Improvements in Molecular Mechanics Sampling and Energy Models

Joseph Bylund

Submitted in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2013

ABSTRACT

Improvements in Molecular Mechanics Sampling and Energy Models

Joseph Bylund

The process of bringing drugs to market continues to be a slow and expensive affair. And despite recent advances in technology, the cost both in monetary terms and in terms of time between target identification and arrival of a new drug on the market continues to increase.

High throughput screening is a first step in the direction of being able to test a large number of possible bioactive compounds very quickly. However the space of possible small molecules is limitless, and high throughput screening is limited both by the size of available libraries and the cost of running such a large number of experiments.

Computational drug design, or computer assisted drug design offers a possible way of addressing some of the shortfalls of conventional high throughput screening. Using computational methods it is possible to estimate parameters such as binding affinity, of any small molecule, even those not currently present in any small molecule library, without having to first invest in the possibly slow and expensive process of finding a synthesis. Computational methods can be used to screen similar molecules, or mutations in small molecule space, seeking to increase binding affinity to the protein target, and thereby efficacy, while simultaneously minimizing binding affinity to other proteins, decreasing cross reactivity, and reducing toxicity and harmful side effects.

Computational biology methods of drug research can be broadly classified in a number of different ways. However; one of the most common classifications is along the lines of the methods used to identify possible drug compounds and later optimize those leads. The first broad category is the informatics or artificial intelligence based approaches. In these approaches artificial intelligence methods such as neural networks, support vector machines and qualitative structure-activity relationships (QSAR) are used to identify chemical or

structural properties that contribute heavily to binding affinity. Ligand based approaches are very useful when a large number of known binders are known for a specific family of proteins. In this case the ligands cluster together in some sort of chemical space and new compounds which occupy a similar chemical space are likely to also bind strongly with the protein of interest. The class explored in this thesis is the diverse class known as structural methods. These methods in the most general sense make use of a sampling method to sample a number of protein, or protein-small-molecule interaction conformations and an energy model or scoring function to measure dimensions which would be very difficult and or expensive to measure experimentally.

In this thesis a number of different sampling methods which are applicable to different questions in computational biology are presented. Additionally an improved algorithm for evaluating implicit solvent effects is presented, and a number of improvements in performance, reliability and utility of the molecular mechanics program used are discussed.

Table of Contents

List of Figures

List of Tables

Chapter 1

Prediction of P450 Sites of Metabolism

1.1 Introduction

The most common method of drug clearance among currently prescribed drugs is metabolism, which is the primary method of clearance for approximately 75% of the top 200 most commonly prescribed drugs in the United States [Williams *et al.*, 2004]. Cytochrome p450 is critical to drug metabolism, being active in approximately 75% of drugs which are cleared in this method [Guengerich, 2007]. Of the human isoforms of P450, Cytochrome P450 2D6 (CYP2D6) is frequently involved metabolism of xenobiotics [Williams *et al.*, 2004], there are also high resolution crystal structures available for CYP2D6 [Rowland *et al.*, 2006] and thus it was used as a test case for this study. As covered in **??**, accurately predicting absorption, distribution, metabolism, and excretion, characteristics of drug compounds can be a critical determining factor in determining drug efficacy, performance in clinical development stages, and the overall costs of bringing new drugs to market. Because of the ubiquity of P450 in metabolic reactions of drugs, there is no other single enzyme family as significant to determining ADME as P450.

The general form of the reaction most frequently catalyzed by P450 is

$$\mathrm{RH} + \mathrm{O_2} + \mathrm{NADPH} + \mathrm{H^+} > \mathrm{ROH} + \mathrm{H_2O} + \mathrm{NADP^+} \tag{1.1}$$

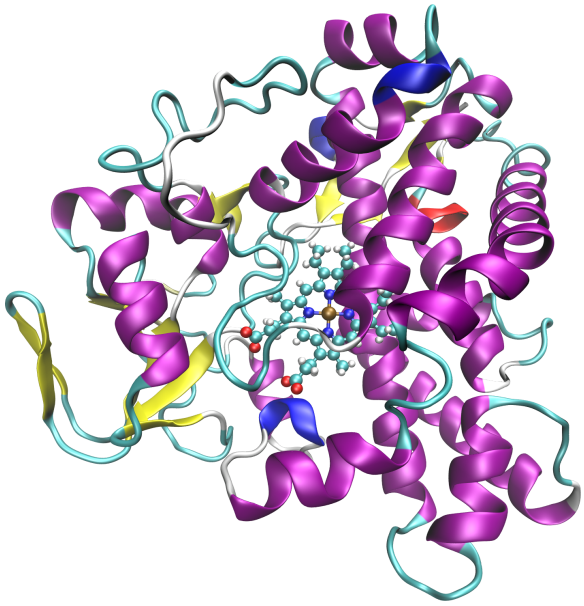

Figure 1.1: The structure of cytochrome P450, taken from PDBid 1JFB, shown in cartoon representation. The bonded heme group, shown as ball and stick model, is visible in the center. The brown iron atom is chelated by four deep blue nitrogen atoms.

The specific locations of sites of metabolism (SOM) on small molecules can have a profound effect on the ADME characteristics of a small molecule. Some cancer drugs such as epipodophyllotoxins, ifosphamide, tamoxifen, taxol and vinca alkaloids, are converted into their active states by oxygenation at specific locations by P450 [Kivisto *et al.*, 1995]. P450 is the body's primary defense against toxicity, usually catalyzing the conversion of toxic compounds into harmless products [Gonzalez, 2005; Guengerich, 2001]. However in certain cases, such as acetaminophen, it is possible for P450 to convert a harmless reactant into a toxic product [Chen *et al.*, 1998], although usually these compounds would be eliminated during the clinical trial stages. Additionally the different metabolites of a compound may be differentially cleared by the body having significant effects on bioavailability. Because of the costs associated with testing ADME parameters in live organisms accurate computational predictions can significantly decrease both costs and times associated with drug development.

Because of its central role in drug metabolism P450 has already been a subject of a number of studies attempting to predict sites of metabolism and chemical metabolites [Afzelius

et al., 2007]. A number of different classes of methods for predicting sites of metabolism by P450 have been developed. Broadly speaking these can be classified into: quantitative structure-activity relationship (QSAR) based, pharmacophore-based, structure-based (docking), reactivity-based, and rule-based methods [Cruciani *et al.*, 2005]. Rule based and pharmacophore based methods make predictions based on a subset of the drug structure, and it is possible for elements of the drug far from a possible site of metabolism to either prevent or promote metabolism at that location. QSAR based approaches work best when the set of reactions being catalyzed are very similar, however P450 catalyzes a very broad range of reactions so these approaches are likewise somewhat limited in the case of P450. Reactivity based methods are both very expensive to compute, being unsuited for screening a large database and do not take into account the structure of the P450 isoform [Singh *et al.*, 2003; Chen *et al.*, 1997; de Visser *et al.*, 2002]. MetaSite, an approach which makes use of structural information of both the ligand and the P450 isoform process has achieved a 84.3% prediction accuracy (296 of 351 total sites of metabolism correctly predicted), and the primary site of metabolism is identified in the top 3 ranked sites in over 90% of cases [Cruciani *et al.*, 2005]. However the sampling of P450 conformations done by MetaSite is quite limited, pre-computing a number of low energy conformations and then docking the substrate into each of those.

We have developed a similar approach which provides significantly more through sampling of the P450 substrate complex. The new method, IDSite, makes use of the structures of both the P450 and the substrate as well as evaluating the intrinsic reactivity of the possible site of metabolism.

1.2 Methods

Prediction of sites of metabolism is a three stage procedure:

1. Initially a number of different ligand conformations are generated, and these are docked into a rigid protein, with soft VDW terms using Glide [Halgren *et al.*, 2004; Friesner *et al.*, 2004].

2. The docked conformations are refined using a Monte Carlo Minimization (MMC)

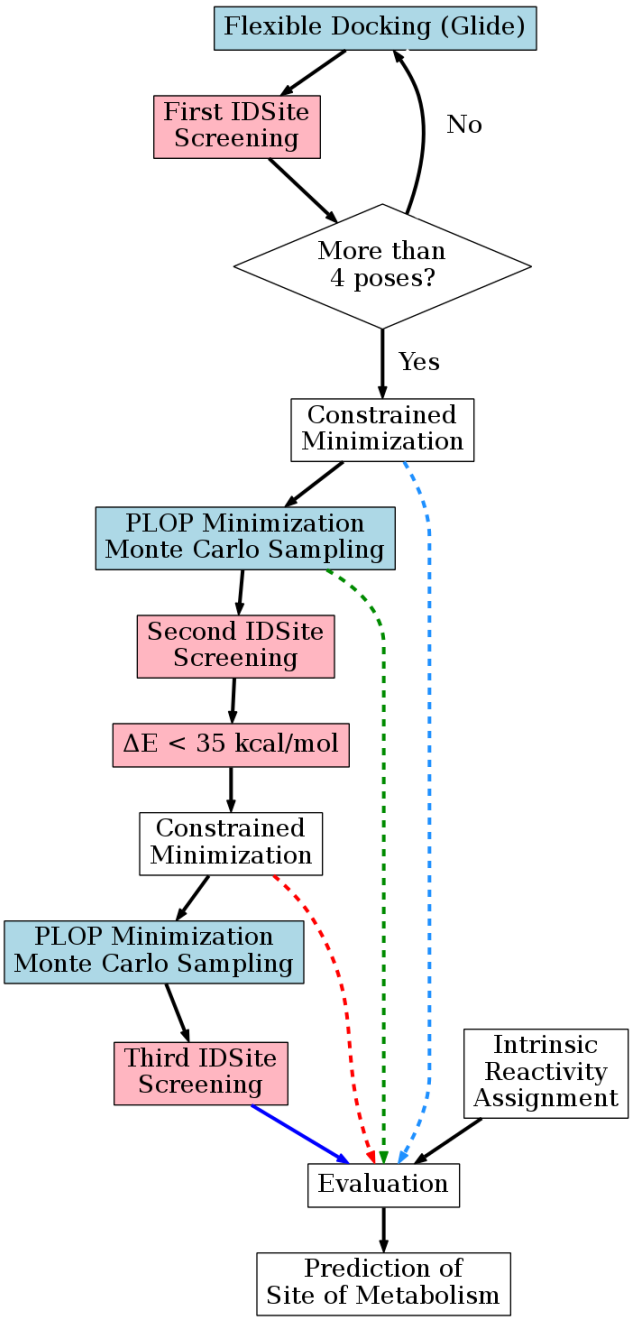

Figure 1.2: An overview of the entire IDSite procedure. The dotted lines represent abbreviated versions of the full procedure. Receiver operating characteristic graphs for the full version, and these abbreviated versions, are presented in 1.12. Series colors on ROC graphs correspond to arrow colors here.

approach which samples degrees of freedom in both the ligand and protein.

3. Refined conformations are classified into reactive site or non-reactive site on the basis of the energy of the refined conformations and the intrinsic reactivity of the site. [Li *et al.*, 2011]

1.2.1 Docking

In the initial docking stage of the IDSite protocol Glide is used to generate a number of proposed docked conformations for each ligand. Glide (standard precision) is used to generate a number of different ligand conformations by sampling conformations of freely rotatable bonds and rings. A bounding box, which will be used for a grid search, is defined centered at the centroid of the ligand with an edge length of 10 angstroms. Because the crystal structure used for CYP2D6 (PDBID: 2F9Q) does not have a ligand, the centroid of residues Glu216, Asp301, Thr309, and Phe483 was used instead in this case. Because the steric clashes present in many proposed docked conformations can be relieved using a simple minimization procedure a reduced Van der Waals (VDW) radii are used in the docking stage for non-polar atoms. The VDW radii used for the P450 are scaled by a factor of 0.4, and the scaling for the ligand starts at 0.8. If an insufficient number of poses, in this case fewer than four, are found using these scaling factors for the radii the scaling of the ligand is stepped down until at least four poses are found. Additional filtering of possible high energy conformations was also skipped in order to ensure the greatest diversity of docked poses reached the refinement stage. The collection of docked poses are then clustered according to the RMSD of the ligand, and each pose is minimized. The top sixty ranked poses according to the Glide SP metric are retained screened using a number of different criteria. A hard sphere overlap criteria is used to remove poses with obvious steric clashes which were not removed during the minimization procedure. A conserved feature of CYP2D6 ligand complexes is a salt bridge with Glu216 or Asp301. In order to reduce sampling cost IDSite only considers structures with at least one hydrogen-bond donor within 4 angstroms of the centroid of these two residues and Ser304. The sphere defined by these residues is illustrated along with the bounding box used for sampling in Figure 1.3 A number of other rule based geometric screens are used to remove structures

which are unlikely to react. Structures meeting any of the following criteria:

1. The distance of the basic nitrogen to the ferryl oxygen is less than 5.0 angstroms;
2. The distance of the basic nitrogen to the negative charged oxygen (in Glu216 or Asp301) is greater than 5.5 angstroms;
3. More than 2 heavy atoms from the ligands are further than 16.0 angstroms away from the heme iron;
4. More than 1 heavy atom from the ligand are closer than 1.0 angstroms to the receptor;
5. More than 6 heavy atoms from the ligand are closer than 1.8 angstroms to the receptor;
6. No heavy atom in the ligand is within 5.0 angstroms to the heme iron;

are removed. If the number of structures at this point is too low, the VDW scaling factors of the non-polar atoms of the ligand are stepped down, and the process is repeated. If four or more poses are found at these point these poses are passed onto the next stage of the IDSite procedure, the Monte Carlo Minimization refinement stage.

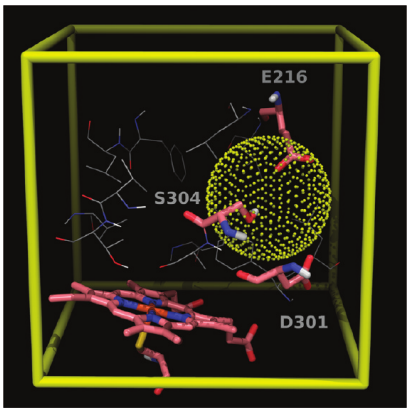

Figure 1.3: The bounding box used by Glide in order to generate the inital set of docked poses. The docking procedure also requires at least one hydrogen bond donor be found within 4 angstroms of the centroid of Glu216, Asp301, and Ser304 is also shown. The sphere representing this constraint is also shown.

1.2.2 Monte Carlo Minimization Refinement

Since the emphasis in IDSite sampling is efficient sampling of low energy conformations, as only the lowest energy conformations are passed on to the next stage of prediction, Monte Carlo Minimization, which provides more efficient sampling of low energy conformations, was used instead of a more traditional Monte Carlo simulation (see **??**). The Monte Carlo Minimization sampling used by IDSite for refinement incorporates three different types of steps: side chain motions, rigid body transformations, and hybrid Monte Carlo simulations. For each Monte Carlo step one of three types of motions is selected according to the weighted probabilities, which are different for the two different PLOP sampling stages, see Table 1.1. Using the chosen method a new conformation is proposed and minimized before

	PLOP Sampling Stage	
	First	Second
Number of Residues Sampled	12	40
Number of Structures Advanced to Next Stage	max(n*8,24)	max(n*20,60)
P(side chain step)	0.5	0.7
P(rigid body step)	0.1	0.2
P(HMC)	0.4	0.2

Table 1.1: The number of residues sampled as well as the number of structures advanced to the next stage from each of the sampling stages. Also, the relative probabilities of selecting each of the different sampling steps during a Monte Carlo minimization sampling stage.

the Metropolis acceptance criteria (equation **??**) is applied to the proposed state, using a temperature of 300 K. All atoms of all residues with any atom within 5 Angstroms of the ligand in the starting crystal structure were allowed to move during Monte Carlo moves, including the ligand itself.

During the minimization Monte Carlo sampling stages of the IDSite procedure artificial constraints are used to guide the sampling towards a transition state like conformation. These constraints create artificial bond or angle potentials which affect the minimization, but are not used in the Monte Carlo acceptance test. For each of the minimization Monte

Carlo sampling stages of the IDSite procedure two different sets of constraints are applied depending on the hybridization of the carbon atom at the possible site of metabolism, for a total of four possible different sets of constraints. In the first minimization Monte Carlo stage two constraints are applied:

1. The sulfur-iron-carbon angle is constrained to 145 degrees, with 20 degrees of "slack", or a flat bottom to the potential well (denoted as 145±20 degrees). The spring constant of this constraint is about 25 kcal/mol/degree^2, or ~40% the strength of a carbon-carbon-carbon angle.

2. A "dummy" oxygen atom is placed above the plane of the heme group, in the same position that it would occupy if an oxygen molecule was bound to the heme. This dummy atom has no interactions with other atoms, but is used as the anchor of a distance constraint for the carbon at the site of metabolism. The carbon-dummy oxygen distance is constrained to 2.5±0.5 angstroms. The spring constant of this constraint is 100 kcal/mol/angstrom^2, approximately 1/3rd the strength of a carbon-carbon bond.

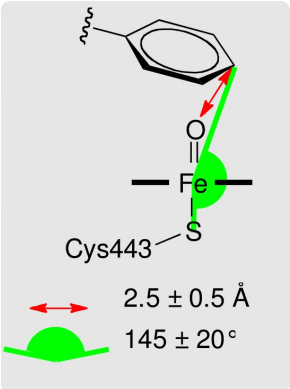

Figure 1.4: The constraints applied to sp^2 atoms during the constrained minimization and first minimization Monte Carlo sampling stage. The spring constant of the bond constraint (red arrow) is 100 kcal/mol/angstrom^2, and that of the angle constraint is 25 kcal/mol/degree^2. The oxygen atom depicted in this figure is a "dummy" atom and does not interact with any other atoms in the structure except through the constraint.

In the second minimization Monte Carlo sampling stage the constraints are different for sp^2 and sp^3 carbons. For sp^3 sites:

1. the hydrogen bound to the carbon at the possible site of metabolism is constrained to a distance of 1.25±0.1 angstroms and a spring constant of 20 kcal/mol/angstrom^2,

2. the carbon in question is constrained to 2.2±0.8 angstroms and a spring constant of 10 kcal/mol/angstrom^2,

3. the heme iron-hydrogen-carbon angle is constrained to 138±5 degrees and a spring constant of 20 kcal/mol/degree^2.

For sp^2 sites:

1. the carbon at the possible site of metabolism is constrained to 1.8±0.1 angstroms and a spring constant of 20 kcal/mol/angstrom^2,

2. both adjacent carbons are also constrained to the dummy oxygen atom, at a distance of 2.5±0.1 angstroms and a spring constant of 20 kcal/mol/angstrom^2, and

3. finally the hydrogen bonded to the carbon at the possible site of metabolism is constrained to the oxygen atom at a distance of 2.0±0.1 angstroms and a 20 kcal/mol/angstrom^2 spring constant.

As CYP2D6 forms a conserved salt bridge with the substrate with either glutamate 216 and aspartate 301 [Paine *et al.*, 2003], this was also incorporated as a constraint during the sampling stages. In the first sampling stage this salt bridge is enforced by introducing a harmonic constraint of 3.0±0.3 angstroms, between the basic nitrogen of the substrate and each of the side chain oxygen atoms in GLU216, ASP301 and SER304. The spring constants of this constraints are 15.0, 8.0 and 4.0 kcal/mol/angstrom^2 for GLU216, ASP301 and SER304 respectively. Additionally, an angle constraint is applied to each of the N-H-O angles, this is set to 150.0±30.0 degrees and has a spring constant of 5.0 kcal/mol/degree^2. In the second sampling stage four separate trajectories are calculated for each of the four carboxalate oxygens of GLU216, ASP301. In each trajectory a constraint of 1.9±0.1 angstroms is applied between the hydrogen attached to the basic substrate nitrogen and one of the

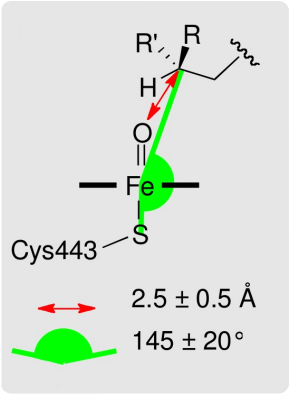

Figure 1.5: The constraints applied to sp^3 atoms during the constrained minimization and first minimization Monte Carlo sampling stage. The spring constant of the bond constraint (red arrow) is 100 kcal/mol/angstrom^2, and that of the angle constraint is 25 kcal/mol/degree^2. The oxygen atom depicted in this figure is a "dummy" atom and does not interact with any other atoms in the structure except through the constraint.

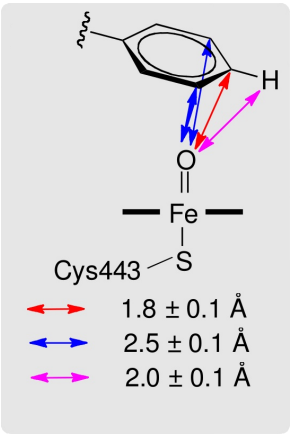

Figure 1.6: The constraints applied to sp^2 atoms during the constrained minimization and second minimization Monte Carlo sampling stage.

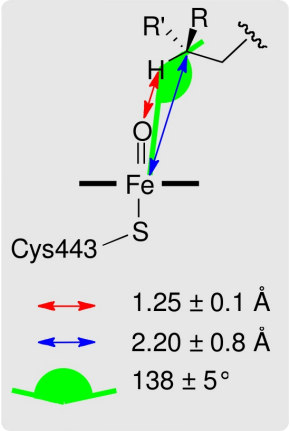

Figure 1.7: The constraints applied to sp^3 atoms during the constrained minimization and second minimization Monte Carlo sampling stage.

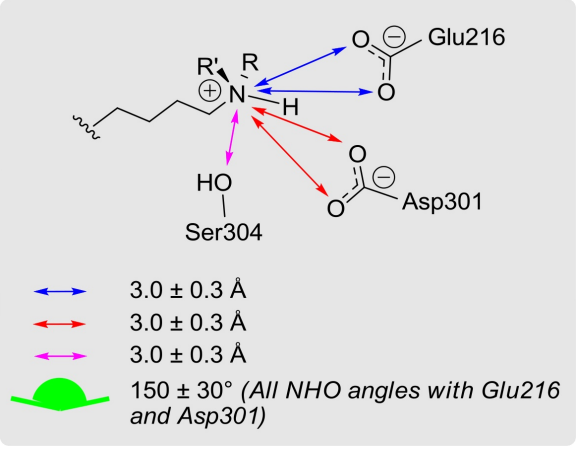

Figure 1.8: The constraints applied to the salt bridge region of CYP2D6 during the *first* minimization Monte Carlo sampling stage.

four carboxalate oxygens. Additionally, the angle of the hydrogen bond is constrained to 168±12 degree, with a spring constant of 5.0 kcal/mol/degree^2.

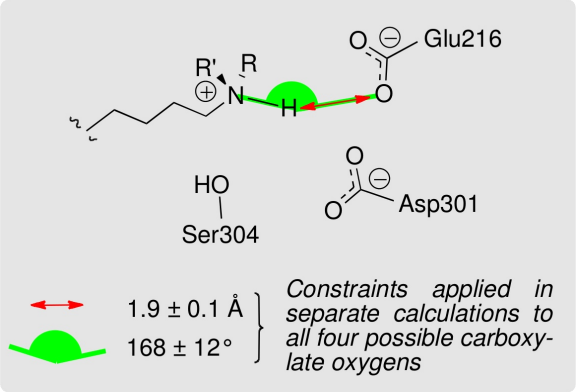

Figure 1.9: The constraints applied to the salt bridge region of CYP2D6 during the *second* minimization Monte Carlo sampling stage.

1. Several types of side chain motions were implemented in PLOP. In all cases, they are defined in a such a way that they can be applied to both ligands and proteins. The same atomic overlap screening function implemented with the rigid body Monte Carlo was implemented with the side chain torsional moves. a. Random torsion angle moves: The first type of move that was implemented is random movement of torsional chi angles. For small torsion moves, a random perturbation of the angle of +/- X is made, where X is a random number with user defined magnitude. For large torsion moves, for each torsion angle that is changed, a random angle is selected in the form 60*Y +/- X, where Y = 1 through 5, and X is the same random number for the small torsion moves. The large move was introduced since positions at the top of rotamer barriers are relatively unlikely to be selected, and efficiency thus can be improved by focusing on the more probable moves. The ratio of small to large torsion moves can be used-adjusted, as can the ratio of probabilities of changing all the torsions in a randomly selected side chain versus changing only one single (randomly selected) torsion among all the free torsions in the simulation can be set as a user-defined parameter. Rotamer side chain moves: A second type of torsional samples implemented is random selection of a new rotamer state for the entire side

chain, plus an optional user defined small noise term for each torsion in the rotamer state. A database of protein rotamer states obtained from crystallographic data are already a part of PLOP [Xiang and Honig, 2001] Rotamer libraries for ligands are generated by examining all possible side chain conformations at 10 degree resolution and screening this set for steric clashes. A Monte Carlo move in this case represents a choice of a new torsional rotamer state for the entire side chain. Monte Carlo moves based on torsional states cannot lead to correct equilibrium distributions, as transitions from non-rotamer states to rotamer states are defined, but not reverse transitions, upsetting detailed balance. However, a pretabulated rotamer state is more likely to be low energy than a randomly generated torsional state, and thus allows for more diverse conformational searching. Correlated torsional moves: Most torsional rearrangements of the side chains in the core of proteins are highly correlated because of the density. In order to attempt to include correlated torsional motion, at each step we examine the distance between all pairs of beta carbons in the ligands that are free to move. At each step, for the set of side chains that are free to move, clusters where beta carbons are all mutually within a user-specified distance are identified. This process takes a trivial amount of time compared to an energy evaluation, so does not slow the simulation at all. Then, with user specified probabilities, clusters of different sizes are selected for the torsional moves, either with random side chain moves, or rotamer selection moves. By selecting only clusters where all residues are mutual neighbors, detailed balanced is observed for simulations where accurate equilibrium sampling is desired. By varying the dihedral angles of the rotatable bonds, IDSite uses side chain MC moves in PLOP to sample the selected side-chain conformations of the protein and of the ligand. Up to three close residues (C beta distance within 6 angstroms) are allowed to rotate collectively, but the moves of the protein residues and those of the ligand are separated. In each attempted movement, the conformations of the selected side chains (from the protein/ligand) are either changed by random perturbations or assigned by the randomly selected rotamers from a library. For an attempt with a random perturbation, the displacement of each dihedral angle is the sum of a large rotation (N times 60 degrees with N as a random integer between

0 and 5) and a random perturbation from 0 to 30 degrees. For a rotamer library attempt, a side-chain conformation is updated with a random rotamer from a high resolution side-chain library for protein residues [Xiang and Honig, 2001], and from a homogeneous library at 10 degree resolution for the ligand. If a structure with tolerable overlaps is generated in an attempt, it is minimized and sent to subsequent stages for judgment of acceptance. Each side-chain move takes less than 15 seconds and is the fastest among all the three move types.

For side chain Monte Carlo, a steric screen with an overlap factor of 0.6 was used. Rotamer torsional moves were selected 75% of the time, with half of the remaining being of random torsions, and the other half random perturbations of all torsions within the randomly selected side chains. Clusters of size 1 (i.e. single side chains), size 2 and size three were selected in equal proportion, and all side chains in the cluster were perturbed with the selected torsion move. A mutual beta carbon distance of 6 Angstroms was used for the clustering size. Small torsion perturbations made +/- 60 degress from the current dihedral angle, and were performed 5% of the time; Large periodic moves were performed 95% of the time. Only outer steps were performed, and each side chain Monte Carlo series consisted in only one move. Minimization was performed after the single step, and acceptance was performed at 1 K.

2. Rigid motion moves. Rigid body translation and rotation were also implemented for noncovalently linked moieties, such as ligands. Random rotations and translations were coupled together, allowing for more concerted movement. Rigid body move implemented in PLOP can optionally include a screening step, where atomic Lennard-Jones overlaps that would lead to energies much higher than would be observed in any conceivably long equilibrium simulation are rejected without further evaluation. A ratio of 0.7 between the distance between the two atoms and the sum of the Lennard-Jones radii of the two atoms yields energies on the order of 10's of thousands of kcal/mol, and is thus reasonable to maintain equilibrium sampling in a Monte Carlo simulation. Translations were implemented in a random direction, with a user-defined magnitude. Rotations were implemented by picking a random quarternion (a random angle around a random axis, through the geometric center of the rigid group) with

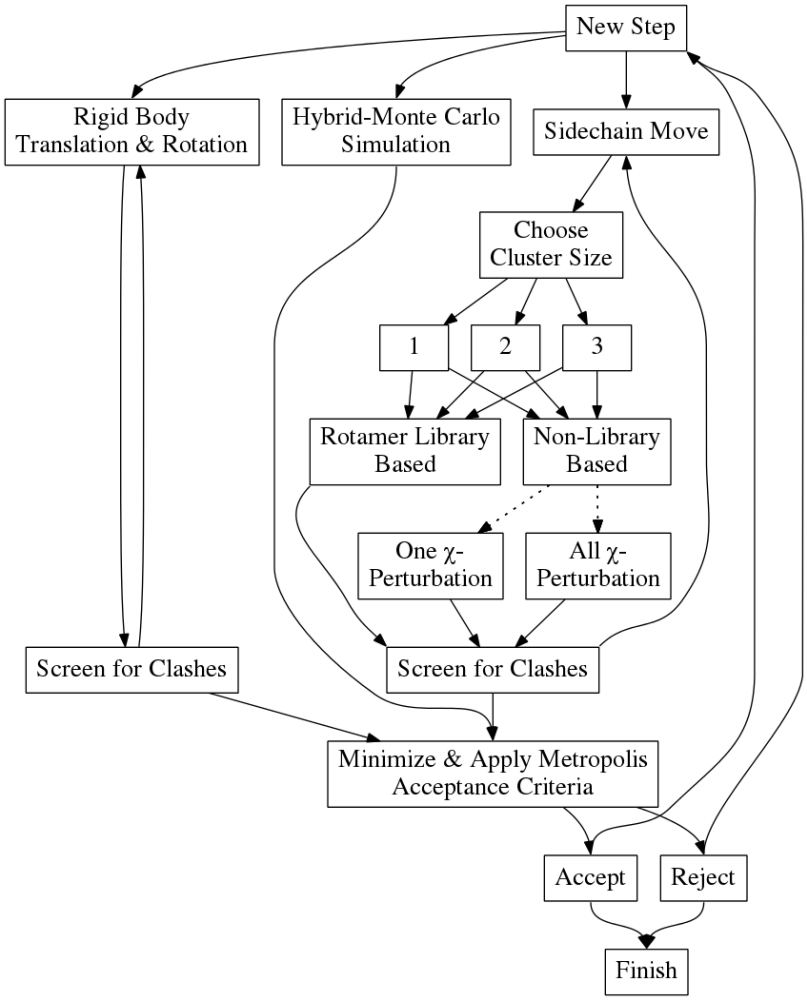

Figure 1.10: An outline of the Monte Carlo minimization refinement stages in PLOP.

a user specified maximum random angle centered around either the current angle, or 180 from the current angle, in the case of a flip. Multiple time scale Monte Carlo sampling was also implemented with rigid body moves, with short range and long range interactions defined as above. In addition, an option to compute the inner Monte Carlo loops with reduced Lennard-Jones radii were also implemented, to increase the ability to escape from tight spacial bottlenecks. In this case, the long time step energies are the full energies with unscaled Lennard-Jones radii. This increases the conformational freedom and therefore sampling for the short, at a cost of decreasing the acceptance probability in the outer loop. Scaled Lennard-Jones radii were also implemented in multiple time dynamics, but yielded very little apparent improvement because of the lack of phase space overlap between dynamics with different scaled Lennard-Jones radii). Rigid body moves are used to sample the translational and rotational space of the ligand. Multiple attempts with reduced VDW radii are applied, as it is quite common to fail in searching for a clash-free conformation in a single rigid body moving attempt (especially when the ligand is large and flexible and the binding pocket is relatively small). Each rigid body move includes 1000 attempts, and each attempt performs a translation along a random vector and a rotation around a random axis, with less than 0.5 angstroms and 60 degree displacement, respectively. In addition, the VDW radii are reduced (scaling factor 0.8) to soften the Lennard-Jones potential, so that mild steric clashes are allowed, which are likely to be resolved by the subsequent minimization. The rigid body move usually takes 20 to 40 seconds per move.

For rigid body Monte Carlo, a steric screen with an overlap factor of 0.7 was used, with a translation size of 0.5 Angstroms and a rotation size of plus or minus 60 degrees. No flip moves were included, as flips were not anticipated with the geometry of the ligand system [Robert, check this is true?] A Lennard-Jones scaling parameter of 0.8 was used during the inner steps. Each rigid MC step consisted of 1000 inner steps, and only one outer step, meaning that only one minimization occurred each time rigid body Monte Carlo was selected as the move step.

3. The Hybrid Monte Carlo (HMC) [Duane *et al.*, 1987] step is a velocity verlet molecular

dynamics simulation. This simulation allows all atoms in both the ligand and residues containing atoms withing 5 angstroms of the ligand to move. Initial velocities are taken from a Maxwell-Boltzmann distribution at 900 K. Bonded and short range interactions evaluated every 1 nanosecond inner time step, and long range potentials are assumed to be fixed over inner steps. Five inner steps compose each outer HMC step. In the outer step the molecular surface, long range interactions and, Born alphas are updated before computing the energy and applying the Metropolis acceptance criteria at a temperature of 900 K after each MD run. Taking up to 15 minutes per move, the HMC is the most expensive among all three types of moves in PLOP.

1.2.3 Evaluation

Both a parameterized and an unparameterized model were used to classify potential sites of metabolism. IDSite makes the assumptions that all intermediates before the rate determining step are at equilibrium [Wang *et al.*, 2007], that hydrogen abstraction is the rate limiting step for hydroxilation of aliphatic carbons and electrophilic attack is the rate limiting step for hydroxilation of aromatic rings [Guengerich, 2001; Shaik *et al.*, 2005]. With these assumptions the rate of metabolism at each possible site of reaction is affected by the free energy of binding in order to put that site in the site of reaction, as well as the free energy barrier of rate determining step, or

$$\Delta G_{\text{total}} = \Delta G_{\text{binding}} + \Delta G_{\text{barrier}} \tag{1.2}$$

The $\Delta G_{\text{binding}}$ above is calculated using a PLOP evaluation of the refined pose. The intrinsic reactivity for the system is computed from DFT calculations on a simplified system, replacing the heme with a methody radical, and using a linear relationship between IR(heme) and IR(methoxy radical) to estimate the true reactivity for the heme system.

$$\text{IR(heme)} = 1.117 * \text{IR(methoxy radical)} + C \tag{1.3}$$

Since this constant C is identical for each state it has no effect on the relative differences in ΔG_{site} or the relative rate of metabolism at possible sites.

$$E = \langle 1.117 * \text{IR(methoxyradical)} + C + E_{\text{TS}} \rangle - kT\ ln(N_H) \tag{1.4}$$

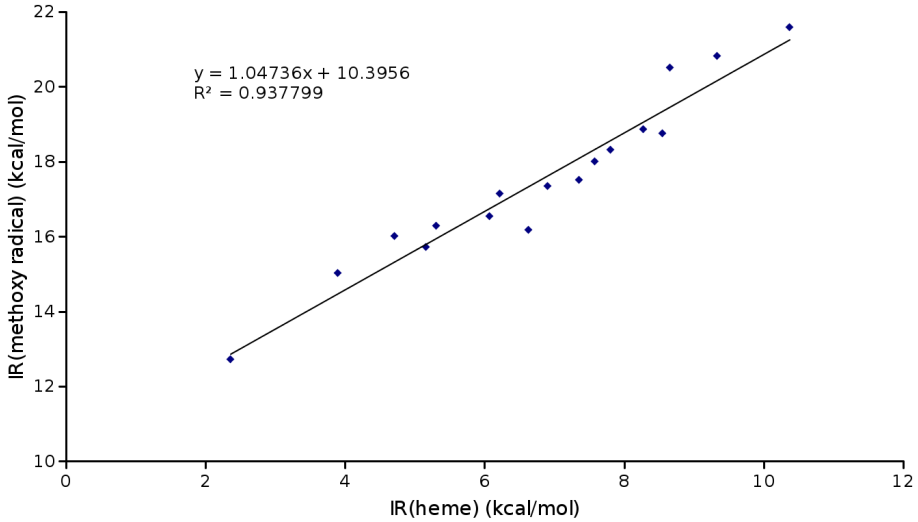

Figure 1.11: The linear relationship between the calculated intrinsic reactivity of the methoxy radical complex and that of the heme complex. Adapted from [Li *et al.*, 2011] with minor correction. In the original manuscript the slope of the regression was reported as 1.117 and that number was used throughout. This difference should not significantly affect the physical IDSite classifier results, and does not affect the results of the fit model. In the rest of this text the value from the original publication of 1.117 will be used.

Since the ligand is forced to assume a different conformation in order to react, the energy of this transition state conformation, E_{TS}, is also computed using PLOP. As the relative abundance of different metabolites is determined by differences in ΔG per site rather than absolute reactivities, the constant in equation 1.4 does not affect which metabolites are produced. A site of possible metabolism is classified as positive if it is observed in greater than 0.1% yield, which corresponds to a $\Delta\Delta G$ of ~4.75 kcal/mol between the most favored state and the cutoff for negative predictions.

The second classifier is similar however:

1. a different constant is used to estimate IR(heme) from IR(methoxy radical), namely 1.071,

2. if the binding energy of the transition state complex of a pose is within 5.26 kcal/mol of the lowest pose, it is set to the binding energy of the lowest pose. Otherwise the

difference is scaled by 0.58,

3. and the cutoff for an active prediction is changed from 4.75 kcal/mol to 1.46 kcal/mol.

These parameters were decided upon by maximizing $\frac{\text{true positives}}{\text{(false positives+false negatives)}}$ on a training set of 36 compounds.

Model compound	Site of Metabolism	Heme model (kcal/mol)	Methoxy model (kcal/mol)
Benzene		20.51	8.66
Anisole	Ortho-	16.29	5.31
	Meta-	18.76	8.55
	Para-	16.01	4.71
	Beta-	16.18	6.63
Dimethylether		15.03	3.9
Dimethylanisole	Meta-	16.54	6.07
	Para-	17.51	7.35
Ethane		21.58	10.37
Ethanol	1	12.73	2.36
	2	17.35	6.9
Propane		18.31	7.8
Toluene	Ortho-	17.15	6.22
	Meta-	18.86	8.27
	Para-	18	7.58
	Alpha-	15.72	5.16
t-Butylebenzene	Beta-	20.82	9.33

Table 1.2: DFT calculated values for internal reactivity of various compounds with either methoxy radical (compound I) or heme system. Correlation between these values is illustrated in Figure 1.11.

1.3 Results

Both physical and fitted IDSite were able to acheive promising results predicting CYP2D6 sites of metabolism.

The physical model correctly identified 68 of 82 active sites of metabolism for a sensitivity of 0.829. For inactive sites this model correctly identified 1054 of 1075 inactive sites with a specificity of 0.980. The fit model performed similarly, and even slighly better identifying 52 of 57 sites of metabolism (sensitivity of 0.912) in the training set and 25 of 25 in the test set (sensitivity of 1.0).

$$\text{sensitivity} = \frac{TN}{TN + FP} = \frac{\#\text{ of true sites of non-metabolism identified}}{\#\text{ experimental sites of non-metabolism}} \tag{1.5}$$

$$\text{sensitivity} = \frac{TP}{TP + FN} = \frac{\#\text{ of true sites of metabolism identified}}{\#\text{ experimental sites of metabolism}} \tag{1.6}$$

The fit model also correctly identified 709 of 717 inactive sites in the training set (specificity of 0.989) and 352 of 358 inactive sites in the test set (specificity of 0.983). As the performance of the fit model is similar to that of the physical model, it does not appear that the fit model is overparameterized to the training set. Specific results for both models are presented in Tables 1.3 and 1.4, and the specific sites identified by both models as well as experiments are illustrated in Figures 1.14 and 1.15.

We believe that the parameters help account for some degree of noise in the molecular mechanics calculations. The scaling of the binding energy difference, either to zero inside a window about the minimum energy pose, or by a factor of 0.58 decreases the relative weight of the molecular mechanics contribution relative the quantum contribution to the classifier. This might imply that some sites are not being classified as active because they are not in the lowest energy conformation around the docked pose, suggesting that additional molecular mechanics sampling might further improve results. However as will be discussed later, the molecular mechanics stage already dominates the total time necessary for an IDSite prediction, and the current molecular mechanics procedure takes about 450 hours.

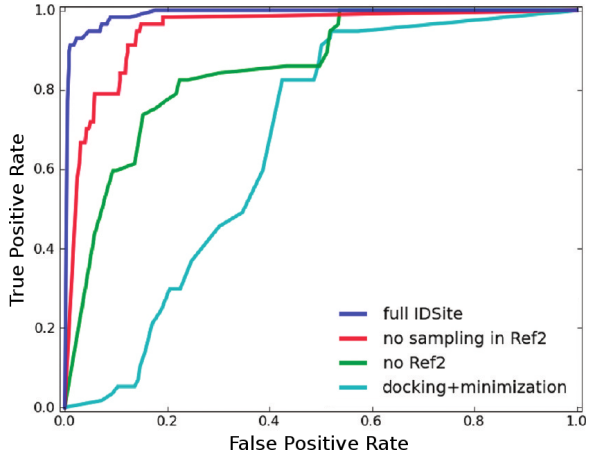

Figure 1.12: The effect of additional sampling on prediction of site of metabolism by P450. The light blue series describes only performing the initial Glide docking stage followed by minimization. The green series is obtained by using the set of structures obtained in the first minimization Monte Carlo sampling stage. The red series is obtained by screening the structures obtained in the first sampling stage, and minimizing these structures using the constraints specified in Figures 1.6 and 1.7. The blue series makes use of the entire IDSite procedure. The color scheme of these series corresponds to the colors of edges in Figure 1.2.

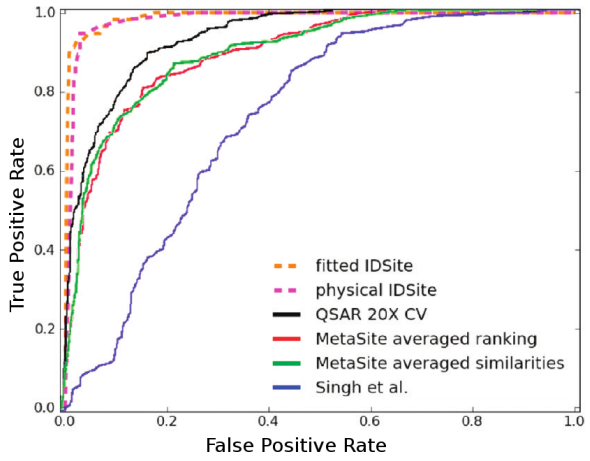

Figure 1.13: A comparison of the performance of IDSite with a variety of other methods of predicting P450 sites of metabolism. IDSite obtains the best performance, followed by a quantitative structure-activity relationship based method [Sheridan *et al.*, 2007]. Adapted from [Sheridan *et al.*, 2007].

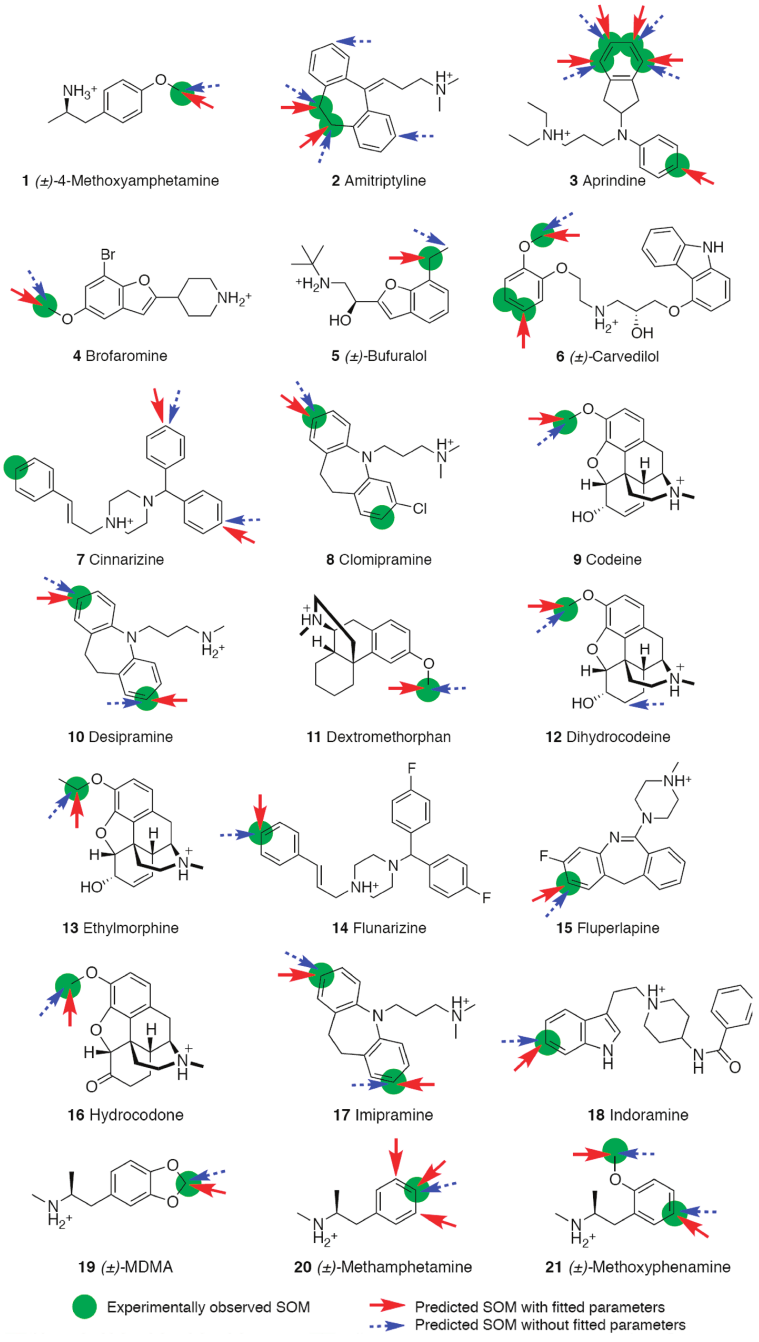

Figure 1.14: Physical and fitted IDSite predictions of sites of metabolism on the training set.

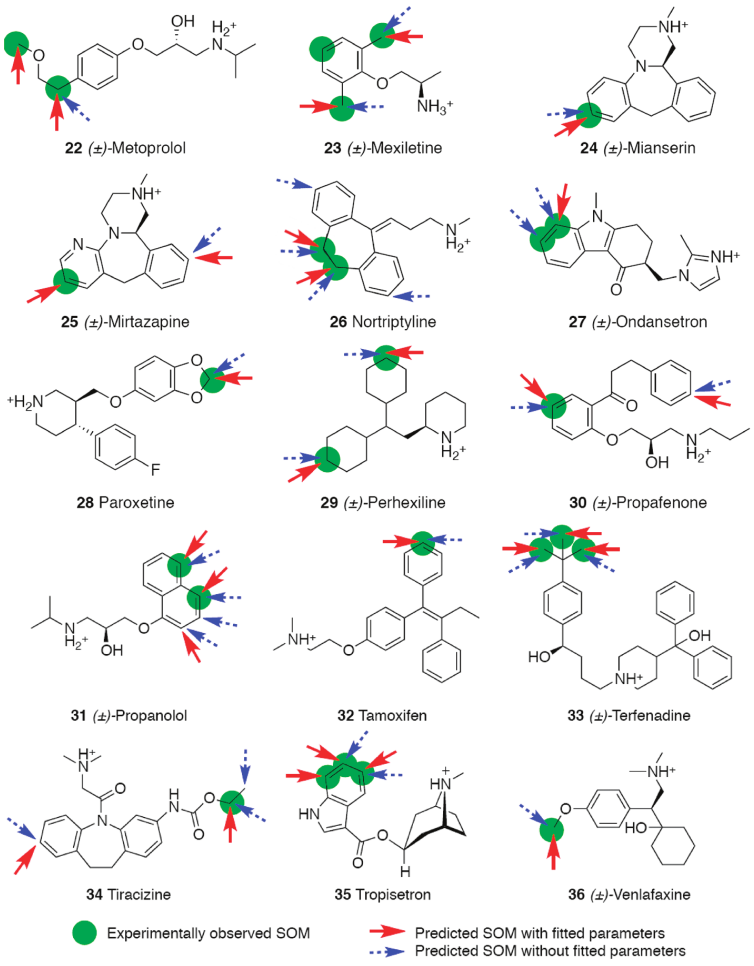

Figure 1.14: (continued)

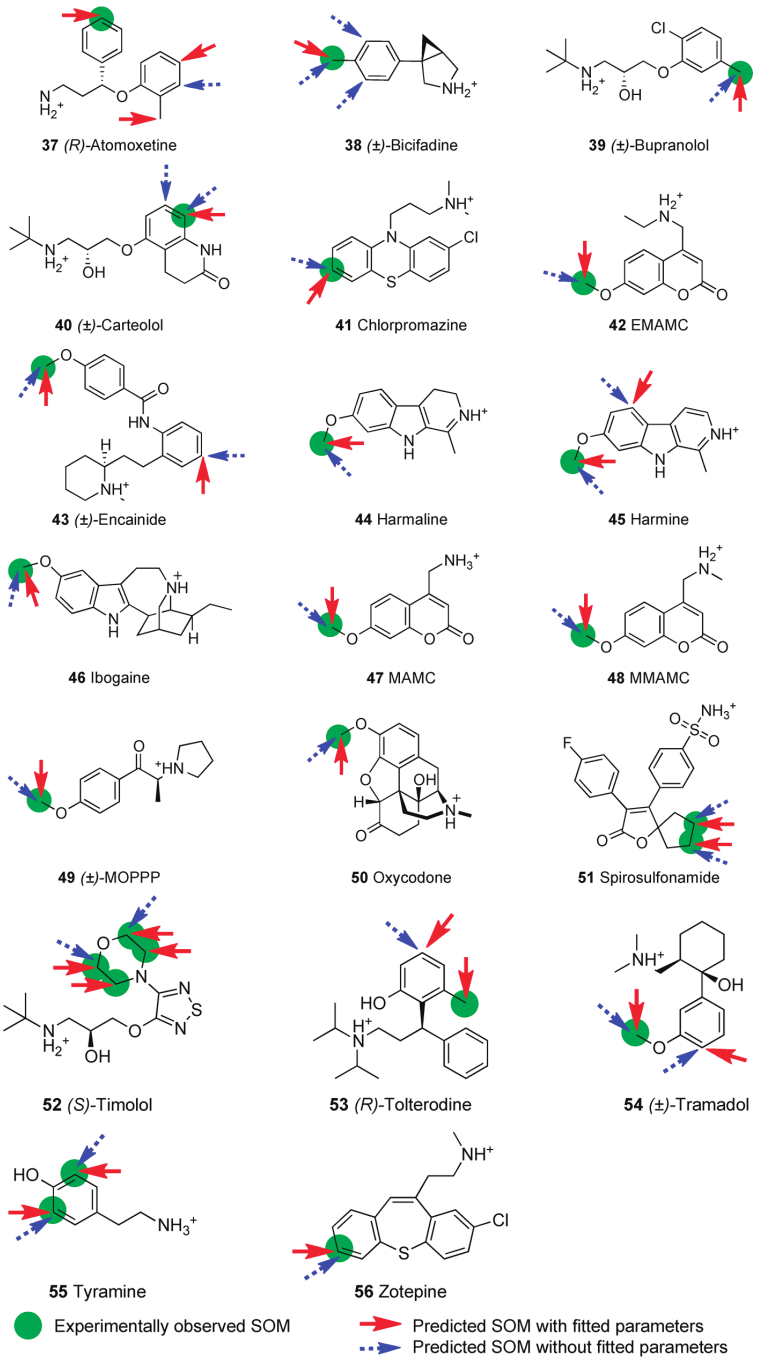

Figure 1.15: Physical and fitted IDSite predictions of sites of metabolism on the test set.

		Physical			Fitted		
Compound #	Compound	TP	FP	FN	TP	FP	FN
1	4-methoxyamphetamine	1	0	0	1	0	0
2	Amitriptyline	2	2	0	2	0	0
3	Aprindine	4	0	1	5	0	0
4	Brofaromine	1	0	0	1	0	0
5	Bufuralol	0	1	1	1	0	0
6	Carvedilol	1	0	2	2	0	1
7	Cinnarizine	0	2	1	0	2	1
8	Clomipramine	1	0	1	1	0	1
9	Codeine	1	0	0	1	0	0
10	Desipramine	2	0	0	2	0	0
11	Dextromethorphan	1	0	0	1	0	0
12	Dihydrocodeine	1	1	0	1	0	0
13	Ethylmorphine	1	0	0	1	0	0
14	Flunarizine	1	0	0	1	0	0
15	Fluperlapine	1	0	0	1	0	0
16	Hydrocodone	1	0	0	1	0	0
17	Imipramine	2	0	0	2	0	0
18	Indoramine	1	0	0	1	0	0
19	MDMA	1	0	0	1	0	0
20	Methamphetamine	1	0	0	1	2	0
21	Methoxyphenamine	2	0	0	2	0	0
22	Metoprolol	1	0	1	2	0	0
23	Mexiletine	2	0	1	2	0	1
24	Mianserin	1	0	0	1	0	0
25	Mirtazapine	0	1	1	1	1	0
26	Nortriptyline	1	1	0	1	0	0
27	Ondansetron	2	0	0	1	0	1
28	Paroxetine	1	0	0	1	0	0
29	Perhexiline	2	0	0	2	0	0
30	Propafenone	1	1	0	1	1	0
31	Propranolol	2	2	0	2	1	0
32	Tamoxifen	1	0	0	1	0	0
33	Terfenadine	3	0	0	3	0	0
34	Tiracizine	1	2	0	1	1	0
35	Tropisetron	2	0	1	3	0	0
36	Venlafaxine	1	0	0	1	0	0
	Total	47	13	10	52	8	5

Table 1.3: Results of physical and fitted IDSite on training set of 36 compounds.

		Physical			Fitted		
Compound #	Compound	TP	FP	FN	TP	FP	FN
37	Atomoxetine	0	1	1	1	2	0
38	Bicifadine	1	2	0	1	0	0
39	Bupranolol	1	0	0	1	0	0
40	Carteolol	1	1	0	1	0	0
41	Chlorpromazine	1	0	0	1	0	0
42	EMAMC	1	0	0	1	0	0
43	Encainide	1	1	0	1	1	0
44	Harmaline	1	0	0	1	0	0
45	Harmine	1	1	0	1	1	0
46	Ibogaine	1	0	0	1	0	0
47	MAMC	1	0	0	1	0	0
48	MMAMC	1	0	0	1	0	0
49	MOPPP	1	0	0	1	0	0
50	Oxycodone	1	0	0	1	0	0
51	Spirosulfonamide	2	0	0	2	0	0
52	Timolol	2	0	2	4	0	0
53	Tolterodine	0	1	1	1	1	0
54	Tramadol	1	1	0	1	1	0
55	Tyramine	2	0	0	2	0	0
56	Zotepine	1	0	0	1	0	0
	Total	21	8	4	25	6	0

Table 1.4: Results of physical and fitted IDSite on a test set of 20 compounds. Note that for the physical model there is no training performed so results in the text are presented in a unified fashion for the training and test set.

1.4 Discussion

Bibliography

[Afzelius *et al.*, 2007] Lovisa Afzelius, Catrin Hasselgren Arnby, Anders Broo, Lars Carlsson, Christine Isaksson, Ulrik Jurva, Britta Kjellander, Karin Kolmodin, Kristina Nilsson, Florian Raubacher, et al. State-of-the-art tools for computational site of metabolism predictions: comparative analysis, mechanistical insights, and future applications. *Drug metabolism reviews*, 39(1):61–86, 2007.

[Chen *et al.*, 1997] Hao Chen, Marcel J de Groot, Nico PE Vermeulen, and Robert P Hanzlik. Oxidative n-dealkylation of p-cyclopropyl-n, n-dimethylaniline. a substituent effect on a radical-clock reaction rationalized by ab initio calculations on radical cation intermediates. *The Journal of Organic Chemistry*, 62(23):8227–8230, 1997.

[Chen *et al.*, 1998] Weiqiao Chen, Luke L Koenigs, Stella J Thompson, Raimund M Peter, Allan E Rettie, William F Trager, and Sidney D Nelson. Oxidation of acetaminophen to its toxic quinone imine and nontoxic catechol metabolites by baculovirus-expressed and purified human cytochromes p450 2e1 and 2a6. *Chemical research in toxicology*, 11(4):295–301, 1998.

[Cruciani *et al.*, 2005] Gabriele Cruciani, Emanuele Carosati, Benoit De Boeck, Kantharaj Ethirajulu, Claire Mackie, Trevor Howe, and Riccardo Vianello. Metasite: understanding metabolism in human cytochromes from the perspective of the chemist. *Journal of medicinal chemistry*, 48(22):6970–6979, 2005.

[de Visser *et al.*, 2002] Sam P de Visser, François Ogliaro, Pankaz K Sharma, and Sason Shaik. What factors affect the regioselectivity of oxidation by cytochrome p450? a dft

study of allylic hydroxylation and double bond epoxidation in a model reaction. *Journal of the American Chemical Society*, 124(39):11809–11826, 2002.

[Duane *et al.*, 1987] Simon Duane, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth. Hybrid monte carlo. *Physics letters B*, 195(2):216–222, 1987.

[Friesner *et al.*, 2004] Richard A Friesner, Jay L Banks, Robert B Murphy, Thomas A Halgren, Jasna J Klicic, Daniel T Mainz, Matthew P Repasky, Eric H Knoll, Mee Shelley, Jason K Perry, et al. Glide: a new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. *Journal of medicinal chemistry*, 47(7):1739–1749, 2004.

[Gonzalez, 2005] Frank J Gonzalez. Role of cytochromes p450 in chemical toxicity and oxidative stress: studies with cyp2e1. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 569(1):101–110, 2005.

[Guengerich, 2001] F Peter Guengerich. Common and uncommon cytochrome p450 reactions related to metabolism and chemical toxicity. *Chemical research in toxicology*, 14(6):611–650, 2001.

[Guengerich, 2007] F Peter Guengerich. Cytochrome p450 and chemical toxicology. *Chemical research in toxicology*, 21(1):70–83, 2007.

[Halgren *et al.*, 2004] Thomas A Halgren, Robert B Murphy, Richard A Friesner, Hege S Beard, Leah L Frye, W Thomas Pollard, and Jay L Banks. Glide: a new approach for rapid, accurate docking and scoring. 2. enrichment factors in database screening. *Journal of medicinal chemistry*, 47(7):1750–1759, 2004.

[Kivisto *et al.*, 1995] Kari T Kivisto, Heyo K Kroemer, and Michel Eichelbaum. The role of human cytochrome p450 enzymes in the metabolism of anticancer agents: implications for drug interactions. *British journal of clinical pharmacology*, 40(6):523–530, 1995.

[Li *et al.*, 2011] Jianing Li, Severin T Schneebeli, Joseph Bylund, Ramy Farid, and Richard A Friesner. Idsite: an accurate approach to predict p450-mediated drug metabolism. *Journal of chemical theory and computation*, 7(11):3829–3845, 2011.

[Paine *et al.*, 2003] Mark JI Paine, Lesley A McLaughlin, Jack U Flanagan, Carol A Kemp, Michael J Sutcliffe, Gordon CK Roberts, and C Roland Wolf. Residues glutamate 216 and aspartate 301 are key determinants of substrate specificity and product regioselectivity in cytochrome p450 2d6. *Journal of Biological Chemistry*, 278(6):4021–4027, 2003.

[Rowland *et al.*, 2006] Paul Rowland, Frank E Blaney, Martin G Smyth, Jo J Jones, Vaughan R Leydon, Amanda K Oxbrow, Ceri J Lewis, Mike G Tennant, Sandeep Modi, Drake S Eggleston, et al. Crystal structure of human cytochrome p450 2d6. *Journal of Biological Chemistry*, 281(11):7614–7622, 2006.

[Shaik *et al.*, 2005] Sason Shaik, Devesh Kumar, Samuël P de Visser, Ahmet Altun, and Walter Thiel. Theoretical perspective on the structure and mechanism of cytochrome p450 enzymes. *Chemical reviews*, 105(6):2279–2328, 2005.

[Sheridan *et al.*, 2007] Robert P Sheridan, Kenneth R Korzekwa, Rhonda A Torres, and Matthew J Walker. Empirical regioselectivity models for human cytochromes p450 3a4, 2d6, and 2c9. *Journal of medicinal chemistry*, 50(14):3173–3184, 2007.

[Singh *et al.*, 2003] Suresh B Singh, Lucy Q Shen, Matthew J Walker, and Robert P Sheridan. A model for predicting likely sites of cyp3a4-mediated metabolism on drug-like molecules. *Journal of medicinal chemistry*, 46(8):1330–1336, 2003.

[Wang *et al.*, 2007] Yonghua Wang, Yan Li, and Bin Wang. Stochastic simulations of the cytochrome p450 catalytic cycle. *The Journal of Physical Chemistry B*, 111(16):4251–4260, 2007.

[Williams *et al.*, 2004] J Andrew Williams, Ruth Hyland, Barry C Jones, Dennis A Smith, Susan Hurst, Theunis C Goosen, Vincent Peterkin, Jeffrey R Koup, and Simon E Ball. Drug-drug interactions for udp-glucuronosyltransferase substrates: a pharmacokinetic explanation for typically observed low exposure (auci/auc) ratios. *Drug Metabolism and Disposition*, 32(11):1201–1208, 2004.

[Xiang and Honig, 2001] Zhexin Xiang and Barry Honig. Extending the accuracy limits of prediction for side-chain conformations. *Journal of molecular biology*, 311(2):421–430, 2001.