

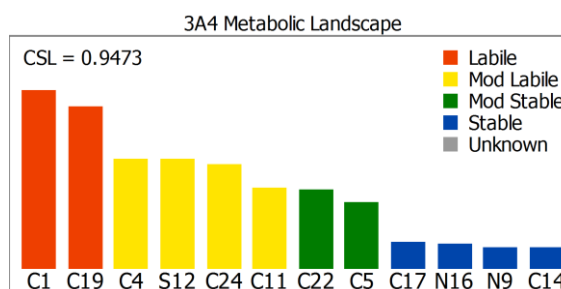
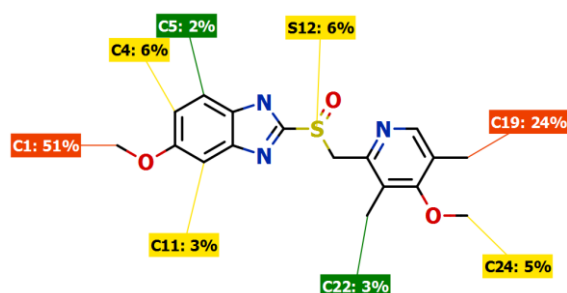
# Predicting Regioselectivity and Lability of Cytochrome P450 Metabolism using Quantum Mechanical Simulations

Jonathan D. Tyzack, Peter A. Hunt, Matthew D. Segall

Optibrium Ltd., 7221 Cambridge Research Park, Beach Drive, Cambridge, CB25 9TL, UK

## Abstract

We describe methods for predicting Cytochrome P450 (CYP) metabolism incorporating both pathway-specific reactivity and isoform-specific accessibility considerations. Semi-empirical quantum mechanical (QM) simulations, parameterized using experimental data and ab initio calculations, estimate the reactivity of each potential site of metabolism in the context of the whole molecule. Ligand-based models, trained using high quality regioselectivity data, correct for orientation and steric effects of the different CYP isoform binding pockets. The resulting models identify a site of metabolism in the top 2 predictions for between 82% and 91% of compounds in independent test sets across seven CYP isoforms. In addition to predicting the relative proportion of metabolite formation at each site, these methods estimate the activation energy at each site, from which additional information can be derived regarding their lability in absolute terms. We illustrate how this can guide the design of compounds to overcome issues with rapid CYP metabolism.



## Introduction

The development of methods to predict the sites, products and rates of metabolism is an important avenue of research and finds application in the development of drugs, cosmetics, nutritional supplements and agrochemicals. It is necessary to understand the pharmacokinetics of a molecule and ensure that it has sufficient exposure at the target to exert its therapeutic effect. In this regard it would be helpful to give an absolute prediction of the rate of, or at least the lability of a compound to, metabolism, rather than just a rank ordering of sites; a factor often neglected by metabolism prediction tools. It is also important to predict the formation of toxic metabolites, which contributes to the high attrition rates experienced in the development of new chemical entities, the imposition of black-box warnings or even the withdrawal of approved pharmaceuticals. Thus, the ability to identify potential toxic metabolites early and make predictions about metabolic stability are of crucial importance in the drug discovery process.

The cytochrome P450s (CYPs) are a family of heme-containing enzymes involved in the phase-I metabolism of over 90% of drugs currently on the market. [1] The CYP family consists of 57 isoforms [2] with the largest contribution to xenobiotic metabolism coming from CYP3A4, the most promiscuous isoform, followed by CYP2D6 and CYP2C9. A comprehensive overview of the structure, reactivity and catalytic cycle of CYPs can be found in the review paper from Shaik et al.. [3]

The catalytic action of CYPs is predominantly that of a monooxygenase:



where RH is the substrate molecule. The most common reactions catalyzed by CYPs involve the insertion of a single oxygen into an organic molecule, such as C=C epoxidation, aromatic C oxidation and aliphatic C hydroxylation, the last example often leading to N-dealkylation or O-dealkylation if oxidation occurs on a suitable leaving group in an amine or ether moiety. The addition of oxygen into the substrate is a precursor to excretion from the body, driving an increase in polarity and hydrophilicity and facilitating Phase II metabolism pathways such as glucuronidation.

The heme moiety at the catalytic center of the CYPs is conserved across all isoforms, where a highly activated oxy-heme, formed by cleavage of molecular oxygen, is generated within the catalytic cycle, as shown in Figure 1. In addition to the main catalytic cycle there are two significant decoupling pathways, labelled D1 and D2, resulting in the formation of hydrogen peroxide and water respectively and returning the active site heme to an inactivated state. The relative rate of decoupling compared to substrate metabolism will influence the observed rate of the conversion of substrate into metabolites and is an important consideration when making absolute assessments of metabolic stability.

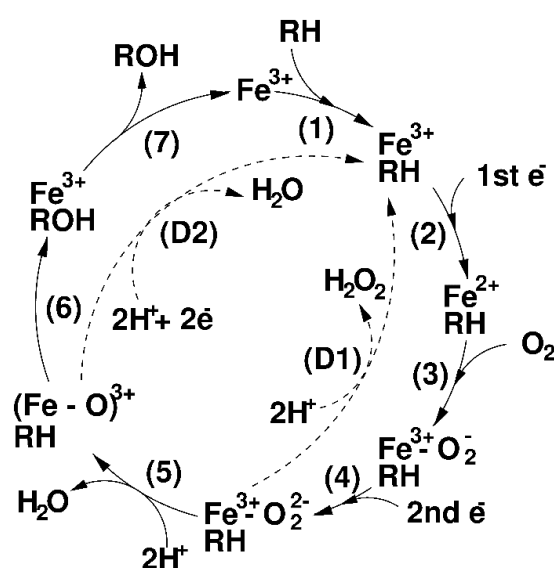



Figure 1 The catalytic cycle of the Cytochrome P450 enzymes. The decoupling pathways to form hydrogen peroxide and water are labelled D1 and D2 respectively.



Experimental investigation of xenobiotic metabolism can be both resource and time consuming, which has encouraged the development of computational techniques. These can be separated into two distinct categories: ligand-based and structure-based. In the first approach, structures and properties of known substrate or non-substrate ligands are modeled to develop structure-activity relationships. The second approach is focused on the structure of the metabolizing CYP enzyme, its known reaction mechanisms and its interactions with substrates. Structure-based methods include docking, to identify potential ligand conformations in the context of the enzyme active site [4]; molecular dynamics simulations, which can estimate the energetics of binding [5] [6] and QM/MM methods, which combine a quantum mechanical description of the catalytic reaction center while capturing the effects of the surrounding protein structure using an empirical molecular mechanics force field [7]. Structure-based methods can provide useful insight into the three-dimensional influence of the protein structure on the site of reaction and binding affinity, but are limited by the availability of good protein structures from crystallography and the flexibility of many P450 isoforms, in particular CYP3A4, which is problematic to take into consideration in a reasonable time-frame. The available evidence also suggests that structure-based methods do not, at present, improve the accuracy of predictions over ligand-based approaches [8] [9]. For a general overview of current computational tools to predict sites of metabolism (SOM) the reader is referred to the many comprehensive review papers. [10] [11] [12] [13] [14]

Most metabolism prediction tools incorporate some form of reactivity and accessibility considerations. The method described herein is no exception: a ligand-based approach is used to model steric and orientation effects whilst the electronic activation energy is modeled using quantum mechanical (QM) simulations to calculate the energies of substrates and reaction intermediates for each potential SOM. This approach offers several advantages:

- QM methods are based on fundamental physical principles and therefore transfer well between chemical classes; they do not rely on specific examples being present within a training set used to fit an empirical model
- Each potential SOM is considered in the context of the whole molecular environment in which it resides, rather than identifying fragments within a substrate and treating each as a discrete uniform entity regardless of neighboring functional groups
- QM methods can estimate the activation energy of the rate-limiting step of the oxidation reaction, allowing comparison of lability on an absolute scale.

The previously published SMARTCyp method [15] also uses a QM-based approach to predict CYP SOM. However, the approach differs from that described herein in that the SOMs are ranked based on a look-up table of small functionalities for which the activation energies have been previously calculated using ab initio density functional (DFT) methods. The use of ab initio QM methods avoids the need for detailed experimental data on which to base estimates of the activation energies. However, these calculations are computationally very expensive and the use of a look-up of pre-calculated results is required to return results in a practical time-frame. Therefore, this approach does not take into account potential long-range effects due to the environment of the whole molecule, an important factor for a medicinal chemist developing a lead series and aiming to predict the likely impact of structural changes on metabolic stability.

The use of semi-empirical QM calculations that estimate the activation energies for each aliphatic and aromatic SOM have previously been described. [16] Semi-empirical methods are significantly faster than ab initio methods and therefore can be applied to an entire substrate on a routine basis. However, they typically require detailed experimental data with which to parameterize a free energy relationship and therefore these models do not include less common, but important, pathways such as epoxidation, or N- and S-oxidation, due to the lack of sufficient experimental data.

The methods described herein build on both of these methods to achieve transferability, application to the whole substrate to explicitly consider the molecular environment of each SOM and computational efficiency, returning results in approximately 1-2 minutes per compound on a single CPU. In the following section, the theory and implementation of the models will be explained in detail. The performance of the models on independent test sets will be presented in the Results section with comparisons made to the SMARTCyp [15]

method. Finally, three example applications will be presented to illustrate how the models can provide valuable guidance to redesign compounds and overcome issues due to rapid CYP metabolism.

## Theory and Implementation

### Modeling Principles

The key factors that determine the SOM are reactivity and accessibility. The models described herein estimate the activation energy at each potential SOM in a substrate using fundamental and transferable QM methods, rather than relying on empirical pattern matching with a limited domain of applicability. The QM models are independent of isoform, reflecting the consistent reaction pathways across isoforms.

However, the binding pockets of the different CYP isoforms differ in size, shape and chemical composition and influence the orientation of the substrate relative to the reactive oxy-heme core. Electrostatic, hydrogen bonding and lipophilic interactions between substrate and CYP binding pocket have varying effects across isoforms and will cause different orientations to be favorable. In addition, the steric bulk within a substrate will influence the accessibility of sites to the reactive oxy-heme core, with those sites embedded towards the center being less accessible than those in open sites on the periphery of the substrate. These steric effects are also isoform specific, as the different sizes of the CYP binding pockets can accommodate different levels of steric hindrance. The models described herein are able to capture these orientation and steric effects with ligand-based models trained on isoform-specific data sets, enabling adjustments to be made to the QM-generated electronic activation energy that are specific to each isoform.

The relative rate of metabolism for a site can be calculated from the activation energy,  $E_a$ , for the rate-limiting step in the reaction pathway, as the rate is proportional to the negative exponential of the activation energy (the Arrhenius equation):

$$k_i \propto e^{\frac{-E_{ai}}{kT}}, \quad (2)$$

where  $k_i$  is the relative rate of metabolism and  $E_{ai}$  is the activation energy of site  $i$ ,  $k$  is the Boltzmann constant and  $T$  is the temperature.

As discussed above, by directly calculating the activation energy, the lability of sites can be compared on an absolute scale between compounds, rather than just a relative ranking of sites within a compound. This is achieved by comparing the rate of product formation with that of a decoupling pathway, to give an absolute assessment of the efficiency of the product formation step in the catalytic cycle. This enables a medicinal chemist to identify likely metabolically vulnerable positions in a molecule and can be used to guide development away from compounds with potentially rapid clearance. In the remainder of this section the various aspects of the CYP metabolism prediction models will be described in detail.

### Quantum mechanical models of electronic effects

The oxidizing species and the chemical mechanisms of oxidation are the same for all CYP isoforms. This allows the intrinsic vulnerability of the sites on a potential substrate to be calculated with reference only to the structure of that molecule.

The oxidation reactions proceed via different pathways depending on the nature of the site of metabolism on the substrate: aliphatic hydroxylation proceeds via an initial hydrogen abstraction followed by reaction of the ferryl oxygen with the alkyl radical, a process known as the rebound mechanism; [17] alkene epoxidation proceeds via activation of the double bond to form an iron alkoxy radical species in a tetrahedral orientation; [18] aromatic C oxidation proceeds via activation of an aromatic bond [19] followed by an intra-molecular hydrogen atom transfer known as "NIH-shift" [20]; and direct oxidation of hetero atoms such as sulphur and nitrogen proceeds via bond formation with the ferryl oxygen, [21] [22] although often dealkylation reactions are favorable over direct oxidation.

The reactivity model performs a QM calculation to estimate the 'electronic' activation energy,  $\Delta H_A$ , for every potential site of metabolism, using knowledge of the reaction pathway for that site.  $\Delta H_A$  is the energy of the transition state relative to that of the reactants for the rate limiting step of the reaction pathway. These calculations are performed using AM1, [23] a quantum mechanical approach based on a semi-empirical

Hamiltonian<sup>1</sup>. While less accurate than a full ab initio simulation, AM1 is many times faster. Ab initio simulations have been used to identify systematic errors due to the use of AM1 and correction factors are applied within the electronic model. [15] [16]

Direct calculation of activation energies is computationally very expensive due to the need to perform a transition-state search. Instead, the heat of reaction,  $\Delta HR$ , is calculated from the heat of formation of the substrate and reaction intermediates and a Brønsted relationship [24] is used to calculate an approximation to  $\Delta HA$  (as a linear relationship has been shown to exist between the activation energy,  $\Delta HA$ , and the heat of reaction  $\Delta HR$ ). The parameters of the Brønsted relationship can be derived from detailed experimental regioselectivity data where this is available. [16] [25] However, in some cases there are insufficient experimental data and, instead, high level ab initio calculations can be used to accurately calculate the activation energies which, in turn, can be used to derive the parameters of the Brønsted relationship for the faster, semi-empirical AM1 calculations [26] [27].

However, due to the differences in the chemical mechanisms and methods for calculation of each of the pathways leading to oxidation, the energy scales of the calculated activation energies will differ. Therefore, in order to compare the rates of reactions that proceed by different pathways, the activation energies must be on the same scale and therefore a normalization must be applied. To achieve this, the activation energy scale relating to the abstraction of Hydrogen from aliphatic carbon sites is used as a reference and calculations performed with other methods are transformed onto this energy scale. Figure 2 shows the linear relationship between the activation energies calculated using the ab initio B3LYP DFT method, as published by Rydberg et al., [15] and those estimated from  $\Delta HR$ , calculated with AM1 and a Brønsted relationship, for hydrogen abstraction sites, as described below.

Further details of the calculations performed for each of the reaction pathways modeled are provided in the next few subsections.

### ***Hydrogen abstraction reactions***

The rate limiting step in the formation of a metabolite by hydrogen abstraction has been identified as the abstraction of the hydrogen from the substrate by the oxy-heme and formation of an alkyl radical intermediate. In the Brønsted relationship used to estimate the activation energy, an additional linear term involving the ionization potential has also been found to be important to capture resonance effects in the transition state. [25] Using detailed experimental measurements of the relative rates of product formation at different sites of the same molecule, this pathway was modelled as described in Korzekwa et al. [25] to estimate  $\Delta HA$  for each potential site of hydrogen abstraction.

### ***Aromatic oxidation***

Aromatic oxidation progresses by formation of a tetrahedral intermediate between the substrate and oxy-heme at the site of metabolism, followed by rearrangement to form a hydroxylated product. The formation of the tetrahedral intermediate is the rate limiting step in this process and the activation energy was also found to be proportional to the heat of reaction. Using experimental measurements of the relative rates of formation of different metabolites on the same molecule, the parameters of this relationship can be determined and  $\Delta HA$  calculated on the same scale as that for hydrogen abstraction, as described in detail in Jones et al.. [16]

---

<sup>1</sup> Initial low-energy substrate geometries are generated with Corina [40] as input to AM1 and further optimized within the AM1 calculations. Although the reaction energies can vary for different conformations of flexible substrates, the overall impact on the accuracy of prediction was found to be small. Therefore, for computational efficiency we consider only a single conformation for each substrate.

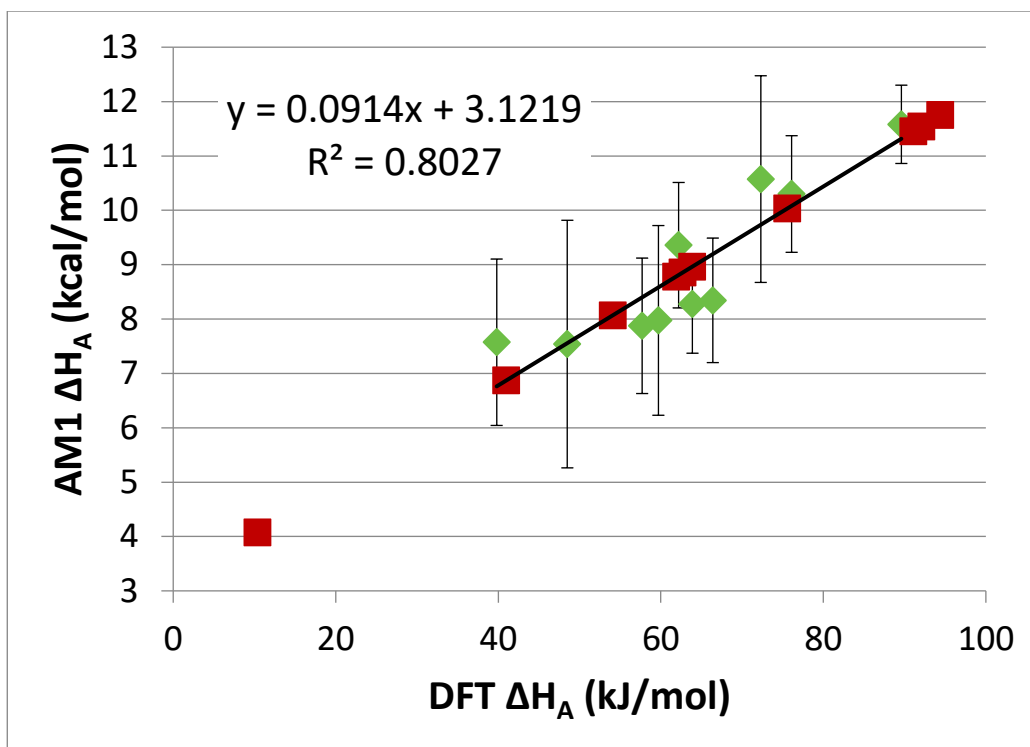


Figure 2 Graph showing the linear relationship between H-abstraction activation energies calculated using DFT and the models based on AM1. The points represented by blue diamonds show the average  $\Delta H_A$ , estimated using a Brønsted relationship based on  $\Delta H_R$  calculated with AM1, plotted against the single activation energy value assigned to the corresponding sites in SMARTCyp, derived from DFT transition state calculations. [15] While only a single activation energy is assigned to each class of site by SMARTCyp, in practice there may be significant variations between similar sites due to different molecular environments in which they occur. To illustrate this, the error bars show one standard deviation in the  $\Delta H_A$  values calculated using AM1 on the full molecules. These averages were calculated over a total of 2252 sites on a wide diversity of compounds. The minimum number of sites in each class for which an average is shown was 18. The transformation of N oxidation and hydroxylation energies from DFT, on the basis of this linear relationship, is represented by the red squares.

### Double bond epoxidation

Similar to aromatic oxidation, the epoxidation of a carbon-carbon double bond proceeds via the formation of a tetrahedral intermediate, followed by rearrangement to form the epoxide. [28] The formation of the tetrahedral intermediate is again the rate limiting step, with the activation energy found to be proportional to the heat of reaction.

There are insufficient experimental data with which to confidently parameterize a Brønsted relationship and, in this case, we rely on activation energies calculated with ab initio DFT calculations that were shown to agree with experimental observations, as described in Kumar et al.. [28] In this case, AM1 calculations of  $\Delta H_R$  exhibit an excellent correlation with the ab initio activation energies, as shown in Figure 3. This enables the estimation of the DFT activation energy from the AM1  $\Delta H_R$ , which, in turn, can be transformed to calculate  $\Delta H_A$  on the same energy scale as that for hydrogen abstraction, using the linear relationship shown in Figure 2.

Epoxidation proceeds by formation of a tetrahedral intermediate with the carbon at either end of the double bond. Therefore,  $\Delta H_A$  is calculated for both potential sites and the lowest value is used to estimate the relative rate of epoxidation of the corresponding double bond.

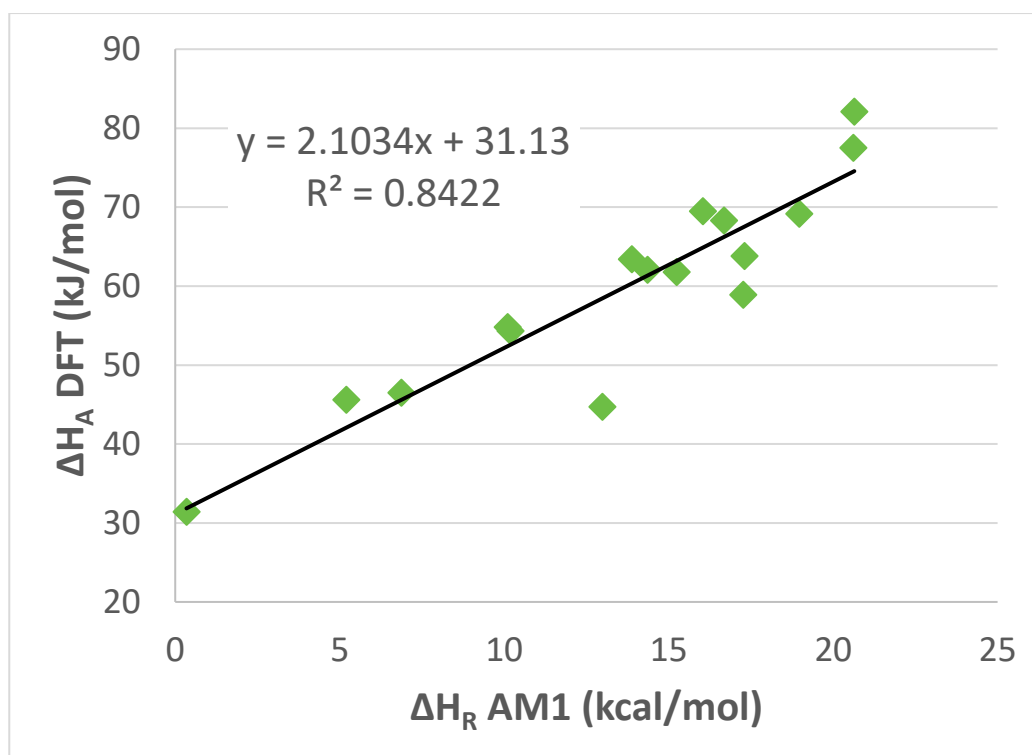


Figure 3 Graph showing the linear relationship between  $\Delta H_R$  calculated with AM1 and  $\Delta H_A$  calculated with DFT for potential sites of double bond epoxidation, as published by Kumar et al.. [28]

#### Other direct oxidation pathways

For S-oxidation, N-oxidation and hydroxylation, and other pathways including desulfurization of phosphothioates, oxidation of disulfides and aldehyde oxidation/deformylation, there are limited experimental data regarding the rates of these reactions relative to other sites on the same compounds. Furthermore, *ab initio* DFT calculations indicate that there is less variation in these rates between similar functionalities. Therefore, for these sites, activation energies derived from *ab initio* DFT calculations published by Rydberg *et al.* [15] were transformed onto the same energy scale as the other sites described previously, using the linear relationship shown in Figure 2. As an illustration, the transformation of N-oxidation and -hydroxylation energies using this linear relationship is also shown in Figure 2.

#### Accessibility: steric and orientation effects

In addition to the intrinsic vulnerability of a site in a molecule to oxidative attack, the accessibility of that site to the oxy-heme core will also influence the relative rate of metabolism. This effect is calculated as a correction to the activation energy due to the orientation of the molecule within the active site and steric hindrance by nearby atoms in the substrate.

Orientation effects are modeled by descriptors representing the topological distance to important functionalities such as acidic, basic, hydrogen bond donor/acceptor and lipophilic groups that interact with key residues in the CYP active site. The steric accessibility of a potential site of metabolism depends on the surrounding atoms in the substrate and will be influenced by nearby bulky functionalities or whether the site is part of a ring, a conjugated system or an aliphatic chain. The steric effects are modeled using descriptors representing the distance to functionalities introducing steric bulk surrounding the SOM. These functionalities are defined as SMARTS patterns [29] and some examples are given in Table 1.



**Table 1** Example definitions of functionalities from which distances are calculated as steric and orientation descriptors used to model accessibility of potential sites of metabolism in ligand-based models.

Descriptor Type	Functionality	SMARTS	Description
Orientation	Basic group	[N;+0;X3]	Non-conjugated amine
Orientation	Acidic group	[\#8;H]-C=O	Carboxylic acid
Orientation	H-bond donor	[\#7,\#8,\#9;!H0]	Intrinsic donor
Orientation	H-bond acceptor	[N;!X4]	Nitrogen acceptor
Steric	Heavy atom	[!#1]	Non-Hydrogen
Steric	Ring atom	[!#1,R]	Non-Hydrogen in a ring

Isoform-specific data sets have been carefully curated from the literature, as described in more detail in the following subsection, with the steric and orientation descriptors calculated for all sites in all molecules. Principal component regression models [30] were trained on these data sets using knowledge about the metabolic fate of each site to set the dependent variable: 0 for a non SOM, 1 for a primary SOM, 0.5 for a secondary SOM and 0.25 for a tertiary SOM. In order to generate an adjustment to the electronic activation energy,  $\Delta H_A$ , it is necessary to calculate the adjustment on the same energy scale as  $\Delta H_A$ . This is achieved by including the  $\Delta H_A$  in the regression and scaling the descriptor regression coefficients relative to the  $\Delta H_A$  regression coefficient. These scaled coefficients can then be applied to descriptors for new molecules and the resulting correction added to the ‘electronic’ activation energy,  $\Delta H_A$ , to calculate the estimate of activation energy,  $E_a$ , adjusted for steric and orientation effects.

### Data Curation

A detailed review of the primary literature was performed to prepare high quality datasets of isoform-specific human CYP substrates annotated with SOM. The papers were manually parsed to extract primary, secondary and tertiary SOM, along with the identity of the major and minor metabolizing CYP isoforms. The emphasis was on high quality data, retaining only human data and excluding data generated with inappropriate experimental conditions, such as un-physiological substrate concentrations. The consequence of this is that the data sets are smaller than some of those previously published. [9] [31] However, analysis of the chemical space covered by the CYP data sets against launched drug space shows that good coverage of drug-like chemical space has been achieved as illustrated in Figure 4 and the higher quality data is expected to result in more accurate models.

Table 2 summarizes the number of compounds in the training set for each isoform, used to fit the contributions of the steric and orientation descriptors, and the independent test sets, used to validate each model. The data sets are available in the supplementary information for inspection, including references to the primary literature from which the SOMs were identified (see Supplementary Information below).

There is an element of judgement to be applied when classifying sites within a molecule as primary, secondary or tertiary and identifying major/minor isoforms, with reliance placed on kinetic data from expressed supersomes and isoform-specific inhibition experiments with human liver microsomes. Variability between assays makes direct comparison of experimental data between publications challenging, but efforts have been made to make classifications across different molecules consistent. This was achieved by taking the consensus of multiple reviewers of the data, in addition to the views of the authors of the original paper, and looking for confirmation of the classifications across multiple assay protocols.

**Table 2** The numbers of compounds in the training and independent test sets of detailed regioselectivity data used to build and validate the models described herein. These data sets are provided as supplementary information.

CYP Isoform	$N_{\text{training}}$	$N_{\text{test}}$
CYP1A2	144	57
CYP2C8	80	27
CYP2C9	145	49
CYP2C19	136	49
CYP2D6	147	56
CYP2E1	76	30
CYP3A4	220	84



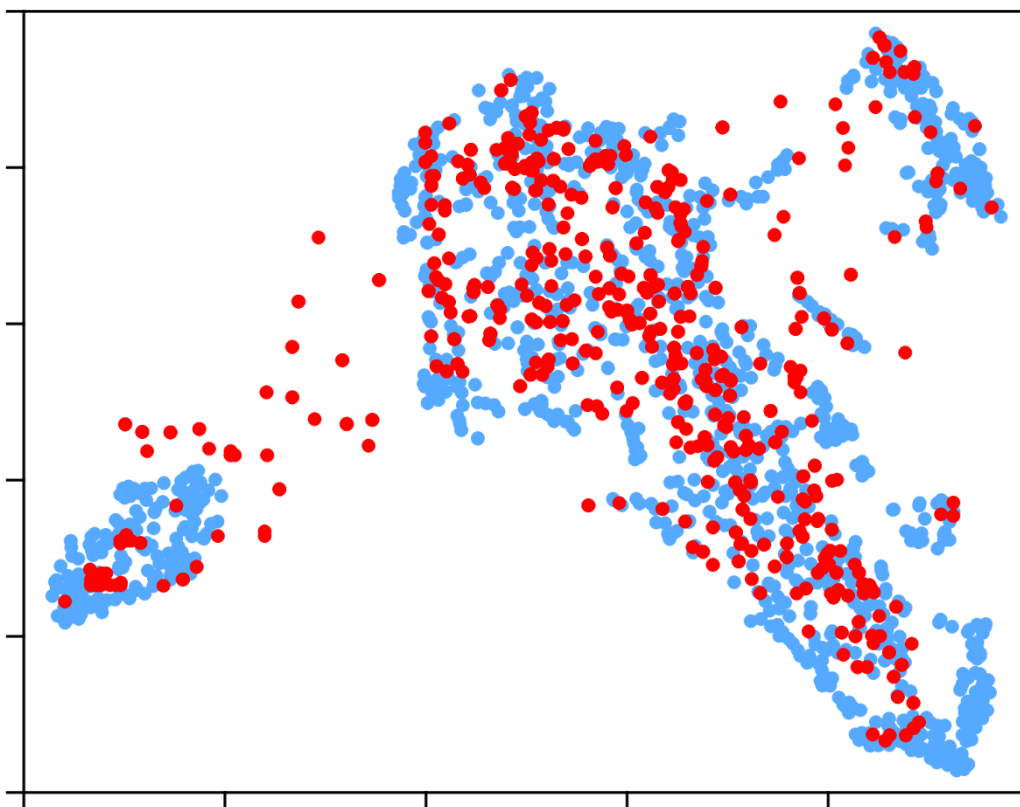


Figure 4 Illustration of the chemical space covered by the CYP data sets (red points) compared with approximately 1,300 launched drugs (blue points), generated using StarDrop. [32] In this chemical space plot, the proximity of two points represents the structural similarity between the corresponding compounds defined using a Tanimoto index based on a 2D path-based fingerprint. The distribution of points is generated using the t-distributed stochastic neighbor embedding algorithm. [33]

### Calculating Regioselectivity

The regioselectivity of metabolism is the proportion of metabolism that occurs at each site. This proportion is given by the rate of metabolism at that site relative to the sum of the rates for all potential sites of metabolism. The advantage of calculating an approximation to the activation energy is that a relative rate can be generated via Equation 2, allowing the regioselectivity of metabolism at site  $i$  to be given by:

$$R_i = \frac{k_i}{k_{total}} \times 100, \quad (3)$$

where  $R_i$  is the predicted proportion of metabolism at site  $i$ , expressed as a percentage, and  $k_{total} = \sum_{\text{all sites}} k_i$ . Therefore, potential SOM can be ranked by their corresponding values of  $R_i$ .

### Calculating Lability

The regioselectivity of metabolism describes the relative rate of metabolism of each potential site on a molecule. However, regioselectivity does not itself provide information on the absolute vulnerability or lability of each site to metabolism.

The lability of each site is derived by comparison of the predicted rate of the product formation step for the site with the water formation decoupling pathway in the catalytic cycle, labeled D2 in Figure 1. The rate of the decoupling pathway is often measured using an intrinsic isotope effect method. [34] If a site on a substrate is metabolized at a significantly higher rate than that of decoupling then metabolite formation will proceed with high efficiency. Conversely, if the rate of water formation is significantly higher than the rate of metabolism of a site then decoupling would dominate and metabolite formation at that site would not be observed. Specifically, the lability of site  $i$  is given by:

$$L_i = \frac{k_i}{k_{water} + k_i}, \quad (4)$$

where  $k_{water}$  is the rate of water formation via the decoupling pathway. The value of  $k_{water}$  has previously been determined experimentally using an intrinsic isotope effect method [34] [35] for metabolism by CYP3A4 and therefore the lability values presented herein should be interpreted with reference to this isoform.

The distribution of the lability of the sites on a compound is conveniently shown on a 'metabolic landscape' histogram using color as a visual guide, based on the efficiency with which metabolism would occur at that site: red for labile ( $>0.80$ ); yellow for moderately labile (between 0.35 and 0.80); green for moderately stable (between 0.05 and 0.35) and blue for stable ( $<0.05$ ). Examples of these representations are shown in the illustrative applications in Figure 5 through Figure 7.

The site labilities of individual sites can be combined to calculate the 'composite site lability' (CSL) reflecting the overall efficiency of product formation for the molecule. This is calculated from the combined estimated rates of metabolism for all sites on the molecule:

$$CSL = \frac{k_{total}}{k_{total} + k_{water}} \quad (5)$$

It should be noted that CSL is not a prediction of rate, but is one important factor influencing the rate amongst others, including reductions rates, which are often rate-limiting in the catalytic cycle, and binding affinity, which itself can be influenced by substrate properties such as size, lipophilicity and pKa. Therefore, due to the impact of changes in these additional, confounding factors between compounds, a direct correlation between small changes to CSL and the CYP3A4 half-life or intrinsic clearance is not necessarily expected.

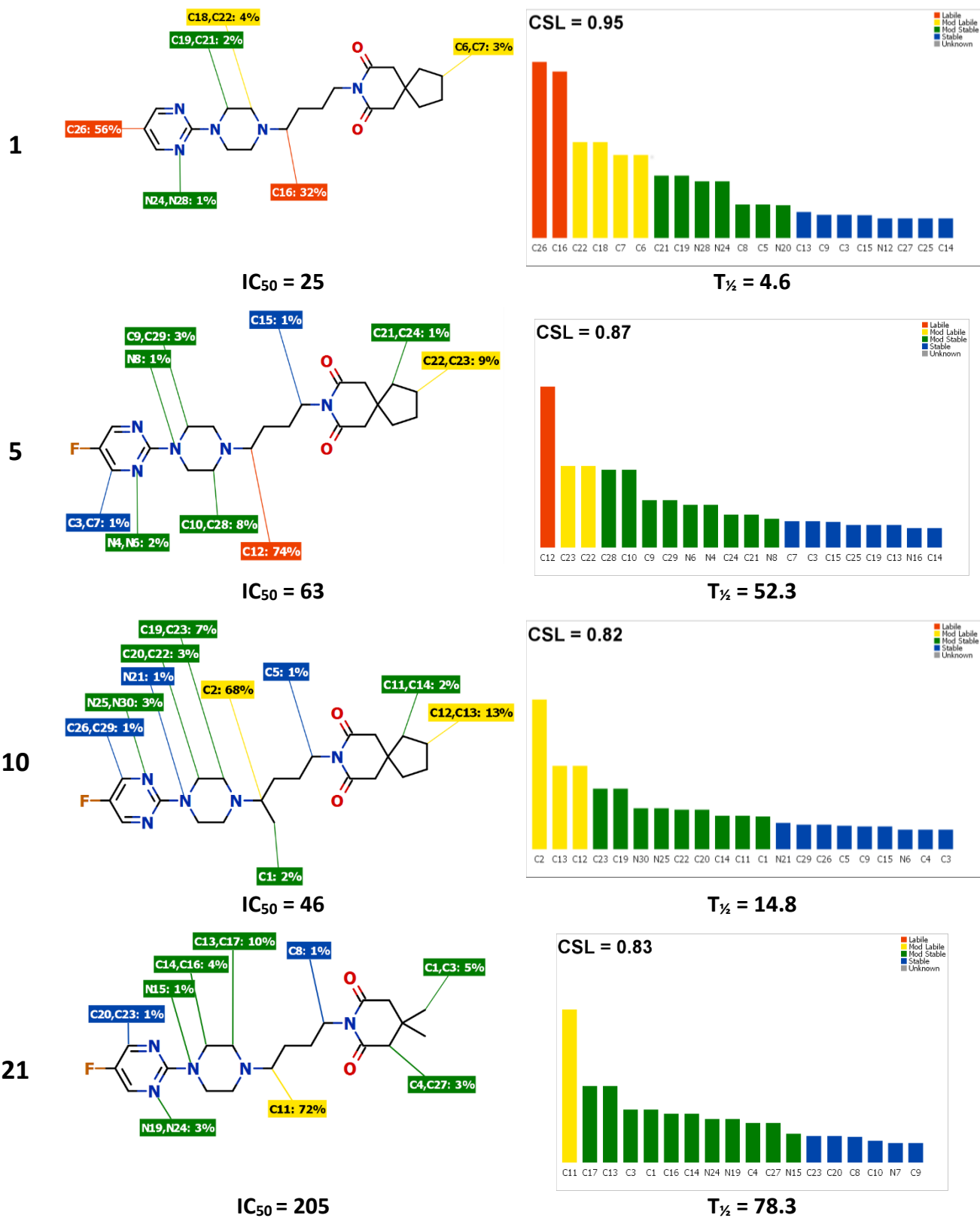
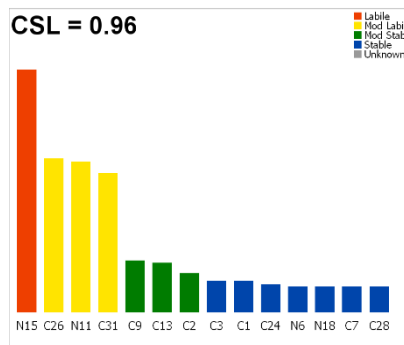
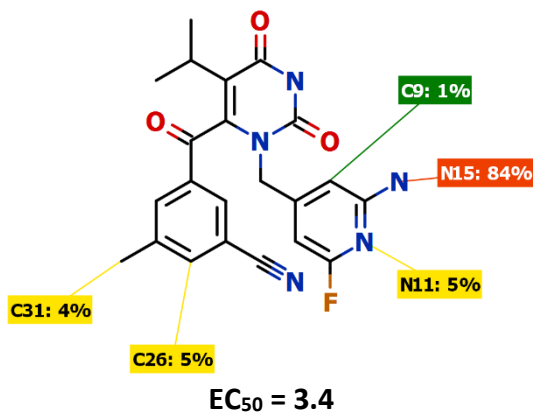
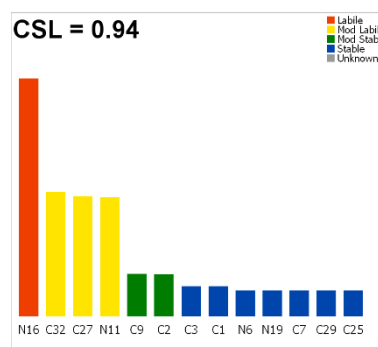
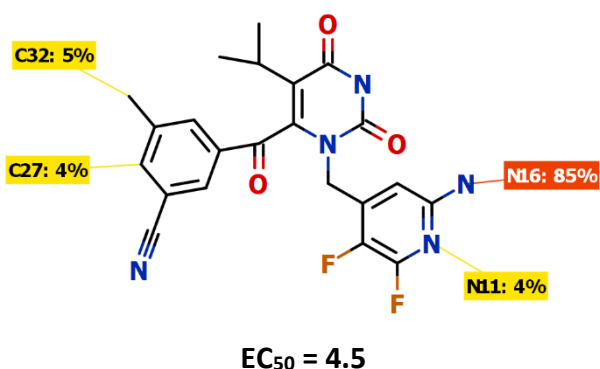


Figure 5 Cytochrome P450 metabolism predictions for example compounds from example 1. The sites of metabolism and predicted regioselectivity are shown for each compound, along with a metabolic landscape illustrating the lability of each site with respect to metabolism by CYP3A4. For each compound the calculated CSL, and experimentally measured half-life ( $T_{1/2}$  in minutes) with respect to *in vitro* metabolism by CYP3A4 and activity ( $IC_{50}$  in nM) against the target 5-HT<sub>1A</sub> are shown. [36]

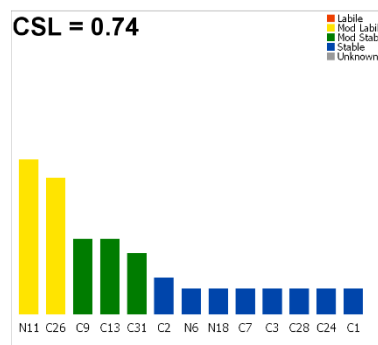
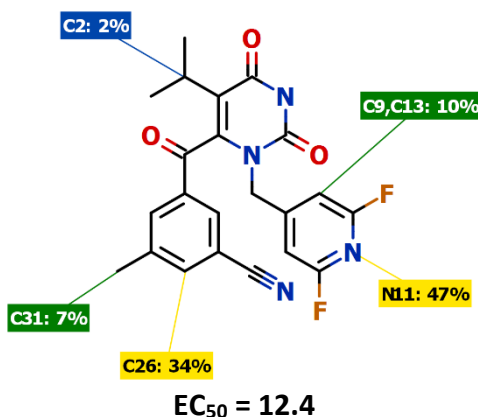
1

T<sub>½</sub> = 45

9

T<sub>½</sub> = 51

10

T<sub>½</sub> = 395

13

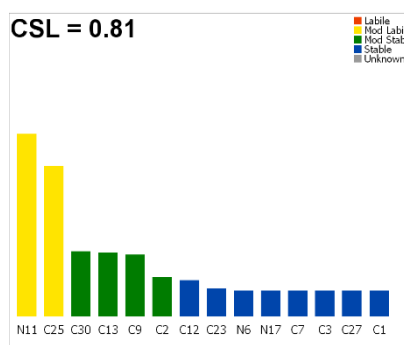
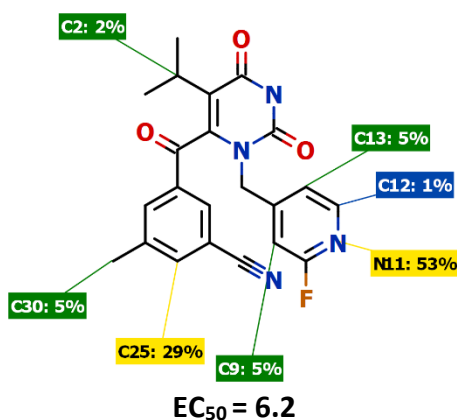
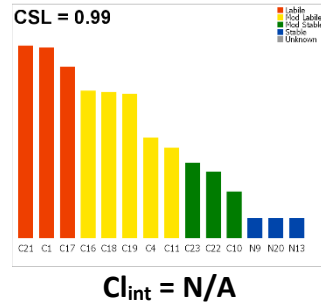
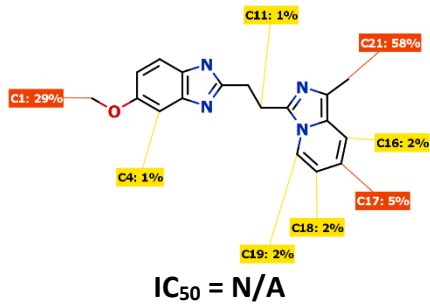
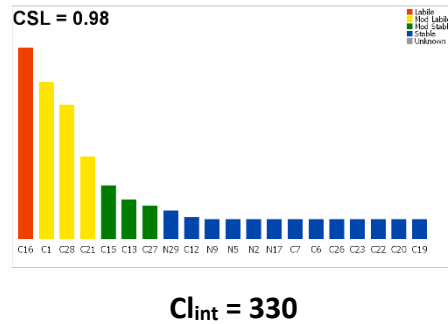
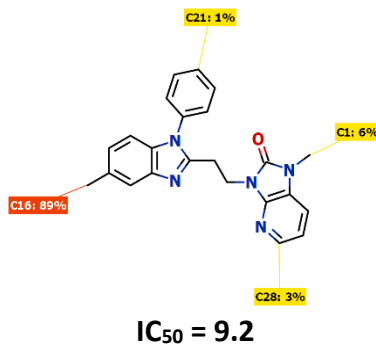
T<sub>½</sub> = 281

Figure 6 Cytochrome P450 metabolism predictions for example compounds from example 2. The sites of metabolism and predicted regioselectivity are shown for each compound, along with a metabolic landscape illustrating the lability of each site with respect to metabolism by CYP3A4. For each compound the calculated CSL, and experimentally measured half-life (T<sub>½</sub> in minutes) in an in vitro human microsomal stability assay and activity (EC<sub>50</sub> in nM) against the target HIV-1 reverse transcriptase are shown. [37]

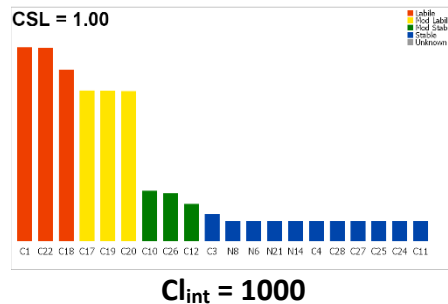
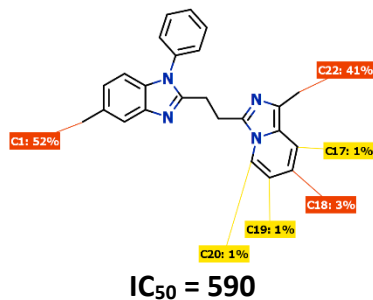
1



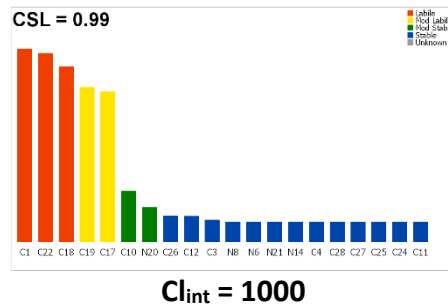
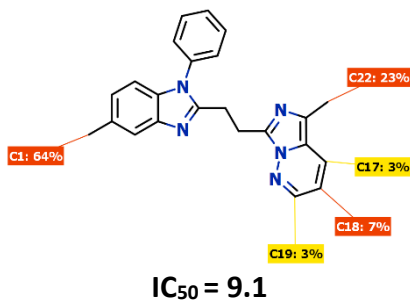
10b



14a



16



24a

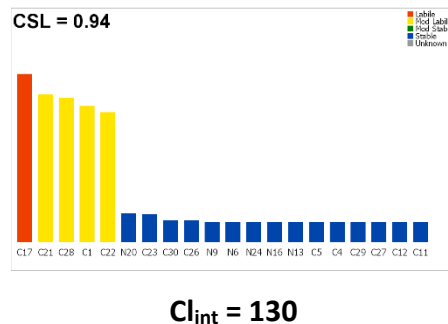
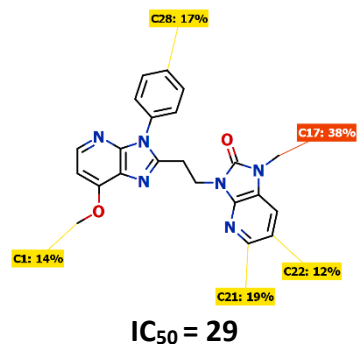


Figure 7 Cytochrome P450 metabolism predictions for compounds from example 3. The sites of metabolism and predicted regioselectivity are shown for each compound along with a metabolic landscape illustrating the lability of each site with respect to metabolism by CYP3A4. For each compound the calculated CSL, and experimentally measured intrinsic clearance ( $Cl_{int}$  in ml/min/kg) in an *in vitro* human microsomal stability assay and activity ( $IC_{50}$  in nM) against the target PDE<sub>10A</sub> are shown. [38]

## Results and Discussion

### Predictive Performance

The results in Table 3 and illustrated in Figure 8 show the predictive performance of the models. The results show the percentage of the independent test sets where a SOM is identified in the top 2 and top 3 predictions, and also the percentage where all SOM are identified in the top 3 predictions.

**Table 3 Site of metabolism (SOM) prediction performance for the independent test sets. Results show the percentage of the compounds in the independent test sets where at least one SOM is correctly identified in the top 2 and 3 predictions, and the percentage where all SOM are identified in the top 3. The average area under the curve of the ROC plots for compounds in the test set is also provided. SMARTCyp comparisons are shown where isoform specific models are available for CYP3A4, CYP2D6 and CYP2C9.**

Isoform	StarDrop				SMARTCyp			
	Top 2 (%)	Top 3 (%)	All Top 3 (%)	AUC	Top 2 (%)	Top 3 (%)	All Top 3 (%)	AUC
3A4	84.5	90.5	53.6	0.87	70.2	84.5	51.2	0.89
2D6	91.1	92.9	71.4	0.91	92.9	96.4	69.6	0.95
2C9	85.7	93.9	75.5	0.91	87.8	91.8	77.6	0.95
1A2	87.7	89.5	64.9	0.87	N/A	N/A	N/A	N/A
2C8	81.5	92.6	70.4	0.86	N/A	N/A	N/A	N/A
2C19	85.7	89.8	69.4	0.89	N/A	N/A	N/A	N/A
2E1	90.0	93.3	80.0	0.84	N/A	N/A	N/A	N/A

In addition, receiver operating characteristic (ROC) plots have been generated for each compound, as illustrated in Figure 9 and the average area under the curve (AUC) for the ROC plots for the compounds in the test set for each isoform are also shown in Table 3. A greater area under the curve for a classifier indicates higher performance; the maximum possible AUC is 1 and a value of 0.5 is equivalent to the performance of random selection. The AUC for each compound and isoform in the training and validation sets is provided in the Supplementary Information.

An alternative measure of performance, Lift, was proposed by Zaretski et al.. [8] This corrects for the fact that it is easier to predict the observed SOM for compounds with a small number of potential sites than for those with a large number. The Lift measures the improvement in accuracy above that expected for random selection. Table 4 shows the Lift achieved by the models described herein.

**Table 4 Lift metric for independent tests sets. Results show the improvement in performance of the models over that expected for a random model for top 2 and 3 predictions. SMARTCyp comparisons are shown where isoform specific models are available for CYP3A4, CYP2D6 and CYP2C9.**

Isoform	StarDrop		SMARTCyp	
	Top 2 (%)	Top 3 (%)	Top 2 (%)	Top 3 (%)
3A4	81	86	67	76
2D6	90	93	94	96
2C9	89	92	91	96
1A2	82	84	N/A	N/A
2C8	78	95	N/A	N/A
2C19	84	87	N/A	N/A
2E1	84	87	N/A	N/A

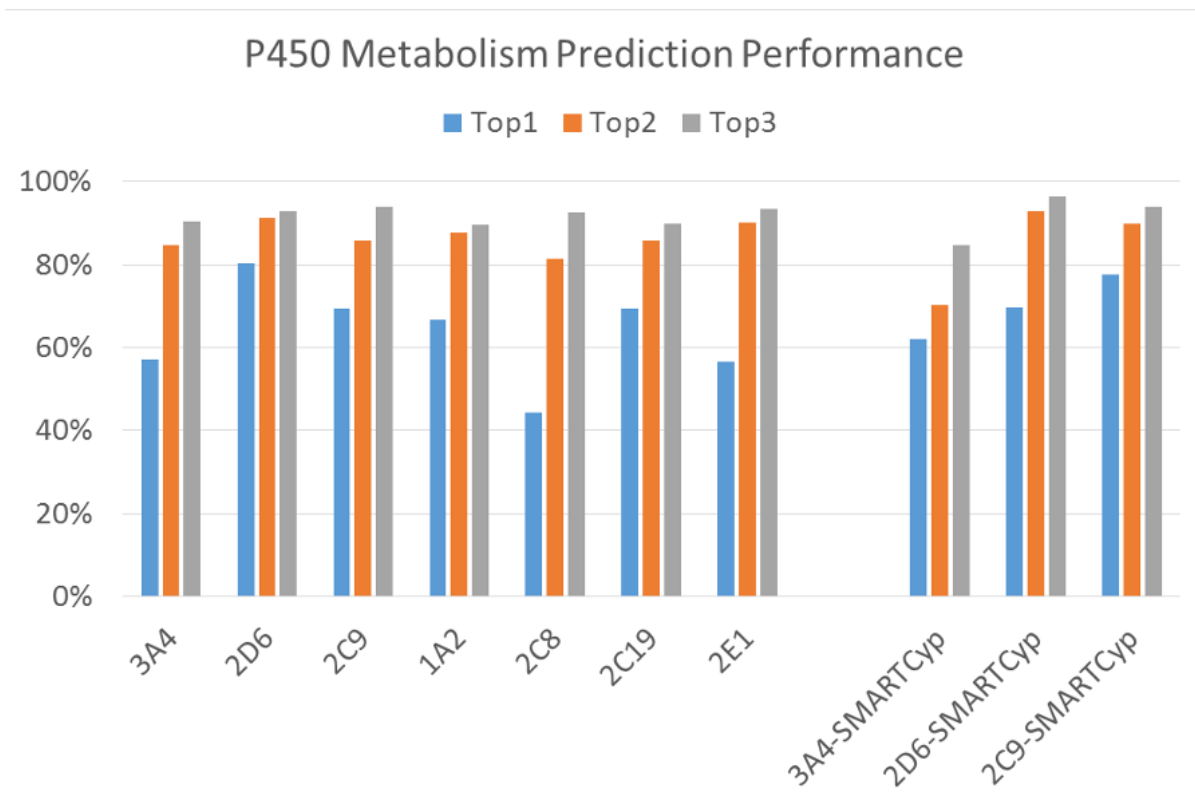
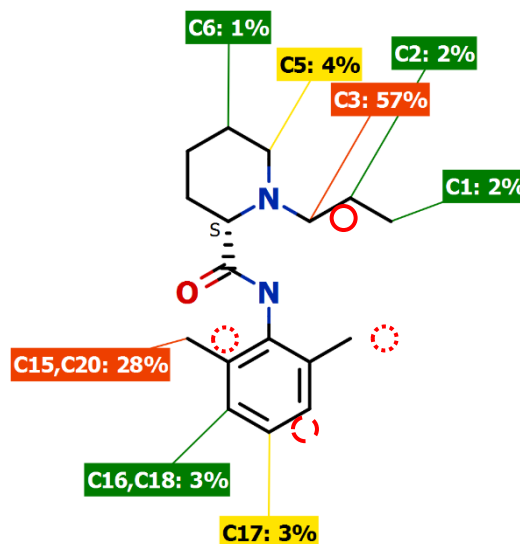
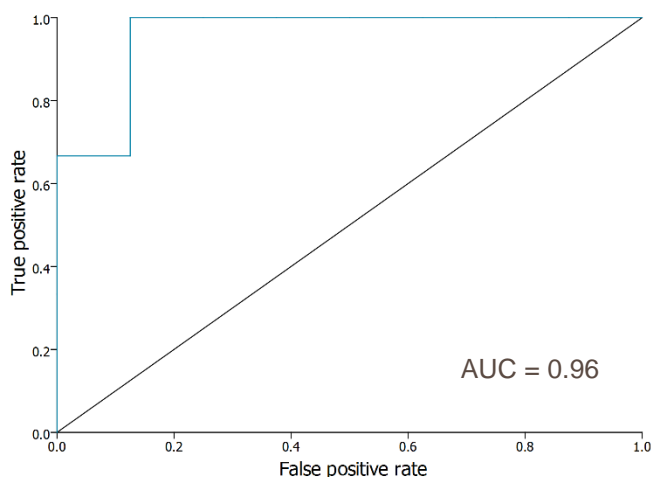
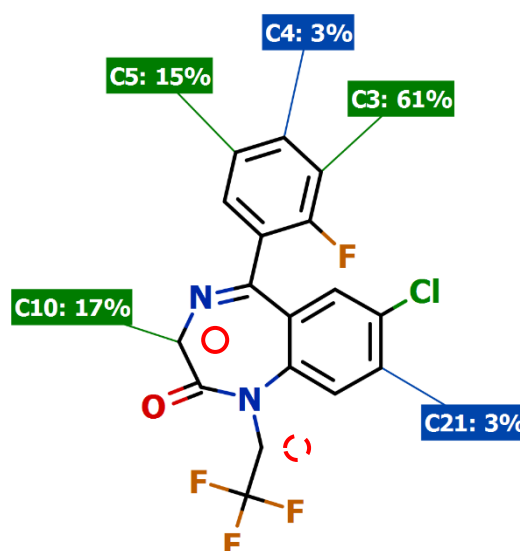
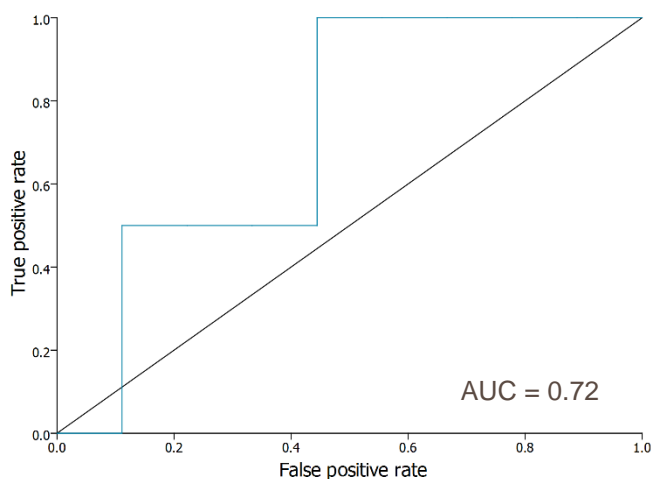


Figure 8 SOM prediction performance of the models described herein on independent tests. The bars labelled top-N show the percentage of an independent data set where at least one observed site of metabolism is identified in the top-N predicted sites. The performance of SMARTCyp on the same sets is shown for comparison for isoforms predicted by SMARTCyp.





(a)



(b)

Figure 9 Receiver operating characteristic (ROC) plots of the true positive rate (TPR (sensitivity)) against the false positive rate (FPR (1 - specificity)) for the prediction of SOM for two compounds. A perfect classifier would be represented by the point in the top left and a performance below the identity line (shown in black) indicates worse performance than a random classification. A greater area under the curve (AUC) for a classifier indicates higher performance; the maximum possible AUC is 1. The corresponding compounds are shown adjacent to each ROC plot, with the predicted regioselectivity indicated by the labels for each site. Primary observed sites of metabolism are highlighted on these structures by a solid red circle, secondary observed sites by a dashed red circle and tertiary observed sites by a dotted red circle. (a) shows an illustrative ROC plot for Ropivacaine which is well, but not perfectly, predicted. (b) shows an example of an ROC plot for a poorly predicted compound, in this case 2-oxo-Quazepam. The AUC is provided in the Supplementary Information for each compound in the data set.

Comparative performance statistics are shown for CYP3A4, CYP2D6 and CYP2C9 from SMARTCyp, which predicts only these isoforms. Similar performance is obtained for CYP2C9 and CYP2D6 but performance on the important CYP3A4 isoform is stronger for the models presented in this paper.

It is informative to examine the contribution to overall predictive performance from the different components of the models: the electronic activation energy,  $\Delta H_A$ , steric hindrance affecting accessibility of each potential site of metabolism and interactions affecting the orientation of the substrate within the CYP binding pocket. The bar charts in Figure 10 compare the performance of different combinations of these components and it is apparent that contribution of the steric component is typically more important than the orientation component. However, the orientation component does have a notable positive influence on the performance of the CYP2D6 models and is able to capture the important interactions between positively charged ligand moieties and negatively charged protein residues (Glu216 and Asp301) that are known to be important for binding. [39]



**Figure 10** SOM prediction performance of different combinations of the three components to this method: elec (the electronic activation energy  $\Delta H_A$ ); steric (the effect of steric hindrance due to the structure of the substrate) and orient (the effect of interactions affecting the orientation of the substrate relative to the oxidizing oxy-heme). The performance of electronic activation energy alone is shown in blue, electronic plus orientation in red, electronic plus steric in grey and all three contributions in yellow. The bars labelled topN show the percentage of compound in an independent data set where at least one observed site of metabolism is correctly identified in the top-N predicted sites. The bars labelled AllTop3 show the percentage of compounds where all SOM are identified in the top 3 predicted sites.

The steric and orientation descriptors capture the influence of functional groups at different bond path distance ranges from a potential site of metabolism, as discussed in the Methods section above. In general, the steric descriptors have a greater contribution to  $E_a$  than the orientation descriptors, consistent with the results in Figure 10 and discussed above. This is especially true close to the potential SOM (less than 6 bonds away) where steric bulk has a positive influence on  $E_a$ , i.e. tend to reduce the relative rate of metabolism. The exception to this is for the descriptors representing SOMs close to planar regions in the CYP2D6 model, where the influence is negative, i.e. they tend to increase the relative rate of metabolism. This is consistent with the observation that CYP2D6 has a tendency to metabolize sites on planar regions of substrates.

The orientation descriptors are subdivided into distances to positively and negatively charged functionalities, with the positive features generally having greater influence on the  $E_a$  than the negative ones. CYPs 1A2 and 2D6 have a similar pattern of orientation descriptor influences that are distinct from the other CYPs. It is also of interest to note that CYP3A4 is not overly influenced by any one set of descriptors, which may be due to its large and flexible active site allowing many different sizes and orientations of molecules to be metabolized.

## Example Applications

We will illustrate the application of the models described above via the following three examples, where example outputs are shown in Figure 5 through Figure 7. These outputs are in the form of a labelled 2D representation of the molecule, summarizing the predicted regioselectivity of metabolism by CYP3A4 (regioselectivity predictions for all isoforms for each compound are provided in the Supplementary Information).

### Example 1 - Developing Buspirone analogues with improved metabolic stability

The feasibility of pursuing a fast-follower for Buspirone, a 5HT<sub>1A</sub> ligand used as an anti-anxiolytic therapeutic, was explored in Tandon et al.. [36] Buspirone experiences rapid metabolism by CYP3A4 leading to low oral bioavailability and a short half-life in humans and this study aimed to identify analogues of Buspirone with greater metabolic stability whilst maintaining receptor affinity. The published study was guided by prospective application of an earlier version of the models described herein, but here we have repeated the calculations with the latest models.

The structure of Buspirone (see compound 1 in Figure 5) can be broken down into 3 regions:

- an arylpiperazine which is a protonatable recognition element important for receptor affinity and is metabolized via hydroxylation of pyrimidine C5
- a tetramethylene linker which is metabolized by N-dealkylation alpha to the piperazine N4
- a piperidinedione which is metabolized via oxidation of the spirocyclopentane ring.

This study explored structural modifications with a view to improving metabolic stability. Here we will compare the experimentally observed changes in in vitro half-life with respect to metabolism by CYP3A4 to predictions from the latest models described in this paper. The predicted metabolic profile of Buspirone and its analogues are shown in Figure 5. The presence of two labile sites and a high CSL is consistent with the observed short half-life observed in vitro (4.6 minutes) and rapid metabolism in vivo.

Blocking the 5 position of the pyrimidine ring (predicted as labile with a regioselectivity of 58%) with fluorine led to compound 5, where activity at the target is maintained but the half-life increased to 52 minutes. In this case, one region of the molecule was modified but other labile and moderately labile sites remain, so only a small change in the overall CSL is observed even though this modification is beneficial. As noted above, other factors also influence the overall rate of metabolism and a direct correlation between the small changes to CSL and the CYP3A4 half-life is not necessarily expected. However, in this case the increase in half-life is reflected by a fall in CSL from 0.957 to 0.885.

An example of the complex relationship between structural changes and metabolic stability is demonstrated by molecule 10, which introduced a methyl substituent alpha to the piperazine in an attempt to hinder N-dealkylation from what is predicted to be a labile site. The models predict a further small decrease in CSL to 0.8458. However, the half-life falls to 14.8 minutes, indicating that the factors mentioned earlier are influential here. In this instance, the addition of the methyl group changes the lipophilicity and basicity of the compound which are likely to increase the binding affinity to CYP3A4 and hence offset the small decrease in CSL to increase the rate of metabolism.

Replacing the spirocyclopentane ring with a gem dimethyl to give compound 21 eliminated the predicted moderately labile site in the five-membered ring which was reflected by an increase in half-life to 78 minutes.

Overall, this example illustrates that the models can be used to guide development of a lead compound towards greater metabolic stability, where in this example compounds 5 and 21 show half-lives of 52 and 78 minutes respectively whilst maintaining activity of 0.2  $\mu\text{M}$  or lower.

### Example 2: Developing HIV-1 reverse transcriptase inhibitors with improved metabolic stability

A series of N1-heterocyclic pyrimidinediones were investigated by Mitchell et al. [37] for application as HIV-1 non-nucleoside reverse transcriptase inhibitors (NNRTIs) with the aim of improving the pharmacokinetic profile whilst maintaining activity.

Compound 1, shown in Figure 6, achieved the required target activity but the half-life of 45 minutes in human liver microsomes was a long way short of the target for once daily dosing. The models predicted the terminal amine group to be a labile site with the results for compound 1 and its analogues given in Figure 6. Compound 1 contains a fluorine substitution ortho to the pyridine N. Further substitution of the pyridine ring with fluorine to give compound 9 did not give a significant improvement in terms of prediction or measured data.

To significantly improve the half-life for this series it was necessary to address the metabolically labile terminal amine. Replacement with a hydrogen led to compound 13 where the CSL falls to 0.8309, corresponding to a large increase in half-life to 281 minutes. Further substitution of the pyridine to block both aromatic sites ortho to the pyridine N with fluorine, to give compound 10, led to a further reduction in the CSL and an increase in half-life to in excess of 395 minutes.

This relatively simple example shows how CYP models can quickly identify labile sites to focus on modifications that are likely to improve metabolic stability. In this case, the replacement of a labile terminal amine and blocking aromatic sites with fluorine, as illustrated by compounds 10 and 13, showed improved metabolic stability and retained good antiviral potency.

### Example 3: Novel benzimidazoles as PDE10A inhibitors with improved metabolic stability

A series of novel benzimidazoles were developed by Chino et al. [38] that show sub-micromolar activity as inhibitors of PDE10A, which is hypothesized to be effective in treating schizophrenia and a wide range of neurological, psychotic, anxiety and movement disorders by increasing levels of cAMP and cGMP in the brain.

Compound 1, as shown in Figure 7, was identified from high throughput screening as a low micromolar PDE10A inhibitor where introduction of a phenyl ring to the N-1 position on benzimidazole was found to improve inhibitory activity. It was noted that compound 14a with a methyl at the 5-position on benzimidazole and a methyl in the 1-prime position of the imidazopyridine was approximately 3 times more active than compound 1, and removal of the methyl in the 1-prime position removed inhibitory activity, indicating the importance of this group (data not shown). This position is predicted by the CYP models to be metabolically labile, along with the 5-methyl on the benzimidazole, with further labile and moderately labile sites in the aromatic positions, as shown in Figure 7. This causes these compounds to have high risk of rapid metabolism by CYPs, as illustrated by the CSL values, and borne out in the experimental results with compound 14a exhibiting clearance of greater than 1000 mL/min/kg.

Introduction of another N into the fused ring system to give the imidazopyridazine in compound 16 gave improved inhibition but did not improve metabolic stability. The methyl substituents were shown to be important for activity so variations to the heterocycles that preserved these groups were made. The imidazopyridazine was replaced with an azabenzimidazolinone to give compound 10b, a change which gave improved metabolic stability, due to the loss of some moderately labile aromatic sites.

Focus then shifted to the benzimidazole part of the molecule to further improve metabolic stability. Variation of the 5 methyl and insertion of N into the benzimidazole ring system at the 7 position led to compound 24a showing improved metabolic stability and reflected by a lower CSL. Whilst these changes are not directly blocking a predicted labile site they do impact on the metabolic stability of sites elsewhere in the molecule and show that sometimes subtler longer range effects come into play.

This example highlights a situation where the lead series was developed by making larger changes to molecular fragments, not simply blocking a labile site, with non-intuitive changes to the metabolic stability. In this example the QM methodology employed by the models was able to capture these trends by considering each fragment in its actual environment and would allow a chemist developing a lead series to gain insights into the likely impact of even quite large changes to their molecules.

## Conclusions

This paper has described the prediction of metabolism by Cytochrome P450 based on an understanding of the catalytic mechanism of P450 metabolism. The approach uses semi-empirical QM methods to estimate the electronic activation energy, based on a strong foundation of experimental data and ab initio DFT calculations. This is coupled with a ligand-based methodology to calculate adjustments for isoform-specific steric and orientation effects due to interactions with the binding pockets. These QM based methods offer generality and transferability since they are derived from fundamental physical principles, rather than relying on empirical pattern matching with a limited domain of applicability. A further advantage of a QM method is that each site is considered in the context of the molecular environment in which it resides rather than splitting a molecule into fragments and treating each as a discrete uniform entity when assessing the likelihood of metabolism.

The resulting models show excellent performance for the prediction of SOM for seven major human drug metabolizing isoforms of CYP: CYP3A4, CYP2D6, CYP2C9, CYP1A2, CYP2C19, CYP2E1 and CYP2C8.

Furthermore, by calculating an approximation to the activation energy for each potential site of metabolism, the lability of each site to metabolism can be estimated on an absolute scale; a unique feature that is beneficial to medicinal chemists optimizing candidate drugs to improve metabolic stability.

Future work in this field will include the coupling of these models with predictions of the CYP isoforms that are most likely to be responsible for metabolism of a compound. This will provide a guide to the most relevant regioselectivity prediction and aid in the prediction of the metabolite profile resulting from CYP metabolism.

## Experimental

All results were generated using StarDrop v6.1 [32].

## Acknowledgements

This paper is dedicated to the memory of Patrik Rydberg, who made major contributions to the field of P450 metabolism modelling and the methods described herein, prior to his untimely passing in November 2013.

This research has received funding from the Union Seventh Framework Programme 2013 under the grant agreement no. 602156.

## Supplementary Information

The data sets described herein are available for download from [http://www.optibrium.com/downloads/Tyzack\\_et\\_al\\_data\\_sets.zip](http://www.optibrium.com/downloads/Tyzack_et_al_data_sets.zip). In addition, the regioselectivity predictions for the compounds in the illustrative examples are provided for all isoforms at [http://www.optibrium.com/downloads/Tyzack\\_et\\_al\\_supporting\\_information\\_regioselectivities.pdf](http://www.optibrium.com/downloads/Tyzack_et_al_supporting_information_regioselectivities.pdf).

## References


- 1 Guengerich PF. Cytochrome P450s and other enzymes in drug metabolism and toxicity. *The AAPS Journal*. 2006;8(1):E101-11.
- 2 Lewis DFV. 57 varieties: the human cytochromes P450. *Pharmacogenomics*. 2004;5(3):305-318.
- 3 Shaik S, Cohen S, Wang Y, Chen H, Kumar D, Thiel W. P450 enzymes: their structure, reactivity, and selectivity-modeled by QM/MM calculations. *Chem. Rev.* 2010;110(2):949-1017.
- 4 Feenstra KA, De Graaf C, Vermeulen NPE. Cytochrome P450 protein modelling and ligand docking. In: Rodrigues AD, editor. *Drug-drug interactions*. New York: Informa Healthcare; 2008. p. 435-470.

- 5 Nair PC, McKinnon RA, Miners JO. Cytochrome P450 structure-function: insights from molecular dynamics simulations. *Drug Metab. Rev.* 2016;48(3):434-452.
- 6 Panneerselvam S, Yesudhas D, Durai P, Anwar MA, Gosu V, S. C. A combined molecular docking/dynamics approach to probe the binding mode of cancer drugs with cytochrome P450 3A4. *Molecules.* 2015;20(8):14915-14935.
- 7 Dubey KD, Wang B, Shaik S. Molecular dynamics and QM/MM calculations predict the substrate-induced gating of cytochrome P450 BM3 and the regio- and stereoselectivity of fatty acid hydroxylation. *J. Am. Chem. Soc.* 2016;138(3):837-845.
- 8 Zaretski J, Bergeron C, Rydberg P, Huang TW, Bennet KP, Breneman CM. RS-Predictor: A New Tool for Predicting Sites of Cytochrome P450-Mediated Metabolism Applied to CYP 3A4. *J. Chem. Inf. Model.* 2011;51(7):1667-1689.
- 9 Zaretski J, Rydberg P, Bergeron C, Bennett KP, Olsen L, Breneman CM. RS-Predictor models augmented with SMARTCyp reactivities: robust metabolic regioselectivity predictions for nine CYP isozymes. *J. Chem. Inf. Model.* 2012;52(6):1637-1659.
- 10 Kirchmair J, Williamson MJ, Tyzack JD, Tan L, Bond PJ, Bender A, Glen RC. Computational prediction of metabolism: sites, products, SAR, P450 enzyme dynamics, and mechanisms. *J. Chem. Inf. Model.* 2012;52(3):617-648.
- 11 Kulkarni SA, Zhu J, Blechinger S. In silico techniques for the study and prediction of xenobiotic metabolism: a review. *Xenobiotica.* 2005;35(10-11):955-973.
- 12 Tarcsay A, Keseru GM. In silico site of metabolism prediction of cytochrome P450-mediated biotransformations. *Expert Opin. Drug Metab. Toxicol.* 2011;7(3):299-312.
- 13 Ekins S, Andreyev S, Ryabov A, Kirillov E, Rakhmatulin E, Bugrim A, Nikolskaya T. Computational prediction of human drug metabolism. *Expert Opin. Drug Metab. Toxicol.* 2005;1(2):303-324.
- 14 Vaz RJ, Zamora I, Li Y, Reiling S, Shen J, Cruciani G. The challenges of in silico contributions to drug metabolism in lead optimization. *Expert Opin. Drug Metab. Toxicol.* 2010;6(7):851-861.
- 15 Rydberg P, Gloriam DE, Zaretski J, Breneman C, Olsen L. SMARTCyp: A 2D method for prediction of cytochrome P450-mediated drug metabolism. *ACS Med. Chem. Lett.* 2010;1(3):96-100.
- 16 Jones JP, Mysinger M, Korzekwa KR. Computational models for cytochrome P450: a predictive electronic model for aromatic oxidation and hydrogen atom abstraction. *Drug Metab. Dispos.* 2002;30(1):7-12.
- 17 Ogliaro F, Harris N, Cohen S, Filatov M, de Visser SP, Shaik S. A Model Rebound Mechanism of Hydroxylation by Cytochrome P450: Stepwise and Effectively Concerted Pathways and their Reactivity Patterns. *J. Am. Chem. Soc.* 2000;122(37):8977-8989.
- 18 Shaik S, de Visser S, Ogliaro F, Schwarz H, Schroder D. Two-state reactivity mechanisms of hydroxylation and epoxidation by cytochrome P-450 revealed by theory. *Curr. Opin. Chem. Biol.* 2002;6(5):556-567.
- 19 Bathelt CM, Ridder L, Mulholland AJ, Harvey JN. Aromatic hydroxylation by cytochrome P450: model calculations of mechanism and substituent effects. *J. Am. Chem. Soc.* 2003;125(49):15004-15005.
- 20 de Visser SP, Shaik S. A proton-shuttle mechanism mediated by the porphyrin in benzene hydroxylation by cytochrome p450 enzymes. *J. Am. Chem. Soc.* 2003;125(24):7413-7424.



- 21 Sharma PK, de Visser SP, Shail S. Can a single oxidant with two spin states masquerade as two different oxidants? A study of the sulfoxidation mechanism by cytochrome p450. *J. Am. Chem. Soc.* 2003;125(29):8698-8699.
- 22 Rydberg P, Ryde U, Olsen L. Sulfoxide, Sulfur, and Nitrogen Oxidation and Dealkylation by Cytochrome P450. *J. Chem. Theo. Comp.* 2008;4(8):1369-1377.
- 23 Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A. Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* 1985;107(13):3902-3909.
- 24 Brønsted JN, Pedersen KJ. Stöchiometrie und Verwandtschaftslehre. *Zeitschrift für Phys. Chemie.* 1924;108:185-235.
- 25 Korzekwa KR, Jones JP, Gillette JR. Theoretical studies on cytochrome P-450 mediated hydroxylation: a predictive model for hydrogen atom abstractions. *J. Am. Chem. Soc.* 1990;112(19):7042-7046.
- 26 Li D, Wang Y, Han K. Recent density functional theory model calculations of drug metabolism by cytochrome P450. *Coord. Chem. Rev.* 2012;256(11-12):1137-1150.
- 27 Rydberg P. Reactivity-Based Approaches and Machine Learning Methods for Predicting the Sites of Cytochrome P450-Mediated Metabolism. In: Kirchmair J, editor. *Drug Metabolism Prediction*. Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA; 2014.
- 28 Kumar D, Karamzadeh B, Sastry GN, de Visser SP. What Factors Influence the Rate Constant of Substrate Epoxidation by Compound I of Cytochrome P450 and Analogous Iron(IV)-Oxo Oxidants? *J. Am. Chem. Soc.* 2010;132(22):7656-7667.
- 29 Daylight Chemical Information Systems Inc. SMARTS Tutorial. [Internet]. [cited 2012 Jul 2]. Available from: <http://www.daylight.com/dayhtml/tutorials/languages/smarts/index.html>.
- 30 Wehrens R, Mevik BH. The pls Package: Principal Component and Partial Least Squares Regression in R. *J. Stat. Soft.* 2007;18(2):1-24.
- 31 Campagna-Slater V, Pottel J, Therrien E, Cantin L, Moitessier N. Development of a computational tool to rival experts in the prediction of sites of metabolism of xenobiotics by P450s. *J. Chem. Inf. Model.* 2012;52(9):2471-2483.
- 32 Optibrium. [Internet]. [cited 2015 January 8]. Available from: <http://www.optibrium.com/stardrop>.
- 33 van der Maaten LPJ, Hinton GE. Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research.* 2008;9:2579-2605.
- 34 Korzekwa KR, Trager WF, Gillette JR. Theory for the observed isotope effects from enzymatic systems that form multiple products via branched reaction pathways: cytochrome P-450. *Biochemistry.* 1989;28(23):9012-9018.
- 35 Korzekwa KR, Jones JP, inventors. Relative rates of cytochrome P450 metabolism. 2000 July 10. US Patent 6,625,547.
- 36 Tandon M, O'Donnel MM, Porte A, Vensel D, Yang D, Palma R, Beresford A, Ashwell M. The design and preparation of metabolically protected new arylpiperazine 5-HT<sub>1A</sub> ligands. *Bioorg. Med. Chem. Lett.* 2004;14(7):1709-1712.



- 
- 37 Mitchell ML, Son JC, Lee IY, Ching K, Kim HS, Guo H, Wang J, Hayes J, Wang M, Paul A, et al. N1-Heterocyclic pyrimidinediones as non-nucleoside inhibitors of HIV-1 reverse transcriptase. *Bioorg. Med. Chem. Lett.* 2010;20(5):1585-1588.
- 38 Chino A, Masuda N, Amano Y, Honbou K, Mihara T, Yamazaki M, Tomishima M. Novel benzimidazole derivatives as phosphodiesterase 10A (PDE10A) inhibitors with improved metabolic stability. *Bioorg. Med. Chem.* 2014;22(13):3515-3526.
- 39 Rydberg P, Olsen L. Ligand-Based Site of Metabolism Prediction for Cytochrome P450 2D6. *ACS Med. Chem. Lett.* 2012;3(1):69-73.
- 40 Molecular Networks. Corina. [Internet]. [cited 18 August 2016]. Available from: <https://www.mn-am.com/products/corina>.