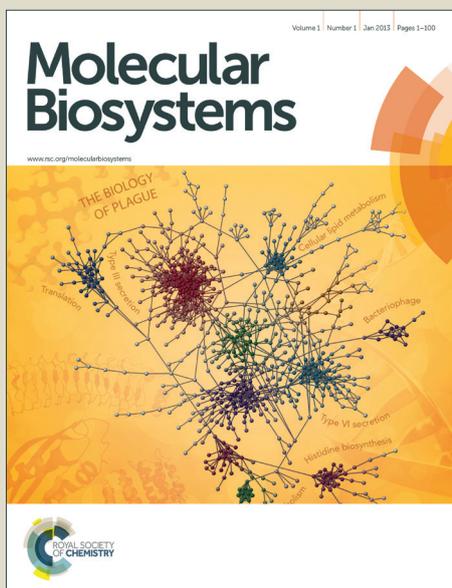


Molecular BioSystems

Accepted Manuscript



This is an *Accepted Manuscript*, which has been through the Royal Society of Chemistry peer review process and has been accepted for publication.

Accepted Manuscripts are published online shortly after acceptance, before technical editing, formatting and proof reading. Using this free service, authors can make their results available to the community, in citable form, before we publish the edited article. We will replace this *Accepted Manuscript* with the edited and formatted *Advance Article* as soon as it is available.

You can find more information about *Accepted Manuscripts* in the [Information for Authors](#).

Please note that technical editing may introduce minor changes to the text and/or graphics, which may alter content. The journal's standard [Terms & Conditions](#) and the [Ethical guidelines](#) still apply. In no event shall the Royal Society of Chemistry be held responsible for any errors or omissions in this *Accepted Manuscript* or any consequences arising from the use of any information it contains.



www.rsc.org/molecularbiosystems

Cite this: DOI: 10.1039/c0xx00000x

www.rsc.org/xxxxxx

ARTICLE TYPE

Predicting the binding modes and sites of metabolism of xenobiotics

Goutam Mukherjee,^{a,b} Pancham Lal Gupta^{a,b} and B. Jayaram,^{*a,b,c}

^aDepartment of Chemistry, ^bSupercomputing Facility for Bioinformatics & Computational Biology, ^cKusuma School of Biological Sciences, Indian Institute of Technology Delhi, Hauz Khas, New Delhi-110016, India. Fax: +91-11-26582037; Tel. No: +91-11-26591505;

⁵E-mail: bjayaram@chemistry.iitd.res.in; Website: www.scfbio-iitd.res.in

Received (in XXX, XXX) Xth XXXXXXXXX 20XX, Accepted Xth XXXXXXXXX 20XX

DOI: 10.1039/b000000x

Metabolism studies are an essential integral part of ADMET profiling of drug candidates to evaluate their safety and efficacy. Cytochrome P-450 (CYP) metabolizes a wide variety of xenobiotics/drugs. Binding modes of these compounds with CYP and their intrinsic reactivities decide the metabolic products. We report here a novel computational protocol, which comprises docking of ligands to heme-containing CYPs and prediction of binding energies through a newly developed scoring function, followed by analyses of the docked structures and molecular orbitals of the ligand molecules, for predicting the sites of metabolism (SOM) of ligands. The calculated binding free energies of 121 heme-containing protein-ligand docked complexes yielded a correlation coefficient of 0.84 against experiment. Molecular orbital analyses of the resultant top three unique poses of the docked complexes achieved a success rate of 87% in identifying the experimentally known sites of metabolism of the xenobiotics. The SOM prediction methodology is freely accessible at www.scfbio-iitd.res.in/software/drugdesign/som.jsp.

Introduction

The success of a drug's journey through the body is measured in terms of its absorption, distribution, metabolism, excretion and toxicity (ADMET) properties.¹ An ideal oral drug should be rapidly and completely absorbed from the alimentary canal (digestive tract) and find its way directly and specifically to its site of action. Absorption can be predicted by the Lipinski's rule of five²⁻³ which sets limits on molecular weight, number of hydrogen bond donors and acceptors and logarithm of partition coefficient. Drug metabolism plays a crucial role in bioavailability of the drug molecule and its side effects. Understanding metabolism in terms of metabolic sites and products (metabolites) is a central aspect in the drug discovery process.⁴⁻⁶ Knowledge of the sites of metabolism (SOM) of a molecule and its biotransformation products can help not only in optimizing the lead molecule with favorable metabolic profile but also in reducing toxicity and enhancing bioavailability and bioactivity.⁶⁻⁸ Most of the drugs undergo metabolic transformations in the human liver, which is the key source of a variety of metabolizing enzymes.⁷ Cytochrome P450s (CYPs), a superfamily of heme-containing enzymes, are the major enzymes involved in drug metabolism. Important oxidative and reductive reactions which are catalyzed by different isoforms of CYPs are summarized in Table 1.

Finding the sites of metabolism (SOM) and the biotransformation products of a molecule through experiment is an expensive process. Hence, there is a need for reliable computational approaches to predict the metabolic fate of a molecule. Numerous *in silico* tools have been developed for the prediction of the metabolic fate of a molecule.^{9-10,17-44}

Table 1: Summary of some important oxidative and reductive reactions catalyzed by different isoforms of CYPs.

Reaction	Example	
	Substrate	Metabolites
Aromatic C-hydroxylation	Phenytoin ⁹	4'-Hydroxyphenytoin
Aliphatic C-hydroxylation	Valproic acid ⁹	5-Hydroxy valproic acid
N-hydroxylation	Dapsone ⁹	Dapsone-hydroxylamine
Olefinic centre epoxidation	Carbamazepine ¹⁰	trans-10,11-Dihydroxy carbamazepine
Aromatic center epoxidation	Benzo(a)pyrene ¹¹	Benzo(a)pyrene-7,8-dihydro-diol
Dehydrogenation	Nifedipine ¹²	Dehydronifedipine
N-dealkylation	Amiodarone ⁹	Desethylamiodarone
O-dealkylation	Codeine ⁹	Morphine
N-oxidation	Quinoline ¹³	Quinoline 1-oxide
S-oxidation	Zaltoprofen ⁹	S-oxide-zaltoprofen
Oxidative Desulfuration	Parathion ¹⁴	Paraoxon
Oxidative Deamination	Amphetamine ¹⁵	Phenylacetone
Oxidative dehalogenation	Halothane ¹⁶	Trifluoroacetyl chloride
Reductive Dehalogenation	Halothane ¹⁶	2-Chloro-1,1,1-trifluoroethane (CTE) and 2-Chloro-1,1-difluoroethylene (CDE)

These techniques, adopt either reactivity-based approaches, fingerprint-based data mining approaches, shape-focused

techniques, protein–ligand docking, and combined methods.³⁷ In reactivity based approach, hydrogen atom abstraction energy is calculated by AM1 or DFT level of theory to predict the likelihood of a metabolic reaction. QMBO,⁴¹ CypScore⁴² tools adopt this methodology to predict reactivity of a molecule toward CYPs. In fingerprint-based data mining approach, search for a query atom having defined atom environment is carried out with a large known biotransformation database. SYBYL atom types are used to encode the atom environment. MetaPrint2D²⁷ software (<http://www-metaprint2d.ch.cam.ac.uk/>) uses this methodology to predict the site of biotransformation of any organic molecule. Shape matching methodology is adopted by ROCS software (<http://www.eyesopen.com/rocs>). The prediction of SOM of a molecule is based on the shape focused alignment of any molecule of interest to a known substrate of CYP. ROCS³⁶ software uses flurbiprofen as a representative molecule for a large number of CYP2C9 substrates. The most powerful technique of prediction of SOM of a molecule is a combined approach. In this approach, a combination of regular docking experiments or solvent accessible surface area (SASA) and pharmacophoric descriptor calculations or topological and quantum chemical atom-specific descriptor calculations and various types of quantum calculations such as, reactivity calculations of any atomic center, DFT activation energy, molecular orbital calculations is carried out. MLite¹⁰ software uses docking followed by quantum chemistry derived reactivity to predict the SOM of a molecule. SMARTCyp,¹⁸ 2D-SMARTCyp⁹ software uses precalculated DFT based activation energies in combination with topological accessibility descriptors to predict the SOM of a molecule. Metasite⁴ software first calculates the accessibility of an atom which is directed towards Fe atom of heme based on the 3D structure of any CYP protein, and GRID derived molecular interaction fields of the protein and ligand. In the next step, calculation of activation energy to produce the reactive radical intermediate (reactivity) is carried out. Finally, predicted SOM of a molecule is the site which possesses a significant score of accessibility and reactivity component compared to other atomic centers of the molecule. However it is not necessary that metabolic reactions always take place at the highly reactive centre of a molecule. Also, a molecule can have multiple sites of metabolism with the same CYPs or different CYPs. For instance, N-methyl-benzodioxolyl-butanamine (MBDB) has two major sites of metabolism with the same isoforms of CYP (CYP1A2),⁴⁵ while Carvedilol has two major sites of metabolism for two different isoforms of CYPs namely CYP1A2 and CYP2C9.⁴⁶ Combination approaches based on docking and reactivity are likely to fail in such cases. Most of the reactivity based approaches calculate AM1 level hydrogen atom abstraction energies. AM1 level calculations may not always predict the correct reactivity order of atomic centers in a molecule which has more than one SOM. Also, majority of the methodologies cite above are trained on a particular class of Cytochrome P450s (CYPs).

Present work is divided into two parts. The first part is related to the development of a scoring function which can consider the heme containing protein for predicting binding free energy of a ligand which is then integrated with the in-house docking software, ParDOCK.⁴⁷ Next part is a new methodology based on

a combination of docking followed by molecular orbital (MO) calculations and knowledge based methods to predict the potential metabolic sites of a molecule.

Dataset description

About 121 heme containing protein-ligand complexes are downloaded from RCSB⁴⁸ with known x-ray structures and experimental binding free energies. These 121 complexes contain 25 unique heme-containing protein targets (Table 2).

Table 2: A list of the 25 unique heme-containing protein targets in the 121 complex dataset considered in this study.

Sl. No.	Protein	Number
1	Cytochrome c peroxidase	28
2	Cytochrome P450-cam	6
3	Nitric oxide synthase	44
4	Cytochrome P450 2C5	4
5	Cytochrome P450 51	6
6	Respiratory nitrate reductase	2
7	Cytochrome P450 2A6	6
8	Cytochrome P450 cyp158A1	2
9	Cytochrome bc1	2
10	Prostacyclin synthase(cyp 450)	1
11	Prostaglandin i2 synthase(cytochrome P450 8A1)	1
12	Cytochrome P450 102	1
13	Cytochrome P450 cyp125	1
14	Cytochrome P450 2B6	1
15	Cytochrome P450 3A4	1
16	Bifunctional P-450:nadph-p450 reductase	1
17	Inducible nitric oxide synthase	2
18	Respiratory nitrate reductase	2
19	Cytochrome P450 158A2	2
20	Cytochrome P450 1A2	1
21	Cytochrome P450 121	1
22	Cytochrome P450 2C8	1
23	Cytochrome P450 46A1	2
24	Putative cytochrome P450 130	2
25	Cytochrome P450(bm-3)	1

A description of the 121 heme-containing protein-ligand complexes with the observed ranges of the various physico-chemical properties is provided in Table 3 which shows the heterogeneous nature of the present dataset with respect to the ligands and complexes.

Table 3: A description of the 121 heme-containing protein-ligand complexes in terms of the observed ranges of various physico-chemical properties.

Sl. No.	Descriptor/ Physicochemical Property	Limits
Ligand		
1	Size Index (Wiener Index)	9 – 5240
2	Hydrogen Bond Donors	0 – 12
3	Hydrogen Bond Acceptors	0 – 10
4	Ligand Net Charge	(-)3 – 1
5	C log P	(-)5.0 – 8.0
6	Molecular Weight	68 – 652
Complex		
7	Number of Unique Proteins in these 121 complexes	25
8	Experimental Binding Energy (kcal/mol)	(-)14.4 – (-)2.9

As case studies, 4 isoforms of CYPs namely CYP1A2, CYP2C9, CYP2C19 and CYP3A4 are considered. About 213 substrate molecules for these target CYPs are obtained from various public databases, such as Pubchem⁴⁹ and Drugbank⁵⁰⁻⁵³ and ZINC database.⁵⁴ Information about the experimental SOM for all the substrates molecules is available in the literature.^{9,10}

Methodology

Development of a scoring function to estimate the binding free energies of ligands in the heme containing proteins.

The molecular assemblies of heme-containing protein-ligand complexes formed by non-covalent interactions mainly comprise three types of non-bonded contributions viz. (i) electrostatic, (ii) van der Waals, (iii) hydrophobic and (iv) entropy. Entropy contribution here refers to a loss in conformational entropy of the protein side chains upon ligand binding. Scoring functions⁵⁵⁻⁵⁶ generally used to calculate binding energies in molecular docking protocols^{47,57-58} capture the above four contributions as a combination of the following mathematical functions (eq. 1). In order to predict the binding free energy accurately, it is required to multiply each of the energy terms in eq. 1 with a suitable scale factor.

$$\Delta G = E_{Tel} + E_{Tvdw} + E_{hydrophobic} + T\Delta S \dots (1)$$

Where,

$$E_{Tel} = \sum (e_{l_{protein-ligand}} + e_{l_{heme-ligand}})$$

$$E_{Tvdw} = \sum (vdw_{protein-ligand} + vdw_{heme-ligand})$$

Here, E_{Tel} and E_{Tvdw} are the total electrostatic and van der Waals contributions to the free energy ΔG of ligand binding with heme-containing protein. Total electrostatic (E_{Tel}) and total van der Waals (E_{Tvdw}) energies in eq. 1 are expressed as the sum of all pair wise atomic interactions in the protein-ligand systems including heme. The electrostatic interaction is computed via Coulomb's law with a sigmoidal dielectric function, while a (12, 6) Lennard-Jones potential function is used for calculating the van der Waals interactions. The hydrophobic contributions are computed via the Gurney parameter approach, a computationally simple approach for treating desolvation effects. Entropy contribution is calculated by an empirical rule. Details of each term are reported previously.⁵⁹⁻⁶³

Hydrophobic contribution is calculated in terms of the net loss in surface area of an atom type A (ΔA_{LSA}) multiplied by σ_A , the

atomic desolvation parameter for the atom type A.⁵⁵ During the calculation of the ΔA_{LSA} , it is observed that some of the atom types (halogen, phosphorus, sp hybridized atom) mentioned in the literature⁵⁵⁻⁵⁶ are not very commonly observed in the ligand molecules. Therefore some modifications are made in defining the atom types. All carbon, nitrogen and oxygen atoms are divided into two categories. One is a planar system (all atoms having sp^2 and sp hybridization) and the other is a non planar system (sp^3 hybridization). Planar systems for carbons and nitrogens are further divided into two categories. One is aliphatic and the other is aromatic. All hydrogens are divided into three categories: first is the hydrogen attached to aliphatic carbons, second is the hydrogen attached to aromatic carbons and the third is the hydrogen attached to a heteroatom other than carbons. Apart from this, halogen, sulfur and phosphorus atoms are combined into a single atom type (as these are not abundant) and iron is treated as a separate atom type. Thus there are a total 13 different atom types whose net loss in the surface area has to be calculated. ΔA_{LSA} of each of the atom types is calculated as reported earlier.⁵⁵

Since ligand is bound to the active site of the protein at the cost of entropy, the loss in conformational entropy (ΔS_{CR}) is an important contributor to the energetics of protein-ligand binding.

In order to calculate the loss in conformational entropy, relative accessibility of the protein side chain ($RA_{binding}$) is calculated. $RA_{binding}$ is defined as the ratio of the accessible surface area of the side chain in the bound form to the accessible surface area of the side chain in the unbound form. Before calculating the $RA_{binding}$, relative accessibility of the folding ($RA_{folding}$) is also calculated. $RA_{folding}$ is defined as the ratio of the accessible surface area of the side chain in the folded state to the accessible surface area of the side chain in the unfolded state. Protein side chains with $RA_{folding} > 60\%$ and $RA_{binding} < 30\%$ are considered to have a loss of conformational entropy. The procedures for calculating $RA_{binding}$ and $RA_{folding}$ are reported earlier.⁵⁵

Using the aforementioned approach, eq. 1 can be represented as follows.

$$\Delta G = \alpha(E_{Tel}) + \beta(E_{Tvdw}) + \sum_{A=1}^{13} \sigma_A(\Delta A_{LSA}) + \gamma(\Delta S_{CR}) + \delta \dots (2)$$

α , β , σ_A , γ and δ are regression coefficients. Their values are obtained via a partial least square fit method using experimental binding free energies. Table 4 gives the values of the regression coefficients.

Table 4: Regression coefficients for electrostatic, van der Waals, desolvation and entropy terms.

Sl. No.	Description	Parameters
1	Electrostatic	0.0115± 0.0277
2	van der Waals	0.0050±0.0266
3	sp ³ Carbon	0.1866±0.1173
4	sp ² aliphatic Carbon	0.0465±0.0252
5	sp ² aromatic Carbon	-0.0056± 0.0168
6	Hydrogen bonded to aliphatic carbon	0.0077±0.0020
7	Hydrogen bonded to aromatic carbon	0.0086±0.0075
8	Hydrogen bonded to heteroatom other than carbon	-0.0143±0.0142
9	sp ³ Nitrogen	0.0551±0.0213
10	sp ² aliphatic Nitrogen	0.0300±0.0093
11	sp ² aromatic Nitrogen	0.0326±0.0198
12	sp ³ Oxygen	0.0216±0.0164
13	sp ² Oxygen	-0.0026±0.0080
14	Phosphorus, Sulphur, Halogen	0.0137±0.0026
15	Iron	-0.7354±4.5862
16	Entropy	0.5312±0.1791
17	Constant (Intercept)	-2.0326±0.1791

Steps involved for estimating the binding free energies of heme containing protein-ligand complexes

Step 1: Derivation of partial atomic charges of the complex

In all the heme-containing protein-ligand complexes, one coordination site of iron is coordinately bound with either sulfur atom of cysteine (CYS) or nitrogen atom of histidine (HIS) and another coordination site is occupied by the ligand atom in a noncovalent bonding interaction. Low-spin (singlet) ferrous heme iron is assumed for the heme-CYS / HIS groups based on previous experimental findings.⁶⁴⁻⁶⁵ Crystallographic water molecules are removed from the structures of the complex. Cysteine residue in the protein which is covalently linked to iron (< 3 Å) in the heme is identified. S atom of cysteine which is covalently bound to iron is deprotonated. To the remaining amino acid residues, ligand and porphyrin group, hydrogen atoms are added. Assignment of partial atomic charges for the protein atoms is done using AMBER force field.⁶⁶⁻⁶⁷ AM1-BCC⁶⁸⁻⁶⁹ partial atomic charges are assigned to the ligand atoms. In the present study, all the heme-containing protein-ligand complexes are divided into two categories where iron of heme is covalently bound with sulfur atom of cysteine or iron is covalently attached to nitrogen atom of histidine. QM calculations are performed using Gaussian 09 software.⁷⁰ Knowledge of spin state and formal charge of iron atom in heme group is required during the quantum mechanical calculations. In the first step, x-ray coordinates of heme with its covalently linked amino acid (cysteine or histidine) are taken. In the next step, side chain of the amino acid which is covalently linked to iron in heme is converted to -CH₃ group. QM optimization is performed on the x-ray geometry with HF level of theory. 6-31G* basis set is

employed for all non-metallic elements and SDD with frozen core electrons for iron. In the next step, Mulliken charges⁷¹ for iron and porphyrin are extracted from the Gaussian output file and used for heme moiety. We have also carried our DFT optimization by using the B3LYP/6-31G*-SDD and B3LYP/6-31+G*-SDD level of theory. However, there is a very little effect on the overall results.

Step 2: Assignment of the force field parameters

Force field parameters for the protein, porphyrin and ligand atoms are adopted from the AMBER ff99SB⁷² and GAFF⁷³ force fields. van der Waals parameters for iron atom are taken from literature.⁷⁴

Step 3: Energy minimization of the complex

The heme-containing protein-ligand complexes are energy minimized in explicit solvent with AMBER 10.⁷⁵ Position of the heme and the covalently bound protein residue (cysteine or histidine) is fixed during energy minimization. Truncated octahedron type solvate box with 12.0 Å cutoff is used during minimization. In the first step, explicit water minimization by keeping complex molecule fixed, is performed with 500 steps of steepest descent and 500 steps of conjugate gradient methods. This step is necessary in order to allow rearrangement of water molecules around the solute. After solvent minimization, an all atom minimization with 1000 steps of steepest descent and 1500 steps of conjugate gradient methods is carried out.

Molecular orbital calculations

All calculations are carried out using Gaussian-09 suite of programs. The geometries of all the ligands are optimized at the B3LYP level of theory using 6-31G** basis set for C, H, N, O, P, S, F, Cl, Br atoms and SDD basis set with frozen core electrons for I atom. Molecular orbital calculations are carried out with B3LYP and BP86 functional using aforementioned basis set in implicit water. Frontier orbital plots are made using the Chemcraft software (www.chemcraftprog.com). During geometry optimization and molecular orbital calculations of a molecule, protonation state of that molecule is taken care of. All carboxylate and amine containing functional groups are converted to neutral carboxylic acid and cationic ammonium ions. Since lone pairs of carboxylate anion and amine functional groups are relatively more reactive (electronic reactivity) than other neutral functional groups (lone pairs of oxygen atom other than carboxylate anion, S, F, Cl, Br, I, π -bond, σ bond and triple bond), these functional groups are converted to less reactive states by adding protons to them. However, amine group which is directed to iron atom of heme moiety (docked pose) is not protonated during QM calculations, since lone pair of this functional group is knocked off during CYP oxidation.

Detection of SOM of a molecule against a particular isoform of CYP

In the first step, a ligand molecule of interest is docked and a minimization run is performed with a particular isoform of CYP. The purpose of minimization is to remove clashes of the ligand with the heme and protein atoms and also to get a proper orientation of the substrate with respect to the heme. ParDOCK docking software provides the user with eight energetically preferred best poses of a ligand molecule in the vicinity of the

protein (heme-containing active site). Binding free energies of each pose are estimated using the scoring function described previously. Top ranked pose based on binding free energies is likely to be the experimental pose. However, in some cases two or more docked poses among the final eight docked structures may be similar where the same atomic center / functional group is directed to the iron atom of the heme moiety. In such cases, unique poses are considered and ranked based on predicted binding energies. The most likely site of metabolism of a molecule can be predicted based on the distance as well as the angle of the atom in the ligand to the iron center in the heme. In the present study, a maximum distance of 7.5 Å as a cutoff and an angle cutoff (S(CYS) / N(HIS)-Fe-ligand atom) of 150° to the iron center of the heme as an oxidizable site of the ligand are considered. If the atoms in a molecule are part of the HOMO and also satisfy the distance plus angle criteria specified, then the closest atom from the catalytic centre (iron atom of the heme moiety) is considered as the real SOM of a molecule. If the distances are nearly indistinguishable, then the reactivity of the atomic centre is considered for identifying the SOM. Also, the criterion adopted for SOM detection is that the minimum HOMO density around an atom has to be greater than or equal to 1% of the total contribution. It is not mandatory that the site of the ligand atom directed to the iron is the real oxidizable site. Oxidation means loss of electrons from the system. Electrons can be lost from the highest occupied molecular orbital (HOMO). However, if the energies of other deep lying molecular orbitals (HOMO-1, HOMO-2, etc.) are very close to that of HOMO, then there may be a possibility of electron loss from other deep lying orbitals. In these cases, extra energy is required to remove the electron from the deep lying molecular orbitals which may be compensated by the formation of energetically favorable transition state (TS)⁷⁶⁻⁷⁷ structure by the heme, ligand and other side chain residues of the protein. Several other factors such as, stability of the radical species, product stability, etc. may also be decisive. Docking and scoring give estimates of the energetics of different TS structures. Therefore for the present study, if the energy gap between deep lying MOs with respect to that of HOMO is less than or equal to that of predicted binding free energy of the pose of interest, then only those deep lying MOs are considered. Thus by combining docking and QM calculations one can predict the site of metabolism of the ligand molecule. For the present work, major reactions involved in drug metabolism with different isoforms of CYPs, such as N-dealkylation, O-dealkylation, aliphatic hydroxylation, aromatic hydroxylation, N-oxidation, S-oxidation, and phosphothionate oxidation are considered. A computational flow chart for SOM prediction of any molecule against any isoform of CYP is depicted in Fig. 1.

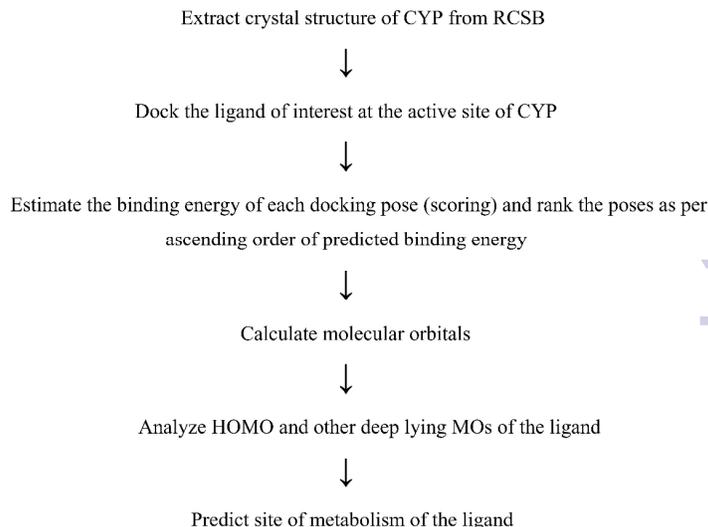


Figure 1: Computational flow chart for SOM prediction for any molecule against any isoform of CYP.

Results

For the prediction of binding free energies, all the 121 complexes are divided into a training set of 58 complexes and a test set of 63 complexes. The data set (Table 3) is diverse in molecular size, physicochemical features, and binding energies. In the first step, electrostatic and van der Waals, hydrophobic and entropic contributions are calculated. Partial least squares fit on the 58 complexes in the training set against experimentally observed binding free energies generated the values of regression coefficients as mentioned in eq. 2. In the next step, binding free energies are predicted for the test of 63 complexes.

Model validation

Next step of model development protocol is to validate the model by various statistical techniques. Model validation establishes the predictive power of the model. For the present model, the following statistical tests are performed.

$$(1) q^2 > 0.5; \quad (2) R^2 > 0.6;$$

$$(3) \frac{(R^2 - R_0^2)}{R^2} < 0.1 \quad \text{and} \quad 0.85 \leq K \leq 1.15;$$

$$(4) \frac{(R^2 - R_0'^2)}{R^2} < 0.1 \quad \text{and} \quad 0.85 \leq K' \leq 1.15;$$

$$(5) |R_0^2 - R_0'^2| < 0.3$$

All the above terms are explained in detail in the Supplementary (Supplementary Table 1).

The model passes all the validation tests (Table 5) indicating its strength and that of its parameters obtained for predicting the binding energies of heme-containing protein–ligand complexes.

Table 5: Statistical tests and their respective values for the training set.

Experiment	Value
q^2	0.725
R^2	0.725
$(R^2-R_0^2)/R^2$	-0.379
$(R^2-R_0'^2)/R^2$	-0.373
K	1.000
K'	0.983
$R_0^2-R_0'^2$	0.005
S_{PRESS} (kcal/mol)	1.191
RMS Error (kcal/mol)	1.002

Table 6: Statistical tests and their respective values for the test set.

Experiment	Value
q^2	0.656
R^2	0.701
$(R^2-R_0^2)/R^2$	-0.412
$(R^2-R_0'^2)/R^2$	-0.426
K	0.972
K'	1.004
$R_0^2-R_0'^2$	-0.010
S_{PRESS} (kcal/mol)	1.277
RMS Error (kcal/mol)	1.091

Y-randomization, Bootstrapping and External Validation

Y-randomization and Bootstrapping on training sets and statistical tests on external test set are performed to establish the statistical significance, robustness and predictive ability of the present model.⁷⁸⁻⁸³ The Y-randomization test consists of unique and repetitive randomization of the dependent variable Y (here Y is the experimental binding energy), followed by generation of a new QSAR equation and estimation of statistical parameter R^2 in each step using original independent parameters.⁷⁸⁻⁷⁹ The highest correlation coefficient obtained when running 100 y-randomization trials is 0.488 (mean = 0.45), which is significantly lower than $R^2 = 0.72$. Bootstrapping analysis (100 runs) is performed to further validate the robustness of the model. Bootstrapping technique consists of generation of new data sets by repeatedly and randomly choosing samples (rows) from the original data (here original data is experimental and predicted binding free energies) followed by calculation of statistical parameter R^2 in each run. These new data sets (bootstrap samplings) must be of the same size as that of original data set. The average value of R^2 after 100 bootstraps experiment is 0.72. Apart from this, statistical tests on test dataset are also done in order to establish the predictive ability of the model.⁷⁸⁻⁷⁹ The external validation set also has to satisfy the conditions mentioned above. The model passes all the statistical tests (Table 6) performed indicating its predictive strength.

As mentioned, the 121 complexes are divided into a training set of 58 complexes, the rest constituting the test set (Supplementary Tables 2 and 3). A correlation coefficient of 0.85 (Fig. 2) and an rms error of ± 1.00 kcal/mol in relation to experimental binding energies is obtained with the training dataset. A graphical analysis (Supplementary Fig. 1) of the residual errors of the 58 complexes of the training set is also performed. In this graph, points are almost equally uniformly distributed above and below the base line.

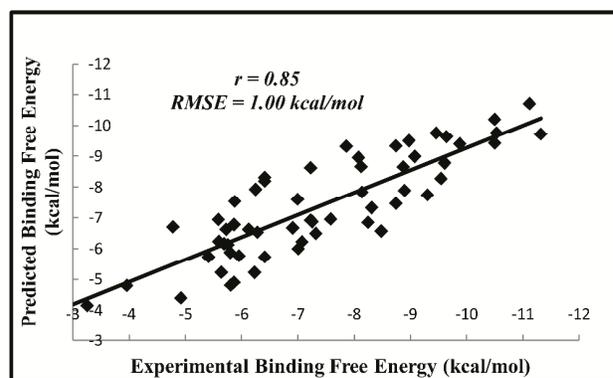


Figure 2: A correlation between experimental and predicted binding energies (kcal/mol) for the training set (58 complexes).

Finally, validation is performed on the external test set of 63 complexes. A correlation coefficient of $r=0.84$ (Fig. 3) and an rms error of ± 1.1 kcal/mol against the experimental binding energy are obtained for the test set. In the Figure 3 a single but significant outlier is observed. This corresponds to the CYP3A4-Bromoergocryptine complex. This is due to very low atom efficiency (binding energy/no. of atoms) of the bromoergocryptine molecule (-0.107 kcal/mol/atom) as compared to that of the average atom efficiency of all the 121 ligand molecules (-0.241 kcal/mol/atom).

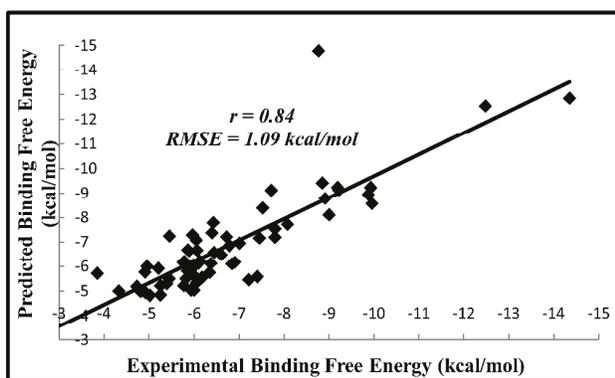


Figure 3: A correlation between experimental and predicted binding energies (kcal/mol) for the the test set (63 complexes).

A comparative study of SOM prediction by the present methodology and other popular methodologies

To evaluate the ability of predicting correct SOM by the present model, a comparative study with other popular techniques like MetaSite version 3.0, 2D SMARTCyp, StarDrop 5.0, MLite is carried out. For the present study, a total 213 substrates for 4 isoforms of CYPs namely, CYP1A2, CYP2C9, CYP2C19 and CYP3A4 are taken, and the accuracy of the present methodology is compared with those of other SOM prediction methodologies. Experimentally verified SOMs of all these substrates are obtained from literature.^{9, 10}

SOM Prediction for CYP1A2-Catalyzed Reactions

CYP1A2 is one of the important metabolizing enzymes which metabolizes 5% of the FDA approved drug molecules. For the present study, 60 substrates are taken whose experimental SOMs are reported in literature. Performance of other methodologies such as 2D SMARTCyp and MetaSite version 3.0, against these 60 substrates for CYP1A2 are also given in the literature.⁹ Among these 60 substrates, 39 molecules show N-dealkylation/O-dealkylation reactions and other 21 molecules show either aromatic or aliphatic hydroxylation reactions. Present docking based protocol followed by MO calculations is able to capture experimentally verified SOMs for 55 substrates (92% accuracy) within top 3 unique docking poses. Reactivity based 2D SMARTCyp protocol gave 87% accuracy in predicting correct experimental SOM while MetaSite server reported 85% accuracy for the same. Information about pose rank which showed experimentally verified SOM with the present methodology and energy gap between deep lying MO to HOMO (where available) for each of these 60 substrates is given in Supplementary Table 4.

SOM Prediction for CYP2C9, 2C19-Catalyzed Reactions

CYP2C9 and CYP2C19 play a vital role in the clearance of xenobiotics. They metabolize about 15% of marketed drugs. For the present study 70 substrates for 2C9 and 36 substrates for 2C19 are taken. Experimental SOMs of these substrates against CYP2C9 and 2C19 are reported in literature.⁹ Accuracies of experimentally observed SOM prediction of 2D SMARTCyp, and StarDrop 5.0 methodologies for the 70 substrates against CYP2C9 are also reported in the literature.⁹ Prediction ability of 2D SMARTCyp methodology on 36 substrates against CYP2C19 is also obtained from the literature.⁹ Among these 106 substrates for CYP2C9 and 2C19, 40 molecules show N-dealkylation / O-

dealkylation reactions, 9 molecules show S-oxidation reactions, 2 molecules show N-oxidation and rest show aromatic or aliphatic hydroxylation reactions. Present methodology successfully predicts experimentally verified SOMs within top three unique docking poses for 60 out of 70 CYP2C9 substrates (86% success rate). Accuracies for 2D SMARTCyp and StarDrop 5.0 are 87% and 77% respectively against the same data set. Zaretski et al.⁸⁴ reported SOM detection methodology based on neural networks with a data set of 631 complexes and an accuracy of 87%. However, results with artificial intelligence techniques in general, are not easily amenable to mechanistic interpretations. Information about pose rank which showed experimentally verified SOM and energy gap between deep lying MO to HOMO where available, for the substrates against 2C9 and 2C19 are given in Supplementary Tables 5 and 6.

SOM Prediction for CYP3A4-Catalyzed Reactions

CYP3A4 is the most important enzyme among all the isoforms of CYPs. It metabolizes 50% of the FDA approved drug molecules. 47 substrates for CYP3A4 are chosen. Experimentally verified SOMs and accuracy of prediction of SOMs on this dataset by MLite methodology, a combined approach based on docking and activation energy prediction, is also available in literature.¹⁰ Among these 47 substrates, 34 molecules show N-dealkylation/O-dealkylation reactions and the rest show either aromatic or aliphatic hydroxylation reactions. In 81% cases, the present methodology is able to predict successfully the experimentally verified SOM within the top three unique docking poses, while MLite methodology achieved an accuracy of 78% within top two sites. Information about pose rank which showed correct SOM and energy gap between deep lying MO to HOMO (where available) for each of the 47 substrates is given in Supplementary Table 7.

In short, overall accuracy of the present methodology is ~ 87% for the 213 substrates considered against the four different isoforms of CYPs. The major advantage of the present approach however is its transferability, i.e. one can dock the substrate by using any docking software and carry out molecular orbital population analysis by any quantum mechanical software, to predict the SOM of a molecule against any isoform of CYP.

Discussion

Based on only docking and scoring studies, it is not always possible to confirm the SOM of a molecule. MO calculations help here. For instance in Nicotine (Fig. 4), real site of oxidation, the -CH₂ group next to the nitrogen of pyrrolidine ring, comes in the 3rd pose after docking and scoring, while in the first two poses, -CH₂ group at the 4th position of the pyrrolidine ring is directed to the iron atom of the heme (Fig. 4a). The MO calculation shows that HOMO is composed of a lone pair from the nitrogen atom and -CH₂ group attached to the nitrogen atom of the pyrrolidine ring. No contribution of -CH₂ group which is directed in pose 1 and pose 2 in the docking and scoring studies, is shown in the HOMO (Fig. 4b).

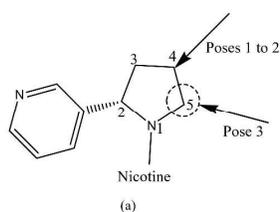


Figure 4: (a) Structure of Nicotine and its accessible sites after docking as indicated by arrows; (b) HOMO of Nicotine.

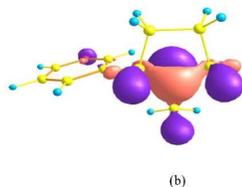


Figure 6: Molecular orbital diagram of Aflatoxin B₁; (a) HOMO, and (b) HOMO-1.

Therefore based on MO calculation, poses 1 and 2 are eliminated, and pose 3 decides the real site of oxidation.

Aflatoxin B₁ is metabolized by CYP1A2 at C₁₆ position to form Aflatoxin M₁ (Fig. 5). Docking pose having 8th rank (after scoring) shows desired orientation of the molecule towards iron atom of the heme. However docking poses for top seven poses are closely similar. Top three poses predict SOM as the C₅ center of the molecule, while according to the next four poses double bond between C₁₈-C₁₉ is the most probable SOM of the molecule (Fig. 5).

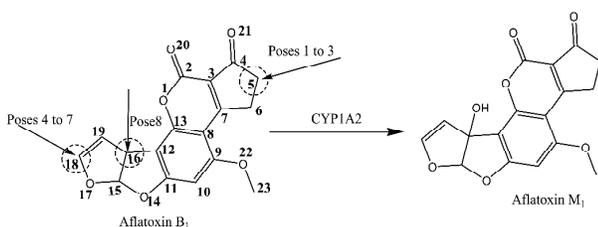


Figure 5: Structure of Aflatoxin B₁ and its metabolites Aflatoxin M₁. Accessible sites after docking are indicated by arrows.

MO calculations confirm that there is no contribution of C₅ atomic center of Aflatoxin B₁ in the HOMO, while double bond between C₁₈ and C₁₉ atoms of the molecule shows up in HOMO (Fig. 6). This result clearly indicates that C₅ center of the molecule is not an oxidizable center. However, C₁₈=C₁₉ double bond of the molecule is definitely affected by the same enzyme or different isoform of CYPs. Therefore based on the MO observations, poses 1 to 3 can be excluded. Real site of oxidation of the molecule is C₁₆ atom which is converted from C-H to C-OH functional group (Fig. 5) via first hemolytic cleavage of the C₁₆-H bond followed by insertion of -OH functional group to C₁₆ atom. At 8th ranked pose the distance between C₁₆ to iron atom is 4.33 Å while C₁₅ to iron atom distance is 4.16 Å. Therefore C₁₅ would be the more probable SOM of the molecule as per distance criterion. Since relative stability of the free radical species at C₁₆ position (allylic type) is more than that of C₁₅ analogue (alkyl type), in spite of the closer distance of C₁₅ to iron atom, C₁₆ atomic center is the likely SOM.

Apart from correct orientation, another prerequisite is the presence of contribution of HOMO to the atomic center of interest. MO calculations at the B3LYP and 6-31G** level in implicit solvent medium confirm that there is no contribution of HOMO at the C₁₆ atom (sigma bond overlap between C₁₆-H bond). However a close lying HOMO-1 molecular orbital (an energy gap of -9.0 kcal/mol with respect to HOMO) shows sigma overlap between C₁₆-H bond (Fig. 6).

45

The predicted binding free energy of 8th pose is -9.2 kcal/mol. As per experimental finding, the loss of electron takes place from C₁₆-H sigma bond (HOMO-1) rather than from C₁₈=C₁₉ double bond (HOMO). This may be due to the formation of a favorable TS through C₁₆ atomic center which is also reflected by the more negative binding free energy value than that of the energy gap between HOMO-1 and HOMO.

p-Isopropoxyacetanilide is metabolized by CYP1A2 at the methyl of acetamide group. HOMO of the molecule (B3LYP/6-31G** as well as BP86/6-31G** level of theory and implicit solvent treatment) is composed of amide bond of acetamide group as well as phenyl ring and oxygen of isopropoxy group (Fig. 7).

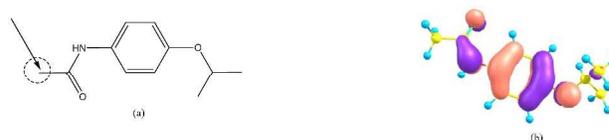


Figure 7: (a) SOM of p-Isopropoxyacetanilide; (b) HOMO of p-Isopropoxyacetanilide.

No contribution of methyl group of acetamide is found. HOMO-1 has the contribution of methyl group of acetamide. The energy gap between HOMO to HOMO-1 is more than 20 kcal/mol, which is much greater in magnitude than the predicted binding free energy of the ligand (-6.7 kcal/mol). Since the mechanism of hydroxylation of alkyl group of acetamide is not well understood, the contribution of acetamide as a whole is considered for the present study, in which case HOMO can offer a plausible explanation for the observed metabolic reaction.

Similarly, in the case of benzylic hydroxylation, contribution of HOMO is analyzed not only on the alkyl group but also on the aromatic center to which the alkyl group is attached. MO analysis predicts alkyl hydroxylation of t-butyl group to a single t-butylhydroxy metabolite which is not necessarily facile. In this type of reactions steric factors rather than electronic factors may play an important role.

Carvedilol is metabolized by CYP2C9 to form O-desmethyl-carvedilol. 3rd pose of the docking and scoring studies showed the desired orientation, i.e. -OCH₃ group of the molecule is directed towards iron atom of the heme (Fig. 8).

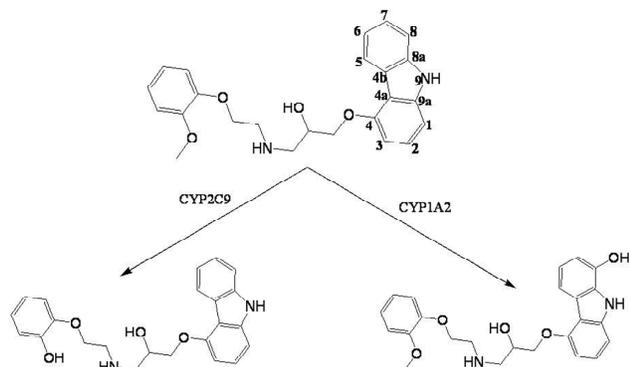


Figure 8: Reactions of Carvedilol by CYP2C9 and CYP1A2 give two different metabolites.

MO calculation at B3LYP level of theory and 6-31G** basis set in implicit solvent medium shows that instead of HOMO, HOMO-2 has contributions of $-CH_3$ group and a lone pair of the oxygen atom (Fig. 9).



Figure 9: Molecular orbital diagram of Carvedilol; (a) HOMO, and (b) HOMO-2.

Energy gap between HOMO-2 to HOMO is -12.2 kcal/mol, while the predicted binding free energy of 2nd pose is -9.6 kcal/mol. However MO calculation at BP86 level of theory and 6-31G** basis set in implicit solvent medium gave an energy gap of -9.0 kcal/mol which is less in magnitude than the predicted binding free energy.

Carvedilol is also metabolized by CYP1A2 at C₈ position to form 8-hydroxycarvedilol. HOMO of this molecule shows contribution of carbazole ring (Fig. 8). Docking pose ranked 2nd gives an orientation where C₈ position of carbazole ring is closer to iron of the heme moiety. The distance between C₈ to iron is 4.21 Å. However the C₇ position of the molecule is the closest atom to that of iron center and the distance is 3.78 Å. The mechanism of hydroxylation of aromatic ring is the formation of aromatic epoxide followed by its cleavage to form hydroxylated product (Fig. 10).

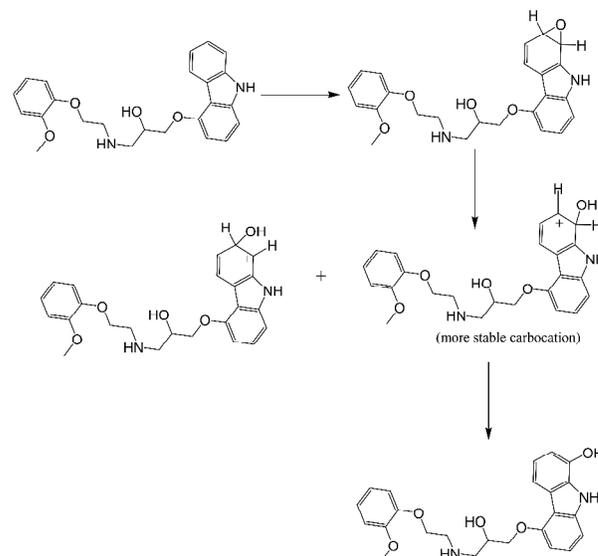


Figure 10: Mechanism of aromatic hydroxylation of Carvedilol by CYP.

Therefore as per epoxidation mechanism both C₇ and C₈ centers could be considered and depending on the stability of the intermediate carbocation, cleavage of the epoxide ring would take place to give the observed metabolite. Similarly, in case of N and O-dealkylation reactions both alkyl groups attached to N and O atoms and lone pairs of N and O atoms have to be considered.

In summary, the rule of finding SOM of a molecule against any isoform of CYPs is as follows. Atomic center/functional group which is directed towards the iron atom of the heme moiety should be part of HOMO or other deep lying MOs. In the latter case, extra energy required for removing the electron from MO other than HOMO of the molecule during CYP oxidation should be less than the binding energy value of the ligand with the particular isoform of CYP. If the atomic center is already in the highest oxidation state, then that center is metabolically inactive for oxidation as for instance the nitro group, carbonyl group ($C=O$), phosphate and sulfate. Allylic/benzylic centers of oxidation are preferred over olefinic/aromatic/aliphatic centers because of the stability of the radical generated through oxidation. For olefinic/aromatic bond, epoxidation is preferred. Oxidation of halogen containing functional groups is not very common owing to their high value of electronegativity. A web-server for predicting the site of metabolism is created and made freely accessible at <http://www.scfbio-iiitd.res.in/software/drugdesign/som.jsp>.

Conclusion

Metabolism studies are an essential part of ADMET profiling of drug candidates to evaluate their safety and efficacy. The aim of the present study is to develop a computational protocol to predict the site of metabolism of any xenobiotic compound. To this end, an empirical free energy function is proposed, which gives a correlation of 0.84 against experimental binding free energies on 121 heme-containing protein-ligand complexes. Present scoring protocol is combined with an in-house developed docking methodology, ParDOCK to suggest the metabolizing atomic center of a molecule. However electronic reactivity of the accessible center is carried out by a visualization of atomic center

of interest in the HOMO of the system. Other deep lying molecular orbitals are also considered only when HOMO does not show any contribution of the accessible center, and energy gap between other deep lying orbitals to HOMO of a molecule

5 does not cross the predicted binding free energy value of the ligand with the particular isoform of CYP. A few rules based on experimental observations and our theoretical analyses for identifying SOMs are proposed.

Combining analyses of MOs and knowledge based approaches with docking and scoring can fine tune the identification of potential sites of metabolism for any molecule. Out of 213 known substrates against 4 different isoforms of CYPs, present combination method gives 87% success rate in determining their experimentally verified SOMs within the top three unique poses

15 of docking. The main feature of the present methodology is the relatively quick estimation of MOs instead of the compute intensive TS or activation energy calculations of any molecule to verify the real SOM. Present methodology of SOM prediction could help drug designers at the stage of lead optimization.

20 Acknowledgements

This work is supported by grants from the Department of Biotechnology, Govt. of India. The authors thank Mr. Manpreet Singh, Mr. Shashank Shekhar and Ms. Vandana Shekhar for help in web-enabling the SOM prediction methodology.

25 Appendix A. Supplementary data

The supplementary includes PDB IDs, and the experimental and predicted binding free energies of the training and test data sets, formulas of statistical tests performed for the validation of the empirical scoring function, information about pose rank which

30 showed experimentally verified SOM and energy gap between deep lying MO to HOMO (where available) for each of the 213 substrates, a graphical analysis of the standard error of training set, 2D representations of 60, 70, 36 and 47 substrates for CYP1A2, CYP2C9, CYP2C19 and CYP3A4 respectively and

35 their experimentally verified sites of metabolism.

References

- 1 S. A. Shaikh, T. Jain, G. Sandhu, N. Latha and B. Jayaram, *Curr. Pharm. Des.*, 2006, **13**, 3454-3470.
- 2 C. A. Lipinski, F. Lombardo, B. W. Dominy and P. J. Feeney, *Adv. Drug Delivery Rev.*, 1997, **23**, 3-26.
- 3 C. A. Lipinski, *Drug Discovery Today: Technologies*, 2004, **1**, 337-341.
- 4 S. D. Nelson, *J. Med. Chem.*, 1982, **25**, 753-765.
- 5 P. C. I. Poggesi, *Eur. J. Med. Chem.*, 2006, **41**, 795-808.
- 6 R. G. Susnow and S. L. Dixon, *J. Chem. Inf. Model*, 2003, **43**, 1308-1315.
- 7 M. Eichelbau and O. Burk, *Nat. Med.*, 2001, **7**, 285-287.
- 8 J. P. Jones, M. Mysinger and K. R. Korzekwa, *Drug Metab. Dispos.*, 2002, **30**, 7-12.
- 9 R. Liu, J. Liu, G. Tawa and A. Wallqvist, *J. Chem. Inf. Model*, 2012, **52**, 1698-1712.
- 10 W. S. Oh, D. M. Kim, J. Jung, K.-H. Cho and K. T. No, *J. Chem. Inf. Model*, 2008, **48**, 591-601.
- 11 M. Moserová, V. Kotrbová, D. Aimová, M. Šulc, E. Frei, and M. Stiborová, *Interdisc Toxicol*, 2009, **2**, 239-244.
- 12 K. C. Patki, L. L. Von Moltke and D. J. Greenblatt, *Drug Metab. Dispos.*, 2003, **31**, 938-944.

- 13 G. Reign, H. McMahon, M. Ishizaki, T. Ohara, K. Shimane, Y. Esumi, C. Green, C. Tyson and S.-i. Ninomiya, *Carcinogenesis*, 1996, **17**, 1989-1996.
- 14 A. M. Butler and M. Murray, *J. Pharmacol. Exp. Ther.*, 1997, **280**, 966-973.
- 15 D. E. Moody, W. Ruangyuttikarn and M. Y. Law, *J. Anal. Toxicol.*, 1990, **14**, 311-317.
- 16 E. D. Kharasch, D. C. Hankins, K. Fenstamaker and K. Cox, *Eur. J. Clin. Pharmacol.*, 2000, **55**, 853-859.
- 17 G. Cruciani, E. Carosati, B. De Boeck, K. Ethirajulu, C. Mackie, T. Howe and R. Vianello, *J. Med. Chem.*, 2005, **48**, 6970-6979.
- 18 P. Rydberg, D. E. Gloriam, J. Zaretski, C. Breneman and L. Olsen, *ACS Med. Chem. Lett.*, 2010, **1**, 96-100.
- 19 StarDrop, version 5.0; Optibrium Ltd.: Cambridge, United Kingdom, 2011.
- 20 G. Klopman, M. Dimayuga and J. Talafous, *J. Chem. Inf. Comput. Sci.*, 1994, **34**, 1320-1325.
- 21 J. Talafous, L. M. Sayre, J. J. Mieyal and G. Klopman, *J. Chem. Inf. Comput. Sci.*, 1994, **34**, 1326-1333.
- 22 G. Klopman, M. Tu and J. Talafous, *J. Chem. Inf. Comput. Sci.*, 1997, **37**, 329-334.
- 23 <http://www.simulations-plus.com/>
- 24 <http://thomsonreuters.com/metadrug/>
- 25 D. D. Stranz, S. Miao, S. Campbell, G. Maydwell and S. Ekins, *Toxicol Mech. Methods*, 2008, **18**, 243-250.
- 26 <http://www-metaprint2d.ch.cam.ac.uk/>
- 27 S. Boyer, C. H. Arnby, L. Carlsson, J. Smith, V. Stein and R. C. Glen, *J. Chem. Inf. Model*, 2007, **47**, 583-590.
- 28 S. B. Singh, L. Q. Shen, M. J. Walker and R. P. Sheridan, *J. Med. Chem.*, 2003, **46**, 1330-1336.
- 29 M. J. Groot de, M. J. Ackland, V. A. Horne, A. A. Alex and B. C. Jones, *J. Med. Chem.*, 1999, **42**, 4062-4070.
- 30 D. Korolev, K. V. Balakin, Y. Nikolsky, E. Kirillov, Y. A. Ivanenkov, N. P. Savchuk, A. A. Ivashchenko and T. Nikolskaya, *J. Med. Chem.*, 2003, **46**, 3631-3643.
- 31 M. Zheng, X. Luo, Q. Shen, Y. Wang, Y. Du, W. Zhu and H. Jiang, *Bioinformatics*, 2009, **25**, 1251-1258.
- 32 B. Testa, A.-J. Balmat and A. Long, *Pure Appl. Chem.*, 2004, **76**, 907-914.
- 33 P. Vasanathanathan, J. Hritz, O. Taboureau, L. Olsen, F. S. Jørgensen, N. P. E Vermeulen and C. Oostenbrink, *J. Chem. Inf. Model*, 2009, **49**, 43-52.
- 34 R. J. Unwalla, J. B. Cross, S. Salaniwal, A. D. Shilling, L. Leung, J. Kao and C. Humblet, *J. Comput.-Aided Mol. Des.*, 2010, **24**, 237-256.
- 35 I. Zamora, L. Afzelius and G. Cruciani, *J. Med. Chem.*, 2003, **46**, 2313-2324.
- 36 M. J. Sykes, R. A. McKinnon and J. O. Miners, *J. Med. Chem.*, 2008, **51**, 780-791.
- 37 J. Kirchmair, M. J. Williamson, J. D. Tyzack, L. Tan, P. J. Bond, A. Bender and R. C. Glen, *J. Chem. Inf. Model*, 2012, **52**, 617-648.
- 38 J. D. Tyzack, M. J. Williamson, R. Torella and R. C. Glen, *J. Chem. Inf. Model*, 2013, **53**, 1294-1305.
- 39 J. Zaretski, C. Bergeron, P. Rydberg, T.-W. Huang, K. P. Bennett and C. M. Breneman, *J. Chem. Inf. Model*, 2011, **51**, 1667-1689.
- 40 Prime, version 3.0.111; Schrödinger: New York, 2011.
- 41 L. Afzelius, C. H. Arnby, A. Broo, L. Carlsson, C. Isaksson, U. Jurva, B. Kjellander, K. Kolmodin, K. Nilsson, F. Raubacher and L. Weidolf, *Drug Metab. Rev.*, 2007, **39**, 61-86.
- 42 M. Hennemann, A. Friedl, M. Lobell, J. Keldenich, A. Hillisch, T. Clark and A. H. Göller, *ChemMedChem*, 2009, **4**, 657-669.
- 43 P. Rydberg, P. Vasanathanathan, C. Oostenbrink and L. Olsen, *ChemMedChem*, 2009, **4**, 2070-2079.
- 44 S. L. C. Moors, A. M. Vos, M. D. Cummings, H. V. Vlijmen and A. Ceulemans, *J. Med. Chem.*, 2011, **54**, 6098-6105.
- 45 M. R. Meyer, F. T. Peters and H. H. Maurer, *Biochemical Pharmacology*, 2009, **77**, 1725-1734.
- 46 H. G. Oldham and S. E. Clarke, *Drug Metab. Dispos.*, 1997, **25**, 970-977.
- 47 A. Gupta, A. Gandhimathi, P. Sharma and B. Jayaram, *Protein and Peptide Letters*, 2007, **14**, 632-646.

- 48 H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Hindyalov and P. E. Bourne, *Nucl. Acids Res.*, 2000, **28**, 235-242.
- 49 PubChem database. <http://pubchem.ncbi.nlm.nih.gov/> (accessed May 11, 2009).
- 50 D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang and J. Woolsey, *Nucl. Acids Res.*, 2006, **34**, D668-D672.
- 51 D. S. Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam and M. Hassanali, *Nucl. Acids Res.*, 2008, **36**, D901-D-906.
- 52 C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, Y. Djoumbou, R. Eisner, A. C. Guo and D. S. Wishart, *Nucl. Acids Res.*, 2011, **39**, D1035-D1041.
- 53 V. Law, C. Knox, Y. Djoumbou, T. Jewison, A. C. Guo, Y. Liu, A. Maciejewski, D. Arndt, M. Wilson, V. Neveu, A. Tang, G. Gabriel, C. Ly, S. Adamjee, Z. T. Dame, B. Han, Y. Zhou and D. S. Wishart, *Nucl. Acids Res.*, 2014, **42**, D1091-D1097.
- 54 J. J. Irwin and B. K. Shoichet, *J. Chem. Inf. Model*, 2005, **45**, 177-182.
- 55 T. Jain and B. Jayaram, *FEBS Letters*, 2005, **579**, 6659-6666.
- 56 T. Jain and B. Jayaram, *Proteins: Struct. Funct. Bioinfo.*, 2007, **67**, 1167-1178.
- 57 T. Singh, D. Biswas and B. Jayaram, *J. Chem. Inf. Model*, 2011, **51**, 2515-2527.
- 58 B. Jayaram, T. Singh, G. Mukherjee, A. Mathur, S. Shekhar and V. Shekhar, *BMC Bioinformatics*, 2012, **13**, S7.
- 59 N. Arora and B. Jayaram, *J. Comput. Chem.*, 1997, **18**, 1245-1252.
- 60 N. Arora and B. Jayaram, *J. Phys. Chem. B*, 1998, **102**, 6139-6144.
- 61 A. Soni, K. Menaria, P. Ray and B. Jayaram, *Curr. Pharma. Des.*, 2013, **19**, 4687-4700.
- 62 Z. I. Hodes, G. Nemethy and H. A. Scheraga, *Biopolymers*, 1979, **18**, 1565-1610.
- 63 A. J. Hopfinger, *Macromolecules*, 1971, **4**, 731-737.
- 64 H. Park, S. Lee and J. Suh, *J. Am. Chem. Soc.*, 2005, **127**, 13634-13642.
- 65 P. A. Williams, J. Cosme, D. M. Vinković, A. Ward, H. C. Angove, P. J. Day, C. Vonnrhein, I. J. Tickle and H. Jhoti, *Science*, 2004, **305**, 683-686.
- 66 Y. Duan, C. Wu, S. Chowdhury, M. C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, J. Caldwell, J. Wang and P. A. Kollman, *J. Comput. Chem.*, 2003, **24**, 1999-2012.
- 67 J. M. Wang, P. Cieplak and P. A. Kollman, *J. Comput. Chem.*, 2000, **21**, 1049-1074.
- 68 A. Jakalian, B. L. Bush, D. B. Jack and C. I. Bayly, *J. Comput. Chem.*, 2000, **21**, 132-146.
- 69 A. Jakalian, D. B. Jack and C. I. Bayly, *J. Comput. Chem.*, 2002, **23**, 1623-1641.
- 70 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, T. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, O. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, D. J. Fox, Gaussian, Inc., Wallingford CT, 2010.
- 71 R. S. Mulliken, *J. Chem. Phys.*, 1955, **23**, 1833-1840.
- 72 V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg and C. Simmerling, *Proteins*, 2006, **65**, 712-725.
- 73 J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman and D. A. Case, *J. Comput. Chem.*, 2004, **25**, 1157-1174.
- 74 K. Shahrokh, A. Orendt, G. S. Yost and T. E. Cheatham, III, *J. Comput. Chem.*, 2012, **33**, 119-133.
- 75 D. A. Case, T. A. Darden, T. E. Cheatham, III, C. L. Simmerling, J. Wang, R. E. Duke, R. Luo, M. Crowley, R. C. Walker, W. Zhang, et al. AMBER 10; University of California: San Francisco, 2008.
- 76 S. Shaik, S. Cohen, Y. Wang, H. Chen, D. Kumar and W. Thiel, *Chem. Rev.*, 2010, **110**, 949-1017.
- 77 N. Taxak, V. A. Dixit and P. A. Bharatam, *J. Phys. Chem. A*, 2012, **116**, 10441-10450.
- 78 A. Tropsha, P. Gramatica and V. K. Gombar, *QSAR & Combinatorial Science*, 2003, **22**, 69-77.
- 79 S. Wold and L. Eriksson, *Chemometric Methods in Molecular Design* Waterbeemd, H. van de Ed Wiley-VCH, Weinheim, 1995, 309.
- 80 S. L. Kinnings, N. Liu, P. J. Tonge, R. M. Jackson, L. Xie and P. E. Bourne, *J. Chem. Inf. Model*, 2011, **51**, 408-419.
- 81 V. D. Mouchlis, T. M. Mavromoustakos and G. Kokotos, *J. Chem. Inf. Model*, 2010, **50**, 1589-1601.
- 82 L. Xing, R. Goulet and K. Johnson, *J. Chem. Inf. Model* **2011**, *51*, 1582-1592.
- 83 W. Deng, C. Breneman and M. J. Embrechts, *J. Chem. Inf. Comput. Sci.* 2004, **44**, 699-703.
- 84 J. Zaretski, M. Matlock, and S. J. Swamidass, *J. Chem. Inf. Comput. Sci.* 2013, **53**, 3373-3383.