Journal of
CHEMICAL INFORMATION
AND MODELING

Article

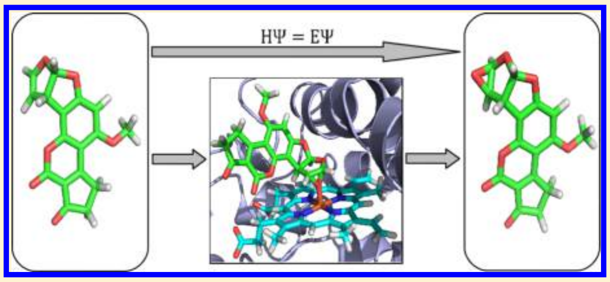
pubs.acs.org/jcim

Prediction of Cytochrome P450 Xenobiotic Metabolism: Tethered Docking and Reactivity Derived from Ligand Molecular Orbital Analysis

Jonathan D. Tyzack, Mark J. Williamson, Rubben Torella, and Robert C. Glen*

Unilever Centre for Molecular Science Informatics, Department of Chemistry, Lensfield Road, Cambridge, CB2 1EW, United Kingdom

Supporting Information



ABSTRACT: Metabolism of xenobiotic and endogenous compounds is frequently complex, not completely elucidated, and therefore often ambiguous. The prediction of sites of metabolism (SoM) can be particularly helpful as a first step toward the identification of metabolites, a process especially relevant to drug discovery. This paper describes a reactivity approach for predicting SoM whereby reactivity is derived directly from the ground state ligand molecular orbital analysis, calculated using Density Functional Theory, using a novel implementation of the average local ionization energy. Thus each potential SoM is sampled in the context of the whole ligand, in contrast to other popular approaches where activation energies are calculated for a predefined database of molecular fragments and assigned to matching moieties in a query ligand. In addition, one of the first descriptions of molecular dynamics of cytochrome P450 (CYP) isoforms 3A4, 2D6, and 2C9 in their Compound I state is reported, and, from the representative protein structures obtained, an analysis and evaluation of various docking approaches using GOLD is performed. In particular, a covalent docking approach is described coupled with the modeling of important electrostatic interactions between CYP and ligand using spherical constraints. Combining the docking and reactivity results, obtained using standard functionality from common docking and quantum chemical applications, enables a SoM to be identified in the top 2 predictions for 75%, 80%, and 78% of the data sets for 3A4, 2D6, and 2C9, respectively, results that are accessible and competitive with other recently published prediction tools.

INTRODUCTION

Understanding the metabolism of xenobiotic and endogenous compounds remains a central challenge to the development and application of drugs, cosmetics, nutritional supplements, and agrochemicals. Undesirable effects such as the emergence and accumulation of toxic metabolites or the elimination of products before they can have their therapeutic effect can result from the actions of metabolizing enzymes. Attrition rates are high in the development of new chemical entities, and the toxicity of metabolites is a major contributor to the withdrawal of new pharmaceuticals or black-box warnings. As a result, being able to predict the sites and products of xenobiotic metabolism has become an important avenue of research.

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.[1,2] The CYP family consists of 57 isoforms with the majority of biotransformations being carried out by 3A4, the most promiscuous isoform, followed by 2D6 and 2C9. A comprehensive overview of the structure and reactivity of CYPs based on QM/MM calculations can be found in the review paper from Shaik et al.[3]

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 C−H 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. This gives rise to the common description of CYP as a monooxygenase. However, the CYPs show their chemical diversity in the other reactions they catalyze such as desaturation, oxidative dehalogenation, reductive dehalogenation, and C−C coupling.[3] The oxidation reactions proceed via different pathways depending on the substrate: hydrocarbon hydroxylation proceeds via an initial hydrogen atom transfer (HAT) followed by reaction of the ferryl oxygen with the alkyl radical, a process known as the rebound mechanism;[4] alkene oxidation proceeds via activation of the double bond to form an iron alkoxy radical species;[5] aromatic C oxidation proceeds via activation of an aromatic bond[6] followed by an intramolecular HAT known as "NIH-shift";[7] terminal alkyne oxidation occurs via activation of the triple bond followed by intramolecular HAT of the terminal hydrogen to give ketene products;[8] and direct oxidation of hetero atoms such as sulfur and nitrogen proceeds via the formation of σ bonds between nonbonding electrons on the heteroatom and the ferryl oxygen,[9,10] although often deal-

Received: January 25, 2013
Published: May 24, 2013

kylation reactions are favorable over direct oxidation. The diverse reaction mechanisms make the generation of a universally applicable SoM prediction tool challenging since different pathways must be included.

Experimental investigation of xenobiotic metabolism can be both resource and time-consuming which has led to the development of a number of computational approaches to predict SoM. Most methods incorporate reactivity and accessibility considerations since chemical reactivity and ligand orientation within the CYP cavity are important factors in dictating SoM. The ferryl oxygen in Compound I (CpdI) of the CYP catalytic cycle is very reactive since it exists as a triradical and can catalyze oxidation reactions with many moieties including inactivated aliphatic groups, hence a ranking of likely SoM in a ligand based on relative accessibility and reactivity is a common approach.

Computational techniques to predict xenobiotic metabolism can be separated into two distinct categories: ligand-based and structure-based. In the first approach, structures and properties of known substrate or nonsubstrate 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 ligands. For a general overview of current computational tools to predict SoM the reader is referred to the many comprehensive review papers,[11−15] with those approaches most relevant to this work summarized in the next section.

Overview of Current, Relevant Computational Approaches. *Reactivity Methods.* Reactivity measures play an important part in many *in silico* SoM prediction tools. The approaches can be broadly divided into two categories: those that make use of reactivity descriptors derived from the electronic structure of the molecule and those that attempt to calculate the activation energy for the formation of reaction intermediates explicitly.

Important concepts for methods based on the former were developed by Fukui[16] who highlighted the importance of the frontier orbitals in describing chemical reactivity. The Fukui function[17] has been defined by Parr and Yang as the rate of change of electron density with respect to adding or removing electrons to assess the ease of nucleophilic, f^+, or electrophilic, f^-, attack, respectively

$$f^{\pm} = \left(\frac{\delta\rho(r)}{\delta N}\right)^{\pm}_{V(r)} \tag{1}$$

where $\rho(r)$ represents the electron density at point r, N represents the number of electrons, and $V(r)$ represents the external potential. This formalized the notion of approximating chemical reactivity with the ease of adding and removing electrons at different points in the molecule. Various methods have been developed in an attempt to approximate these derivatives, and a condensed form[18] was proposed where the Mulliken population of a molecule is compared after adding or removing an electron, a method that suffers from the drawback that the multiplicity changes. An approximation to the Fukui function[19] at atomic site A has been defined as

$$f^{i}_{A} = \sum_{\alpha\in A}\sum_{\beta} c_{\alpha i}c_{\beta i}S_{\alpha\beta} \tag{2}$$

where α represents basis functions centered on atom A, $c_{\alpha i}$ and $c_{\beta i}$ are basis function coefficients for α and β in molecular orbital i, molecular orbital i is either the HOMO or LUMO for electrophilic or nucleophilic attack, respectively, and $S_{\alpha\beta}$ is the overlap integral for α and β

$$S_{\alpha\beta} = \int_{-\infty}^{\infty}\chi_{\alpha}\chi_{\beta}\,dxdydz \tag{3}$$

where again α and β are used to index the basis functions, and χ_{α} is a function describing the basis function α in spatial coordinates.

The definition of superdelocalizability[20] extended these approaches by considering all molecular orbitals, weighting the contribution from each molecular orbital i by its energy. Electrophilic superdelocalizability, L^{-}_{A}, is a measure of the ability to remove electrons from occupied orbitals at atomic site A and is defined as

$$L^{-}_{A} = 2\sum_{i=1}^{N_{occ}}\frac{\sum_{\alpha\in A}(c_{\alpha i})^2}{|\varepsilon_i|} \tag{4}$$

where ε_i is the energy of molecule orbital i. Nucleophilic superdelocalizability, L^{+}_{A}, is a measure of the ability of atom A to accommodate additional electron density in the unoccupied orbitals and is given by

$$L^{+}_{A} = 2\sum_{i=N_{occ}+1}^{N_{MO}}\frac{\sum_{\alpha\in A}(c_{\alpha i})^2}{|\varepsilon_i|} \tag{5}$$

Both of these definitions only consider the one center terms.

Other methods that have been used as a gauge of the ease of removing electron density from different parts of a molecule include the average local ionization energy[21] which calculates an isosurface at a particular electron density around the molecule and is defined as the average energy needed to ionize an electron at a particular point r on this isosurface

$$I(r) = \frac{\sum_i \rho_i(r)|\varepsilon_i|}{\rho(r)} \tag{6}$$

where ε_i represents the energy of molecular orbital i.

HATs are the rate determining step for a large proportion of CYP catalyzed biotransformations with hydrogen bond order,[22] BO_{H_A}, commonly used to assess the ease of HAT

$$BO_{H_A} = \sum_{\alpha\in H_A}\sum_{\beta\notin H_A}(DS)_{\alpha\beta}(DS)_{\beta\alpha} \tag{7}$$

where α represents the basis functions centered on hydrogen atom H_A, β represents all other basis functions in the molecule, D represents the density matrix, and S represents the overlap matrix. The density matrix D is defined as

$$D_{\alpha\beta} = \sum_{i}^{N_{MO}}(n_i c_{\alpha i}c_{\beta i}) \tag{8}$$

where i indexes the occupied molecular orbitals; n_i refers to the number of electrons in orbital i; α and β are used to index the basis functions, and $c_{\alpha i}$ refers to the coefficient of basis function α in molecular orbital i. In this way the hydrogen bond order is defined over all two center terms involving basis functions centered on the hydrogen atom and is a measure of the strength by which the hydrogen is bound.

The QMBO method[23] makes SoM predictions based on this type of bond order descriptor coupled with an accessibility measure in the form of a solvent accessible surface area (SASA) metric. CypScore[24] creates multiple models for the different CYP reaction pathways based on a range of atomic reactivity descriptors calculated using AM1 semiempirical molecular orbital theory. Multiple linear regression is used to generate the descriptor weights for each reaction type, with some of the most important descriptors including SASA, atomic charge, ionization energy, and C−H bond order. Other approaches such as RS-Predictor[25,26] calculate a plethora of reactivity and topological descriptors, including the reactivity descriptors described above, and use machine learning techniques to produce a prediction tool validated by cross-validation within the data set.

The second category of reactivity methods attempts to calculate the activation energy explicitly by modeling the reaction between a molecular fragment and a surrogate for the heme-iron oxygen species using *ab initio* or semiempirical methods.[27−30] A database of activation energies for predefined molecular fragments can then be created, and, when used in conjunction with docking approaches to model accessibility, successful SoM prediction tools[31,32] have been produced. A further study[33] used machine learning to combine descriptors based on the electronic structure of the molecule, including the Fukui function and atomic charge descriptors, with explicitly calculated activation energies[29,30] and SASA descriptors, with the activation energy shown to be the most important in determining SoM. Another approach called E_A*MEAD* (Activation energy of Metabolism reactions with Effective Atomic Descriptors)[34] also modeled the transition state of CYP oxidation explicitly, with a separate model for aromatic hydroxylation and aliphatic hydroxylation, but used descriptors such as bond dipole moment, atomic charges, and atomic polarizabilities to generate an approximation to the activation energy.

A large body of work has been generated by the Rydberg group which includes an open-source Java-based SoM predictor called SMARTCyp.[35] A database of activation energies, calculated using Density Functional Theory[36] (DFT), has been generated for various predefined ligand fragments and used to assign reactivity estimates to matching moieties in a query ligand, with an accessibility descriptor used to tune the ranking.

MetaSite[37] uses fragment matching to precomputed energies of reaction intermediates coupled with GRID-derived Molecular Interaction Fields (MIFs) between CYP and ligand to make SoM predictions, while the commercial application StarDrop[38] uses AM1 HAT energy calculations combined with accessibility descriptors.

Docking Methods. Other methods take the accessibility consideration further and employ docking techniques to refine the SoM predictions from reactivity approaches. Crystal structures are available for the major human CYP isoforms enabling structure-based techniques such as docking to be applied.

A docking experiment generally proceeds by sampling conformational space of the ligand within the binding site and optimizing the pose based on a measure of the affinity or fit of the ligand for the receptor. Current ligand-docking methods have many limitations[39] including the extent to which conformational space is sampled; approximate descriptions of desolvation and entropic and enthalpic components; modeling the presence of solvent/water; and the extent to which protein and ligand flexibility is modeled.

Issues specific to CYP protein−ligand docking relate to the open active site, particularly for the 3A4 isoform,[40] and the extensive hydrophobic surface areas lacking strong, directed hydrogen bonding interactions. This makes scoring functions highly dependent on the interpretation of weak van der Waals interactions and steric clashes between the ligand and enzyme surface. The docking score could also be dependent on the state of the CYP protein in the catalytic cycle, i.e. whether CpdI or the resting state is being considered.

Also, it is known that many CYP isoforms, particularly 3A4, exhibit high levels of flexibility[41] and can accommodate ligands of a wide range of shapes and sizes. It would thus be desirable to incorporate protein flexibility to model induced fit effects or use an ensemble based approach[42] to sample protein conformational space as applied in a 3A4 study.[43]

Examples of methods that feature docking include MLite,[44] which makes use of AutoDock[45] to orient ligands in the binding site, and IDSite,[46] a combined docking and reactivity approach for the 2D6 isoform using Schrödinger software: GLIDE[47] is used to orient the ligand in the CYP binding site; protein side chain and ligand refinements are carried out using Protein Local Optimization Program (PLOP);[48] and Jaguar[49] is used to estimate the intrinsic chemical reactivity. IDSite has a constraint that forces protonated amines to bind at a pharmacophore based position, hence it is unsuitable for predicting N-dealkylations.

However, CYP metabolism depends not only on noncovalent ligand binding within the protein cavity but also on ligand orientation relative to the reactive ferryl oxygen in CpdI allowing the formation of the transition state at the putative SoM. To address this requirement the docking of transition state structures has been reported,[50] with this concept developed further in IMPACTS[51] where a linear combination of the noncovalent and covalent bound substrate is taken to model accessibility of potential SoM. This was achieved using the FITTED docking program[52,53] applying the ACE (Asymmetric Catalyst Evaluation) transition state force field[54,55] and, when coupled with reactivity considerations in the form of a fragment based activation energy database, competitive SoM prediction performance was reported.

■ THEORY AND IMPLEMENTATION

The approach documented here adopts the format of a reactivity measure coupled with an accessibility measure.

The reactivity measure is based on a molecular orbital analysis obtained from a ground state DFT geometry optimization for the entire ligand, rather than a fragment based approach. The molecular orbital energy levels and molecular orbital basis function coefficients have been combined to give a novel implementation of the average local ionization energy enabling the reactivity of different sites in the molecule to evaluated.

Accessibility is considered via a docking approach that involves covalently binding the ligand at each potential SoM to the reactive heme center and assessing the success from the docking score. The docking score is enhanced by the insertion of constraints to simulate the electrostatic attraction of important charged residues in the CYP protein environment to oppositely charged ligand moieties. Each part of the process is discussed in more detail below.

Accessibility. In order to run docking experiments and obtain docking scores for the various ligand poses within the binding site, it is necessary to obtain a representative structure of the CYP. Molecular dynamics simulations were performed for each of the CYPs under investigation, and the trajectories were analyzed using cluster analysis to select consensus frames for ligand docking.

Molecular Dynamics of CYP. Crystal structures have been solved for the three CYP isoforms under investigation in this study, and structures were obtained from the Protein Data Bank (www.pdb.org).[56] The crystal structures used were 1TQN[57] for 3A4, 2F9Q[58] for 2D6, and 1R9O[59] for 2C9. All simulations have made use of explicit water followed by a standard molecular dynamics (MD) protocol (see the Experimental Section).

Cluster analysis was performed on the resulting MD trajectories using a GROMOS method[60] within GROMACS[61] which generated a main cluster representative of over 70% of the 100 ns trajectory for all isoforms. The representative structure generated from the top cluster was inspected for suitability, and for 3A4 it was noted that arginine residue Arg212 was located only *ca.* 5 Å above the CpdI ferryl oxygen and likely to have a strong influence on docking. The separation of the CpdI ferryl oxygen and the charged nitrogen in Arg212 over the course of the 100 ns trajectory is shown in Figure 1. It can be seen that Arg212 moves to a separation of *ca.* 12 Å for the interval 60−70 ns, and therefore a frame belonging to the 65th ns of the trajectory was also selected to assess the impact of this relatively mobile arginine residue. Close interactions of this nature were not observed in the trajectories for the 2D6 and 2C9 isoforms, hence only the representative structures from cluster analysis were used for docking.

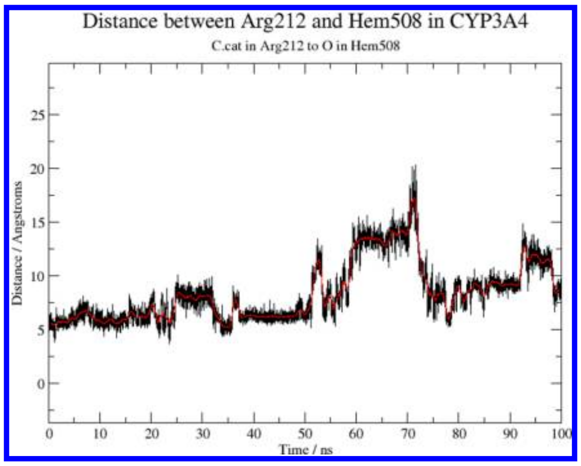


Figure 1. Separation of Arg212 and Hem508 from the 100 ns MD trajectory of 3A4 in the CpdI state.

Docking. For 2D6 and 2C9 the most representative structure from the principal cluster of structures explored during the MD simulations has been used, while for 3A4, two structures, representative of different side-chain conformations of Arg212, have been taken into consideration for further analysis. Using these structures, docking analysis was performed for each ligand in the data sets using GOLD.[62] In order for a reaction to occur the ligand must not only bind within the CYP active site but must also orient with the SoM in proximity to the CpdI ferryl oxygen. The ease with which different potential SoM in a molecule can approach the CpdI ferryl oxygen will be key to determining the likelihood of metabolic transformations occurring. Steric clashes, electrostatic interactions, van der Waals interactions, and hydrogen bond formation between ligand and protein will all contribute to the success or otherwise of the potential SoM approaching the CYP, all of which can be modeled using a docking algorithm.

In order to constrain a docking simulation to model ligand orientation at the reaction site rather than just ligand placement within the cavity it was decided to use a tethered docking approach. This was achieved using the covalent docking functionality within GOLD by inserting an oxygen into the ligand at the potential SoM to act as the link atom to the CpdI ferryl oxygen (see the Experimental Section for further details). Tethering has the combined effect of reducing the degrees of freedom for the docking experiment and focusing the search on conformational space where the ligand is oriented in the desired way for the site under investigation to be metabolized.

Docking was performed considering residues within a radius of 12 Å and 20 Å of the reactive heme center to assess the influence of protein radius on SoM prediction performance. The length of the tether was also varied to model different points in the approach of the ligand to the protein although a tether length of 2.0 Å was used for the comparison of scoring functions and parameter sets. The scoring functions available within GOLD are Goldscore, Chemscore, ASP (Astex Statistical Potential), and PLP (Piecewise Linear Potential). For Goldscore and Chemscore some CYP specific scoring parameters have been published[63] which were also assessed for performance.

The scoring functions documented above do not explicitly consider electrostatic interactions as such although a significant proportion of the ligand data sets are charged, and electrostatic interactions may help to orient and stabilize the ligands in the active site. Residues Glu216 and Asp301 have been documented as being important to orientate ligands containing positively charged moieties in the 2D6 isoform,[64] and an improvement was seen in SMARTCyp based on distance from a protonated amine to a potential reaction site. Similarly, Arg108 has been documented as performing a similar role for the 2C9 isoform[65] interacting with ligand carboxylic acid groups, and improvements to SMARTCyp predictions were again reported. Manual inspection of the MD frames selected for docking confirmed that these residues were indeed likely to be important for docking since they are all located within 15 Å of the CpdI ferryl oxygen and border the cavity. Further inspection of the 2C9 isoform resulted in the additional selection of charged residues Glu300 and Arg97, while for the 3A4 isoform charged residues Asp301 and Glu216 were chosen.

The electrostatic attraction between these important charged residues and oppositely charged ligand moieties was modeled by using the spherical constraint feature within GOLD to define a series of spherical constraints around the charged residues at radii calculated to model the Coulombic ($1/(r^2)$) relationship (see the Experimental Section for further details).

Reactivity. Oxidation reactions catalyzed by CYPs proceed either via an initial HAT mechanism for hydrocarbon hydroxylation; via bond-breaking and bond-making pathways for alkene, alkyne, and aromatic carbon oxidations; or via direct bond-formation with nonbonding electrons for sulfur oxidations. The latter two pathways involve the transfer of electrons from bonding or nonbonding regions in the ligand to form new bonds with the CpdI ferryl oxygen, and these will be defined as single electron transfer (SET) pathways. The SoM prediction approach documented below combines the concept of hydrogen bond order, to assess the ease of HAT, with a novel implementation of average local ionization energy, to assess the ease of SET from different regions in the molecule.

The average local ionization measure is derived from an energy-weighted density matrix,[67] D^w, defined over valence molecular orbitals as

$$D^w_{\alpha\beta} = \sum_{i}^{N_{valenceMO}} (n_i c_{\alpha i} c_{\beta i} \sqrt{|\varepsilon_i|}) \tag{9}$$

where i indexes the valence molecular orbitals, $c_{\alpha i}$ represents the coefficient of basis function α in molecular orbital i, ε_i represents the energy of molecular orbital i, and a square root is used to ensure that eq 10 returns a reactivity score in units of energy. The number of valence molecular orbitals is obtained by calculating the number of valence electrons contributed by the constituent atoms and dividing by 2 and are identified by counting back from the *HOMO*.

The product of D^w and the overlap matrix S can be used as the data source for the SET reactivity measure, where, by interrogating different parts of the product matrix D^wS, the ease of removing electrons from different regions of the molecule can be assessed. This approach is a novel implementation of the average local ionization energy defined in eq 6 where instead of calculating a value at a particular point r on an isosurface of electron density, a region of the molecule is defined in terms of basis functions. For instance, alkene oxidation proceeds via the breaking of a π bond and can be modeled by the two center terms formed from basis functions on the sp^2 C atoms involved in the double bond; similarly alkyne oxidation proceeds via the breaking of a π bond and can be modeled by the two center terms formed from basis functions on the sp^1 C atoms involved in the triple bond; and oxidation of S atoms occurs without the requirement to break bonds and so only one center terms are considered. Aromatic C oxidation proceeds via initial activation of an aromatic bond followed by an intramolecular HAT, thus involving both SET and HAT pathways. It was found that modeling HAT was a better indicator of SoM for aromatic C, and so this approach was applied to aromatic C in this study.

Thus an average local ionization energy reactivity score, R_{AB}, relating to SET from the bonding region between atom A and atom B can be defined by considering all two center terms formed by basis functions α on atom A and β on atom B

$$R_{AB} = \frac{\sum_{\alpha \in A} \sum_{\beta \in B} (D^wS)_{\alpha\beta} (D^wS)_{\beta\alpha}}{\sum_{\alpha \in A} \sum_{\beta \in B} (DS)_{\alpha\beta} (DS)_{\beta\alpha}} \tag{10}$$

where D and D^w are calculated over the valence molecular orbitals.

For HAT it is necessary to consider all two center terms because all bonding regions must be broken to abstract the hydrogen, thus a reactivity score for HAT, R_{H_A}, can be calculated as

$$R_{H_A} = \frac{\sum_{\alpha \in H_A} \sum_{\beta \notin H_A} (D^wS)_{\alpha\beta} (D^wS)_{\beta\alpha}}{\sum_{\alpha \in H_A} \sum_{\beta \notin H_A} (DS)_{\alpha\beta} (DS)_{\beta\alpha}} \tag{11}$$

The approach to calculate a reactivity score for sites that follow a SET pathway is summarized in the pseudocode in Algorithm 1, where different parts of the product matrix D^wS can be interrogated depending on the nature of the site.

Algorithm 1 Pseudo Code for Calculating the Reactivity Score for SET from Atom A.

```
Set of basis functions on atom $A = F_A$
Set of basis functions to overlap with $F_A = F_B$

if atom $A$ is sp2 carbon or sp1 carbon then
   $F_B$ = basis functions on adjacent partner in $\pi$ bond
end if
if atom $A$ is sulfur or boron then
   $F_B = F_A$
end if

Populate the lists reactivityScoreList, bondOrderList, where $D$ and $D^w$ are calculated over
valence orbitals

$j = 1$
for $\alpha \in F_A$ do
   for $\beta \in F_B$ do
      $reactivityScoreList_j = (D^wS)_{\alpha\beta} (D^wS)_{\beta\alpha}$
      $bondOrderList_j = (DS)_{\alpha\beta} (DS)_{\beta\alpha}$
      $j = j + 1$
   end for
end for

Calculate reactivity score for one electron equivalent

$bondOrderTotal$ = sum $bondOrderList$
$reactivityScoreTotal$ = sum $reactivityScoreList$
$reactivityScore = \frac{reactivityScoreTotal}{bondOrderTotal}$
```

It was found that for SoM involving HAT, hydrogen bond orders calculated using eq 7 gave stronger predictive performance compared to reactivity scores calculated using eq 11, thus it is desirable to use hydrogen bond order to model HAT pathways and the reactivity score for SET pathways. In order to combine these two measures so that all potential SoM in a molecule can be compared and ranked it is necessary to scale one of the measures relative to the other. This was achieved for each molecule in turn by calculating an average hydrogen reactivity score, $R_H^{average}$, and an average hydrogen bond order, $BO_H^{average}$, for all hydrogens in the molecule attached to a potential SoM. The reactivity scores for sites involving a SET pathway, R_{SET}, can then be scaled using the following equation to give a scaled reactivity score, R_{SET}^{scaled}:

$$R_{SET}^{scaled} = \frac{R_{SET} \times BO_H^{average}}{R_H^{average}} \tag{12}$$

Combining Reactivity and Docking. Low reactivity scores but high docking scores correspond to more likely SoM, hence a ratio or subtractive method would be suitable to combine these two measures. In this approach a ratio method has been presented with the *Reactivity* measure as the numerator and the *Docking* measure as the denominator (although a suitably parametrized subtractive method was found to give virtually identical results)

$$Score_A = \frac{Reactivity_A}{(Docking_A)^P} \tag{13}$$

where P is a parameter to balance the contributions from *Reactivity* and *Docking* with values of 0.5, 0.7, and 0.6 found to be suitable for isoforms 3A4, 2D6, and 2C9, respectively.

Data Sets. The combined docking and reactivity approach outlined above has been applied to the publicly available 3A4, 2D6, and 2C9 data sets.[51] These data sets originate from those initially released in the Supporting Information of the RS-Predictor paper[26] but further curated with reference to the primary literature to identify and eliminate conflicting information. Open Babel v2.3.1[68] was used to generate the protonation state of each ligand at pH 7.4.

RESULTS AND DISCUSSION

Potential Sites of Metabolism. The results presented in this section document the percentage of the data sets where a SoM is identified in the top 2 or top 3 predictions (top 1 predictions are given in the Supporting Information). Initially a

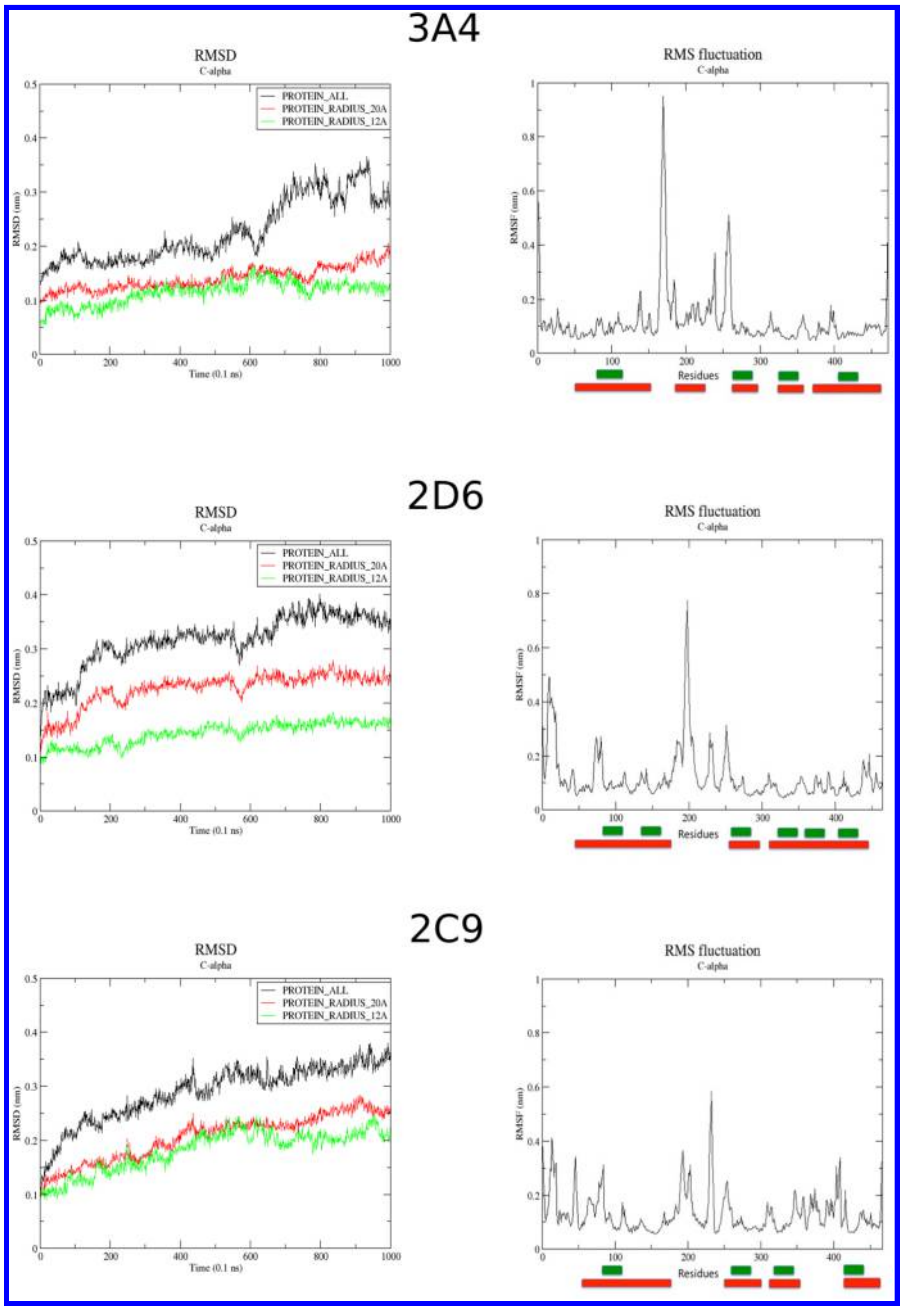


Figure 2. RMSD (left) and RMSF (right) plots calculated from 100 ns of CYP production MD using the positions of the carbonyl C, C-α, and N atoms (backbone) of the protein (black). Additional plots were calculated using a subset of these atoms at a radius 12 Å (green) and 20 Å (red) from the iron atom within the heme moiety. Green and red bars on the RMSF plots indicate residues within the cutoff radius 12 Å (green) and 20 Å (red).

script was run to identify all the potential SoM in a ligand using the mol2[69] atom type, with those sites deemed not to be potential SoM removed from predictions. A summary of the rules used to identify potential SoM is given below:

1. Carbon atoms that are sp^3 hybridized (SYBYL atom type C.3) and aromatic carbons (SYBYL atom type C.ar) must have a hydrogen attached to be a potential SoM

2. Carbon atoms that are sp^2 hybridized can be a SoM provided that they are not part of a ketone or carboxylic acid group

3. Carbon atoms that are sp^1 hybridized can only be a SoM if they are part of a terminal alkyne group

4. Sulfur can be a SoM so long as it is not fully oxidized (i.e., not SYBYL atom type S.O2)

5. Boron, although not common in the data sets, can be a SoM

6. Nitrogen atoms are excluded from predictions (i.e., direct oxidation of nitrogen is ignored (there are very few examples in the training set) but N-dealkylations can still be predicted since the carbon on the leaving alkyl group is still included in the predictions)

Additionally, steps are taken to identify atoms in equivalent sites and ensure that such sites are only included once in predictions. This is achieved by generating circular fingerprints[70,71] for each atom site, based on the occurrence frequency of SYBYL atom types at different topological distances spanning the entire ligand (i.e., a circular fingerprint is defined for each atom site based on the topological distance to all other atoms in the molecule), with equivalent atom sites defined as those with identical circular fingerprints. The best reactivity score for the equivalent sites is then used in the final predictions.

The treatment of nitrogen is problematic since it occurs frequently in the data sets and the reactivity score usually deems SET pathways to be favorable at these sites leading to overprediction. Many occurrences of nitrogen in the data sets are as alkyl substituted amines where the reaction proceeds via N-dealkylation and the current protocol allows these to be predicted by assigning the SoM to the carbon in the departing alkyl group that was attached to the amine. A possible explanation to the overprediction of nitrogen has recently been published.[72]

The drawback to excluding nitrogens is that there are instances, albeit only a few, of direct nitrogen oxidation in the data sets that will not be predicted using this protocol, although since these are limited in number better predictive performance is obtained by excluding nitrogens, and we therefore define the applicability domain of the method at present to exclude nitrogen.

Molecular Dynamics. Analysis of the 100 ns trajectories in terms of RMSD and RMSF is given in Figure 2. All three isoforms show reasonable stability, and as expected the RMSD decreases as the radius within which protein residues are considered decreases. Inspection of the RMSF charts shows that the more mobile residues are confined to regions on the periphery of the protein at a radius of greater than 20 Å.

The protocol here only makes use of a single frame from the MD trajectories since significant conformational transitions within the cavity were not observed. However, it must be emphasized that the CYP proteins are highly flexible[73] in the presence of ligands, particularly the 3A4 isoform, which is illustrated by the wide range of ligand shapes and sizes that can be accommodated. Therefore the selection of only one frame from these *apo* trajectories can be justified (through performance criteria) although it should be emphasized that this is a significant simplification when considering the wide range of conformations the protein could adopt in the presence of ligands.

When only using one MD frame for docking care must be taken to ensure that the structure is representative of the overall trajectory. However, consideration must also be given to the flexibility of protein side chains, as illustrated by the relatively mobile Arg212 group close to the heme center in 3A4, see Figure 1. Using the representative structure from the top cluster for 3A4 gave a placement of Arg212 *ca.* 5 Å from the ferryl oxygen in Hem508, impairing SoM predictive performance when compared to the frame with a 12 Å separation of these residues, see Table 1. It is likely that with the Arg212 at only 5 Å from Hem508 the reactive center is sterically crowded and some potential SoM will be discounted. Therefore the frame with a separation of 12 Å was selected in the further docking simulations.

It is possible that the close interaction of Arg212 with the ferryl oxygen in Hem508 for 3A4 is an artifact of running dynamics in the CpdI state allowing electrostatic attraction between the ferryl oxygen and positively charged Arg212 to predominate, whereas in reality the CpdI state exists only fleetingly[74] before reaction with a substrate is initiated.

Table 1. Docking Results for 3A4 with Arg212 at Radius 5 Å and 12 Å from Hem508[a]

				3A4		
Arg212 - Hem508 separation (Å)	scoring function	parameter set	tether length (Å)	top 5%	top 3%	top 2%
5	PLP	default	1.5	81	64	55
12	PLP	default	1.5	84	69	59

[a] Results show the % of the 3A4 data set where a SoM is identified in the top 5, 3, and 2 predictions, respectively.

Docking. Docking results using protein radii of 12 Å and 20 Å from the heme center are shown in Table 2, and it can be seen that generally the larger radius gives stronger SoM predictive performance. This is to be expected since more protein residues are being considered for docking and subsequent scoring, thus a protein binding site radius of 20 Å was used for the remaining docking simulations.

The docking results obtained using different scoring functions and parameter sets are shown in Table 3. It is apparent that the PLP scoring function is superior to ASP, Chemscore, and Goldscore for the purpose of identifying SoM using the tethered docking approach. The Goldscore and Chemscore CYP specific parameter sets[63] did not give SoM prediction performance benefits, although it is important to note that these parameter sets were not designed specifically for tethered docking or the CYP in CpdI state.

The tethered docking approach shows a clear performance improvement over the results obtained without tethering with the best SoM prediction performance obtained with the lowest tether length of 1.5 Å, although since the C−O single bond length is *ca.* 1.43 Å reducing the tether length lower than this value cannot be justified.

The default PLP scoring function does not explicitly consider electrostatic interactions, and the use of spherical constraints to model electrostatic attractions between charged moieties in the protein and ligand gave improved performance for the 2D6 and 2C9 isoforms in terms of the top 2 and top 3 classification. The 3A4 isoform showed a slight deterioration on a top 2 basis although incorporating electrostatic interactions does give improvements after combining with the reactivity scores, see Table 5.

Reactivity. The results in Table 4 show that using the scaled average local ionization energy from an energy-weighted density matrix for potential SoM involving SET gives improved performance over using bond orders from an unweighted density matrix for 3A4 and 2D6. However, after combining with docking results improved performance is obtained across all isoforms, see Table 5. The bond order metric calculated from an unweighted density matrix, see eq 7, is unable to differentiate between sites where the reaction pathway occurs via SET rather than HAT, thus the benefit of using the energy-weighted density matrix to enable the reactivity of sites involving SET to be ranked becomes apparent.

Correlation of Reactivity Scores with SMARTCyp. A comparison of the reactivity scores generated using this approach for the 3A4 data set were compared to those generated from SMARTCyp. Only potential SoMs as determined by this approach were used, and atom sites where SMARTCyp generates a score of 999 were ignored. Using RStudio[75] to analyze the correlation between the two approaches, a Spearman rank correlation of 0.495 was obtained indicating a reasonable correlation between the two approaches

Table 2. Docking Results for 3A4, 2D6, and 2C9 with a Protein Radius of 12 Å and 20 Å[a]

				3A4		2D6		2C9	
protein radius (Å)	scoring function	parameter set	tether length (Å)	top 3%	top 2%	top 3%	top 2%	top 3%	top 2%
12	PLP	default	1.5	67	55	82	67	78	69
20	PLP	default	1.5	69	59	80	70	78	70

[a]Results show the % of the CYP data sets where a SoM is identified in the top 3 and 2 predictions, respectively.

Table 3. Docking Results for 3A4, 2D6, and 2C9 Using Tethered and Untethered Approaches with a Protein Radius of 20 Å[a]

			3A4		2D6		2C9	
scoring function	parameter set	tether length (Å)	top 3%	top 2%	top 3%	top 2%	top 3%	top 2%
Chemscore	PDB	2.0	34	23	56	41	63	54
Chemscore	CSD	2.0	32	23	58	39	61	55
Chemscore	default	2.0	38	27	61	41	65	57
Goldscore	PDB	2.0	53	38	70	57	70	60
Goldscore	CSD	2.0	51	36	69	55	71	61
Goldscore	default	2.0	52	37	71	55	71	62
ASP	default	2.0	57	42	66	53	71	64
PLP	default	no tether	48	37	72	54	68	60
PLP	default	2.0	64	51	80	69	73	65
PLP	**default**	**1.5**	**69**	**59**	**80**	**70**	**78**	**70**
PLP + constraints	default	no tether	48	37	69	54	71	60
PLP + constraints	**default**	**1.5**	**69**	**57**	**83**	**73**	**80**	**71**
PLP + constraints	default	2.0	62	48	80	69	78	68
PLP + constraints	default	2.5	54	39	73	60	74	64
PLP + constraints	default	3.0	47	32	69	52	77	66

[a]Results show the % of the CYP data sets where a SoM is identified in the top 3 and 2 predictions, respectively.

Table 4. Reactivity Results for 3A4, 2D6, and 2C9 Representing the % of the CYP Data Sets Where a SoM Is Identified in the Top 3 and 2 Predictions, Respectively

		3A4		2D6		2C9	
SET method	HAT method	top 3%	top 2%	top 3%	top 2%	top 3%	top 2%
bond order	bond order	72	63	61	48	80	64
scaled average local ionization energy	bond order	78	67	66	51	80	60

Table 5. Overall Results Combining Reactivity and Docking Representing the % of the CYP Data Sets Where a SoM Is Identified in the Top 3 and 2 Predictions, Respectively

		3A4[a]		2D6[b]		2C9[c]	
density matrix	docking method	top 3%	top 2%	top 3%	top 2%	top 3%	top 2%
unweighted	PLP + constraints	77	68	80	75	80	76
unweighted	PLP	75	66	78	71	80	74
energy-weighted	PLP + constraints	83	75	85	80	85	78
energy-weighted	PLP	81	71	83	75	85	76

[a]Tether = 1.5, P = 0.5. [b]Tether = 1.5, P = 0.7. [c]Tether = 1.5, P = 0.5.

although further work would be required to investigate the correlation more fully.

Combining Reactivity and Docking. The SoM prediction performance on combining the docking and reactivity approaches is shown in Table 5, yielding results that are competitive with those obtained from other approaches.[51,76] For all isoforms the docking scenario that gave the best results when combined with reactivity used the shortest tether length of 1.5 Å and included the modeling of electrostatic interactions.

■ CONCLUSIONS

A SoM prediction methodology has been created using a combined reactivity and docking approach that is competitive with other tools[51,76] and illustrates that strong results can be obtained by making use of standard functionality from common QM and docking applications. A SoM is identified in the top 2 predictions for 75%, 80%, and 78% of the 3A4, 2D6, and 2C9 data sets, respectively.

A novel implementation of the average local ionization energy has been presented that can be applied to CYP oxidation pathways involving SET, which can subsequently be combined with hydrogen bond order information to provide a ranking of all potential SoM in the ligand. This methodology has the advantage of being derived from a typical molecular orbital analysis of a ground state ligand that can be generated using standard functionality within most common QM packages, albeit at the not insignificant computational cost of performing a geometry optimization using DFT. Each potential SoM is considered in the context of the whole ligand as opposed to discretizing into fragments and matching to a predefined database where each fragment is considered to be a uniform entity. Potentially, there may be scope for applying this method to other electrophilic reactions where it is important to locate the site of attack and could possibly be extended to nucleophilic reactions by considering unoccupied, instead of occupied, orbitals.

This study is one of the first to make use of recently published CpdI parameters[77] in an MD simulation and shows that, in the absence of a ligand, the CYP protein structure is largely conserved with variation restricted to loop regions on the periphery of the protein. This supports the success that was obtained with docking into only one representative structure from the MD trajectory. However, this approach does neglect the flexibility of the CYPs when exposed to ligands of a variety of shapes and sizes, and further enhancements to this work could include ensemble docking into a selection of different structures generated from MD with a variety of different template ligands.

The results from this study support the contention that charged residues within the cavity of CYPs are important in determining the orientation and thus SoM of ligands bearing a charge. It has been shown that electrostatic interactions can be modeled using a series of spherical constraints within GOLD to model the Coulombic electrostatic interaction between charged moieties. Further enhancements could include extending the modeling of electrostatics to include functional groups bearing a partial charge, using a docking tool with built-in electrostatic functionality, or running MD simulations of the ligand in the CYP active site with an electrostatically aware force field to sample the amount of time potential SoM spend in proximity to the CpdI ferryl oxygen.

The docking results illustrate the benefits of using tethered docking to model the accessibility of the CpdI ferryl oxygen to different sites in the ligand, and a clear boost in SoM predictive performance is obtained when compared to untethered docking. Tethering forces the ligand into close proximity to the reactive heme center at the correct orientation to facilitate metabolism allowing the scoring function to evaluate any interactions favorable or otherwise, with a short tether length of 1.5 Å being the most effective for this purpose.

PLP was found to be the most successful scoring function used in this work, being notably better than Goldscore, Chemscore, and ASP, while the CYP specific parameters[63] available within GOLD for Goldscore and Chemscore did not give a performance boost. In summary, it has been shown that with only a limited amount of intervention in the form of defining spherical constraints to model electrostatic interactions, the GOLD program, in conjunction with reactivity descriptors derived from an energy-weighted density matrix defined over valence molecular orbitals and an overlap matrix, can be successfully applied to the problem of determining SoM for CYP xenobiotic metabolism.

■ EXPERIMENTAL SECTION

MD Simulations. All simulations were carried out using AMBER11[78] (with patches up to 19 applied) and AMBERTOOLS version 1.5 suite of programs. Production simulations were performed using the Darwin Supercomputer of the University of Cambridge High Performance Computing Service (http://www.hpc.cam.ac.uk/), provided by Dell Inc. using Strategic Research Infrastructure Funding from the Higher Education Funding Council for England and funding from the Science and Technology Facilities Council. Specifically, the Darwin Tesla GPU subcluster was used, with each simulation being run over four Tesla S1070 GPUs cards, with CUDA API version 3.2. The compute nodes were running Scientific Linux release 6.3 and IFORT version 12.1 in conjunction with the OS's default gcc compiler to compile the parallel version of PMEMD, with MVAPICH2 1.6 for the MPI layer.

Common System Preparation Protocol. The webservice PDB2PQR, version 1.8,[79] was used to calculate the protonation state of titratable residues at physiological pH (7.4) and assign missing hydrogen atoms within the prepared PDB files. The AMBER force field and naming scheme was used, and pK_a calculations were performed by PROPKA at pH 7.4. The resulting PQR file was loaded into tLEaP using the FF99SB[80] force field parameters, and parameters for the heme moiety and the covalently bonded cysteine, in the CpdI state, were taken from Shahrokh et al.[77] Periodic boundaries (truncated octahedron) were applied, and the system was solvated with TIP3P[81] water with a minimal distance of 9 Å from the protein unit. Counter ions were added respectively to ensure system neutrality. This yielded a system with ∼200k atoms.

3A4 Specific Preparation. The 1TQN[57] crystal structure was used, and the location of four missing residues 282−285 (LYS, GLU, THR, GLU), in a loop region, were determined using the following protocol:

1. The coordinates of α carbon atoms of the four missing residues were manually added in a line between the ends of residues 281 and 286 to the PDB.
2. The PDB was loaded into xLEaP, which automatically reconstructed the rest of the residues from their backbone atoms.
3. xLEaP's relax function was used on these four residues.
4. The system was solvated with TIP3P water, and a corresponding prmtop/inpcrd file pair was generated.
5. 1,000 steps of minimization using the XMIN method were carried out.
6. VMD 1.9's RMSD align tool[82] was used to align backbone atoms of this structure with the original 1TQN PDB.
7. The coordinates of the C, CA, and N atoms of the four missing residues were extracted.

2D6 Specific Preparation. The 2F9Q[58] crystal structure was used; specifically chain A.

2C9 Specific Preparation. The 1R9O[59] crystal structure was used, and the loopmodel class of MODELER 9.8[83] was used to determine a model for the coordinates of the missing residues 214−220 (GLN, ILE, CYS, ASN, ASN, PHE, SER).

Common MD Protocol. Initial minimization of the system over a maximum of 1,000 cycles was performed using a steepest descent method, to remove any close contacts from initial system building. A nonbonded cutoff of 8 Å was used for both the electrostatic and VdW terms; particle mesh Ewald (PME)[84] was used for the treatment of long-range electrostatic terms with a spline interpolation order of 4 and a tolerance of 1×10^{-5}.

50 ps of NVT thermalization dynamics using a Langevin thermostat[85] with a collision frequency of 2 ps^{-1} was carried out. Initial velocities were assigned from a Maxwellian distribution at 298 K, and the same temperature was used as the reference temperature for the system to be kept at. All trajectories were stored using the NetCDF format. All MD from this point forward used a time step of 2 fs, and the SHAKE[86] algorithm was applied to all bonds containing hydrogen with a geometric tolerance of 1×10^{-5} Å during the coordinate reset. 70 ps of NPT dynamics were carried out to obtain the correct density. Periodic boundary conditions with constant pressure were used with a reference pressure of 1 bar and a 2 ps pressure relaxation time.

100 ns of NPT dynamics were carried out; constant temperature was maintained using a weakly coupled Berendsen thermostat[87] with a reference temperature of 298 K. During production runs, the translational and rotational center-of-mass motion of the entire system was removed every 1,000 MD steps, and molecules were wrapped to the original unit cell. The production was run in blocks of 2.5 ns, restarting from the

previous block using an ASCII restart file. This performed at ~3 ns/day.

Cluster Protocol. Post-trajectory analysis was carried out using GROMACS.[61] Carbon alpha backbone RMSD calculations were carried out using the first frame of the 100 ns production run as a reference. Cluster analysis and selection of representative frames for each trajectory were performed using the GROMOS method[60] using a cutoff of 0.2.

Docking. Docking was performed using GOLD v5.1 run as single processor jobs on a cluster of 16 dual-processor six core Intel Xeon X5650 based servers, each machine with twelve cores and 24GB of RAM. BASH scripts were written to implement the docking protocol as described below.

The docking protocol proceeds for C.3, C.ar, and C.2 aldehyde carbons by replacing each attached hydrogen in turn with an oxygen and overlaying this inserted oxygen atom with the CpdI ferryl oxygen, i.e. labeling the inserted oxygen as a link atom via the covalent docking option within GOLD to introduce a tether between ligand and CYP. However, some potential SoM, such as direct oxidation at a sulfur atom or alkene epoxidation, do not have a hydrogen attached, and so the docking script inserts an oxygen at the correction orientation. For alkenes and alkynes a bridging oxygen is inserted with attack from both faces for the planar alkene system. For sulfur the situation is more complicated since the number of neighbors can vary. For a terminal sulfur with one bond the oxygen is inserted parallel to the original sulfur bond pointing away from the molecule; for a sulfur with two neighbors an oxygen is inserted in the same plane as the current neighbors bisecting the largest angle between them; and for a sulfur with 3 neighbors the sulfur sits at the top of a trigonal pyramid, and an oxygen is inserted perpendicular to the plane of the base of the trigonal pyramid from above and below separately.

The electrostatic interactions between important charged residues in the CYP and oppositely charged moieties in the ligand are modeled by defining a series of 10 spheres around each protein charged residue, centered on the carboxylic acid oxygens (SYBYL[88] atom type O.co2) for aspartic acid (Asp) and glutamic acid (Glu) and the carbocation (SYBYL atom type C.cat) for arginine (Arg), with the shortest radius defined as 4 Å. Location of oppositely charged ligand moieties within each sphere contributes a boost to the docking score of 0.5, with the radii of the spheres calculated so that the cumulative boost to the docking score reflects the Coulombic $(1/(r^2))$ relationship, see Table 6.

Table 6. Spherical Constraint Radii To Model Electrostatic Attraction

radius r (Å)	$(1/(r^2)) \times 10^3$	incremental score boost	cumulative score boost
4.000	62.50	0.5	5.0
4.216	56.25	0.5	4.5
4.472	50.00	0.5	4.0
4.781	43.75	0.5	3.5
5.164	37.50	0.5	3.0
5.657	31.25	0.5	2.5
6.325	25.00	0.5	2.0
7.303	18.75	0.5	1.5
8.944	12.50	0.5	1.0
12.649	6.25	0.5	0.5

For each ligand, constraints were defined for protonated amines (SYBYL atom type N.4) and carbocations (C.cat) for attraction to Asp and Glu, and, correspondingly, constraints were defined for ligand carboxylic acid oxygens (O.co2) for attraction to Arg. In this way each electrostatic interaction between protein residue and oppositely charged ligand moiety is defined by the interaction of one positively charged entity with two negatively charged entities yielding a maximum boost to the docking score of 10 if both interactions are within 4 Å.

Prior to combination with the reactivity scores the GOLD docking results for each ligand have a score of 300 added to remove negative scores and were then "normalized" by dividing through by the maximum score for that ligand. In this way a docking score is obtained in the range zero to one in a suitable format for combination with the reactivity scores. For sites where more than one docking score is available, e.g. multiple hydrogens attached to C.3 or epoxidation from both faces of a planar alkene system, an average docking score is taken.

For the untethered docking comparatives each heavy atom in the ligand was constrained to be within 1.5−2.5 Å of the CpdI ferryl oxygen with a penalty of 100.0 and the docking score again used for ranking purposes.

QM Calculations. The NWChem v6.0[89] software package has been used throughout running in parallel over 12 cores on a cluster of 16 dual-processor six core Intel Xeon X5650 based servers, each machine with twelve cores and 24GB of RAM.

A geometry optimization was carried out using B3LYP/STO-3G[90] converging to the NWChem-defined loose criteria, followed by a further optimization using the B3LYP/6-31G** (6-311G** for bromine and iodine) again using the NWChem-defined loose criteria. On the optimized structure a single point calculation using B3LYP/6-31G** (6-311G** for bromine and iodine) was used to generate the final molecular orbital analysis and the Mulliken population analysis.

The default functionality within NWChem only outputs the molecular orbital basis function coefficients in excess of 0.15 and limits the output to a maximum of ten. In order to generate the density matrix a full molecular orbital analysis is required, and the source code was amended in src/ddscf/movecs_pr_anal.F to ensure all molecular orbital basis function coefficients were output.

The overlap matrix was generated from the total Mulliken overlap population analysis (obtained by using the keyword Mulliken within NWChem) and the density matrix by dividing each element of the Mulliken overlap population matrix by the corresponding element in the density matrix.

For sites with multiple hydrogens attached the lowest reactivity score for the attached hydrogens was used as the measure for HAT.

SoM Prediction Tool. The data sets in the form of mol2[69] files and the output files from NWChem and GOLD were analyzed in an internally written program using Java J2SE-1.5[91] and Eclipse v3.5.2[92] in the manner described in this paper to produce the SoM predictions.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Top 1 predictions. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: rcg28@cam.ac.uk.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors would like to thank Unilever for funding, members of the Unilever Research & Development team at Vlaardingen for useful discussions on the development of this work, and Andreas W. Götz for detailed review of the manuscript.

■ REFERENCES

(1) Guengerich, F. P. *AAPS J.* **2006**, *8*, E101−11.

(2) Lewis, D. F. V. *Pharmacogenomics* **2004**, *5*, 305−18.

(3) Shaik, S.; Cohen, S.; Wang, Y.; Chen, H.; Kumar, D.; Thiel, W. *Chem. Rev.* **2010**, *110*, 949−1017.

(4) Ogliaro, F.; Harris, N.; Cohen, S.; Filatov, M.; de Visser, S. P.; Shaik, S. *J. Am. Chem. Soc.* **2000**, *122*, 8977−8989.

(5) Shaik, S.; Devisser, S.; Ogkiaro, F.; Schwarz, H.; Schroder, D. *Curr. Opin. Chem. Biol.* **2002**, *6*, 556−567.

(6) Bathelt, C. M.; Ridder, L.; Mulholland, A. J.; Harvey, J. N. *J. Am. Chem. Soc.* **2003**, *125*, 15004−5.

(7) de Visser, S. P.; Shaik, S. *J. Am. Chem. Soc.* **2003**, *125*, 7413−24.

(8) De Montellano, P. R.; Kunze, K. L. *Arch. Biochem. Biophys.* **1981**, *209*, 710−712.

(9) Sharma, P. K.; De Visser, S. P.; Shaik, S. *J. Am. Chem. Soc.* **2003**, *125*, 8698−9.

(10) Rydberg, P.; Ryde, U.; Olsen, L. *J. Chem. Theory Comput.* **2008**, *4*, 1369−1377.

(11) Kirchmair, J.; Williamson, M. J.; Tyzack, J. D.; Tan, L.; Bond, P. J.; Bender, A.; Glen, R. C. *J. Chem. Inf. Model.* **2012**, *52*, 617−48.

(12) Kulkarni, S. A.; Zhu, J.; Blechinger, S. *Xenobiotica; the fate of foreign compounds in biological systems* **2005**, *35*, 955−73.

(13) Tarcsay, A.; Keseru, G. M. *Expert Opin. Drug Metab. Toxicol.* **2011**, *7*, 299−312.

(14) Ekins, S.; Andreyev, S.; Ryabov, A.; Kirillov, E.; Rakhmatulin, E. a.; Bugrim, A.; Nikolskaya, T. *Expert Opin. Drug Metab. Toxicol.* **2005**, *1*, 303−24.

(15) Vaz, R. J.; Zamora, I.; Li, Y.; Reiling, S.; Shen, J.; Cruciani, G. *Expert Opin. Drug Metab. Toxicol.* **2010**, *6*, 851−61.

(16) Fukui, K. *Angew. Chem., Int. Ed. Engl.* **1982**, *21*, 801−809.

(17) Parr, R. G.; Yang, W. *J. Am. Chem. Soc.* **1984**, *106*, 4049−4050.

(18) Yang, W.; Mortier, W. J. *J. Am. Chem. Soc.* **1986**, *108*, 5708−5711.

(19) Contreras, R. *Chem. Phys. Lett.* **1999**, *304*, 405−413.

(20) Fukui, K.; Yonezawa, T.; Nagata, C. *J. Chem. Phys.* **1957**, *27*, 1247.

(21) Sjoberg, P.; Murray, J. S.; Brinck, T.; Politzer, P. *Can. J. Chem.* **1990**, *68*, 1440−1443.

(22) Mayer, I. *Chem. Phys. Lett.* **1983**, *97*, 270−274.

(23) Afzelius, L.; Arnby, C. H.; Broo, A.; Carlsson, L.; Isaksson, C.; Jurva, U.; Kjellander, B.; Kolmodin, K.; Nilsson, K.; Raubacher, F.; Weidolf, L. *Drug Metab. Rev.* **2007**, *39*, 61−86.

(24) Hennemann, M.; Friedl, A.; Lobell, M.; Keldenich, J.; Hillisch, A.; Clark, T.; Göller, A. H. *ChemMedChem* **2009**, *4*, 657−69.

(25) Zaretzki, J.; Bergeron, C.; Rydberg, P.; Huang, T.-W.; Bennett, K. P.; Breneman, C. M. *J. Chem. Inf. Model.* **2011**, *51*, 1667−89.

(26) Zaretzki, J.; Rydberg, P.; Bergeron, C.; Bennett, K. P.; Olsen, L.; Breneman, C. M. *J. Chem. Inf. Model.* **2012**, *52*, 1637−59.

(27) Yin, H.; Anders, M. W.; Korzekwa, K. R.; Higgins, L.; Thummel, K. E.; Kharasch, E. D.; Jones, J. P. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 11076−11080.

(28) Jones, J. P. *Drug Metab. Dispos.* **2002**, *30*, 7−12.

(29) Olsen, L.; Rydberg, P.; Rod, T. H.; Ryde, U. *J. Med. Chem.* **2006**, *49*, 6489−6499.

(30) Rydberg, P.; Ryde, U.; Olsen, L. *J. Phys. Chem. A* **2008**, *112*, 13058−65.

(31) Rydberg, P.; Vasanthanathan, P.; Oostenbrink, C.; Olsen, L. *ChemMedChem* **2009**, *4*, 2070−9.

(32) Jung, J.; Kim, N. D.; Kim, S. Y.; Choi, I.; Cho, K.-H.; Oh, W. S.; Kim, D. N.; No, K. T. *J. Chem. Inf. Model.* **2008**, *48*, 1074−80.

(33) Hasegawa, K.; Koyama, M.; Funatsu, K. *Mol. Inf.* **2010**, *29*, 243−249.

(34) Kim, D. N.; Cho, K.-H.; Oh, W. S.; Lee, C. J.; Lee, S. K.; Jung, J.; No, K. T. *J. Chem. Inf. Model.* **2009**, *49*, 1643−54.

(35) Rydberg, P.; Gloriam, D. E.; Zaretzki, J.; Breneman, C.; Olsen, L. *ACS Med. Chem. Lett.* **2010**, *1*, 96−100.

(36) Parr, R. G.; Yang, W. *Annu. Rev. Phys. Chem.* **1995**, *46*, 701−728.

(37) Cruciani, G.; Carosati, E.; De Boeck, B.; Ethirajulu, K.; Mackie, C.; Howe, T.; Vianello, R. *J. Med. Chem.* **2005**, *48*, 6970−9.

(38) *StarDrop, version 5.0*; Optibrium: Cambridge, U.K., 2011.

(39) Sun, H.; Scott, D. O. *Chem. Biol. Drug Des.* **2010**, *75*, 3−17.

(40) Otyepka, M.; Skopalík, J.; Anzenbacher, P. *Biochim. Biophys. Acta* **2007**, *1770*, 376−389.

(41) Ekroos, M.; Sjögren, T. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 13682−7.

(42) Amaro, R. E.; Li, W. W. *Curr. Top. Med. chem.* **2010**, *10*, 3−13.

(43) Teixeira, V. H.; Ribeiro, V.; Martel, P. J. *Biochim. Biophys. Acta* **2010**, *1804*, 2036−45.

(44) Oh, W. S.; Kim, D. N.; Jung, J.; Cho, K.-H.; No, K. T. *J. Chem. Inf. Model.* **2008**, *48*, 591−601.

(45) Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. *J. Comput. Chem.* **2009**, *30*, 2785−91.

(46) Li, J.; Schneebeli, S. T.; Bylund, J.; Farid, R.; Friesner, R. A. *J. Chem. Theory Comput.* **2011**, *7*, 3829−3845.

(47) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. *J. Med. Chem.* **2004**, *47*, 1739−49.

(48) Zhu, K.; Pincus, D. L.; Zhao, S.; Friesner, R. A. *Proteins* **2006**, *65*, 438−52.

(49) *Jaguar, Suite 2012: version 7.9*; Schrödinger, LLC: New York, NY, 2012.

(50) Rydberg, P.; Hansen, S. M.; Kongsted, J.; Norrby, P.-O.; Olsen, L.; Ryde, U. *J. Chem. Theory Comput.* **2008**, *4*, 673−681.

(51) Campagna-Slater, V.; Pottel, J.; Therrien, E.; Cantin, L.-D.; Moitessier, N. *J. Chem. Inf. Model.* **2012**, *52*, 2471−83.

(52) Corbeil, C. R.; Englebienne, P.; Moitessier, N. *J. Chem. Inf. Model.* **2007**, *47*, 435−49.

(53) Corbeil, C. R.; Moitessier, N. *J. Chem. Inf. Model.* **2009**, *49*, 997−1009.

(54) Corbeil, C. R.; Thielges, S.; Schwartzentruber, J. A.; Moitessier, N. *Angew. Chem.* **2008**, *120*, 2675−2678.

(55) Weill, N.; Corbeil, C. R.; De Schutter, J. W.; Moitessier, N. *J. Comput. Chem.* **2011**, *32*, 2878−89.

(56) Berman, H. M. *Nucleic Acids Res.* **2000**, *28*, 235−242.

(57) Yano, J. K.; Wester, M. R.; Schoch, G. A.; Griffin, K. J.; Stout, C. D.; Johnson, E. F. *J. Biol. Chem.* **2004**, *279*, 38091−4.

(58) Rowland, P.; Blaney, F. E.; Smyth, M. G.; Jones, J. J.; Leydon, V. R.; Oxbrow, A. K.; Lewis, C. J.; Tennant, M. G.; Modi, S.; Eggleston, D. S.; Chenery, R. J.; Bridges, A. M. *J. Biol. Chem.* **2006**, *281*, 7614−22.

(59) Wester, M. R.; Yano, J. K.; Schoch, G. A.; Yang, C.; Griffin, K. J.; Stout, C. D.; Johnson, E. F. *J. Biol. Chem.* **2004**, *279*, 35630−7.

(60) Daura, X.; Gademann, K.; Jaun, B.; Seebach, D.; van Gunsteren, W. F.; Mark, A. E. *Angew. Chem., Int. Ed.* **1999**, *38*, 236−240.

(61) Christen, M.; Hünenberger, P. H.; Bakowies, D.; Baron, R.; Bürgi, R.; Geerke, D. P.; Heinz, T. N.; Kastenholz, M. A.; Kräutler, V.; Oostenbrink, C.; Peter, C.; Trzesniak, D.; van Gunsteren, W. F. *J. Comput. Chem.* **2005**, *26*, 1719−51.

(62) Jones, G.; Willett, P.; Glen, R. C. *J. Mol. Biol.* **1995**, *245*, 43−53.

(63) Kirton, S. B.; Murray, C. W.; Verdonk, M. L.; Taylor, R. D. *Proteins* **2005**, *58*, 836−44.

(64) Rydberg, P.; Olsen, L. *ACS Med. Chem. Lett.* **2012**, *3*, 69−73.

(65) Rydberg, P.; Olsen, L. *ChemMedChem* **2012**, *7*, 1202−9.

(66) Field, M. *A Practical Introduction to the Simulation of Molecular Systems*; Cambridge University Press: 2007; Vol. 2nd ed.

(67) Reference 66, p 77.

(68) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. *J. Cheminf.* **2011**, *3*, 33.

(69) *Accelrys*; Accelrys Inc.: 10188 Telesis Court, Suite 100, San Diego, CA, 92121, USA.

(70) Xing, L.; Glen, R. *J. Chem. Inf. Model.* **2002**, *42*, 796−805.

(71) Xing, L.; Glen, R. C.; Clark, R. D. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 870−879.

(72) Rydberg, P.; Jørgensen, M. S.; Jacobsen, T. A.; Jacobsen, A.-M.; Madsen, K. G.; Olsen, L. *Angew. Chem.* **2013**, *125*, 1027−1031.

(73) Yu, X.; Cojocaru, V.; Wade, R. C. *Biotechnol. Appl. Biochem.* **2013**, *60*, 134−145.

(74) Rittle, J.; Green, M. T. *Science (New York, N.Y.)* **2010**, *330*, 933−7.

(75) *RStudio, version 0.96.316*. http://http://www.rstudio.com/ (accessed May 19, 2013).

(76) Rydberg, P.; Rostkowski, M.; Gloriam, D. E.; Olsen, L. *Mol. Pharmaceutics* **2013**, *10*, 1216−1223.

(77) Shahrokh, K.; Orendt, A.; Yost, G. S.; Cheatham, T. E. *J. Comput. Chem.* **2012**, *33*, 119−33.

(78) Case, D. *AMBER11*; 2010.

(79) Dolinsky, T. J.; Czodrowski, P.; Li, H.; Nielsen, J. E.; Jensen, J. H.; Klebe, G.; Baker, N. A. *Nucleic Acids Res.* **2007**, *35*, W522−5.

(80) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins* **2006**, *65*, 712−25.

(81) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926.

(82) Humphrey, W. *J. Mol. Graphics* **1996**, *14*, 33−38.

(83) Mart-Renom, M. A.; Stuart, A. C.; Fiser, A.; Sánchez, R.; Melo, F.; Sali, A. *Annu. Rev. Biophys. Biomol. Struct.* **2000**, *29*, 291−325.

(84) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089.

(85) Wu, X.; Brooks, B. R. *Chem. Phys. Lett.* **2003**, *381*, 512−518.

(86) Jean-Paul Ryckaert, G. C.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327−341.

(87) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684.

(88) *SYBYL Molecular Modeling Software*; Tripos Associates Inc.: St. Louis, MO, USA.

(89) Valiev, M.; Bylaska, E.; Govind, N.; Kowalski, K.; Straatsma, T.; Van Dam, H.; Wang, D.; Nieplocha, J.; Apra, E.; Windus, T. *Comput. Phys. Commun.* **2010**, *181*, 1477−1489.

(90) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785−789.

(91) *Java, J2SE-1.5*. http://www.java.com/en/ (accessed May 19, 2013).

(92) *Eclipse, version 3.5.2*. http://wiki.eclipse.org/Platform (accessed May 19, 2013).