either there is an observed SOM within the top k predicted SOMs, 1, or there are none, 0. When topologically distinct sites receive the same rank (as can be the case with SMARTCyp and StarDrop), gauging the prediction accuracy of each competitive method is no longer trivial.

Examples of the difficulties associated with evaluating StarDrop or SMARTCyp prediction results can be seen in the molecules illustrated in Figure 7, where multiple topologically distinct sites of a given substrate have been assigned identical predicted rank-positions. Example 7a is an actual prediction made by StarDrop for 1_nitropyrene, in which two topologically distinct SOMs are both ranked 1^{st} but only one of them is an observed oxidation site. Given that observation, should the molecule be considered correctly predicted by StarDrop using a Top-1 metric? What should the value of P_1 be? The predictions made by SMARTCyp on 1_nitropyrene are also ambiguous; a single nonoxidized SOM is predicted in the first rank-position, and four distinct SOMs are predicted in the second rank-position — only two of which are sites of 3A4-mediated oxidation. How should a Top-2 metric be applied in this case? To the casual user of these methods, each of these outcomes might considered a success even though they may be of limited value to a medicinal chemist. Examples 7b and 7c represent different possible site rankings for the same molecule. In each of these hypothetical cases, two distinct SOMs are ranked 1^{st} and 2^{nd}, while two other SOMs are equivalently ranked 3^{rd}, but only one out of the four predicted SOMs is an experimentally observed oxidation site. How should a Top-3 metric be applied to these two cases? Should it make a difference to P_3 if the observed SOM is one of the sites distinctly ranked 1^{st} and 2^{nd} (7b), or if it was one of the two sites ranked 3^{rd} (7c)? To answer these questions, we have applied domain knowledge and basic probability theory in an attempt to develop a framework that fairly gauges the prediction power of each method.

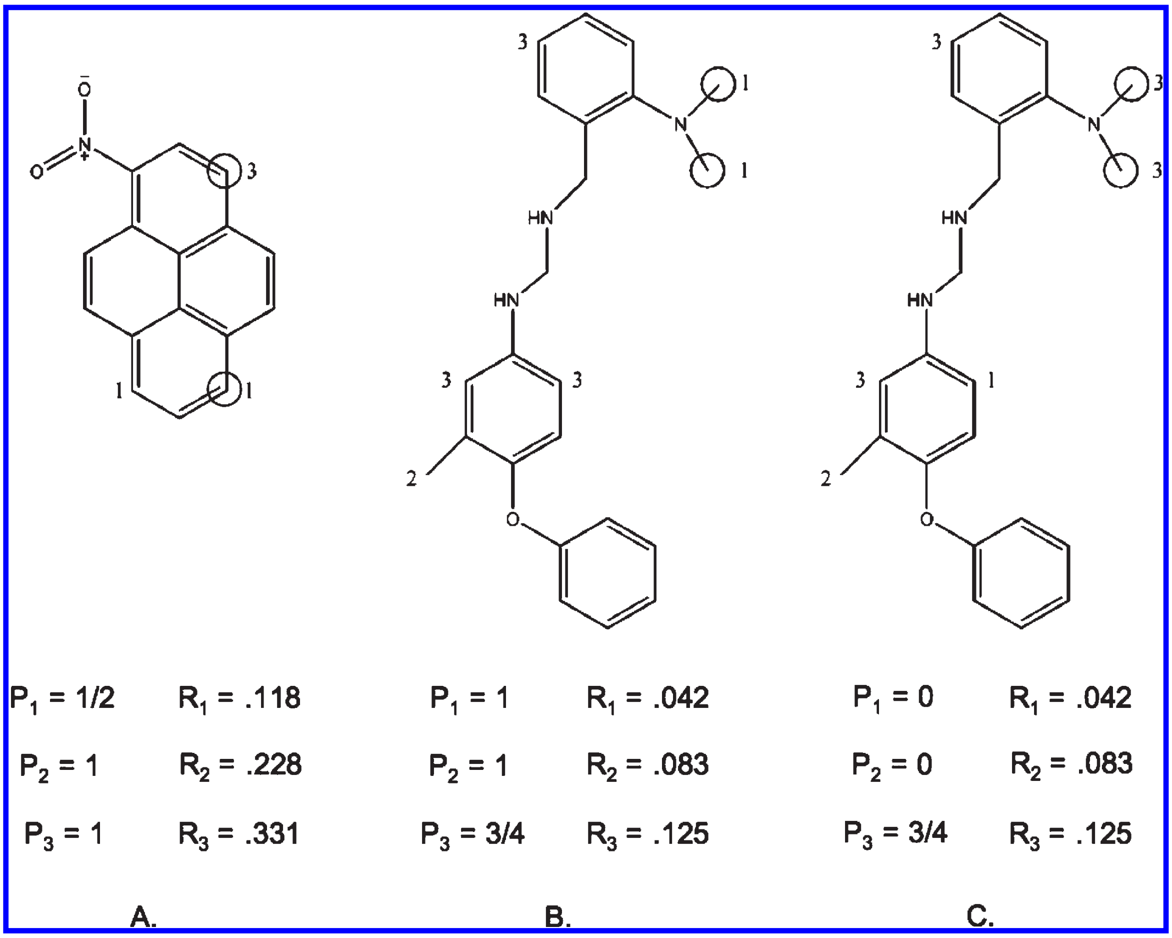


Figure 7. Example substrates having topologically distinct SOMs with the same predicted rank-position. 3A4-oxidized SOMs are circled, while number labels indicate rank-position of putative SOMs according to a predicted regioselectivity. Example a is an actual prediction of StarDrop on 1_nitroprYene, while examples b and c are artificial. Below each substrate are corresponding P_k and R_k values. As detailed in eq 13, P_k is the prediction accuracy of the given top k ranked SOMs using a Standard Top-k metric; in similar fashion, R_k (eq 10) is the statistical likelihood of randomly "picking" at least one observed SOM from all putative SOMs of the substrate within k guesses. Each substrate is assigned a lift weight of $(1)/(R_k)$ when calculating the substrate prediction accuracy with a Lift metric (eq 11).

The prediction accuracy of an input rank-ordering (RO) of putative SOMs of substrate i with a Top-k metric is determined in a similar fashion to the random model

$$P_k(i, RO) = 1 - \frac{\begin{pmatrix} NOM \\ k \end{pmatrix}}{\begin{pmatrix} M \\ k \end{pmatrix}} \qquad (13)$$

the only difference being the set of M metabolophores (marbles) is not composed of all potential sites of the substrate but of all metabolophores that are predicted within the top k rank-positions of the input RO. To ensure fairness, in all cases where x metabolophores are predicted to have the same rank-position, subsequently ranked metabolophores have their corresponding rank-positions incremented by $x - 1$, thereby not allowing the RO as many additional predicted SOMs. In 7a for example, since two topologically distinct sites are ranked 1^{st}, the next predicted site is considered to be ranked in the 3^{rd} position instead of the 2^{nd} position. After rank-positions are adjusted, all metabolophores within the top k rank-positions are placed into a prediction set M, thereby losing the distinction of whether a given metabolophore was ranked 1^{st} or 2^{nd} or 3^{rd} in the original scheme. The rationale for this decision comes from the Standard Top-k metrics of prediction quality; a substrate is considered correctly predicted if *any* CYP-oxidized SOM is predicted in *any* of the top k rank-positions; therefore, the exact rank-position ($\leq k$) an oxidized metabolophore occupies is not relevant when gauging the accuracy of a prediction.

It is important to remember that the determination of prediction set M is directly dependent upon the value of k. The same RO will have potentially different prediction sets for different Top-k metrics, with corresponding differences in NOM, M, and $P_k(i,RO)$. Also, when M is equal to k, $\left(\begin{smallmatrix} M \\ k \end{smallmatrix}\right)$ becomes 1, and the resultant P_k value will be either 0 or 1 — the binary Top-k metric traditionally applied to gauge regioselectivity prediction rates. To satisfy this constraint, each putative SOM must have a distinct rank-position, such as any rank-ordering produced by RS-Predictor, or the value of the Top-k metric employed must equal the number of topologically distinct SOMs ranked in positions < k, which is the case when calculating P_2 in Figure 7a.

The determination of P_k through the statistical likelihood of randomly "choosing" an observed oxidation site from a given prediction set within k guesses gives credit for an accurate

Table 3. Standard Prediction Rates for the Complete, Calibration, and External Data Sets[b]

metric	RS Minimum Consensus	RS Minimum Average	RS Boltzmann Consensus	RS Boltzmann Average	SMARTCyp Version 1.1	StarDrop Version 4.2.1	Random Model
Top-1(Com.)	59.6	55.1 ± 2	60.7	55.2 ± 2	**63.4**	58.8	10.1
Top-2(Com.)	77.9	75.7 ± 1	**78.2**	75.2 ± 2	73.2	73.8	19.3
Top-3(Com.)	86.8	84.5 ± 1	**87.3**	84.4 ± 1.3	79.8	82.7	27.5
Top-1(Cal.)	57.5	56.5 ± 1.7	60.3	58.6 ± 1.7	**63.1**	61.3	9.7
Top-2(Cal.)[a]	76.7	75.9 ± 1.6	**78.0**	76.2 ± 2.2	72.8	76.3	18.5
Top-3(Cal.)	85.4	85.1 ± 1.3	**87.6**	85.3 ± 1	79.1	84.3	26.7
Top-1(Ext.)	54.2	53.5 ± 3.6	55.6	54.2 ± 3.5	**64.6**	47.2	11.7
Top-2(Ext.)	**75.0**	73.1 ± 3.5	73.6	72.8 ± 3.2	74.5	62.5	22.5
Top-3(Ext.)	**84.7**	81.6 ± 3.3	83.3	82.0 ± 2.3	83.3	75.3	31.2

[a] 77.4 and 61.8 respectively for Merck method and MetaSite Version 2.7.5 for nonupdated Calibration set. [b] Optimal model in **bold**.

Table 4. Lift Prediction Rates for the Complete, Calibration, and External Data Sets[a]

metric	RS Minimum Consensus	RS Minimum Average	RS Boltzmann Consensus	RS Boltzmann Average	SMARTCyp Version 1.1	StarDrop Version 4.2.1
Top-1(Com.)	55.4	50.1 ± 2.5	**57.8**	51.1 ± 2.6	57.2	53.8
Top-2(Com.)	74.4	71.9 ± 1.4	**75.0**	70.9 ± 2.4	66.4	69.6
Top-3(Com.)	83.9	81.2 ± 1.7	**84.3**	80.6 ± 1.6	73.9	79.5
Top-1(Cal.)	54.3	53.3 ± 1.7	**57.8**	55.2 ± 2	57.1	55.8
Top-2(Cal.)	72.9	72.1 ± 1.5	**74.4**	72.9 ± 2.5	65.4	71.5
Top-3(Cal.)	81.7	81.5 ± 1.1	**84.8**	82.0 ± 1.3	73.0	80.9
Top-1(Ext.)	46.0	45.9 ± 4.3	46.9	46.8 ± 4.1	**61.7**	43.1
Top-2(Ext.)	**74.5**	70.9 ± 4.5	73.1	70.3 ± 4.3	72.0	59.2
Top-3(Ext.)	**83.4**	79.3 ± 3.8	81.4	79.7 ± 2.8	82.2	72.0

[a] Optimal model in **bold**.

prediction, while penalizing the prediction of more than k distinct SOMs.

Another way in which QSAR models are traditionally judged is through ROC curves that assess true-positive and false-positive prediction rates. Regioselectivity predictions do not fit into this paradigm, as Top-k prediction accuracy is gauged solely through relative rankings of different SOMs of the same substrate. An oxidized SOM may be predicted in the second rank-position, below a nonoxidized SOM, and the substrate would still be considered correctly predicted using a Top-2 or Top-3 metric. That SOM would be considered a true-positive with respect to other SOMs of the substrate and a false-negative relative to the SOM in the first rank-position. A different true-positive/false-positive gauge was designed to both address this type of issue and to present information potentially relevant to both medicinal chemists and future regioselectivity modelers.

As illustrated earlier in Table 1 and Figure 2, the substrates of a given data set may be broken down into composite SOMs, which are then classified into pathway sets based upon the reaction mechanism that they have the potential to undergo. A reaction pathway set is thereby composed of a certain number of SOMs that undergo CYP-mediated oxidation, and a majority of SOMs that do not undergo oxidation. Ratios of (observed)/(potential) SOMs within each set are used in the top panel of Figure 10 to gauge the pathway-based catalytic propensities CYP 3A4 for different substrate sets. The predicted rankings of a given model for the composite observed and potential SOMs of a set are used as a means of gauging true-positive and false-positive rates in the lower panel of Figure 10. This type of analysis gives insights into the strengths and weaknesses of a given method at identifying SOMs with the potential to undergo specific oxidation pathways; when applied to the predictions made by different methods upon the same set of substrates, chemistry-based insights as to where and when a given model has the best application are illustrated.

RESULTS AND DISCUSSION

Overall prediction rates of RS-Predictor, SMARTCyp, and StarDrop for the Complete data set of 394 substrats are shown for Standard metrics in Table 3 and Lift metrics in Table 4. RS-Predictor models were generated from substrates characterized using descriptors from either a single minimum energy conformation or a Boltzmann average of the descriptor values of up to 25 conformers per compound. Average results of RS-Predictor represent the overall prediction rates with each compound having 10 independently predicted rank-orderings of putative SOMs generated through 5-fold cross-validation (CV). Since multiple predictions are made for each molecule, the overall average prediction rate has a corresponding distribution and standard deviation. Consensus RS-Predictor results were obtained by rank aggregating 10 predictions for each compound into a single predicted rank-ordering of potential SOMs, which then determine overall performance. Performances are also given for two other regioselectivity prediction methods, SMARTCyp and StarDrop. The prediction enrichment of each method may be gauged through the baseline performances of the Random Model. This model is simply an average of the R_k (eq 10) values for each substrate within the given data set for the given Top-k

metric. There is no Random Model for the Lift metrics since these metrics incorporate R_k on a substrate by substrate basis.

To ascertain whether results generated from 5-fold CV models are truly representative of how RS-Predictor would perform on potential lead candidates, additional tests were performed. The updated data set of 322 substrates released by *Sheridan et al.* was used as a Calibration set for RS-Predictor, with the additional 72 compounds being used as an External test set. Consistent with RS-Predictor implementation, each molecule from the Calibration set has 10 independent predictions, obtained through 10 iterations of 5-fold CV. The 50 different Calibration models were then applied to each compound of the External test set. When these models were applied to a second external test set, 20 proprietary 3A4 substrates from a partnering major pharmaceutical company, 85% of the substrates were correctly predicted by RS-Predictor with a Top-2 metric.

Knowledge of the potential CYP-mediated regioselectivity of oxidation is an important aspect of drug discovery. While determination of metabolites by experimental means is accurate, such procedures are neither quick nor cheap enough to be applied to all compounds under development. Consequently there has been an emphasis on the development of new *in silico* regioselectivity prediction methods in recent years. The usefulness and quality of each method depends mainly on overall accuracy of regioselectivity prediction and the speed with which the method is able to predict the likely metabolites of new compounds. These were the two factors used to gauge relative benefits of using descriptors from a single-minimum energy conformation versus Boltzmann averaging of descriptor values of up to 25 conformations. Comparing overall results between different RS-Predictor Consensus models for the full 394 data set with a Standard Top-2 metric reveals that Boltzmann models correctly predicted 12 substrates not predicted by Minimum models, while Minimum models predicted 11 molecules that were missed by Boltzmann models. Differences in overall performance across metric and data set as well as pathway-specific prediction rates were therefore found to be minimal. Model differences are likely due to variance in different CV partitions, as opposed to differences in descriptor information content. In light of the fact that calculating Boltzmann averaged descriptors takes on an average of 22.50 s per compound, while using a single conformation takes on an average 1.5 s per compound, future implementations of RS-Predictor will utilize only a single energy-minimized structure.

A single calibrated weight vector is a set of 540 descriptor coefficients that implicitly represents the oxidative regioselectivity trends of a given subset of substrates. The 148 topological descriptors are quickly determinable from 2D structure, while the 392 quantum chemical descriptors are quickly derived from the output of MOPAC. The rate-limiting step of descriptor generation for a specific conformation is the AM1 MOPAC wave function calculation. Once performed, the time needed to extract all quantum chemical descriptor values from the MOPAC output is negligible. To lessen the risk of developing overdetermined models from a large number of potentially correlated descriptors, regularization was employed within MIRank. A complementary technique that could be used to lessen the likelihood of model overtraining is feature selection. Unsupervised feature selection was not employed, because the removal of features based upon trends within a given calibration set could have eliminated significantly correlated features with different relative regioselectivity signal for a different set of substrates; rank aggregation is specifically designed to take advantage of signal of this nature. Supervised feature selection can be challenging since descriptors that have little signal for one isozyme may in fact have greater relevance for the regioselectivity of other isozymes. Consequently, a comprehensive supervised feature selection study should be performed when creating RS-predictor model for multiple substrate sets of different isoforms.

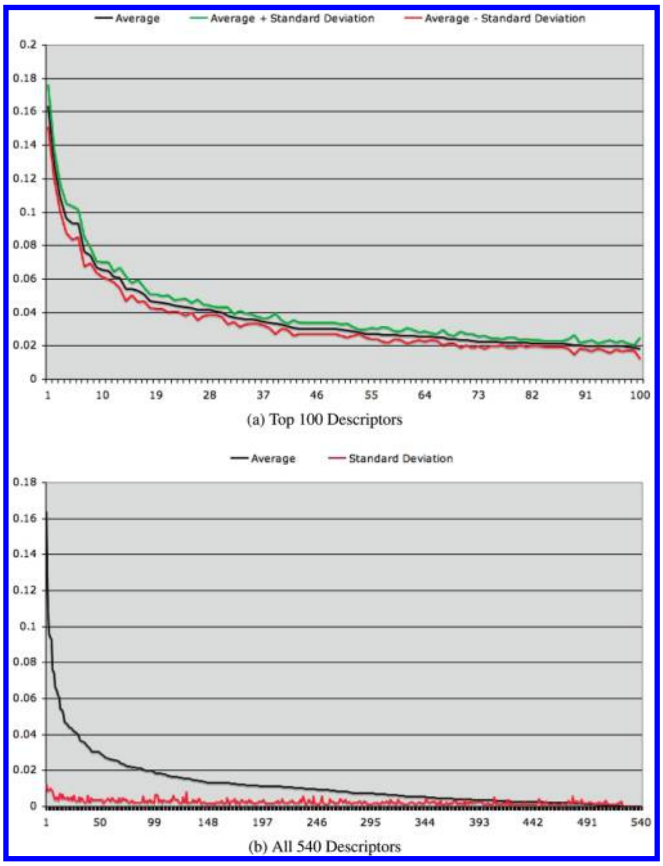


Figure 8. The weights and standard deviations of the top 100 descriptors (a) and all 540 descriptors (b) given in decreasing order according to absolute weight value. Values were averaged from 10 iterations of MIRank employing 5-fold CV upon 394 substrates of 3A4 quantified with descriptors from a single energy-minimized conformation.

Model interpretation, gauged in this case through the relative weights of 540 unique descriptors, is an important aspect of QSARs. Unfortunately the same factors that make RS-Predictor models robust, large numbers of potentially correlated descriptors with 10 independent iterations of 5-fold CV, make model interpretability more difficult. The relative importance of different descriptors averaged across 10 CV iterations is illustrated in Figure 8a through the 100 descriptor weights with the highest absolute values. Absolute values were used because some descriptors are positively correlated with a metabolophore representing an observed SOM, while some are negatively correlated. The interested reader may find weights and correlation for each descriptor and each CV model within the Supporting Information.

The ten descriptors with the highest absolute weight across all models, with corresponding positive or negative correlation to P450-mediated regioselectivity, are listed in order of decreasing weight as follows: NA_0_S(+), Span(+), atom_area(+), N(+), AR(-), NA_0_O(-), BL(+), MR(-), SC_1_sum(-),

Table 5. Percentage Performance Increases of Consensus Prediction Rates over Average Prediction Rates

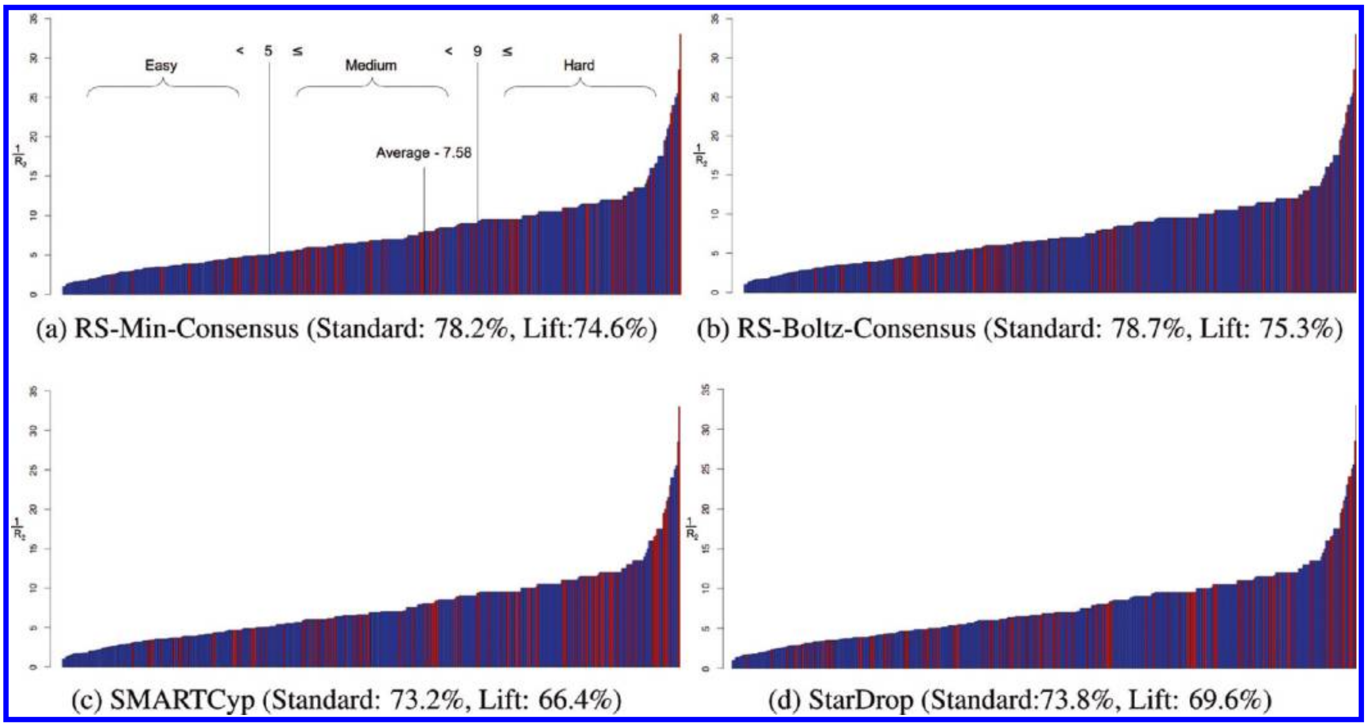
metric	Minimum Complete	Boltzmann Complete	Minimum Calibration	Boltzmann Calibration	Minimum External	Boltzmann External
Standard Top-1	4.5%	5.5%	1.0%	1.7%	0.7%	1.4%
Standard Top-2	2.2%	3.0%	0.8%	1.8%	1.9%	0.8%
Standard Top-3	2.3%	2.9%	0.3%	0.3%	3.1%	1.3%
Lift Top-1	5.3%	6.7%	1.0%	2.6%	0.1%	0.1%
Lift Top-2	2.4%	4.1%	0.8%	1.5%	3.6%	2.8%
Lift Top-3	2.7%	3.7%	0.2%	2.8%	4.1%	1.7%

NA_0_C(-). While it is understandable why an atom being a sulfur, or having a high nucleophilicity or average exposed surface area, would indicate oxidative potential, reverse engineering a simple regioselectivity ranking-function based upon smaller descriptors subsets would be a challenging problem. The smaller signal inherent to having fewer input descriptors would likely give interpretable models, but such models would not necessarily extrapolate well toward predicting the mediated regioselectivity of other isozymes. Meanwhile, the average weights and standard deviations of all 540 descriptors illustrated in Figure 8b represent a large amount of variance and signal that may be exploited when applying the RS-Predictor algorithm to any set of known substrates.

Statistical descriptor trends may vary slightly between different calibration partitions, but their broad applicability to validation or external compounds has been demonstrated. The standard deviation in Average Complete model prediction rates is ±2%, a maximal difference in correct/incorrect predictions of 16 molecules between individual models. The variance in results when applying the 50 Calibration models to the External data set is ±3.6%, a maximal difference of 6 molecules between individual models. Choosing a single model may not result in optimal, or even average, performance, which is why rank aggregation was employed. Even though the primary motive for model aggregation was the simplification of substrate prediction representations for a medicinal chemist end-user, it is encouraging to see that Consensus performance rates are never below those of the Average performance rates. Indeed, Consensus model performance bumps are often near the top end of the standard deviation range of Average performance rates. Rank aggregation circumvents difficulties in averaging weight vectors across multiple models by first applying those individual statistical regioselectivity trends to each candidate substrate and only then merging the response. While increases in performance is matched by a corresponding increase in difficulty for Consensus model descriptor based interpretation, other conclusions may be drawn.

In Table 5, increases in Consensus prediction rates over Average prediction rates do not appear correlated to the different Top-k metrics, Standard versus Lift metrics, data sets, or descriptor sets. Since rank aggregation is a generic technique for merging multiple sets of signals into a single signal set, this is not unexpected. It is, in fact, encouraging that even though the aggregation function was optimized using the predictions for 100 molecules from the Calibration set, performance increases for both External and Complete data sets were were even greater than those observed for the Calibration set. This leads us to believe the rank aggregation function trained on 3A4 substrates may safely be used to make consensus predictions for other isozyme models. Comparing consensus ranked prediction rates to average rates obtained from either 10 or 50 predictions per molecule in Table 5 reveals an average Consensus performance increase of 3% across all metrics for the Complete data set; the regioselectivity predictions of 12 molecules improve when consensus ranking is applied. The fluctuations of performance increases observed between Minimum and Boltzmann models support the previously mentioned hypothesis — variance in model performance most likely arises from the differences between random CV partitions, as opposed to differences in descriptor quality. We extend this hypothesis further by incorporating the findings of *Sculley*, who proposed that no rank aggregation function is optimal across all problems.[45] Differences in performance bumps between different aggregation functions (each with their own hyperparameters) will likely be more dependent upon the random partitions used to generate CV models, rather than by the specific data set, descriptor set, or performance metric associated with those models. Investigation into rank aggregation has yielded a systematic framework whereby the signal of independently generated regioselectivity rank-ordering of substrate SOMs may be combined into a single consensus SOM ordering by predicted regioselectivity. While the *regioselectivity rank aggregation* function proposed in this work is not guaranteed to be the optimal aggregation function for any CYP substrate set, metric, descriptor set, or set of CV partitions, benefits in improved overall performance as well as simplification for medicinal chemists being able to look at a single set of predicted SOMs for a given substrate justify its use.

Direct comparison between Merck and MetaSite (version 2.7.5) results with those of RS-Predictor, SMARTCyp, and StarDrop is difficult. Top-2 metric performance rates from *Sheridan et al.* were presented using separate substrate sets of size 316 (structures were only released for 305) and 19, which have been merged as follows: Merck = $77 \times (316)/(335) + 84 \times (19)/(335) = 77.39\%$, MetaSite = $62 \times (316)/(335) + 58 \times (19)/(335) = 61.77\%$. Calibration set results for RS-Predictor, SMARTCyp, and StarDrop were calculated upon the updated Merck data set, which had 2 molecules removed and 85 compounds with fixed structures or updated responses. Despite the data set discrepancies, it is encouraging to see that Consensus model results statistically match previously reported efforts. It is also revealing to compare External results with Calibration results. Performances for RS-Predictor remain statistically equivalent between sets. At worst, the External Boltzmann predictions only miss three molecules more than an equivalent Calibration model might miss. This indicates that Consensus models for the Complete data set are not likely to be overfit. Calibration model performance on the 20 proprietary compounds further demonstrates the robustness of RS-Predictor; the oxidation sites of certain compounds, which in-house methods of the partnering pharmaceutical company were unable to identify, were blindly predicted by Calibration RS-Predictor models. As should be expected with a first-principles approach,



(a) RS-Min-Consensus (Standard: 78.2%, Lift:74.6%)

(b) RS-Boltz-Consensus (Standard: 78.7%, Lift: 75.3%)

(c) SMARTCyp (Standard: 73.2%, Lift: 66.4%)

(d) StarDrop (Standard:73.8%, Lift: 69.6%)

Figure 9. Predictions for the 394 Complete data set using a Top-2 metric. Substrates are sorted by increasing $(1)/(R_2)$ value, with the average being 7.58. Molecules are also classified according to $(1)/(R_2)$ value, with $(1)/(R_2) < 5$, $5 \leq (1)/(R_2) < 9$, and $9 \leq (1)/(R_2)$ denoting easy, medium, and hard sets, respectively. These values were chosen to obtain the most even distribution between sets, with each set having 130, 136, and 128 substrates, respectively. Blue column color indicates the method predicted the substrate correctly, red column indicates the method was unsuccessful.

Table 6. Percentage Performance Differences between Standard Prediction Rates and Lift Prediction Rates

data set	metric	RS Minimum Consensus	RS Minimum Average	RS Boltzmann Consensus	RS Boltzmann Average	SMARTCyp	StarDrop
Complete	Top-1	4.2%	5.0%	2.9%	4.1%	6.2%	5.0%
	Top-2	3.5%	3.8%	3.2%	4.3%	6.8%	4.2%
	Top-3	2.9%	3.3%	3.0%	3.8%	5.9%	3.2%
Calibration	Top-1	3.2%	3.2%	2.5%	3.4%	6.0%	5.5%
	Top-2	3.8%	3.8%	3.6%	3.3%	7.4%	4.8%
	Top-3	3.7%	3.6%	2.8%	3.3%	6.1%	3.4%
External	Top-1	8.2%	6.6%	8.7%	7.4%	2.9%	4.1%
	Top-2	0.5%	2.2%	0.4%	2.5%	2.5%	3.3%
	Top-3	1.3%	2.3%	1.9%	2.3%	1.1%	3.4%

SMARTCyp is also robust across sets, with an External rate 1.7% higher than the Calibration rate. Meanwhile the Top-2 External rates of StarDrop are 13.8% lower than Calibration rates. In general it appears RS-Predictor does not perform as well as SMARTCyp using a Top-1 metric. However, when Standard and Lift Top-2 and Top-3 metrics are applied, RS-Predictor outperforms both SMARTCyp and StarDrop on all data sets.

As defined earlier, Lift rate calculations are based on the idea that specific molecular predictions should be weighted according to R_k — the statistical likelihood randomly picking a substrate's oxidation site(s) out of all potential substrate SOMs within k guesses. In Figure 9, each substrate in the Complete data set is represented through corresponding Lift weight of $(1)/(R_2)$; sorting substrates by weight gives a visualization of the prediction difficulty distribution of the data set. Larger, "harder", more statistically difficult to predict molecules receive higher weights than smaller, statistically easier to predict molecules when calculating Lift prediction rates. To interpret the meaning of a Lift rate, it is helpful to consider its corresponding Standard prediction rate and difficulty distribution. Any Standard Top-k metric would give all bars an equal height of 1. Meanwhile, the average $(1)/(R_2)$ lift value for the Complete data set is 7.58. The Standard rate may be thought of as a useful baseline, all molecules being considered equally, while Lift rates reflect the $(1)/(R_k)$ difficultly distribution of the molecules correctly predicted for the given data set by a given prediction method. Consider the case where two different methods correctly predict the same number of molecules from the same data set, but one method produces more correct predictions on "difficult" molecules than the other one. Each method would then have the same Standard prediction rate, but there would be an important difference in the utility of the two methods, reflected through their respective Lift rates. Methods with higher Lift rates, such as RS-Predictor, are likely to be more accurate at predicting larger substrates, whereas

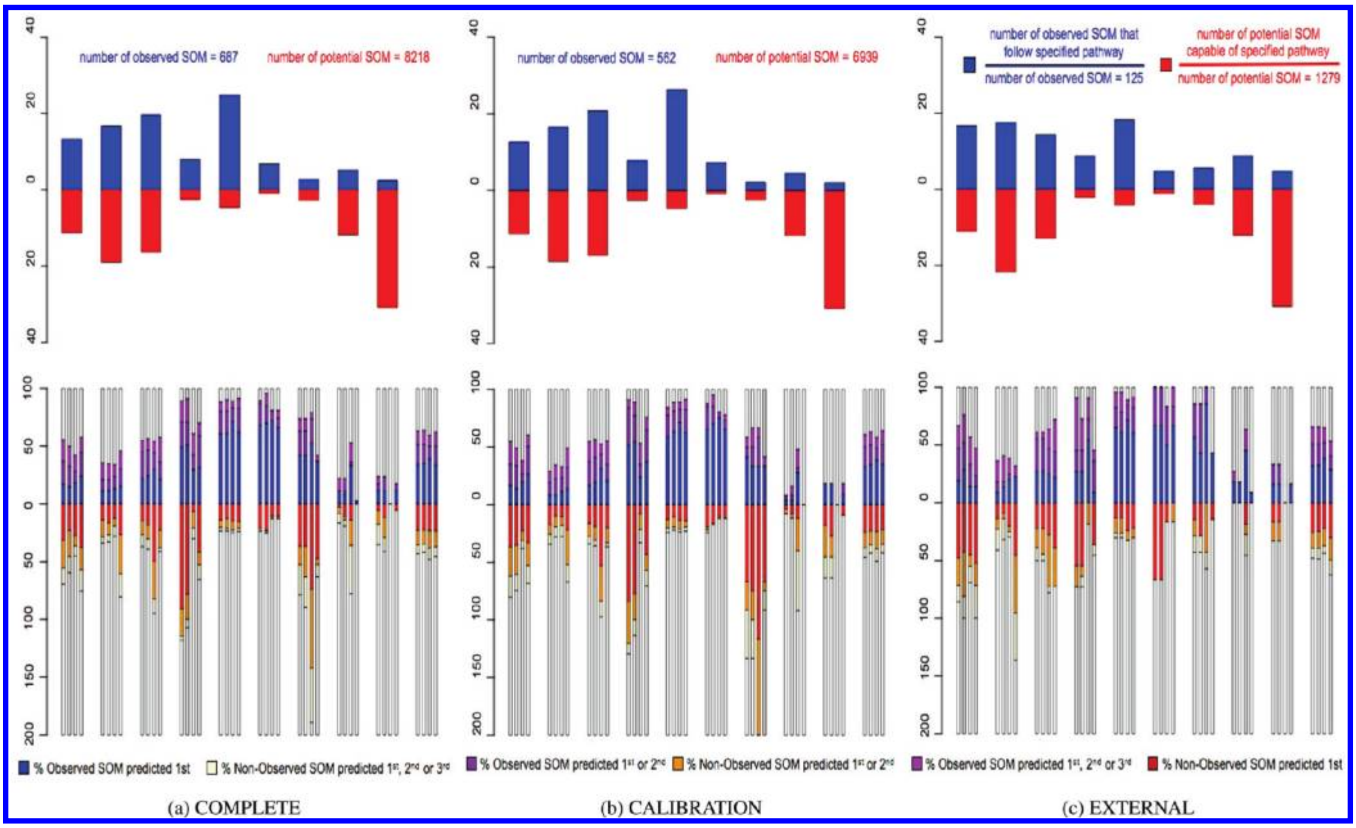


Figure 10. Metabolic propensities and method predictions rates broken down by CYP-mediated pathway and data set. Major column designations from left to right correspond to the biotransformations initially presented from top to bottom in Figure 2: Csp^3 Hydroxylation, Aromatic Hydroxylation, Ring Hydroxylation, O-dealkylation, N-dealkylation, Sulfur-based reactions, Csp^2-based reactions, Nitrogen-based reactions, Uncommon reactions, and summations over all reactions. Pathway prediction columns in the lower graphs are composed of 4 subcolumns, representing from left to right the prediction rates of RS Minimum Consensus, RS Boltzmann Consensus, SMARTCyp, and StarDrop. The *y*-axis percentage of each major column of the lower graphs is based upon the total number of oxidized SOMs for the given pathway for the given set of substrates; this scaling was chosen to ensure that the visualization of true-negatives would not overshadow more interesting results.

those with a lower Lift rates, such are SMARTCyp, may be better suited toward predicting smaller substrates. The greater number of incorrectly predicted red bars at the "harder" end of the difficulty spectrum for both SMARTCyp(9c) and StarDrop(9d) relative to those of RS-Predictor(9a, 9b) are illustrative of this.

In Table 6 the differences between Standard and Lift prediction rates are shown for each metric, method, and data set. In all cases, Lift performance rates were found to be below the corresponding Standard rates. This is not an unexpected event, as any regioselectivity prediction method will be more likely to successfully predict a greater number of "easier" substrates than "harder" ones. The results indicate that for Complete and Calibration sets, RS-Predictor does better on molecules on the "harder" end of the difficulty distribution than StarDrop, which in turn shows better performance than SMARTCyp in this domain. It also appears that differences between Standard and Lift rates decrease as the *k* metric increases, illustrating that when a larger number of predictions are considered, a corresponding increase in the correct prediction of more difficult molecules emerges. Dramatic differences in RS-Predictor performance rates upon the External set between Top-1 and Top-2 metrics demonstrate this. RS-Predictor exhibits a high prediction rate decrease between Standard and Lift Top-1 metrics, indicating that the substrates that are correctly predicted using the highest ranked SOM are at the "easy" end of the difficulty distribution. However, when the second highest rank-position is brought into consideration, the drop in Lift performance becomes negligible; this indicates the oxidation sites of a large number of statistically difficult External substrates are predicted in only the second rank-position of RS-Predictor. Similar differences between Top-1 and Top-2 metrics were not observed for either Complete or Calibration RS-Predictor results; one possible explanation is that Complete and Calibration results come from the aggregation of 10 independent models, while External results come from 50 independent models, which may have a greater degree of variance within the signal to be aggregated. Neither SMARTCyp nor StarDrop exhibited such a significant rate difference between between Top-*k* metrics of the same data set. However, when considering rate differences between Calibration and External sets, SMARTCyp displayed similar Standard rates across sets with a significantly higher Lift rate on the External set. A partial explanation for this is provided by the Random Model, which indicates that the External set is statistically "easier" to predict than Calibration set. Meanwhile, StarDrop showed equivalent differences in Standard and Lift rates for each data set, but External rates were found to be approximately 10% lower than Calibration rates. Extending the *Sheridan et al.* 2D6 and 2C9 data sets and running similar calibration/external comparisons between Standard and Lift rates of multiple methods would likely be revealing.

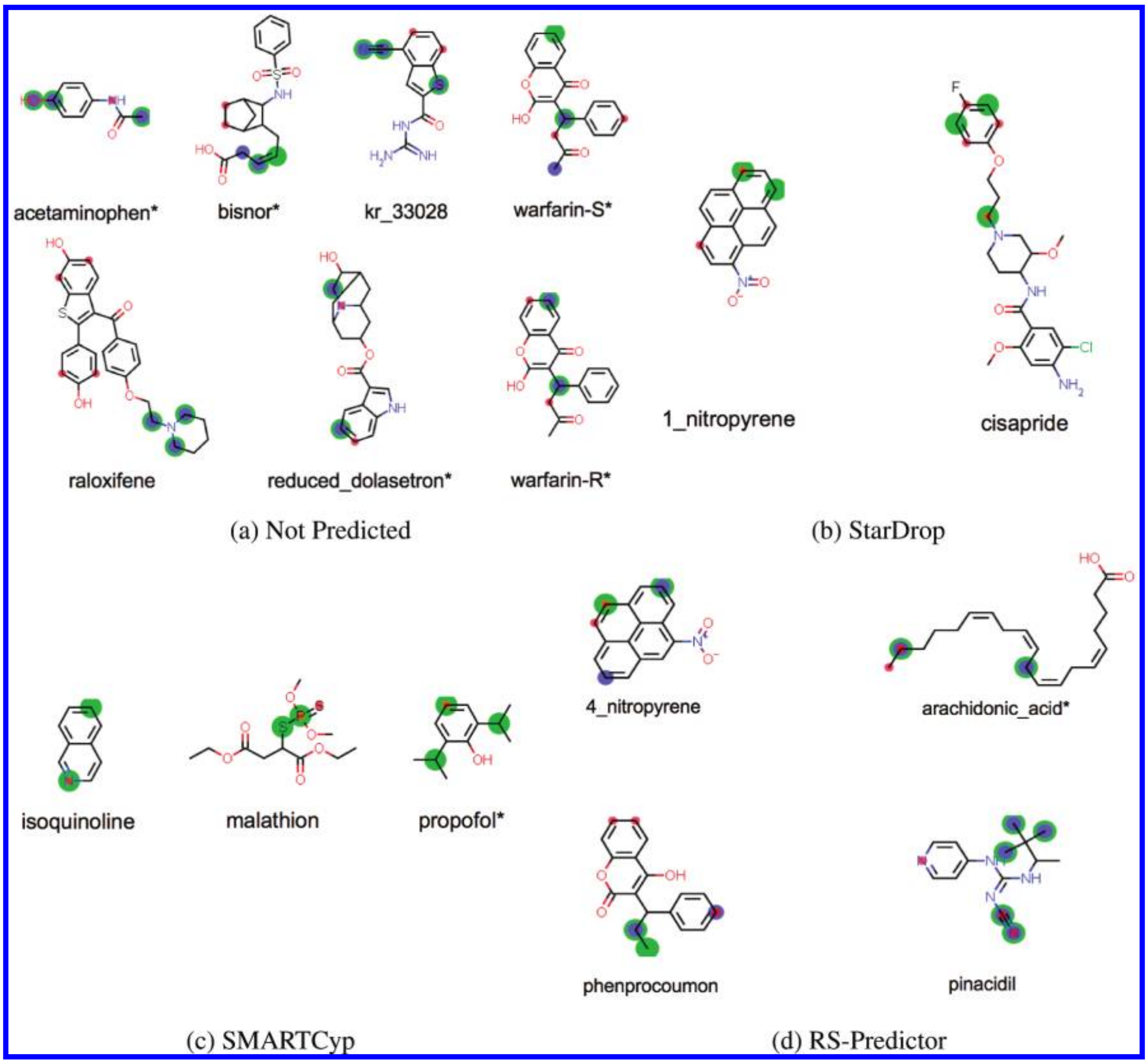


Figure 11. "Easy to predict" ((1)/(R_2) < 5) substrates that were not predicted by any method or only predicted by a single method using a Top-2 metric. Substrates are only considered correctly predicted if their calculated P_1 or P_2 value is 1. Small red circles indicate sites of 3A4-mediated metabolism, while a * identifies substrates from the External set. For panels (b) and (c) green circles denote all SOMs predicted in the top two rank-positions of the labeled method. For (a) and (d) blue and green circles respectively represent the top two predicted sites of Boltzmann and Minimum Consensus models.

It should be noted that representing substrate prediction difficultly solely through ratios of (# oxidized)/(# potential) SOMs makes sense mathematically but misses important chemical information. Every putative SOM has the potential to undergo one or more specific CYP-mediated biotransformations; the classification of all SOMs from a substrate data set into pathway based sets reveals the regioselectivity pathway propensities for a given isozyme, represented as ratios of (# oxidized)/(# potential) SOMs. As shown in Figure 2 and the top panel of Figure 10, CYP 3A4 exhibits different propensities to catalyze different pathways. To assess substrate prediction difficulty without consideration of putative SOM oxidation paths, or the catalytic propensities of the metabolizing isozyme, is to ignore important pieces of chemical and statistical information.

In the current investigation, pathway regioselectivity propensities between Calibration and External sets were found to be similar, as should be expected for two sets composed of 3A4 substrates. The amount of positive signal, in terms of (# observed)/(# potential) SOMs, in the External set (9.77%) is slightly higher than that of the Calibration set (8.10%). A higher percentage of Calibration set oxidation sites undergo N-dealkylation, nonaromatic ring hydroxylation, or sulfur oxidation than for the External set, which in turn has greater propensities for Csp^3 hydroxylation, O-dealkylation, and Csp^2 and nitrogen-based oxidation reactions. These small differences in biotransformation propensities do not appear to affect overall RS-Predictor prediction rates between data sets, though some corresponding affects may be observed from the pathway-based prediction analysis in the lower panel of Figure 10.

There appears to be broad correlation between the regioselectivity propensity, in terms of (observed)/(potential) SOM ratio, and RS-Predictor (true-positive)/(false-positive) prediction rates for a particular data set of substrates. This is especially apparent for biotransformations with high relative propensities such as O-,N-dealkylations as well as S-oxidations, where the average Top-3 (true-postive)/(false-positive) rates are (~90%)/(~20%). High Csp^2 propensity ratios are similar to those of O-dealkylation, though not as high as N-dealkylation or S-oxidation. As a result, RS-Predictor true-positive rates are high,

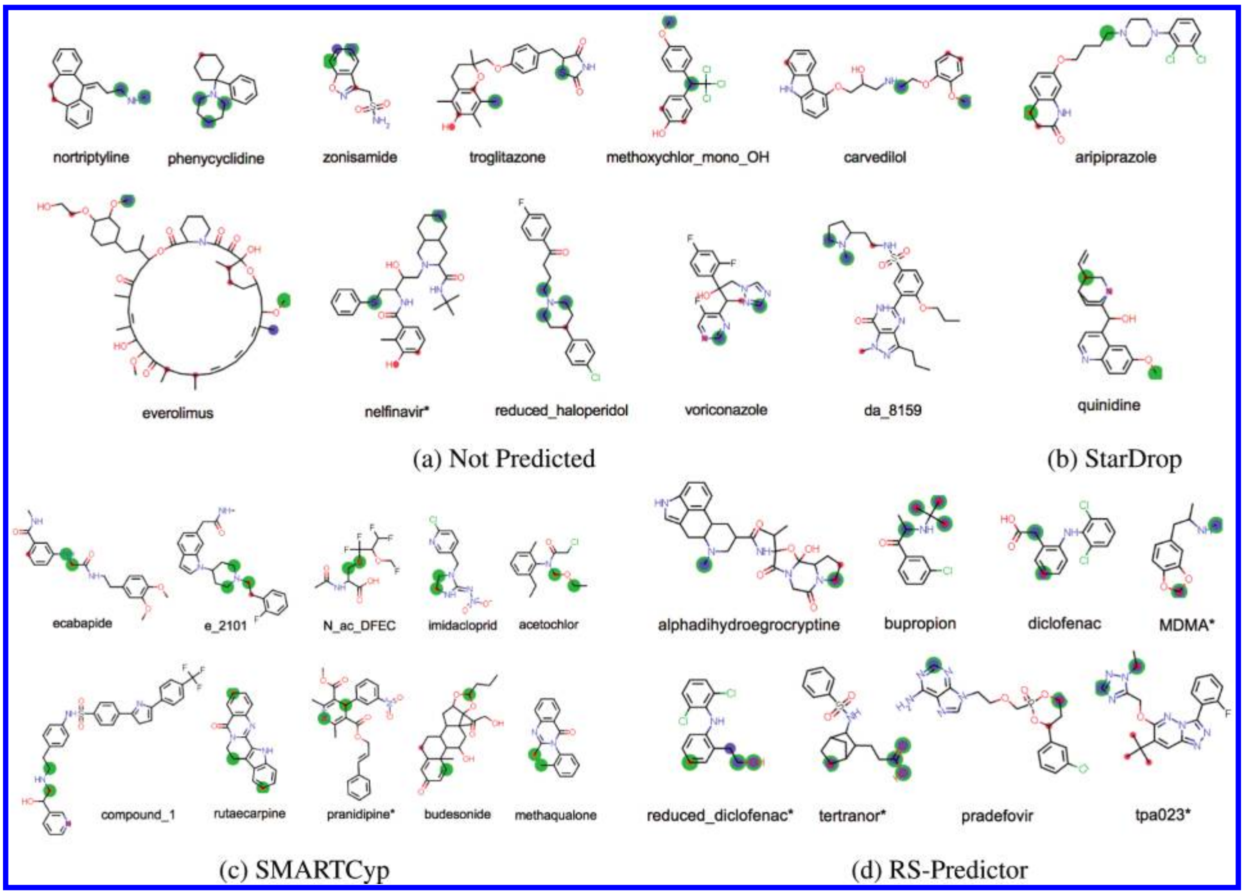


Figure 12. "Moderately difficult to predict" ($5 \leq 1/R_2 < 9$) substrates not predicted by any method or only predicted by a single method.

suggesting that MIRank models successfully utilize topological descriptors to identify which SOMs are capable of undergoing Csp^2 hydroxylation or O-dealkalytion while also predicting that they are more likely to undergo metabolism than SOMs that follow different potential pathways. False-positive rates are also high, indicating that current models have difficulty distinguishing which of these SOMs actually undergo CYP-mediated oxidation and which ones do not. On the other hand, propensity ratios of aromatic ring hydroxylation, nitrogen-based reactions and uncommon reactions are quite low, with correspondingly low RS-Predictor (true − positive)/(false − positive) rates. It is expected that MIRank models utilize regioselective pathway propensities implicitly represented by topological descriptors to make predictions with corresponding pathway tendencies; high descriptor weights for the NA_0_S(+) and AR(-) descriptors corroborate this view. SMARTCyp and StarDrop have similar correlations in (observed)/(potential) SOM ratio to (true − positive)/(false − positive) prediction rates. However, there are visible differences in predicted pathway "preferences" between RS-Predictor, SMARTCyp, and StarDrop, differences greater than those between Minimum and Boltzmann models. Comparisons of pathway preferences across Calibration and External sets for different methods is revealing, especially when viewed in conjunction with Figure 11, Figure 12, and Figure 13, which show every substrate that was successfully predicted by only a single method or not predicted by any method. These figures were respectively created from easy, medium, and difficult to predict substrates (defined in Figure 9 through $(1)/(R_2)$ value) on the basis of a Top-2 metric, where molecules were only considered to be correctly predicted if the P_1 or P_2 value for the predicted SOM rank-ordering was 1.

Slight increases in Csp^3 hydroxylation propensities between data sets are reflected by increases in the true-positive and false-positive prediction rates of both RS-Predictor and SMARTCyp. The slightly higher true-positive rate of RS-Predictor likely comes from these instances [RS-Predictor: arachidonic_acid(E*), buproprion(M), sdz_imm_125(H), cyclosporine(H), ivermectin(H), bms_275183-(H), docetaxel(H)]. The only Csp^3 hydroxylation reaction correctly predicted only by SMARTCyp is methaqualone(M). StarDrop, on the other hand, does quite well at predicting sites of aliphatic hydroxylation on statistically difficult substrates from the Calibration set [StarDrop: capasaicin(H), salmeterol(H), and finasteride(H)] but has a significantly worse (true − positive)/(false − positive) ratio for the External set.

Similar Calibration/External discrepancies are observed in StarDrop predictions of aromatic ring hydroxylations, while smaller improvements in data set propensity ratios (relative to Csp^3 hydroxylation) helps to explain corresponding rises in RS-Predictor and SMARTCyp true-positive rates. The high false-negative rates are reflected in the fact that $(1)/(3)$ of all substrates that are not predicted by any method undergo aromatic hydroxylation [Not Predicted: kr_330328(E), warfarin-R/S(E*), reduced_dolasetron(E*), raloxifene(E), zonisamide(M), methoxychlor(M), carvedilol(M), nelfinavir(M*), 4_hydroxytamoxifen(H), licofelone(H), fluvastatin(H), and py74(H)]. Clearly there is room for all methods to improve. Trazadone(H) represents a surprising case where two potential aromatic hydroxylation sites had a higher RS-Predictor ranking than the

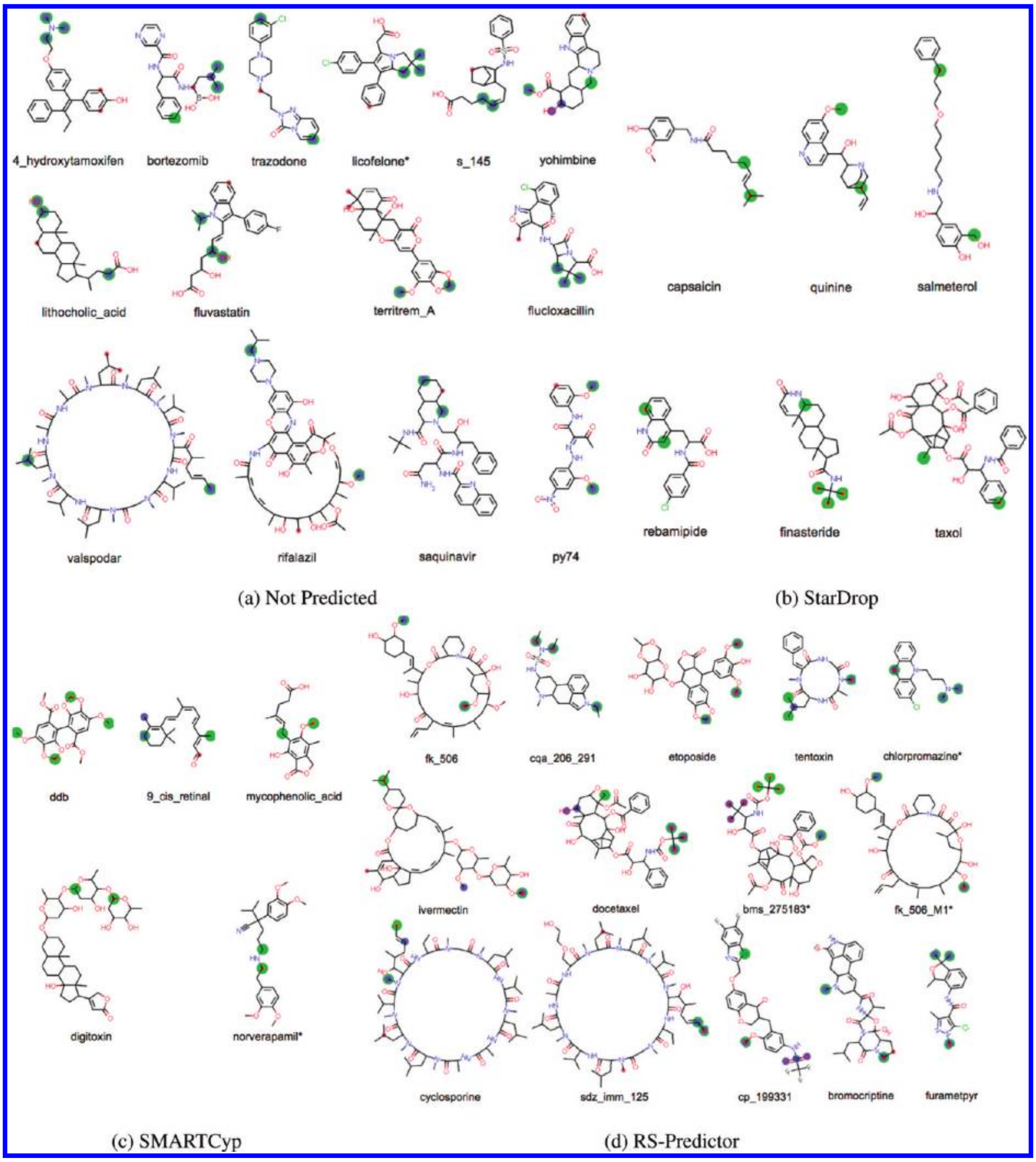


Figure 13. "Difficult to predict" ($9 \leq 1/R_2$) substrates not predicted by any method or only predicted by a single method.

actual, statistically more likely, N-dealkylation reaction. Examining Rutaecarpine(M), reveals that three topologically distinct sites are predicted within the top two rank-positions of SMARTCyp. Since two of those three are sites of 3A4-mediated aromatic hydroxylation, the calculated P_2 value is 1. Interestingly, each method uniquely predicts a number of different observed aromatic hydroxylations [StarDrop: 1_nitropyrene(E), taxol(H), rebamipide(H), SMARTCyp: propofol(E*), pranidipine(M*), rutaecarpine(M), RS-Predictor: phenprocoumon(E), 4_nitropyrene(M), pupropion(M), diclofenac(M), reduced_diclofenac(M*)], providing justification for the integration of different methods.

Each method predicts unique sites that undergo nonaromatic ring hydroxylation as well [StarDrop: aripiprazole(M), quinidine(M), quinine(H), SMARTCyp: budenoside(M), imidacloprid(M), ddb(H), digitoxin(H), RS-Predictor: alphadihydroegrocryptine(M), MDMA(M), pradefovir(M),tertranor(M*), bromocriptine(H)], while false-negative rates are slightly lower [RS-Predictor: phenprocoumon(E), 4_nitropyrene(M), pupropion(M), diclofenac(M), reduced_diclofenac(M*)] than those of aromatic hydroxylation. Performances between Calibration and External sets remained similar. One explanation for this is that the relative ratio between (oxidized)/(potential) sites of nonaromatic ring hydroxylations remain equivalent between data sets, with the External set having less of both.

Performances of each method upon O-dealkylation reactions provide one of the more interesting pathway analyses of this work. The Calibration set prediction propensities indicate that RS-Predictor emphasizes sites capable of undergoing O-dealkylation to a greater extent than StarDrop and to a much greater degree than SMARTCyp. When applied to the External set, which has a higher (oxidized)/(potential) ratio, it was expected that RS-Predictor would have similar true-positive rates, and decreased false-positive rates. What was not expected was the dramatic improvement in true-positive rate for SMARTCyp, nor the relative drop in StarDrop performance. Examining the large number of substrates correctly predicted solely by RS-Predictor [RS-Predictor: MDMA(M*), fk_506(H), fk_M506_M1(H*), ivermectin(H), cp_12199331(H)], compared to just a single substrate, mycophenolic_acid(H), for SMARTCyp, and none for StarDrop provides some rational for this and other previous observations. All but two of the RS-Predictor substrates come from the Calibration set, and (5)/(8) of them are at the hard end of the difficulty spectrum. These molecules help to explain why RS-Predictor has lower Top-1 Standard Calibration rates than SMARTCyp and StarDrop but higher Lift rates. However, RS-Predictor false-positive rates are also high. An examination of incorrectly predicted substrates show four cases where pathways of O-dealkylation were predicted above observed sites of 3A4-mediated aromatic ring hydroxylation [Not Predicted: carvedilol-(M), methoxychlor_mono_OH(M), py74(H), yohimbine(H)] and two where potential O-dealkylations sites were ranked above observed Csp^3 hydroxylations sites [Not Predicted: territrem_A(H), rifalazil(H)]. Everolimus(M), represents a case where RS-Predictor ranked two nonobserved sites of O-dealkylation above the actual site of O-dealkylation. On the other hand for the Everolimus analog fk_506(H), RS-Predictor not only correctly identified the observed site of O-dealkylation but also that of its primary metabolite fk_506_M1(H*).

N-dealkylation and S-oxidation have the highest ratio of (observed)/(potential) SOMs. It is unsurprising therefore to see high true-positive and low false-positive prediction rates for all methods. The number of N-dealkylation reactions correctly predicted by only a single method is surprising however [StarDrop: cisapride(E), SMARTCyp: acetochlor(M), ecabapide(M), e_2101(M), compound_1(M), norverapamil(H*), RS-Predictor: tpa023(M*), cqa_206_291(H), tentoxin(H), furametpyr-(H)]. Perhaps since most of the substrates come from the Calibration set, and the greatest number of observed Calibration set oxidations occur via 3A4-mediated N-dealkylations, some variance between methods should be expected. Certainly the variance in correct predictions between methods is much smaller for sulfur oxidations. SMARTCyp uniquely predicts the desulfuration of malathion and S-oxidation of N_ac_DFEC(M), while RS-Predictor alone predicts the S-oxidation of chlorpromazine-(H). An analysis of the molecules not predicted correctly reveals that RS-Predictor chose the wrong site of N-dealkylation for da_8159(M) over the observed one. A single occurrence of a boron atom within the Complete data set helps to explain why bortezomib(H) was not identified by any method.

External (observed)/(potential) ratios are greater than Calibration ratios for Csp^2 and nitrogen-based reactions as well as uncommon reaction pathways. Correspondingly, each method has increased true-positive and decreased false-positive rates across data sets. SMARTCyp appears to "prefer" Csp^2 and nitrogen-based reactions more strongly than RS-Predictor or StarDrop, though isoquinoline(E) was the only substrate undergoing N-oxidation that was correctly predicted solely by SMARTCyp. In contrast, the N-oxidation of acetaminophen(E), reduced_dolasetron(E), and voriconazole(M) were not predicted correctly by any method. SMARTCyp was the sole method to correctly predict Csp^2 based oxidation of 9_cis_retinal(H*). StarDrop does not consider N-oxidations, Csp^2 hydroxylations, or aldehyde oxidations, which helps to explain lower prediction rates for the External set. The formulation of RS-Predictor, which involves consideration of every potential 3A4-mediated oxidation path within the same model, explains its higher true and false-positive rates for uncommon P450-mediated pathways relative to SMARTCyp, which does not appear to contain parametrized reactivity results for many of the uncommon biotransformations. One example of an uncommon reaction is amide formation, which was correctly predicted for pinacidil(E) by RS-Predictor but incorrectly predicted over the actual 3A4-mediated aromatic ring hydroxylation of kr_33028(E).

■ CONCLUSIONS

This work describes RS-Predictor, an *in silico* method for creating and implementing cytochrome P450 regioselectivity models. RS-Predictor models may be generated from any set of known P450 substrates and metabolites, and work is currently proceeding on eight other isozymes in addition to CYP 3A4. By quantifying potential SOMs as metabolophores, direct regioselectivity comparisons can be made between different oxidative pathways or pathways of the same type that occur within different electronic and steric environments. The MIRank modeling procedure exploits this framework by learning which distinct topological and electronic environments exhibit a greater propensity for oxidation than others on a molecule by molecule, reaction by reaction basis, to ultimately create a trend-based regioselectivity QSAR. The descriptor weight vector that optimizes regioselectivity prediction for an entire calibration set quickly, and often accurately, predicts the oxidation sites of external substrates. In one of the first applications of rank aggregation within the cheminformatics community, a systematic framework was developed to merge multiple predicted SOM rankings from independent cross-validated models into a single consensus prediction, resulting in improved overall performance. Consensus regioselectivity models predict at least one oxidation site within the top two rank-positions for 78% of 394 substrates of 3A4. Models are robust and quickly applicable, taking on average under 3 s to make a prediction on a new compound (using a 2.6 GHz Opteron Linux workstation to make predictions for the external set).

Another contribution of this work is the definition of a new Lift metric that uses the statistical likelihood of making a randomly correct prediction for a given substrate as a gauge of the difficulty of making an accurate prediction on that same substrate. Overall Lift rates demonstrate that RS-Predictor is able to identify the observed SOMs of larger and more complex molecules with a higher degree of accuracy than other methods. Lift assessment of substrate prediction difficulty through its ratio of (#oxidized)/(#potential) SOMs makes sense mathematically but misses important chemical information about the relative propensities of a given CYP isozyme to catalyze different reaction pathways (ex. N-dealkylation versus aromatic ring hydroxylation). For this reason new data visualization techniques were created to assess the true-positive and false-positive prediction rates of different methods on a pathway-by-pathway basis. Pathways with lower (observed)/(potential) SOM ratios, such as aromatic ring, nonaromatic ring, or Csp^3 hydroxylation reactions

Table 7. Same SOM Predicted in the First Rank-Position by Multiple Methods

RS-Min	SMARTCyp	StarDrop	# compounds	% compounds	Standard Top-1
X	X	X	114	29	88.6%
X	X	-	33	8	72.7%
X	-	X	61	15	50.8%
-	X	X	62	16	79.0%

were shown to be predicted with poorer accuracy by multiple regioselectivity prediction methods than other, higher ratio pathways, such as S-oxidation or N-dealkylation. We propose the creation of a new "Chemical Lift" metric, which would incorporate mathematical Lift in some fashion with isozyme-mediated pathway propensities, would be a viable way to expand current techniques used to gauge the prediction accuracy of regioselectivity models.

Performance analysis illustrates that no single method is optimal across all pathways or substrates, suggesting that integration of different methods could be valuable. Indeed it is common practice within the pharmaceutical industry to apply different regioselectivity predictions methods to each compound under development. As a preliminary gauge of the potential of method integration, the number and prediction accuracy of substrates having the same SOM predicted within the first rank-position by multiple methods is shown in Table 7. When multiple methods predict the same SOM, the Top-1 prediction accuracy is significantly higher than the prediction rates of individual methods (Standard Top-1 Rates: RS-Minimum Consensus − 59.6%, SMARTCyp − 63.4%, StarDrop − 58.8%) in all but one case. Low RS-Predictor and StarDrop overlap rates could be explained by the fact that both methods use MOPAC. Similarity between method inputs likely correlates to similarity in predictions, both correct and incorrect. Meanwhile, the DFT based reactivity models of SMARTCyp have a different theoretical basis than RS-Predictor or StarDrop, capturing complementary information. It should be expected that if two theoretically distinct methods predict the same SOM, then that SOM is more likely to be an actual site of CYP-mediated oxidation; ergo, when the SOM predictions of SMARTCyp correspond to those of either RS-Predictor or StarDrop, the overall overlap prediction accuracy is quite high. Such findings provide justification for the incorporation of SMARTCyp transition state energies into other methods, especially RS-Predictor, where smaller prediction overlap indicates that the two methods are capturing mutually exclusive information.

While RS-Predictor generates isozyme-specific models from a set of known substrates, a current limitation of the method is that no explicit information from the enzymatic structure is used. However, a significant strength of the RS-Predictor algorithm is the ease with which new descriptors that capture different aspects of regioselectivity may be incorporated. To capture information directly from CYP enzymatic structure, docking algorithms could be employed to generate bound poses of a candidate substrate within the metabolizing P450 isozyme. Quantum chemical descriptors for each metabolophore could then be calculated from the pose that placed the metabolophore closest to the catalytic heme. Another approach would be to generate descriptors directly from the terms used by the docking algorithm to calculate predicted binding energies. Docking operations would take more time, justifiable only through improvement in prediction rates.

The currently available public metabolite data[19] allows for the creation of 2C9, 2D6, and 3A4 regioselectivity models, but more information is available to be culled. While important, the substrates of these isoforms do not represent the sum total of all P450-mediated metabolic reactions. The extension of 72 molecules to the current set of 3A4 substrates represents an initial work; compilation of additional substrate data sets and extension of older 2C9, 2D6, and 3A4 sets is currently ongoing. Once a sufficient number of substrates are compiled, we plan to investigate the creation of isozyme specific RS-Predictor models utilizing different and additional sets of descriptors.

ASSOCIATED CONTENT

S Supporting Information. Structural and metabolite data for all 394 compounds, including all predictions and calibrated descriptor weight vectors, are made available. The relative importance of each descriptor averaged across all Complete models is also given. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: brenec@rpi.edu.

ACKNOWLEDGMENT

The authors like to thank NIH grants (1P20HG003899-01) and (R01LM009731), ONR grant (N00014-06-1-0014) the Alfred Benzon foundation and the Danish Medical Research Council for funding. We also thank the RPI Center for Biotechnology and Interdisciplinary Studies for being an excellent collaborative work environment. We would also like express our gratitude to Optibrium for providing access to StarDrop and permission to publish results, Dr. Gregory Moore for contributions to the MIRank subgradient algorithm, and Michael Krein for systems support; MOE, Molecular Operating Environment.

ABBREVIATIONS:

RS-Predictor, Regioselectivity-Predictor; SOM, site of metabolism; MIRank, multiple instance ranking; SVM, support vector machines; OM, oxidized metabolophores; NOM, nonoxidized metabolophores; CV, cross-validation; RO, rank-ordering

REFERENCES

(1) Nebert, D. W.; Dieter, M. Z. The Evolution of Drug Metabolism. *Pharmacology* **2000**, *61*, 124–135.

(2) Nebert, D. W.; Russell, D. W. Clinical importance of the cytochromes P450. *Lancet* **2002**, *360*, 1155–1162.

(3) Guengerich, F. P. Cytochrome P450s and other enzymes in drug metabolism and toxicity. *AAPS J.* **2006**, *8*, E101–E111.

(4) Czodrowski, P.; Kriegl, J. M.; Scheuerer, S.; Fox, T. Computational approaches to predict drug metabolism. *Expert Opin. Drug Metab. Toxicol.* **2009**, *5*, 15–27.

(5) Korolev, D.; Balakin, K. V.; Nikolsky, Y.; Kirillov, E.; Ivanenkov, Y. A.; Savchuk, N. P.; Ivashchenko, A. A.; Nikolskaya, T. Modeling of human cytochrome P450-mediated drug metabolism using unsupervised machine learning approach. *J. Med. Chem.* **2003**, *46*, 3631–3643.

(6) Zhou, D.; Afzelius, L.; Grimm, S. W.; Andersson, T. B.; Zauhar, R. J.; Zamora, I. Comparison of methods for the predictions of the metabolic sites for CYP3A4-mediated metabolic reactions. *Drug Metab. Dispos.* **2006**, *34*, 976–983.

(7) Zimmerman, H. J.; Maddrey, W. C. Acetaminophen (paracetamol) hepatotoxicity with regular intake of alcohol: Analysis of instances of therapeutic misadventure. *Hepatology* **1995**, *22*, 767–773.

(8) Terfloth, L.; Bienfait, B.; Gasteiger, J. Ligand-based models for the isoform specificity of cytochrome P450 3A4, 2D6, and 2C9 substrates. *J. Chem. Inf. Model.* **2007**, *47*, 1688–1701.

(9) Yap, C. W.; Chen, Y. Z. Prediction of cytochrome P450 3A4, 2D6, and 2C9 inhibitors and substrates by using support vector machines. *J. Chem. Inf. Model.* **2005**, *45*, 982–992.

(10) de Graaf, C.; Vermeulen, N. P. E.; Feenstra, K. A. Cytochrome P450 in silico: an integrative modeling approach. *J. Med. Chem.* **2005**, *48*, 2725–2755.

(11) 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.

(12) Fox, T.; Kriegl, J. M. Machine learning techniques for in silico modeling of drug metabolism. *Curr. Top. Med. Chem.* **2006**, *6*, 1579–1591.

(13) Schuster, D.; Steindl, T. M.; Langer, T. Predicting drug metabolism induction in silico. *Curr. Top. Med. Chem.* **2006**, *6*, 1627–1640.

(14) Zhou, S. Drugs behave as substrates, inhibitors and inducers of human cytochrome P450 3A4. *Curr. Drug. Metab.* **2008**, *9*, 310–322.

(15) Kontijevskis, A.; Komorowski, J.; Wikberg, J. E. S. Generalized proteochemometric model of multiple cytochrome P450 enzymes and their inhibitors. *J. Chem. Inf. Model.* **2008**, *48*, 1840–1850.

(16) Bazeley, P. S.; Prithivi, S.; Struble, C. A.; Povinelli, R. J.; Sem, D. S. Synergistic use of compound properties and docking scores n neural network modeling of CYP2D6 binding: predicting affinity and conformational sampling. *J. Chem. Inf. Model.* **2006**, *46*, 2698–2708.

(17) Jones, J. P.; Mysinger, M.; Korzekwa, K. R. Computational models for cytochrome P450: a predictive electronic model for aromatic oxidation and hydrogen atom abstraction. *Drug Metab. Dispos.* **2002**, *30*, 7–12.

(18) Singh, S. B.; Shen, L. Q.; Walker, M. J.; Sheridan, R. P. A model for predicting likely sites of CYP3A4-mediated metabolism on drug-like molecules. *J. Med. Chem.* **2003**, *46*, 1330–6.

(19) 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.

(20) Smith, J.; Stein, V. SPORCalc: A development of a database analysis that provides putative metabolic enzyme reactions for ligand-based drug design. *Comput. Biol. Chem.* **2009**, *33*, 149–159.

(21) Zheng, M.; Luo, X.; Shen, Q.; Wang, Y.; Du, Y.; Zhu, W.; Jiang, H. Site of metabolism prediction for six biotransformations mediated by cytochromes P450. *Bioinformatics* **2009**, *25*, 1251–1258.

(22) Hennemann, M.; Friedl, A.; Lobell, M.; Keldenich, J.; Hillisch, A.; Clark, T.; Göller, A. H. CypScore: Quantitative prediction of reactivity toward cytochromes P450 based on semiempirical molecular orbital theory. *ChemMedChem* **2009**, *4*, 657–669.

(23) Kim, D. N.; Cho, K.; Oh, W. S.; Lee, C. J.; Lee, S. K.; Jung, J.; No, K. T. E_aMEAD: Activation energy prediction of cytochrome P450 mediated metabolism with effective atomic descriptors. *J. Chem. Inf. Model.* **2009**, *49*, 1643–1654.

(24) 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.

(25) de Groot, M. J.; Ackland, M. J.; Horne, V. A.; Alex, A. A.; Jones, B. C. A novel approach to predicting P450 mediated drug metabolism. CYP2D6 catalyzed N-Dealkylation reactions and qualitative metabolite predictions using a combined protein and pharmacophore model for CYP2D6. *J. Med. Chem.* **1999**, *42*, 4062–4070.

(26) Park, J.; Harris, D. Construction and assessment of models of CYP2E1: predictions of metabolism from docking, molecular dynamics, and density functional theoretical calculations. *J. Med. Chem.* **2003**, *46*, 1645–60.

(27) Cruciani, G.; Carosati, E.; Boeck, B. D.; Ethirajulu, K.; Mackie, C.; Howe, T.; Vianello, R. MetaSite: Understanding metabolism in human cytochromes from the perspective of the chemist. *J. Med. Chem.* **2005**, *48*, 6970–6979.

(28) Oh, W. S.; Kim, D. N.; Jung, J.; Cho, K.; No, K. T. New combined model for the prediction of regioselectivity in cytochrome P450/3A4 mediated metabolism. *J. Chem. Inf. Model.* **2008**, *48*, 591–601.

(29) *StarDrop*, version 4.2.1; Optibrium Ltd.: Cambridge, United Kingdom, 2009.

(30) MetaSite user manual. http://www.moldiscovery.com/docs/metasite/background.html (accessed March 23, 2011).

(31) Bergeron, C.; Zaretzki, J.; Breneman, C.; Bennett, K. P. Multiple instance ranking. In *Proceedings of the 25th ICML*, Helsinki, Finland, 2008; ACM: New York, NY, 2008; Vol. 307, pp 48−55.

(32) Rendic, S. Summary of information on human CYP enzymes: human P450 metabolism data. *Drug Metab. Rev.* **2002**, *34*, 83–448.

(33) Brown, C. M.; Reisfeld, B.; Mayeno, A. N. Cytochromes P450: a structure-based summary of biotransformations using representative substrates. *Drug Metab. Rev.* **2008**, *40*, 1–100.

(34) Daher, W.; Pelinski, L.; Klieber, S.; Sadoun, F.; Meunier, V.; Bourrié, M.; Biot, C.; Guillou, F.; Fabre, G.; Brocard, J.; Fraisse, L.; Maffrand, J.; Khalife, J.; Dive, D. In vitro metabolism of ferroquine (SSR97194) in animal and human hepatic models and antimalarial activity of major metabolites on plasmodium falciparum. *Drug Metab. Dispos.* **2006**, *34*, 667–682.

(35) *MOE*, version 2009.10; Chemical Computing Group: Montreal, Canada, 2009.

(36) Wildman, S. A.; Crippen, G. M. Prediction of physicochemical parameters by atomic contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868–873.

(37) Bush, B. L.; Sheridan, R. P. PATTY: A programmable atom type and language for automatic classification of atoms in molecular databases. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 756–762.

(38) Stewart, J. J. P. MOPAC: A semiempirical molecular orbital program. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 1–103.

(39) Bergeron, C.; Moore, G.; Zaretzki, J.; Breneman, C. M.; Bennett, K. P. Fast bundle algorithm for multiple instance learning. *PAMI* **Under Revision, 2011.**

(40) Bi, J.; Bennett, K.; Embrechts, M.; Breneman, C.; Song, M. Dimensionality reduction via sparse support vector machines. *J. Mach. Learn. Res.* **2003**, *3*, 1229–1243.

(41) Shao, L.; Wu, L.; Fan, X.; Cheng, Y. Consensus Ranking Approach to Understanding the Underlying Mechanism With QSAR. *J. Chem. Inf. Model.* **2010**, *50*, 1941–1948.

(42) Datta, S.; Pihur, V.; Datta, S. An adaptive optimal ensemble classifier via bagging and rank aggregation with applications to high dimensional data. *BMC Bioinf.* **2010**, *11*, 427.

(43) Schalekamp, F.; Zuylen, A. Rank Aggregation: Together We're Strong. In *Proceedings of the 11th ALENEX*, New York, New York, 2009; SIAM: Philadelphia, PA, 2009; pp 38−51.

(44) Dwork, C.; Kumar, R.; Naor, M.; Sivakumar, D. Rank Aggregation Methods for the Web. In *Proceedings of the 10th International Conference on the WWW*. Hong Kong, Hong Kong, 2001; ACM: New York, NY, 2001; pp 613−622.

(45) Sculley, D. Rank Aggregation for Similar Items. In *Proceedings of the Seventh SIAM International Conference on Data Mining*. Minneapolis, Minnesota, 2007; SIAM: Philadelphia, PA, 2007.

(46) Klementiev, A.; Roth, D.; Small, K. Unsupervised Rank Aggregation with Distance-Based Models. In *Proceedings of the 25th ICML*, Helsinki, Finland, 2008; ACM: New York, NY, 2008; Vol. 307, pp 472−479.

(47) Liu, Y.; Liu, T.; Qin, T.; Ma, Z.; Li, H. Supervised Rank Aggregation. In *Proceedings of the 16th International Conference on WWW*, Banff, Alberta, 2007; ACM: New York, NY, 2007; pp 481−490.